



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

2021-05-08

西北农林科技大学

第二章 数据收集、整理和清洗

2.1 数据来源与形式

2.5 数据整理

2.2 数据收集

2.6 数据清洗

2.3 抽样设计

2.7 数据的数据库化

2.4 抽样分布和抽样误差

2.8 数据质量

2.1 数据来源与形式

数据来源

数据载体和形态



数据来源

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



数据载体和形态

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



数据载体和形态

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



数据载体和形态

- 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



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。



课堂思考

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。



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。

2.2 数据收集

陈述现实难点 (problem)

提出研究问题 (question)

数据搜集步骤

收集二手数据

收集一手数据



陈述现实难点 (problem)：基本内容

在开始数据收集过程之前，您需要明确难点并确定要实现的目标。

陈述难点，就是说明你要解决的实际或科学难点是什么，为什么重要？

明确的陈述难点需要具备如下几个要素：

- 将难点放在特定背景之中（我们已经知道什么？）
- 描述研究将要解决的确切难点（我们需要知道什么？）
- 显示难点的相关性（为什么我们需要知道它？）
- 设定研究目标（您将做什么以找出答案？）
- 提出研究难点：精确定义你要聚焦或解决的难点。



陈述现实难点 (problem)：如何陈述？

步骤1：将难点具体化。

- 对于实际研究难点，具体关注情况：难点何时何地出现？该难点影响谁？为了解决这个难点已经做了什么尝试？

例子：与该国其他地区相比，过去十年中X区的选民投票率一直在稳定下降。根据组织Y进行的调查，在25岁以下的人群和低收入人群中，投票率最低。已经进行了一些有效的尝试，以使这些团体参与其他地区的活动，在最近的两次选举中，政党A和B增加了在X区域的竞选活动，但是这些干预措施尚未对投票率产生任何重大影响。



陈述现实难点 (problem)：如何陈述？

- 对于理论研究难点，具体关注和考虑科学，社会，地理或历史背景等方面：关于该问题的已知信息是什么？难点是否仅限于特定时间段或地理区域？这个难点在学术文献中是如何定义和辩论的？

例子：在过去的十年中，“零工经济”（Zero hour）已成为劳动力市场中越来越重要的部分。30岁以下的年轻人更有可能从事自由、合同或零小时工作安排，而不是传统的全职工作（full time）。关于这种转变的原因和后果的研究集中在收入、工作时间和就业条件的客观衡量上，但是很少有研究探索年轻人对零工经济的主观经验。



陈述现实难点 (problem)：如何陈述2

步骤2：说明其重要性。

对于实际研究难点，重要性往往与特定难点直接相关，这个特定难点如何更广泛地影响组织、机构、社会团体或社会。因此，可以问：如果不解决难点将会怎样？谁会感到后果？难点是否具有更广泛的相关性（例如，在其他情况下是否也发现了类似的难点）？

例子：低投票率与社会凝聚力和公民参与度之间存在负相关关系，在许多欧洲民主国家中，这一点正日益引起人们的关注。当特定的公民群体缺乏政治代表权时，随着时间的流逝，他们很可能会被更多地排斥在外，从而导致人们对民主制度的信任度下降。解决该问题将为X区域带来实际好处，并有助于理解这一普遍现象。



陈述现实难点 (problem)：如何陈述2

对于理论研究难点，有时理论难点会产生明显的实际后果，但有时它们的相关性并不那么明显。要确定难点为何重要，可以问：解决难点将如何增进对议题的理解？它对未来的研究有什么好处？这个难点对社会有直接或间接的影响吗？

例子：在零工经济的文献中，这些新形式的就业有时被称为灵活的积极选择，有时被视为剥削性的不得已而为之。为了更全面地了解年轻人为何从事零工经济，需要进行深入的定性研究。关注工人的经验可以帮助建立更稳健的灵活性和不稳定性的理论，同时也为未来的政策目标提供信息。



陈述现实难点 (problem)：如何陈述3

步骤3：设定目的和目标。

难点陈述应说明您打算如何解决难点。您的目标不应是找到最终的解决方案，而应找出难点背后的原因，并提出解决或理解难点的更有效方法。

目的 (aim) 是你研究的总体目的，通常以不定式形式编写：

- 这项研究的目的是确定……
- 该项目旨在探索……
- 我打算调查……

目标 (objectives) 是你将要实现该目的的具体步骤：

- 定性方法将用于识别……
- 我将使用调查来收集……
- 使用统计分析，该研究将测量……



陈述现实难点 (problem)：如何陈述3

步骤3：设定目的和目标。

示例1（实际难点研究的目的和目标）：这项研究的目的是调查有效的参与策略，以增加X区域的投票人数。它将通过调查和访谈来确定不参与投票的最重要因素，并进行实验以衡量不同策略的有效性。

示例2（理论难点研究的目的和目标）：该项目旨在更好地了解年轻人在零工经济中的经验。定性方法将用于深入了解各个行业中从事自由职业和零小时工作的30岁以下青少年的动机和看法。这些数据将通过对演出经济的最新文献进行回顾，并对劳动力中的人口变化进行统计分析，从而进行背景分析。



(示例) 如何有效地陈述现实难点

现实难点：与该国其他地区相比，在过去十年中，X区的选民投票率一直在下降。

背景情形（已知的内容）：根据组织Y进行的调查，在25岁以下的人群和低收入人群中，投票率最低【引用具体数字】。有关Z国投票模式的文献表明，这反映了更广泛的趋势，但是该地区的人口统计信息使其成为一个更重要的问题【请使用来源进行扩展和解释】。已经进行了一些成功的尝试来提高其他区域的投票率，但是类似的干预措施尚未在X区域产生任何重大影响【引用来源】。需要就该地区参与的具体障碍以及影响青年和低收入人群的有效战略开展更多研究。



(示例) 如何有效地陈述现实难点

相关性（为什么重要）：低投票率与社会凝聚力和公民参与度之间存在负相关关系，在许多欧洲民主国家，政党和公民社会组织正在日益关注这一领域【提供实例和引用来源】。当特定的公民群体缺乏政治代表权时，随着时间的推移，他们很可能被更多地排斥在外，从而导致对民主机构的信任受到削弱，并在治理上带来困难【扩大并提供资料解释】。解决这一难点将使各政党有洞察力来调整其政策和竞选策略，改善X地区居民的民主包容性，并有助于对选民行为的当前趋势有更细微的了解。



(示例) 如何有效地陈述现实难点

目的和目标（您将要做什么）：这项研究的目的是调查有效的参与策略，以增加X区投票者的投票率。它将通过调查和访谈来确定不参与投票的最重要因素，并进行实验以衡量不同策略对投票意图的影响。



提出研究问题（question）：基本要求

一个好的研究问题对于指导您的研究、项目或论文至关重要。它将精确地指出了你要查找的内容，并为你的工作提供了明确的重点和目标。

所有研究问题应为：

- 专注于单个问题
- 可使用主要和/或次要来源进行研究
- 在时限和实际限制条件下可行回答
- 具体到足以彻底回答
- 足够复杂，可以在论文或论文的范围内得出答案
- 与你的学习或社会有广泛相关性



提出研究问题（question）：基本要求

下面给出两个示例：

（案例1）学校教育：



- 现实难点：X学校的老师没有能力识别或正确指导教室里的有天赋的孩子。
- 研究问题：X学校的老师可以使用哪些实用技术来更好地识别和指导有天赋的孩子？

（案例2）零工经济：



- 现实难点：30岁以下的年轻人越来越多地从事“零工经济”而不是传统的全职工作，但是很少有人研究年轻人在此类工作中的经历。
- 研究问题：影响年轻人参与零工经济的决定的主要因素是什么？工人认为它的优缺点是什么？年龄和受教育程度对人们体验这种工作的方式有影响吗？



提出研究问题 (*question*) : 常见类型

下表显示了一些最常见的研究问题类型。

研究问题类型	公式
描述性研究	X的特征是什么？
比较研究	X和Y之间有什么区别和相似之处？
相关研究	变量X和变量Y之间有什么关系？
探索性研究	X的主要因素是什么？ Y在Z中的作用是什么？
解释性研究	X对Y有影响吗？ Y对Z的影响是什么？ X的原因是什么？
评估研究	X的优缺点是什么？ Y工作得如何？ Z有多有效或理想？
行为研究	如何实现X？ 改善Y的最有效策略是什么？



提出研究问题 (question)：什么是好的研究问题？

好的研究问题应该：专注性和可研究

- 专注于单个主题和问题：您的主要研究问题应从您的研究问题开始，以使您的工作重点突出。如果您有多个问题，那么所有这些问题都应与该中心目标明确相关。
- 不要求主观价值判断。避免使用诸如“好/不好/更好/更糟”的主观用语，因为这些主观用语没有给出回答问题的明确标准。如果您的问题正在评估某些内容，请使用具有更可衡量的定义的术语。
 - (bad) X或Y是更好的策略吗？
 - (good) X和Y策略在降低Z率方面的效果如何？



提出研究问题 (question)：什么是好的研究问题？

好的研究问题应该：专注性和可研究

- 可使用主要或次要数据。您必须能够通过收集定量和/或定性数据或通过阅读有关该主题的学术资料来论证来找到答案。如果无法访问此类数据，您将不得不重新考虑您的问题，并提出更具体的问题。
- 不问为什么。为什么问题通常过于开放而不能充当良好的研究问题。通常有很多可能的原因导致研究项目无法给出详尽的答案。尝试询问什么或如何提问。
 - (bad) 为什么会出现X?
 - (good) 造成X的主要因素是什么?
 - (good) X如何受到Y的影响?



提出研究问题 (question)：什么是好的研究问题？

好的研究问题应该：可行而具体

- 在限定条件下完成：确保您有足够的时间和资源来进行回答问题所需的研究。如果您认为您可能难以获得足够的数据访问权限，请考虑缩小问题的范围，使其更加具体。
- 使用明确的特定概念：您在研究问题中使用的所有术语均应具有明确的含义。避免使用模糊语言和广阔的思路，清楚你的问题指向是什么、谁、哪里和何时出现？
 - (bad) 社交媒体对人们的思想有什么影响？
 - (good) 每天使用Twitter对16岁以下青少年的注意力范围有什么影响？
- 不要求最终的解决方案/政策或行动方案：研究是在告知而非指导。即使您的项目专注于实际问题，它也应该旨在增进理解并提出可能性，而不是寻求现成的解决方案。
 - (bad) 政府应如何处理投票率偏低的问题？



提出研究问题 (question)：什么是好的研究问题？

好的研究问题应该：复杂而有争议

- 无法回答“是”或“否”：封闭的是/否问题太简单，无法像优秀的研究问题一样工作——它们没有提供足够的调查和讨论范围。
 - (×) 在过去的十年中，英国的无家可归人数有所增加吗？
 - (✓) 在过去十年中，经济和政治因素如何影响英国的无家可归者模式？
- 无法用容易找到的事实和数字来回答：如果您可以通过Google搜索或阅读一本书或一篇文章来回答问题，那么它可能还不够复杂。一个好的研究问题需要原始数据，多种来源的综合，解释和/或论据才能提供答案。
- 提供辩论和审议的范围：问题的答案不仅应是简单的事实说明，还需要有空间供您讨论和解释所发现的内容。在论文或研究论文中，这一点尤其重要，因为对于您的问题的答案通常采取有争议的论文陈述的形式。



提出研究问题 (*question*)：什么是好的研究问题？

好的研究问题应该：与现实相关且具有原创性！

- 解决与您所在领域或学科相关的问题：研究问题应基于围绕您的主题的初步阅读而提出，并且应专注于解决现有知识中的问题或差距。
- 有助于进行话题性的社会或学术辩论：这个问题的目的应该是促进现有的辩论——理想情况下是您所在领域或整个社会当前的辩论。它应该产生将来的研究人员或从业人员可以依靠的知识。
- 问题尚未得到回答：您不必问以前从未有人想到的突破性问题，但是这个问题应该具有独创性（例如，通过专注于特定位置或对长期辩论进行新的思考）。



(示例) 提出好的研究问题 (question)

示例1：社交媒体

- (bad) 社交媒体对人们的思想有什么影响?
- (good) 每天使用Twitter对16岁以下青少年的注意力范围有什么影响?



第一个问题还不够具体：什么类型的社交媒体？哪个人 有什么样的效果？

第二个问题更清楚地定义了其概念。它是可研究通过定性和定量数据收集。



(示例) 提出好的研究问题 (question)

示例2：荷兰住房危机

- (bad) 为什么荷兰出现住房危机?
- (good) 大学国际化政策对荷兰住房的可用性和可负担性产生了哪些影响?



以“为什么”开头通常意味着您的问题不够集中：可能的答案太多，没有明确的研究起点。通过仅针对问题的一个方面并使用更具体的术语，第二个问题为找到答案提供了一条清晰的途径。



(示例) 提出好的研究问题 (question)

示例3：医疗保健系统

- (bad) 美国或英国有更好的医疗保健系统吗？
- (good) 美国和英国如何比较慢性病低收入人群的健康结局和患者满意度？



第一个问题过于笼统和过于主观：什么才是“更好”没有明确的标准。第二个问题要研究得多。它使用明确定义的术语，并将其关注范围缩小到特定人群。



(示例) 提出好的研究问题 (question)

示例4：地区投票率

- (bad) 政党应该如何应对X区的低投票率？
- (good) 在X区30岁以下的年轻人中，提高选民投票率的最有效的交流策略是什么？



对于学术研究来说，回答关于“应该做什么”的广泛问题是不可行的。第二个问题更具体，目的是了解可能的解决方案，以便提出明智的建议。



研究数据怎么来？

“老师，我要做一个研究”

“你的数据从哪里来？”

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。



研究数据怎么来？

根据研究数据的持续性，来源有：

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。



研究数据怎么来？

根据研究数据是否为研究者产生，来源有：

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。



研究数据的获取权限

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？”



数据搜集步骤1：定义研究目标

在开始数据收集过程之前，您需要准确确定要实现的目标。

您可以从编写现实难题（problems）陈述开始：您要解决的实际或科学难题是什么，为什么重要？接下来，提出一个或多个研究问题（questions），以精确定义您要查找的内容。根据您的研究问题，您可能需要收集定量或定性数据：

- 如果您的目的是检验假设，精确测量某些东西或获得大规模的统计见解，请收集定量数据。
- 如果您的目的是探索想法，了解经验或获得对特定环境的详细见解，请收集定性数据。
- 如果您有多个目标，则可以使用混合方法来收集两种类型的数据。



(示例) 如何定义研究目标

(示例)：您正在研究员工对大型组织中直接经理的看法。



- 您的首要目标是评估不同部门和办公室地点对经理的看法是否存在显著差异。
- 您的第二个目标是从员工那里收集有意义的反馈意见，以探索有关管理人员如何改进的新想法。

数据搜集步骤2：选择数据收集方法

根据您要收集的数据，确定最适合您的研究的方法。

```
df_tools <- read_delim("../data/data-collection-tool.txt", delim = "\t", header = "infer")
#DT::datatable(df_tools, options = list(pageLength = 6, dom = "t"))
df_tools %>%
  add_column("序号"=1:nrow(.), .before = "方法") %>%
  kable(align = "l") %>%
  kable_styling(full_width = T) %>%
  column_spec(1, width = "2em")
```



数据搜集步骤3：规划资料收集程序

当您知道正在使用哪种方法时，您需要准确计划如何实现它们。您将遵循什么程序对您感兴趣的变量进行准确的观察或测量？

- 如果您要进行调查或访谈，请决定将采取何种形式的问题；
- 如果您要进行实验，请对实验设计做出决定。
- 实现可操作化。可操作化意味着将抽象的概念转变为可测量的观察结果。在计划如何收集数据时，需要将要学习的内容的概念定义转换为实际要测量的操作定义。
- 设计采样方式。您可能需要制定抽样计划以系统地获取数据。这涉及到定义总体，您要得出结论的组以及要从中实际收集数据的组的样本。
- 编写标准化程序。如果涉及多个研究人员，请编写详细的手册以标准化研究中的数据收集程序。
- 制定数据管理计划。在开始收集数据之前，您还应该决定如何组织和存储数据。



数据搜集步骤4：动手收集资料

最后，您可以实现所选方法来测量或观察您感兴趣的变量。

为确保以系统的方式记录高质量数据，以下是一些最佳做法：

- 在获取数据时记录所有相关信息。例如，记下在实验研究期间是否或如何重新校准实验室设备。
- 仔细检查手动数据输入是否有错误。
- 如果收集定量数据，则可以评估可靠性和有效性，以表明您的数据质量。



收集二手数据(搜索引擎类)

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 data

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

电车统计数据 + 美国

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 or selected vehicles on the U.S. market 2015.

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

Charge Durations: 0-15, 1-20, 2-25, 3-40, 4-45

Chart: Sales of Hybrid Electric Vehicles, 1999-2015

Month	REV	PREV	REV	TOTAL
January	21,776	10,033	10,341	32,150
February	30,222	11,023	6,078	37,303
March	30,461	11,118	7,753	40,722
April	30,164	11,109	7,031	40,905
May	32,164	11,109	7,031	40,905
June	32,815	12,057	8,774	44,646
July	32,815	12,057	8,774	44,646
August	36,035	4,920	7,300	40,255
September	36,035	4,920	7,300	40,255
October	33,260	4,060	7,000	40,320
November	33,260	4,060	7,000	40,320
December	40,600	4,900	7,754	50,254
Total	474,848	38,923	14,291	488,072

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，订阅者会通过 wsc_j2008@gmail.com 收到电子邮件通知

- 二宝！谷歌学术
-
-  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 Economics, 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 ...
-
-  Todd Clark

三宝！谷歌图书

图书

我的书架

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

样章	样章	样章	样章	样章	样章	样章
CSS: The Definitiv... Eric A. Meyer, Este...	CSS Pocket R... Eric A. Meyer	中国农业政策的...	The Economics o...	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(国内文献和统计数据)

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(国外文献和统计数据)

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(一个小项目!)

- 中国旱区农业科技资源配置研究





收集二手数据(一个小项目2)





典型的学术数据来源（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\(Understanding Society\)](#)。



典型的学术数据来源（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。



学术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。



收集调查数据I：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2：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3：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：总结

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

| 有时大家可以先看看“调查问卷设计”和“市场调研”相关图书！

西北农林科技大学
NORTHWEST A&F UNIVERSITY



收集实验数据

此处略！



收集数据的几点忠告

学生：“既然有这么多的数据，这门课是不是可以不学了？”

回答：“这门课你不仅要学，而且要认认真真地学”

掌握数据采集的知识与能力，是用好数据的基础。如果不了解数据是怎么获得的，就没有能力甄别已有的数据到底可不可靠、可不可用，甚至都不知道上哪儿去找数据。

- 第一，研究数据有多种、多重的来源，好好运用既有的数据是研究者的第一选择；
- 第二，获取已经存在的数据有很多个方法，也有多种途径
- 第三，万一没有办法获取需要的研究数据，那就只好自己动手。

2.3 抽样设计

抽样的逻辑

抽样的要素

概率/非概率抽样

抽样方案和实施

抽样误差



什么是抽样

假设我们希望通过自己调查来获得一手数据，就需要回答一系列问题：

1. 抽样的基本原理是什么？
2. 抽样的基本要素有哪些？
3. 抽样的逻辑是什么？
4. 什么条件该要采用概率抽样方法？怎么样做概率抽样设计？
5. 什么条件下要采用非概率抽样方法？又如何做非概率抽样？
6. 一个抽样方案应该包括哪些内容？
7. 怎么样去实施抽样工作？



抽样的要素（总体）

总体：是研究问题指涉对象的集合体，也就是研究问题涉及的全部对象。

- CFPS的总体是中国所有的家庭户
- CGSS的总体是中国所有的个体
- 入学机会的地区不平等研究的总体，就是某年所有的高中毕业生。

问题是：

- 什么叫中国所有家庭户，中国所有个体？
- 什么叫所有，台湾算不算？香港和澳门算不算？
- 住在中国的还是有中国户籍的？住在中国的外国人算不算？
- 长期出国却依然有着中国户籍的人算
不管？

- 什么叫家庭户？没有生活在一起，户口在一起算不算？生活在一起，
- 户口不在一起的，算不算？怎么才算某个地方的家庭户？
- 户口在甲地，却很少在甲地居住，算不算甲地的家庭户？
- 什么叫所有高中毕业生？没有参加高考的算不算？因非主观原因有参加高考的算不算？



抽样的要素（研究总体）

研究总体，是指可操作的研究对象，或称为可及总体。

CFPS把总体定义为中国的家庭户，是指有中国户籍的家庭户，指住在一起的，不管户籍是不是在一起的家庭户。

提问：



- CFPS中，家庭户指居住在二十五个省级单位内的家庭户吗？
- 户籍不在本地的算不算？
- 住又是什么意思呢？
- 住多长算是住？
- 一个人打工住在本地算是一户吗？



抽样的要素（抽样框和抽样单位）

抽样框：又叫抽样总体、框总体，是从研究总体中获得的用于抽取样本的研究对象的集合。

- CFPS的总体是中国所有的家庭户；
- CFPS的研究总体是二十五个省、市、自治区的常住户；
- CFPS的 抽样框是二十五个省、市、自治区在一个地方连续居住六个月或以上常住户。
- 从覆盖面和覆盖的对象数量出发， $\text{总体} \geq \text{研究总体} \geq \text{抽样框}$ 。



抽样的要素（抽样单位和样本）

抽样单位：是抽样指涉的基本单位，或包括基本单位的单位集合体。

- CFPS在抽到家庭户之前还要抽样本区县，样本村居。每一次抽样面对的基本单位就是抽样单位。

样本：是从抽样框中运用抽样策略和抽样方法获取的样本单位的集合。

- CFPS的样本是，从25个省市自治区抽取的160个区县样本，从160区县样本中抽取的640村居样本，从640村居样本中抽取的16000个家庭户样本。



抽样的逻辑（样本代表性）

抽样的基本逻辑1：选择一定数量的样本，来拟合总体中个体变异性的分布，进而代表总体。

抽样的基本逻辑2：用尽量少的样本，在可接受的误差范围内，来代表总体的研究特征。

柯西的思想：用代表性的样本就可以估计总体的研究特征。

柯西曾去美国国会作证，他反对在美国实施人口普查，认为每十年一次的人口普查，耗费太多的资源，实在没有必要。

如果个体在总体的分布是随机的，根据随机性原则抽取的样本就能代表总体，就是代表性样本。

在研究实践中，样本与总体之间总是有差异的。即时在随机条件下，尽管每个抽样单位被抽中的概率是相等的。由样本特征与总体特征之间，总是有差距的。



抽样的逻辑（抽样误差）

抽样误差：是样本研究特征与总体研究特征之间的差异。误差的大小一般取决于样本的代表性。样本对总体的代表性越好，误差就越小，否则误差就会越大。

依据误差的来源环节，可分为：

1. 随机误差：误差就是由抽样环节造成的误差。随机误差是希望尽量避免的误差。
2. 系统误差：误差具有规律性，主要是由抽样设计造成的。系统误差是我们最应当避免的。因为一旦出现的系统误差，几乎就没有补救的余地，
 - 假设希望知道性别与成就之间的关系。严格按照抽样方案完成的抽样，抽到的样本却都是男性的，没有女性。



抽样的逻辑（抽样误差）

依据抽样活动涉及的对象，可以把误差来源分为：

1. 覆盖性误差：是抽样活动没有正确的覆盖需要覆盖的总体，要么对总体覆盖过度，要么覆盖不住，过度和不足都会导致误差。
 - 假设界定的总体为参加高考的高中毕业生
 - 如果在抽样中把自愿或者是因为其他原因没有参加高考的毕业生都纳入到了抽样的范围，这就是覆盖过度。
 - 如果我们把复读并参加了高考的学生排除在了抽样的范围，这就是覆盖不足。
2. 选择性偏差：在设计与执行中，因偏好或者抽样活动而导致某个特定类型的样本的分布出现问题。
 - 某一类人群过多或者过少或者缺失。
 - 某个人群不在抽样框，被选机会就没了。



抽样的逻辑（样本分布）

抽样的逻辑3：利用重复多次抽样，提高抽样代表性，减小抽样误差。

抽样分布：又称统计量分布，指样本估计值的分布。抽样分布可以用来测量抽样方法的稳定性。

总体分布：是指总体特征值的分布。总体分布并不总是可得的，即使可得，也不满足经济性原则。



抽样方式（总览）

概率抽样：就是运用等概率原则进行抽样的总称。等概率原则，是指总体中每一个研究对象被抽中的概率是相等的。包括：简单随机抽样、系统抽样、整群抽样、与规模成比例的概率抽样、分层抽样以及隐含的分层抽样、多阶段混合抽样。

非概率抽样：抽取样本时不是依据随机原则，而是根据研究目的对数据的要求，采用某种方式从总体中抽出部分单位对其实施调查。包括：方便抽样、判断抽样、自愿样本、滚雪球抽样、配额抽样。

从抽样方式的具体运用，又可分为：

- 直接抽样：一次抽样或独立抽样。简单随机抽样、系统抽样和整群抽样都是直接抽样
- 半截抽样：通常不可以独立地用，要结合前直接抽样来使用。规模成比例的概率抽样、分层抽样以及隐含的分层抽样、多阶段混合抽样都属于半截抽样。



概率抽样I：简单随机抽样（步骤方法）

简单随机抽样(simple random sampling)：从总体N个单位中随机地抽取n个单位作为样本，每个单位入抽样本的概率是相等的。它是最基本的抽样方法，也是其它抽样方法的基础。

实施方法：

- 第一步，制备抽样框
- 第二步，对要素进行编码
- 第三步，根据抽样的要求抽取样本。
 - 直接抽选法
 - 抽签法
 - 随机数码表法（或kish table）
 - 软件抽取法



概率抽样I：简单随机抽样（优缺点）

优点：

- 简单、直观，在抽样框完整时，可直接从中抽取样本
- 用样本统计量对目标量进行估计比较方便

缺点：

- 当N很大时，不易构造抽样框
- 抽出的单位很分散，给实施调查增加困难
- 没有利用其它辅助信息以提高估计的效率
- 使用随机数表抽样的效率往往比较低下，即使用到，也会使用随机数表的一些变体如 kish table



概率抽样I：简单随机抽样 (kish table)

Kish, L. (1949). A Procedure for Objective Respondent Selection Within the Household, Journal of the American Statistical Association, 380–387.

- 第一步，制备末端抽样框，将样本家户所有符合要素资格的成员，按照规则顺序编号，依据性别也好，年龄也好，逆序也好，顺序也好，怎么排都行，要求是不重，不漏。
- 第二步，拿出事先准备好的kish表，根据指引，抽取样本。抽样的约定是不管家里有几个要素，只抽取其中的一个要素作为样本。



概率抽样I：简单随机抽样 (kish table)

Household	Kish Table
1	A
2	A
3	B1
4	B2
5	C
6	C
7	D
8	D
9	E1
10	E2
11	F
12	F
13	A
14	A
15	B1
16	B2
17	C
18	C
19	D
20	D
21	E1
22	E2
23	F
24	F
25	A
26	A
27	B1
Etc.....	

Sequentially work down the list

Selection table D	
If the number of adults in household is:	Select adult numbered:
1	1
2	2
3	2
4	3
5	4
6 or more	4

Overall Selection Probabilities

Adult numbered	If the number of adults in household is:					
	1	2	3	4	5	6 or more
1	1	1/2	1/3	1/4	1/6	1/6
2		1/2	1/3	1/4	1/6	1/6
3			1/3	1/4	1/4	1/6
4				1/4	1/6	1/6
5					1/4	1/6
6						1/6
7 or more						0



概率抽样I：简单随机抽样（软件实现）

利用统计软件能快速实现简单随机抽样：

- SPSS
- excel
- R

简单随机抽样的两点忠告：

- 简单随机抽样是不得已的办法，不是最先选用的办法
- 只有在总体的信息所知甚少的情况下，才用它。



概率抽样I：简单随机抽样（R示例）

任务：从教学班上随机抽取6人。

- 第一步，确认当前的班级是样本班级，制作抽样框。
- 第二步，对班级的83位同学从1到83实行顺序编码。编码顺序可以按学号、按座位等，只要是有规则，并且保证每一位同学只有一个唯一的编号就行。
- 第三步，选择一个随机数表，大家可以找到很多的随机数表（教材附录）。在查阅随机数表之前，说出第一个样本的行列位置作为起点。
- 第四步，在随机数表上找到上面的起点，然后取一组随机数的固定位置，按照事先制定的规则，依次选中随机数字中的一位。



概率抽样I：简单随机抽样（R示例）

全体学生名单（按班级和学号排序）：共83人。

序号	学号	姓名	班级
1	2017014588	马丽	农管1801
2	2018011379	安舒心	农管1801
3	2018013874	刘照润青	农管1801
4	2018014553	李铮	农管1801
5	2018014556	张晓旭	农管1801

Showing 1 to 5 of 83 entries

Previous

1

2

3

4

5

...

17

Next



概率抽样I：简单随机抽样（R示例）

不放回-简单随机抽样：

从1-83中产生6个随机数：

```
choice <- base::sample(1:n, size = 6, replace = FALSE )  
choice  
[1] 34 45 59 11 19 46
```



概率抽样I：简单随机抽样（R示例）

不放回-简单随机抽样：

抽取到的6个学生：

序号	学号	姓名	班级
11	2018014594	赖波妮	农管1801
19	2018014645	谢一雪	农管1801
34	2018014694	许贝贝	农管1802
45	2018014759	梁艳	农管1802
46	2018014762	刘怿冷	农管1802
59	2018014823	呼恬	农管1802



概率抽样I：简单随机抽样（kish table）R示例

下面是一张kish随机数表：

一份kish随机数表

adults	A	B1	B2	C	D	E1	E2	F
1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2	2
3	1	1	1	2	2	3	3	3
4	1	1	2	2	3	3	4	4
5	1	2	2	3	4	3	5	5
>=6	1	2	2	3	4	5	5	6



概率抽样1：简单随机抽样 (kish table) R示例2

假设需要调查共30户家庭，并对每户的成年人进行了编号：

id	adults. num
1	7
2	7
3	3
4	6
5	3
6	2

Showing 1 to 6 of 30 entries

Previous [1](#) [2](#) [3](#) [4](#) [5](#) Next



概率抽样I：简单随机抽样 (kish table) R示例3

按家庭成年人总数做分类，结合kish表可以得到：

id	adults.num	adults	A	B1	B2	C	D	E1	E2	F
1	7	>=6	1	2	2	3	4	5	5	6
2	7	>=6	1	2	2	3	4	5	5	6
3	3	3	1	1	1	2	2	3	3	3
4	6	>=6	1	2	2	3	4	5	5	6
5	3	3	1	1	1	2	2	3	3	3
6	2	2	1	1	1	1	2	2	2	2

Showing 1 to 6 of 30 entries

Previous

1

2

3

4

5

Next



概率抽样I：简单随机抽样 (kish table) R示例4

进一步地，每户都可以在8张表（A-F）中做出随机选择：

id	adults.num	adults	table	select
1	7	≥ 6	A	1
1	7	≥ 6	B1	2
1	7	≥ 6	B2	2
1	7	≥ 6	C	3
1	7	≥ 6	D	4
1	7	≥ 6	E1	5

Showing 1 to 6 of 240 entries

Previous

1

2

3

4

5

...

40

Next



概率抽样1：简单随机抽样 (kish table) R示例5

最后随机抽取kish表的结果如下

id	adults.num	adults	table	select
1	7	≥ 6	A	1
2	7	≥ 6	E2	5
3	3	3	B1	1
4	6	≥ 6	E1	5
5	3	3	B1	1
6	2	2	B1	1

Showing 1 to 6 of 30 entries

Previous

1

2

3

4

5

Next



概率抽样2：系统抽样（实施方法）

系统抽样(systematic sampling): 将总体中的所有单位按一定顺序排列(变量要素), 在规定的范围内随机地抽取一个单位作为初始单位, 然后按事先规定好的规则确定其它样本单位。系统抽样也称为等距抽样。

应用情景:

- 总体要素与抽样对象一致
- 总体通常规模也不大
- 变量异质性没有大到需要分层处理的程度
- 要素的特征在排列中没有周期性变化

实施方法是:

1. 把抽样框的要素按照规则进行编码。
2. 用要素总体数除以样本数, 得到抽样距(不是整数怎么办)。
3. 选择任何一个随机起点, 依照抽样距或者顺序抽样或者循环抽样。

假设一个班级有50个人, 男生25位, 女生25位, 在排列时, 每位男生的后面或者前面都是女生。这样男生跟女生之间的排列就是周期性的排列。万一要素的排列的周期与抽样距吻合了, 抽到的就只有一类样本, 从而引起选择性偏差。



概率抽样2：系统抽样（示例）

任务：用系统抽样方法从16名学生中随机抽取3人：

- 第一步，把班内所有的的学生名单按照按照学号进行排列。
- 第二步，把排列好的学号，从1开始顺序编号。
- 第三步，假设我们要在16位学生中，抽出3个样本，抽样距为5，样本量为3。
- 第四步，假设我们把要素排列成一个循环圈，选择一个随机起点为编号8，编号8就是第一个样本。顺时针数第5个也就是编号11，就是第二个样本，以此类推。

在排列要素的时候，我们不仅可以排列成循环圈，也可以排列为直线。
数到16不够测量距了怎么办？回头接着数到编号2，就是第三个样本。



概率抽样2：系统抽样（优缺点）

优点：

- 操作简便，可提高估计的精度

缺点：

- 系统抽样的框不能太大。太大了就很费事，仅就要素编号 就比较费事，
- 要素的排列特征不能呈现周期性变化。



概率抽样3：整群抽样（应用情景）

整群抽样 (cluster sampling): 将总体中若干个单位合并为组(群)，抽样时直接抽取群，然后对中选群中的所有单位全部实施调查。

应用情景：

- 是群内具有异质性，不过异质性还没有还没有大到需要专门处理的程度。
- 群之间具有差异性，但也没有大到需要专门处理的程度。
- 通常不作为独立抽样的方法使用，而是用于多阶段、多层次抽样的末端。

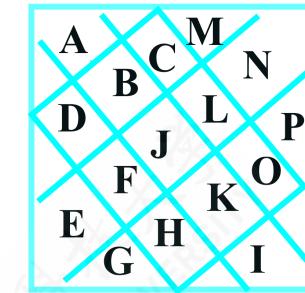


概率抽样3：整群抽样（步骤方法）

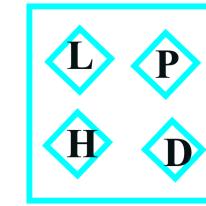
实施步骤：

1. 确定抽样框
2. 根据变量或辅助变量把总体分成若干子群
3. 确定样本容量和样本子群数
4. 依照简单随机抽样方法随机抽取子群

总体群数R=16



样本群数r=4



样本容量

$$n = n_d + n_p + n_l + n_h$$



概率抽样3：整群抽样（示例）

任务：为了分析教学法在16个班级上的效果，请按照整群抽样法，抽取90个学生。

注意：

- 采用整群抽样法，是假设了班与班之间特征的差异不大。每个班级同学的学习成绩有一个分布，在班与班之间，具有相似性。
- 整群抽样法实施中，分群过程非常重要。分群的基本原则是：
 - 在选择研究变量或者辅助变量时，让它在群间具有相似性，在群内具有异质性。
 - 如果群内同质群间非常异质，那就不适合用整群抽样了。
 - 相似的可以用做分群标准的辅助变量，比如说行政区划、组织、行业、班级、年龄、性别等等之类。



概率抽样3：整群抽样（优缺点）

优点：

- 抽样时只需群的抽样框，可简化工作量
- 调查的地点相对集中，节省调查费用，方便调查的实施

缺点：

- 估计的精度较差
- 在分群中有一点需要注意，群的规模不宜过大，否则就有可能出现内部同质性。影响抽样的效率，操作起来也很麻烦。



概率抽样4：成比例抽样（原理）

成比例的概率抽样(Probability Proportionate to Size Sampling)：又称按规模大小成比例的概率抽样或PPS抽样。

原理：

- 如果总体的要素之间在研究变量上有异质性，不同规模要素群体之间异质性的分布不是随机的。在这样的条件下，就要考虑把规模因素纳入抽样的考量了。
- PPS抽样理论上运用了等概率原理，希望让每一个抽样单位被抽中的概率与抽样单位的规模成比例。



概率抽样4：成比例抽样（实施方案）

实施方案：

- PPS抽样是一种按概率的比例抽样，在多阶段抽样中，尤其是二阶段抽样中，初级抽样单位被抽中的机率取决于其初级抽样单位的规模大小
- 初级抽样单位规模越大，被抽中机会就越大，初级抽样单位规模越小，被抽中机率就越小。
- PPS抽样也可以运用软件工具执行
 - Stata工具下ADO模块的gsample或者samplepps。



概率抽样4：成比例抽样（示例）

海淀区西北旺乡有100个社区，4万户。假设抽样要求是，要抽取10个社区，每个样本社区抽取20户，一共抽200户。假设针对海淀区西北旺乡A、B两个社区抽样，其中A社区有2000户，B社区只有500户。

在社区层次来看：

- A社区的总户数占西北旺乡的总户数的比例为 $2000/40000 = 0.05$ 。
- B社区被抽中的概率为 $500/40000 = 0.0125$ 。

从整个西北旺乡来看：

- A社区家庭户在西北旺乡的被选概率就是 $0.05 * 0.01 = 0.0005$ 。
- B社区家庭户的被选概率是 $0.0125 * 0.04 = 0.0005$ 。

从社区家户来看：

- A社区家户的被选概率就是 $20/2000 = 0.01$ ；
- B社区家户的被选概率为 $20/500 = 0.04$ 。



概率抽样4：成比例抽样（规模度量）

概率抽样基本的条件是具有抽样大小规模的辅助变量，又叫规模度量。

规模度量如何选择：

- 主要是代表规模的，比如说社区的家庭户数。
- 规模量度的变量可以有多个，最常用的方法是依据研究变量相关程度来挑选。
- 选择规模度量的影响因素还有获取资料的难易程度、可靠程度等。
- 在两阶段/多阶段抽样中，每一阶段使用的规模度量一定要相同。

西北旺乡案例中：第一阶段是社区抽样，被选概率计算还是采用了家庭户数。第二阶段是家庭户抽样，备选概率的计算也采用了家庭户数。



概率抽样4：成比例抽样（特点）

成比例抽样PPS的特点：

- 第一，PPS抽样常常会考虑抽样面对的现实，一般是进行多阶段抽样，不是抽一回。
- 第二，有些信息，抽样时并不知道，常常要步步为营，充分利用已经知道的信息。
- 第三，每一个阶段的抽样概率不一定相等。
- 第四，总的原则是总体要素的被选概率一定要相等。



概率抽样5：分层抽样（实施步骤）

分层抽样(stratified sampling): 将抽样单位按某种特征或某种规则划分为不同的层，然后从不同的层中独立、随机地抽取样本。

实施步骤：

1. 把研究总体按照研究特征变量进行分层。
2. 在每一层采用合适的方法来抽样
 - 简单随机抽样或者等距抽样、整群抽样
 - 等比例或者不等比例的抽样，甚至pps抽样都行。
3. 把每个层的样本合起来加总，计算得到对总体进行推断的样本容量。



概率抽样5：分层抽样（应用情景）

决定是否采用分层抽样，需要：

- 对研究总体同质性程度有一定了解，知道总体的同质性、异质性如何。
- 了解了总体的异质性的程度是不是大到了必须分层的程度。总体在研究变量上的同质性越高，对分层的要求就越低。
- 分层抽样通常不会独立使用，通常用来构造子抽样框、子总体，它不是独立抽样的方法，也不是末端抽样的方法。
- 对研究变量了解越充分，采用合适的分层方式，就越有利于降低抽样误差。



概率抽样5：分层抽样（示例）

任务：一项研究拟讨论教育模式对高校学生能力的影响，研究者打算采用分层抽样方法抽取n名学生。

- 从学校到院系，简单起见可以先分文和理两个大类的院系。
- 从院系到班，可以采用任何简单抽样的方法。
- 从班抽到学生呢，就可以采用整群抽样的办法。
- 把文和理两类样本加起来，就是一所学校的样本。
- 如果文理之间学生的数量相差的太大，也可以考虑按学生数量的比例分配样本。



概率抽样5：分层抽样（分层依据）

分层依据是分层抽样中关键的环节：

- 分层依据的变量通常与研究目标有关，与研究变量有关系。
- 分层并不就是分等级，大多数情况下是分类别（提问）。
- 研究目的越复杂，分层变量越多，要区分的层数也就越多。
- 实践中一般希望尽可能地选取主要的分层变量，因为分层越多，看起来越精准。
- 在抽样实践中，有些分层明显，有一些分层则不太明显，可能实际上还携带着层变量的分层，称之为内隐分层 或者叫隐含分层。



概率抽样5：分层抽样（示例）

学生教育模式对学生能力的影响研究案例。

有的院系一个年级有多个班，如经济管理学院，有的学院只有一个班，如农学院。

如果有多个班的学院用平均能力对班进行排序，再抽取班级样本，则抽到的班样本不仅携带了院系信息，也携带了能力信息。

不仅按文理院系在分层，也在按照能力进行分层，只是按能力分层被隐含在了按文理院系分层之中。

大学里的院系，院系之间是平行的，不是层级关系。同一个院系的不同年级之间的分层，实际上是垂直的序列关系，但也叫分层。



概率抽样5：分层抽样（优缺点）

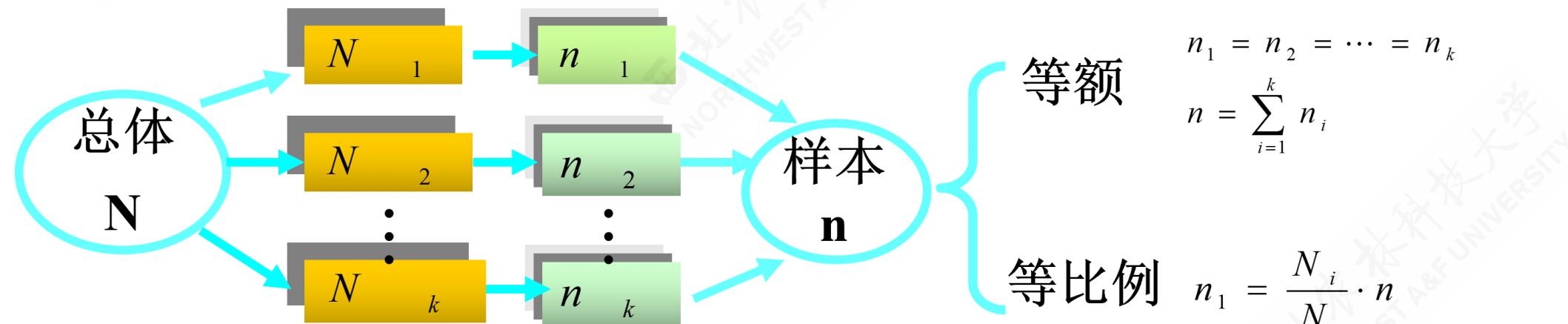
- 保证样本的结构与总体的结构比较相近，从而提高估计的精度
- 组织实施调查方便
- 既可以对总体参数进行估计，也可以对各层的目标量进行估计



概率抽样5：分层抽样（样本分配）

在各层次中样本量的分配有两种基本的方法，

- 等比例分层抽样：各层的样本量与要素的规模成比例。
- 不等比例分层抽样：依据经验或者既有的研究结论减少或增加特定群体的样本量比例。





概率抽样5：分层抽样（CFPS分层和样本分配）

CFPS的分层和分配示例：

- 第一个层：区分大省（5个）和小省（20个）。
- 第二个层1：大省（5个
 - 子层1：4个省（辽宁、甘肃、河南、广东），各省为一个抽样框，但遵循相同的抽样策略。
 - 子层2：1个省（上海），为一个独立的抽样框
- 第二个层2：小省（20个
 - 20个省级行政区，按照人均社会经济指标降序排列
 - 每一个省级行政区内，地级市按照人均GDP指标降序排列
 - 地级市内，分为区、县级市和县三个层。层内按人均GDP降序排列。



概率抽样5：分层抽样（CFPS分层和样本分配）

CFPS的分层和分配示例：

- 分配样本数。
 - 抽取样本县、区的初级抽样单位（PSU），分配各层样本数。
 - 实现既有发达的，也有不发达的，既有城市，也有县，人多的地区有样本，人少的地区也有样本。



概率抽样6：多阶段抽样（复习）

- 如果总体规模不大，要素在研究变量上的异质性分布具有随机性，则我们可以采用简单随机抽样、系统抽样。
- 如果不同群之间的异质性不大，群内的异质性对总体具有代表性，就可以采用整群抽样。
- 如果总体规模比较大，总体要素的异质性也比较大，且与不同特征群体的规模无关，研究变量在要素中呈现出某种非随机的分布，则需要采用分层抽样。
- 如果总体规模比较大，总体要素的异质性也比较大，且与不同特征群体的规模有关，那么至少要采用两个阶段的抽样，并且采用与群体规模成比例的概率抽样。
- 如果遇到搜集数据的范围非常大，要素的异质性分布也很复杂，那么采用上述任何一种方法都不足以解决抽样问题，而应该采用多阶段抽样。



概率抽样6：多阶段抽样（实施方案）

多阶段抽样(multi-stage sampling): 先抽取子群，但并不是调查群内的所有单位，而是再进行一步抽样，从选中的群中抽取出若干个单位进行调查。在多阶段抽样的每个阶段，采用的抽样方法也不一定相同。

实施方法：

- 先抽大单位（可用分层抽样）
- 再在大单位中抽小单位（可用成比例抽样）
- 小单位中再抽更小的单位（可用简单随机抽样）



概率抽样6：多阶段抽样（示例）

CGSS调查中基本要素是家庭中年满18岁或以上的个体。

假设研究者希望一次直接抽到个体，就需要编制一份有18岁或以上中国常住人口的抽样框。一个差不多有10亿人口的列表，这是不可能的

CGSS 2010年的抽样方案：

- 第一阶段，采用了分层抽样（覆盖全国区、县级市、县）。
- 第二阶段，抽到了村居，采用PPS抽样。
- 第三阶段，抽到了家户，采用了简单随机抽样。
- 末端抽样，抽到个体，采用了Kish表抽样。



概率抽样6：多阶段抽样（抽样单位）

- 初级抽样单位 (PSU)：初级阶段样本框的抽样单位。
 - CGSS的PSU就有160个区县
- 次级抽样单位 (SSU)：次级阶段样本框的抽样单位。
 - 对于上海，每个PSU只抽两个村居，也就是32乘2等于64，总的SSU的数量与其他大省一致。
- 末端抽样单位 (USU)：最末端阶段样本框的抽样单位。
 - CGSS的末端抽样单位是在样本户中抽到个人，是由调查员去抽取的



概率抽样6：多阶段抽样（优缺点）

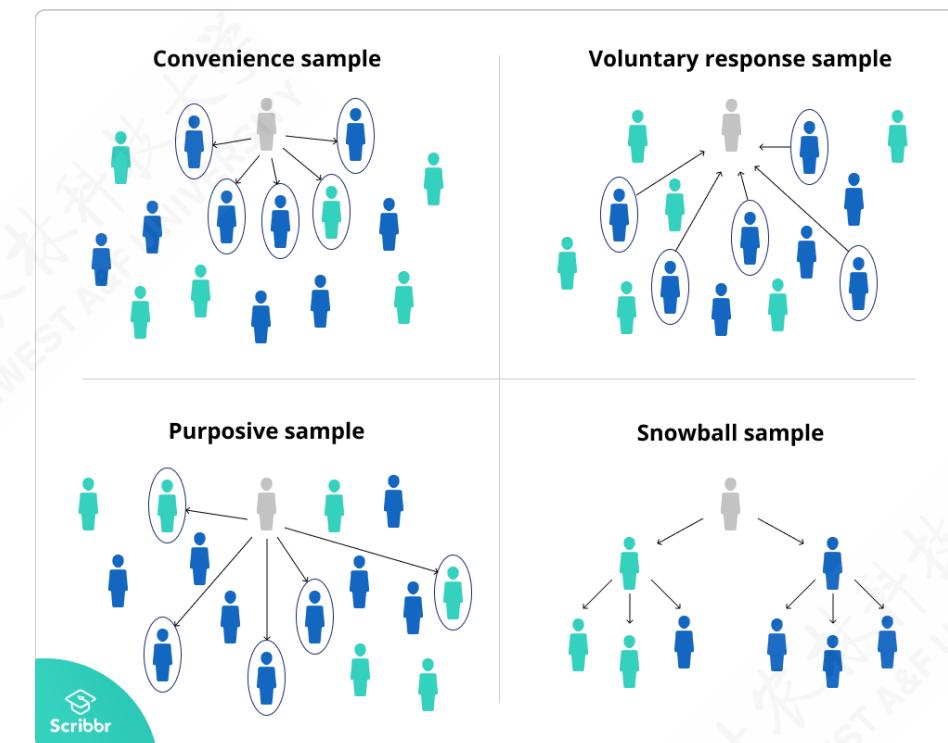
- 具有整群抽样优点，保证样本相对集中，节约调查费用
- 需要包含所有低阶段抽样单位的抽样框；同时由于实行了再抽样，使调查单位在更广泛的范围内展开
- 在大规模的抽样调查中，是经常被采用的方法



非概率抽样

非概率抽样：抽取样本时不是依据随机原则，而是根据研究目的对数据的要求，采用某种方式从总体中抽出部分单位对其实施调查。

1. 方便抽样 (convenience sample)
2. 判断抽样 (purposive sample)
3. 自愿样本 (voluntary response sample)
4. 滚雪球抽样 (snowball sample)
5. 配额抽样





非概率抽样I：方便抽样

方便抽样：调查过程中由调查员依据方便的原则，自行确定入抽样本的单位。

- 调查员在街头、公园、商店等公共场所进行拦截调查
- 厂家在出售产品柜台前对路过顾客进行的调查

优点：容易实施，调查的成本低

缺点：

- 样本单位的确定带有随意性
- 样本无法代表有明确定义的总体
- 调查结果不宜推断总体



非概率抽样2：判断抽样

判断抽样：研究人员根据经验、判断和对研究对象的了解，有目的选择一些单位作为样本。具体方式有：

- 重点抽样
- 典型抽样
- 代表抽样

缺点：

- 判断抽样是主观的，样本选择的好坏取决于调研者的判断、经验、专业程度和创造性
- 样本是人为确定的，没有依据随机的原则，调查结果不能用于推断总体

优点：抽样成本比较低，容易操作



非概率抽样3：判断抽样

自愿样本：被调查者自愿参加，成为样本中的一分子，向调查人员提供有关信息

- 参与报刊上和互联网上刊登的调查问卷活动
- 向某类节目拨打热线电话等

特点：

- 自愿样本与抽样的随机性无关
- 样本是有偏的
- 不能依据样本的信息推断总体



非概率抽样4：滚雪球抽样

滚雪球抽样：先选择一组调查单位，对其实施调查，再请他们提供另外一些属于研究总体的调查对象；调查人员根据所提供的线索，进行此后的调查。持续这一过程，就会形成滚雪球效应。

特点：

- 适合于对稀少群体和特定群体研究
- 容易找到那些属于特定群体的被调查者
- 调查的成本也比较低



非概率抽样5：配额抽样

配额抽样：先将总体中的所有单位按一定的标志(变量)分为若干类，然后在每个类中采用方便抽样或判断抽样的方式选取样本单位。

特点：

- 操作简单，可以保证总体中不同类别的单位都能包括在所抽的样本之中，使得样本的结构和总体的结构类似
- 抽取具体样本单位时，不是依据随机原则，属于非概率抽样



抽样方案（抽样方法选择）

为了保证抽样过程的严谨，也需要一个文本，用来指导抽样活动，这就是抽样方案。

在抽样方案中，抽样方法的选择是核心内容。一般情况下，抽样方法的取舍取决于三个基本因素：要素的同质性、总体的规模、变量的多少。

- 第一，如果总体规模很大，异质性很强，研究变量很多，通常会采用多阶段、分层的PPS抽样。
- 第二，如果总体规模很大，异质性也很强，研究变量很少，通常采用多阶段抽样，末端通常采用整群抽样或者配额抽样。
- 第三，如果总体规模也很大，同质性也很强，这个时候，变量的多少没有太大关系一般的情况下会采用非概率抽样，比如说末端采用就近抽样、判别抽样。
- 第四，如果总体规模很小，异质性很强，变量多少都没关系，通常会采用滚雪球、RDS抽样或者是知情人抽样。



抽样方案（文本内容I）

抽样方案一般需要说明，采用什么方法，采用哪些步骤，获得用于收集数据的样本。一份抽样方案在内容上至少要有以下的内容，

- 第一，总体。不仅要把总体界定清楚，还要明确地界定研究总体、框总体，或者叫抽样总体。如果采用多阶段抽样、分层抽样的，还要说明每一个阶段、每一层的抽样框。
- 第二，研究对象。包括调查对象或者研究对象，就是收集数据的对象、受访者。比如说CGSS调查对象是家庭中的个人，CFPS调查对象是家庭中的所有成员。
- 第三，样本量。尤其是末端抽样单位的数量要做明确的说明。
- 第四，抽样方法。如果采用复杂设计，例如多阶段混合抽样，那么每一个阶段的抽样方法都要做说明。



抽样方案（文本内容2）

- 第五，如果采用多阶段混合抽样，或者多阶段抽样，每一个阶段的抽样单位、抽样框、抽样方法、样本量的配置以及末端抽样的方法也需要写清楚。否则读者就无法知道每一个阶段的权重。
- 第六，如果不是采用大量熟悉的抽样框，自己在制备抽样框，还需要说明抽样框的制备方法。大型调查中的抽样框制备也是一项复杂的工程。
- 第七，还包括估计量的计算方法。比如说，权重到底怎么算，怎么配权重，如果是多阶段抽样，等概率又怎么保证。



抽样实施（工作安排）

即使有很好的抽样方案，如果不落到实处还是没有样本。抽样的实施一般来讲，根据抽样方案按照研究设计做就行了。听起来很简单，不过千万别大意。获得样本真的是一个非常艰难的过程。

- 第一，正确地理解方案，制定每一个环节的实施方案。抽样方案只是指引，指南，索引，在实践中操作中还需要实施方案。
- 第二，组织资源。比如人力，社会关系，设备，后勤保障等等。稍稍大一点的调查就得请人，请学生，请朋友，怎么计酬，怎么支付这就是后勤问题。后勤对社会调查与研究也非常重要。
- 第三，培训抽样人员，督导人员，后勤人员。把实施中可能遇到的问题讲透彻，把合作与分工讲透彻，让每一个人明确的知道自己到底要干什么。
- 第四，逐步实施。一般来讲，前三步工作做完以后就一步一步地实施，先从制作抽样框开始，再抽样，最后再做质量检验和误差估计。



抽样实施（经验建议）

在抽样的实施中，多问自己几个问题：

- 第一，总体到底有多大？到底多大范围的调查？
- 第二，研究总体在哪里？有哪些会影响到对调查对象的识别？
- 第三，有没有可用的抽样框？比如说，有没有可能让执行人提供一个抽样框？如果没有怎么制备抽样框？
- 第四，选择什么样的抽样方法可以减少误差？
- 第五，执行的难点到底是什么？怎么样去组织资源能够使得花最少的钱最有效地办事？
- 第六，最重要的一条经验就是多沟通。与相关各方尽可能就抽样设计，抽样实施的目标达成一致。



抽样误差（误差来源1）

抽样方法搜集数据，误差来源可能会出现在多个阶段：

- 第一，在发起阶段，由研究者带来的误差。理论假设不好，概念界定不清，对样本要求不明确。
- 第二，在设计阶段，由设计者带来的误差。比如测量工具选的不对，实施策略选的不对，抽样设计也有问题。
- 第三，在抽样阶段，由抽样员带来的误差。如果末端抽样框的界定不明确，抽样过程监管也不明确，就有可能产生随机性误差。
- 第四，在访问阶段，由访问员带来的误差。如果访员作弊、作假，轻易地接受拒访，诱导性提问，不规范的提问，也会造成随机误差、应答误差，甚至系统误差。



抽样误差（误差来源2）

- 第五，在访问阶段，由受访者带来的误差。
 - 如果受访者拒绝访问，或者没有能力作答，作假、作弊、随意作答、回忆误差，也会造成随机误差、应答误差。
- 第六，在数据清理阶段，由数据管理者带来的误差。
 - 如果数据的管理者编制的数据录入程序有问题，编码有问题，清理程序有问题，管理程序也有问题，也就有可能会前功尽弃，既可能产生随机误差，也可能产生系统误差。
- 第七，在数据分析阶段，由分析者带来的误差。
 - 如果分析者分析工具选择不当，模型建构不当，对数据有误读，也会造成研究误差。



抽样误差（误差类型）

涉及到调查活动的有三个阶段，也就是设计阶段、抽样阶段和访问阶段。这三个阶段涉及到的误差主要有：

- 第一，覆盖性误差，与抽样设计和抽样活动有关。
- 第二，抽样性误差，是抽样活动造成的误差。
- 第三，应答性误差，指访问阶段产生的误差。
- 第四，测量性误差，指测量、测量工具产生的误差。



抽样误差（ Δ 覆盖性误差）

覆盖性误差，主要指因抽样方制作不当带来的误差。它属于抽样设计和抽样活动有关一类误差。

- 如果抽样方与研究总体不一致，就会产生误差。
 - 假定CGSS使用电话号码作为抽样框，就会出现覆盖性误差。
 - 覆盖不足产生的误差：比如有些人没有电话，就会被抽样方忽略，太穷的、太富的都有可能没有电话，或者呢，有电话，却不在电话簿的列表中。
 - 覆盖过度产生的误差：比如很多人有多部电话，这些人就有可能被过度代表
- 任何抽样方法都不可避免地会带来误差。
 - 忽略样本特征而随意选择抽样方法，就会直接带来误差。
 - 即使让抽样方正确地反映了研究总体，抽样活动不可避免地也会带来误差。



抽样误差（B主要变量的抽样误差）

主要变量的抽样误差，由变量特征带来的。

- 每一个变量 都有自己的抽样误差
- 主要变量的抽样误差一般是指的均值的误差，用均值的标准误 $\sigma_{\bar{x}}$ 来代表误差。
- 主要变量的抽样误差也能用相对误差来表示，比如说均值的变异系数 $V_{\bar{x}}$ 。



抽样误差（应答性误差）

访问阶段的误差也会涉及到抽样误差，尤其是应答性误差。它属于访问阶段活动有关一类误差。

- 样本无应答：又叫单人无应答，是指如果受访者对整个访问无论是问卷，还是访谈，都不回答的情形。简言之，就是无法从样本得到任何应答，比如说受访人拒访，或者根本联系不上。
- 选项无应答：又叫访题无应答，指受访者接受了访问，可能对某些访题不提供应答。

看起来这样的误差属于纯粹的访问误差，实际上不一定，也可以被认为是抽样误差的一种，比如，某些访题涉及到稀有应答，在抽样设计中，就需要予以考虑。

2.4 抽样分布和抽样误差

离散和连续随机变量

总体和样本特征

抽样误差计算



离散随机变量：离散事件

六点骰子的样本空间 (sample space) 为: $\{1, 2, 3, 4, 5, 6\}$, 随机摇一次骰子结果可能是:

```
sample(1:6, 1)
```

```
[1] 6
```

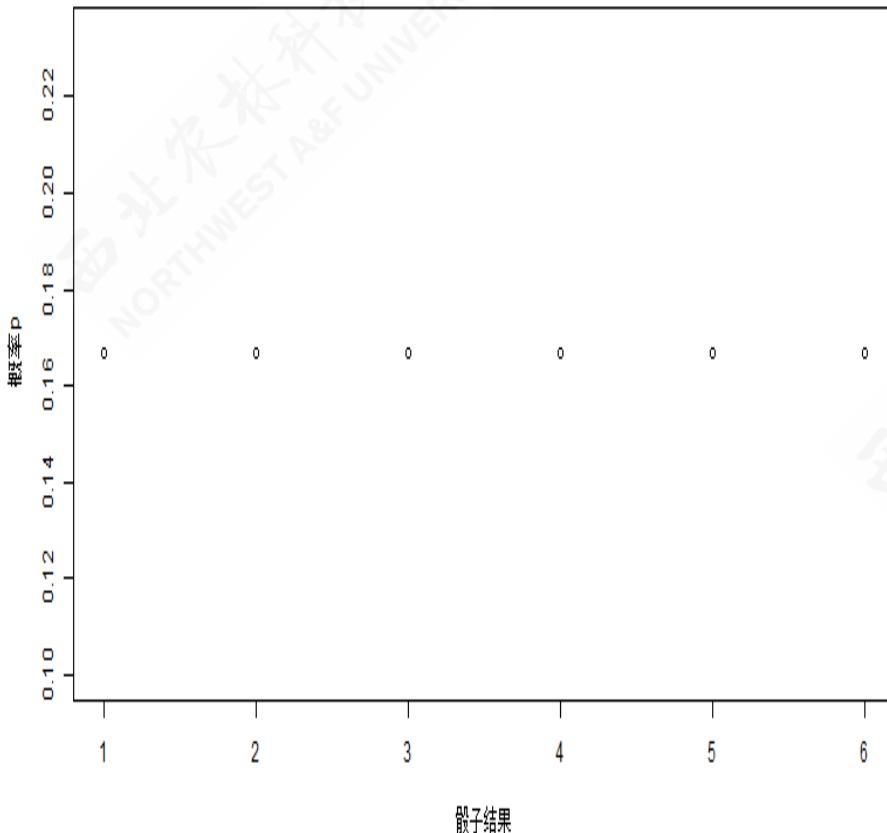
打六点骰子的PDG和CDG

骰子结果	1	2	3	4	5	6
概率	1/6	1/6	1/6	1/6	1/6	1/6
累积概率	1/6	2/6	3/6	4/6	5/6	1

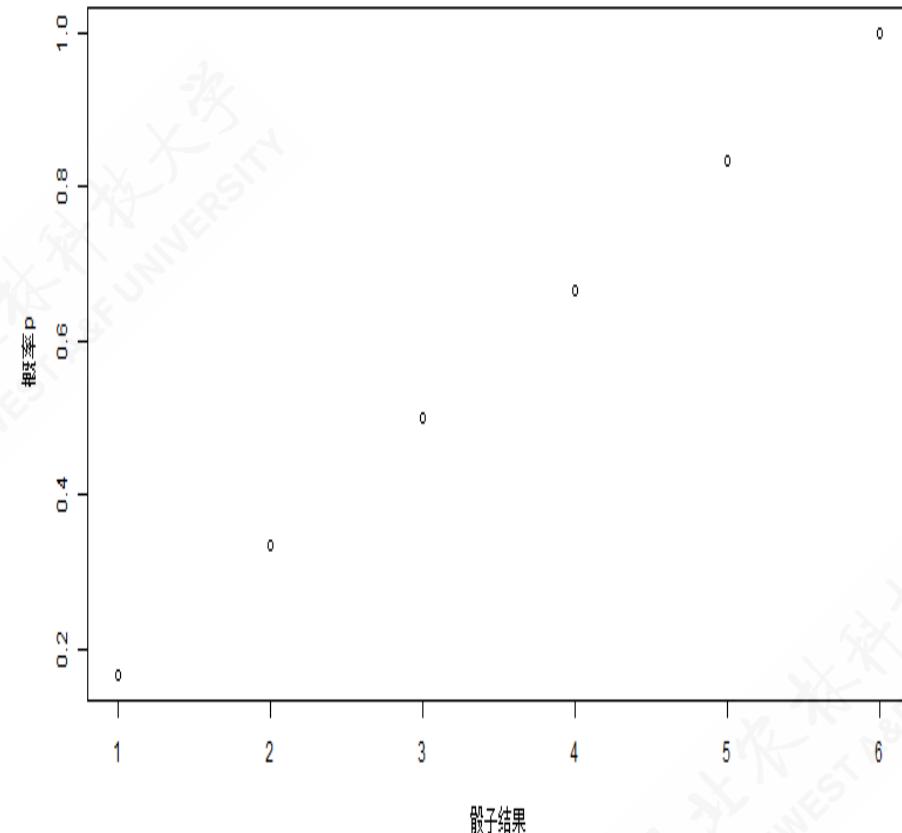


离散随机变量：概率分布

六点骰子的概率分布pd



六点骰子的累积概率分布





离散随机变量：伯努利事件（概率分布）

抛硬币事件 k 有两种可能结果： H （头像）和 T （花案）。我们随机抛一次硬币的结果可能是：

```
sample(c("H", "T"), 1)
```

```
[1] "T"
```

对于连续 n 次抛硬币，事件 k 服从伯努利 $k \sim B(n, p)$ 分布，其概率为：

$$f(k) = P(k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} = \frac{n!}{k!(n - k)!} \cdot p^k \cdot (1 - p)^{n-k}$$



离散随机变量：伯努利事件（概率分布）

例如，连续抛10次硬币且其中5次为头像朝上的伯努利概率记为
 $P(k = 5|n = 10, p = 0.5)$ ，具体计算R函数及其结果为：

```
[1] dbinom(x = 5, size = 10, prob = 0.5)
```

```
[1] 0.25
```

连续抛10次硬币（ $n = 10$ ）且其中5次为头像朝上（ $k = 5$ ）的伯努利事件（ $p = 0.5$ ）出现的概率为

24.61%。

例如，连续抛10次硬币且其中头像朝上次数在 4 ~ 7 次之间的伯努利概率记为

$P(4 \leq k \leq 7) = P(k \leq 7) - P(k \leq 3)$ ，具体计算R函数及其结果为：

```
[1] pbinom(size = 10, prob = 0.5)
```

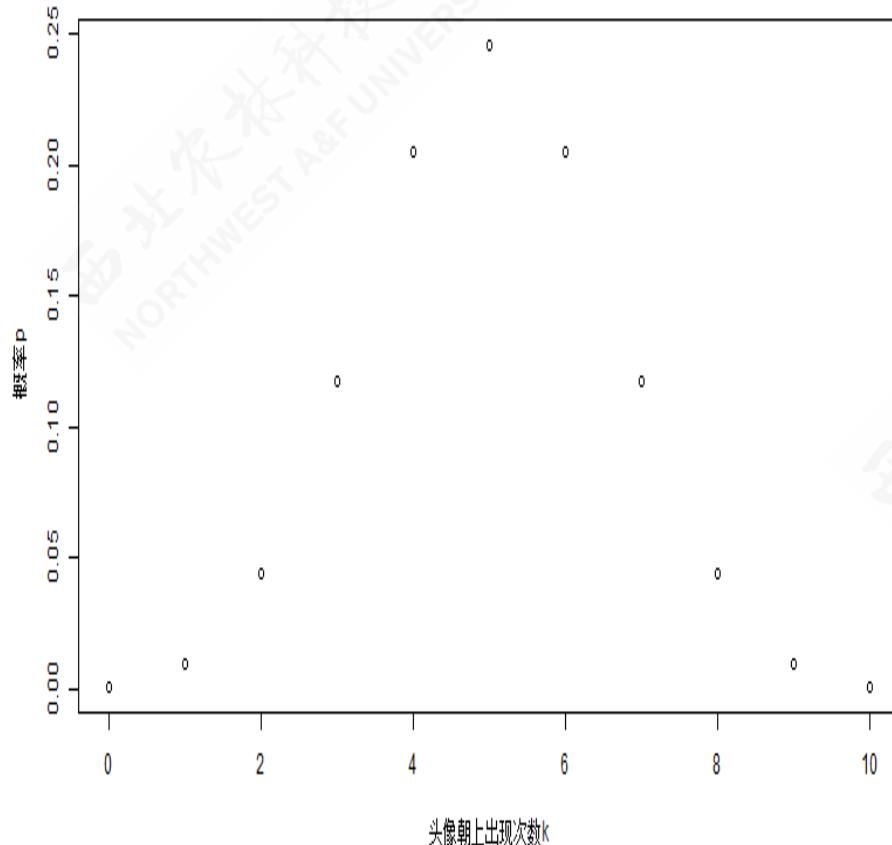
```
[1] 0.77
```

连续抛10次硬币（ $n = 10$ ）且其中头像朝上次数（ $4 \leq k \leq 7$ ）的伯努利

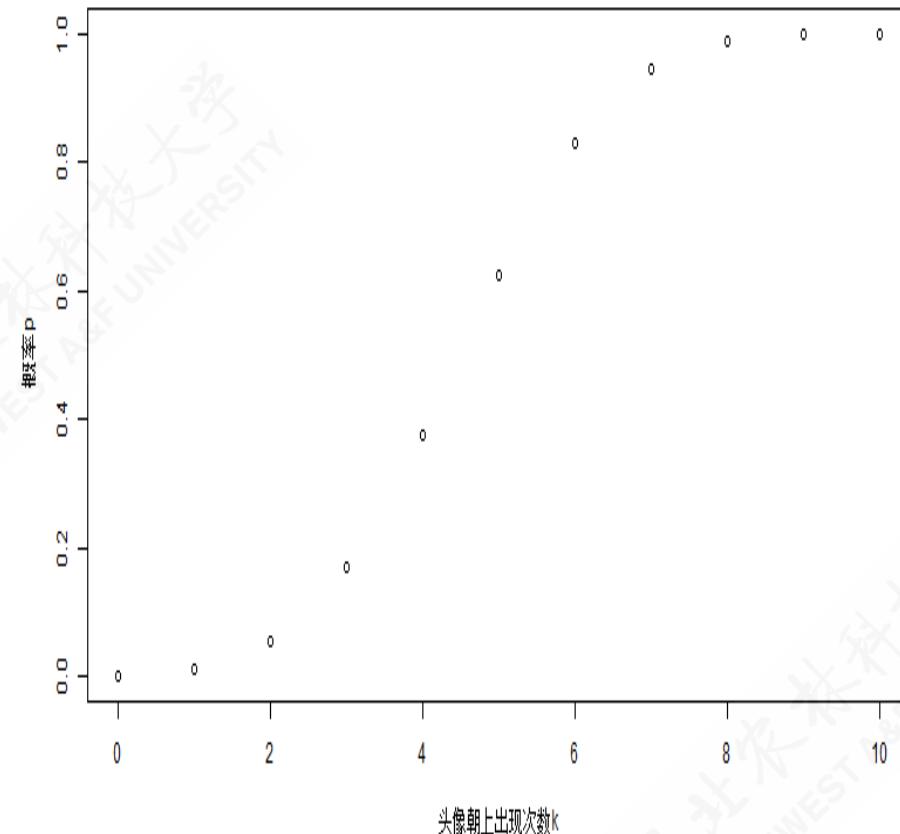


离散随机变量：伯努利事件（概率分布）

不同头像朝上出现次数的概率($n=10$)



头像朝上出现次数的累积概率($n=10$)





连续随机变量：概率、期望和方差

对于一个连续分布事件 X , 给定其概率密度函数 (PDF) 为 $f_X(x)$, 那么:

- 其累积概率密度函数 (CDF) : $P(a \leq X \leq b) = \int_a^b f_X(x)dx$
- 而且有完全概率密度为: $P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f_X(x)dx = 1$ 。
- 进一步, 其期望为: $E(X) = \mu_X = \int Xf_X(x)dx$ 。
- 其方差为: $\text{Var}(X) = \sigma_X^2 = \int (X - \mu_X)^2 f_X(x)dx$



连续随机变量：正态分布（PDF、CDF）

一个变量 X 若服从正态分布，则由两个参数来确定，一个是期望 μ ，另一个是方差 σ^2 ，并记为： $Y \sim \mathcal{N}(\mu, \sigma^2)$ 。正态分布的分布密度函数（PDF）的理论表达式为：

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp [-(X - \mu)^2 / (2\sigma^2)].$$

其中，标准正态分布属于一种特殊形式的正态分布，其期望为0，方差为1，一般记为： $X \sim Z(0, 1)$ ，其概率密度函数（PDF）一般记为 ϕ ，其累积概率密度函数（CDF）一般记为 Φ ，也即有：

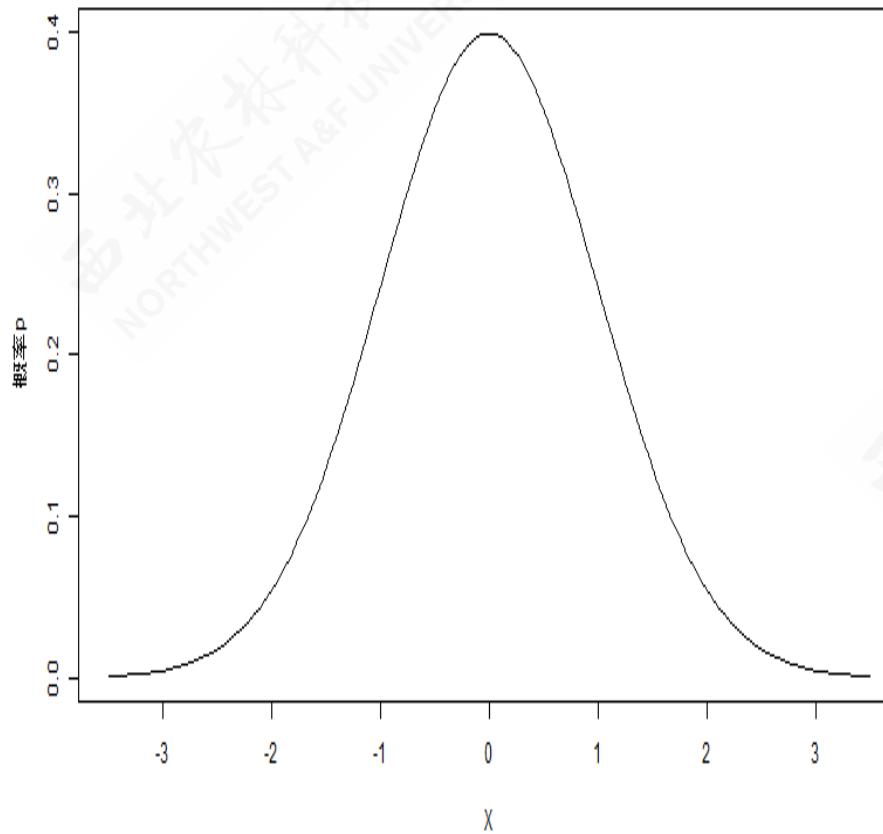
$$\phi(c) = \Phi'(c), \quad \Phi(c) = P(Z \leq c), \quad Z \sim \mathcal{N}(0, 1).$$

而且还有：若 $Y \sim \mathcal{N}(\mu, \sigma^2)$ ，则 $Y^* = \frac{Y-\mu}{\sigma} \sim Z(0, 1)$ 。

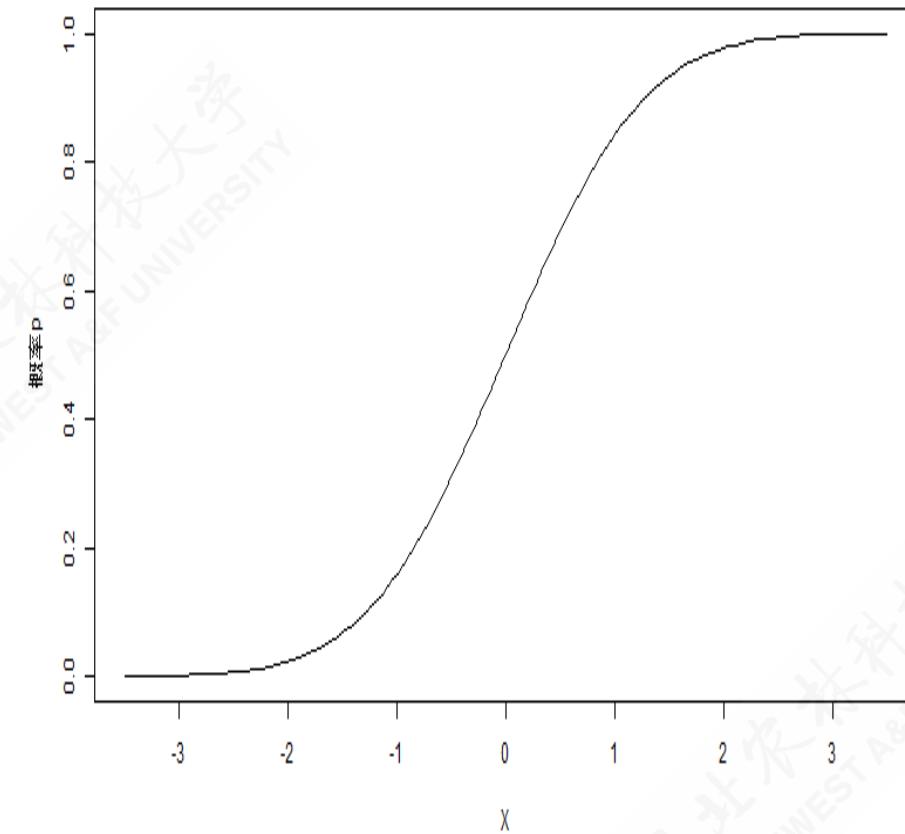


连续随机变量：正态分布（PDF、CDF）

标准正态分布的PDF



标准正态分布的CDF





连续随机变量：正态分布（二元联合正态）

WebGL is not supported
by your browser - visit
<https://get.webgl.org> for
more info



ρ (Coefficient of correlation between X and Y)

0



E(X)

1

Var(X)

1

E(Y)

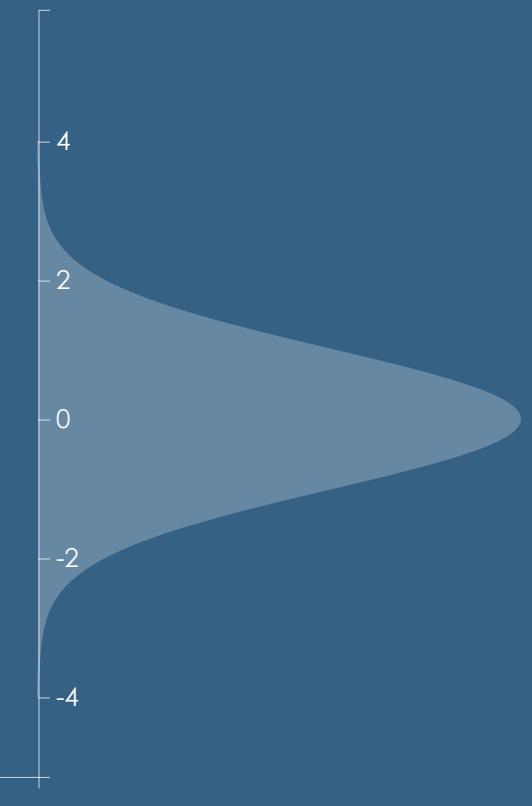
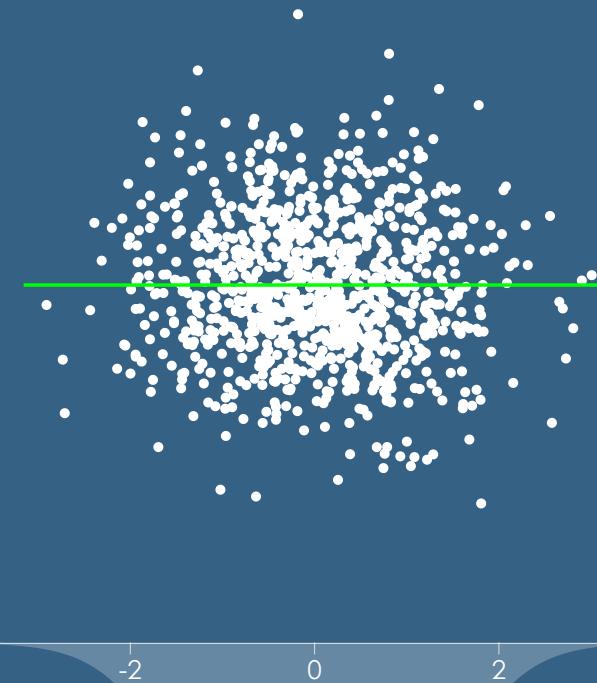
1

Var(Y)

1

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right]$$

$$E[Y|X] = 0 + 0 \cdot (X + 0)$$





连续随机变量：卡方分布 (PDF、CDF)

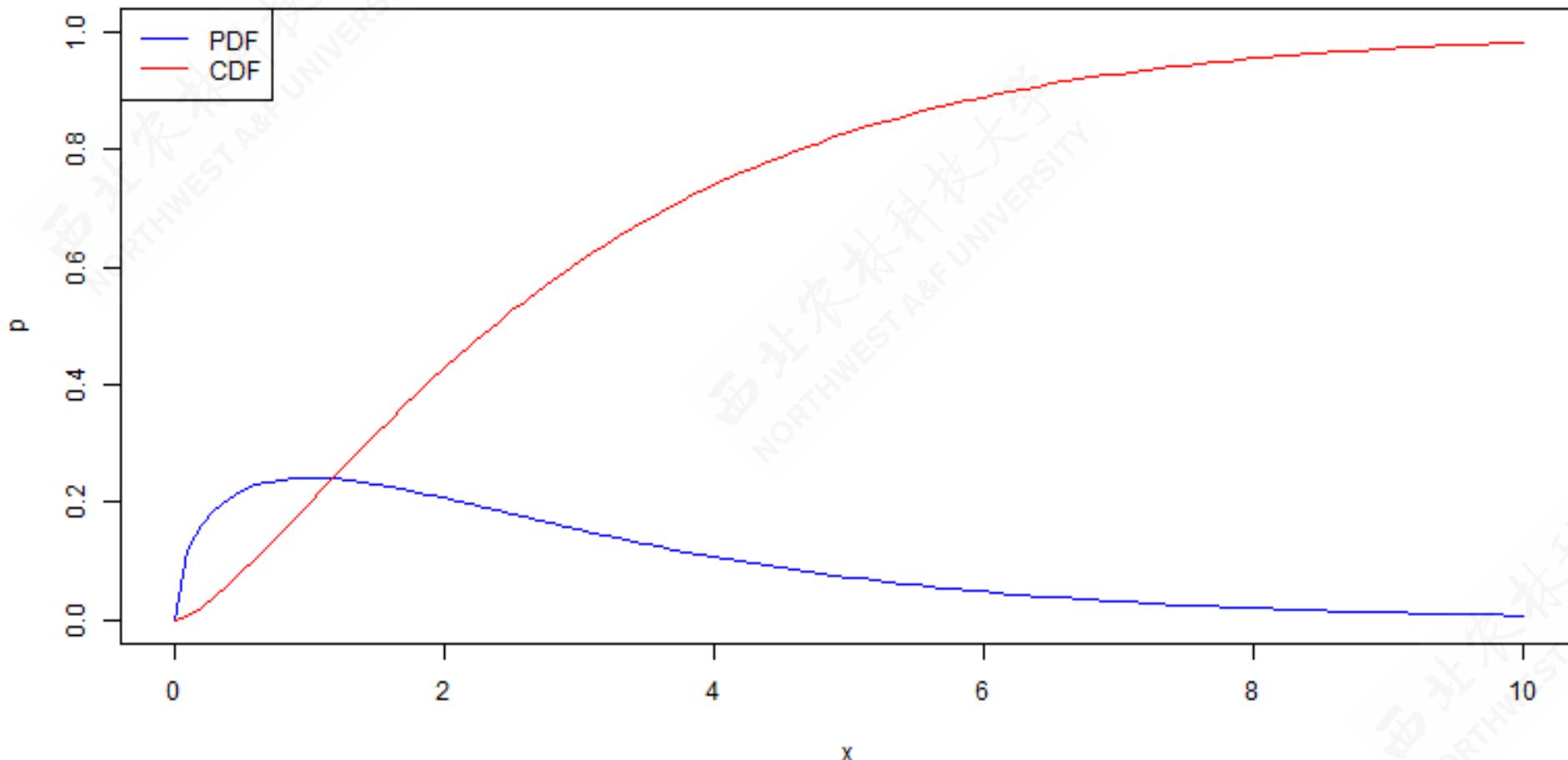
if, $Z_m \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$

then, $Z_1^2 + \cdots + Z_M^2 = \sum_{m=1}^M Z_m^2 \sim \chi^2(M)$



连续随机变量：卡方分布（PDF、CDF）

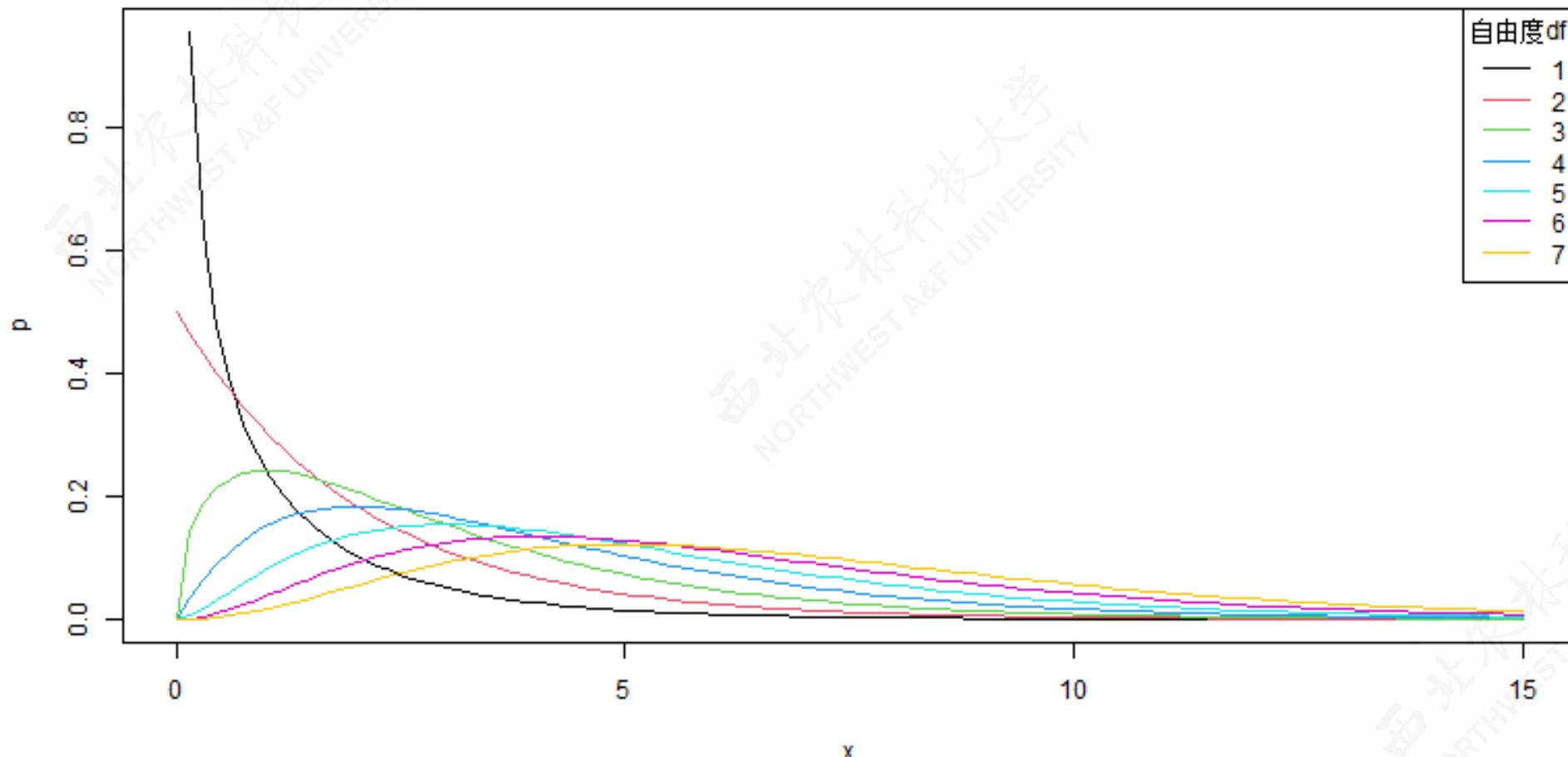
卡方分布的PDF和CDF ($M = 3$)





连续随机变量：卡方分布（PDF、CDF）

不同自由度df下卡方分布的概率密度函数





连续随机变量：t分布 (PDF、CDF)

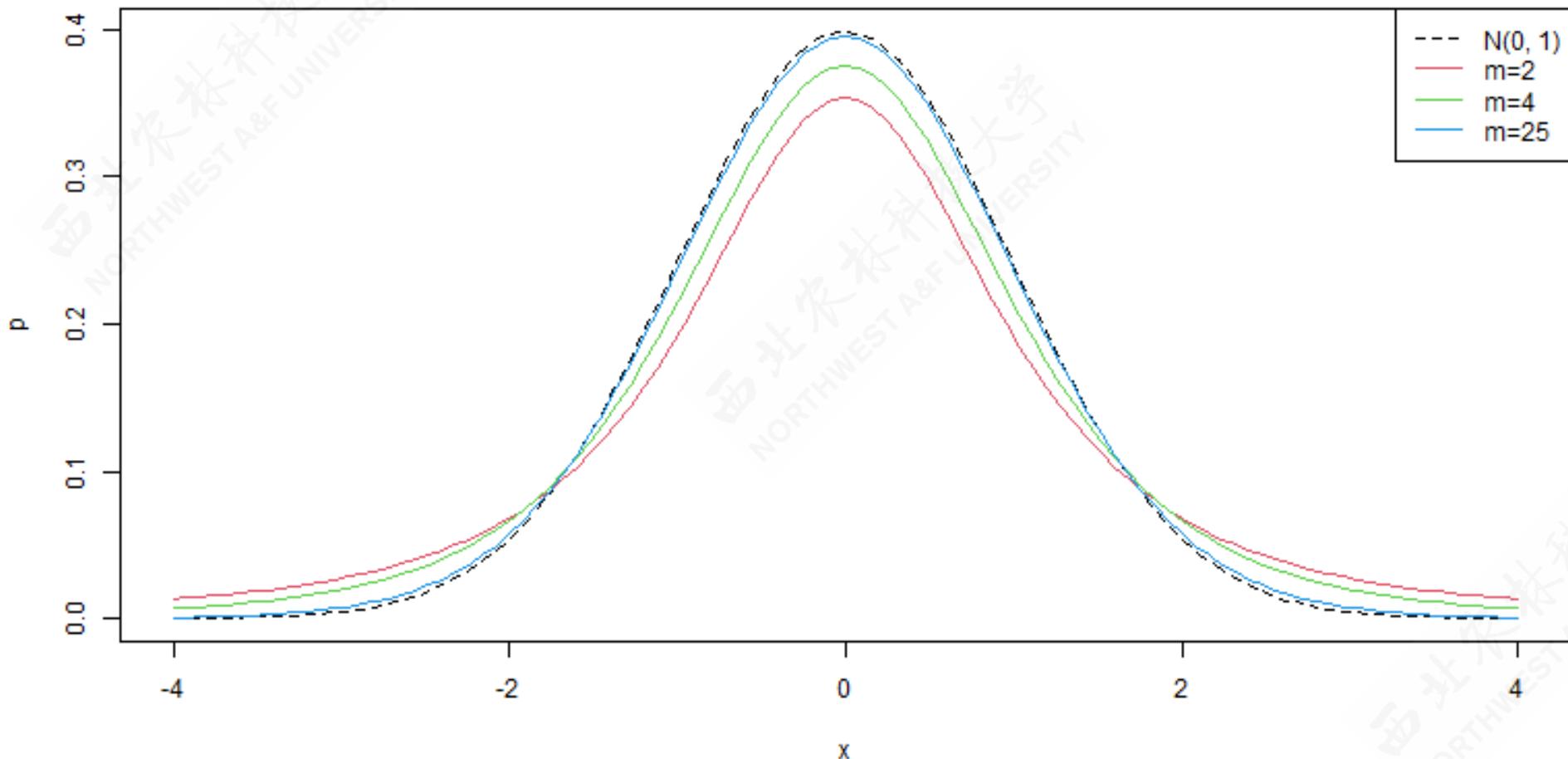
假定随机变量 $Z \sim \mathcal{N}(0, 1)$ 服从标准正态分布，随机变量 $W \sim \chi^2(m)$ 卡方分布，而且二者互相独立，那么可以构造出一个如下的新随机变量 T ，它将服从 t 分布：

$$T = \frac{Z}{\sqrt{W/m}} \sim t(m)$$



连续随机变量：t分布 (PDF、CDF)

不同自由度下t分布的概率密度





总体的特征：总体期望和总体方差

随机变量 Y 有 6 种可能取值 $\{1, 2, 3, 4, 5, 6\}$ ，那么每种可能取值的概率分别为：

事件 Y_i	1	2	3	4	5	6
概率 p	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

总体期望 μ_Y 和总体方差 σ_Y^2 ：

$$E(Y) \equiv \mu_Y = \sum_1^6 (Y_i \cdot p(Y_i)) = \frac{1}{6} \sum_1^6 Y_i = \frac{1}{6} \times 21 = 3.50$$

$$\begin{aligned} Var(Y) \equiv \sigma_Y^2 &= E(Y_i - E(Y))^2 = E(Y_i - \mu)^2 = \sum_1^6 ((Y_i - \mu)p(Y_i)) \\ &= \frac{1}{6} [(1 - 3.5)^2 + (2 - 3.5)^2 + \cdots + (6 - 3.5)^2] \\ &= 3.89 \end{aligned}$$



样本的特征：样本均值和样本方差

从上述总体中有放回地随机抽选8次，得到1份样本容量 $n = 8$ 的如下样本数据：

```
[1] 5 1 5 6 3 5 5 3
```

样本均值 \bar{y} 和样本方差 s_y^2 分别表达并计算为：

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{8}[5 + 1 + \dots + 3] = 4.125$$

$$\begin{aligned}s_y^2 &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1} \\&= \frac{1}{8 - 1} \times [(5 - 4.125)^2 + (1 - 4.125)^2 + \dots + (3 - 4.125)^2] \\&= 2.6964\end{aligned}$$



样本的特征：样本均值和样本方差

因此，我们可以不断从前述总体 $Y \in \{1, 2, 3, 4, 5, 6\}$ 中进行有放回的随机抽样。下表展示了10份随机样本 y ，每份样本的容量都相同 $n = 8$ 。每份样本的均值 \bar{y} 见列 `bar_y`，样本方差 s_y^2 见列 `s2_y`。

sample	y	bar_y	s2_y
1	c(6, 6, 6, 2, 1, 3, 6, 1)	3.88	5.5536
2	c(6, 3, 1, 5, 1, 3, 3, 6)	3.50	4.0000
3	c(1, 4, 5, 1, 4, 4, 3, 4)	3.25	2.2143
4	c(4, 3, 2, 4, 5, 3, 1, 2)	3.00	1.7143
5	c(2, 2, 5, 3, 5, 3, 5, 4)	3.63	1.6964
6	c(3, 5, 5, 4, 6, 1, 1, 1)	3.25	4.2143

Showing 1 to 6 of 11 entries

Previous 1 2 Next



总体和样本特征的关系

根据中心极限定理和大数定理，我们可以推导得到总体与样本特征的如下关系：

$$E(\bar{y}) = \mu_Y \quad (\text{eq.1})$$

$$Var(\bar{y}) = \frac{\sigma_Y^2}{n} \quad (\text{eq.2})$$

$$E(Var(\bar{y})) = \widehat{Var}(\bar{y}) \equiv \frac{s^2}{n} \quad (\text{eq.3})$$

其中， $s^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{n-1}$ 表示随机样本的样本标准差。以上方程蕴含着如下结论：

- 方程1表明：随机变量 \bar{y} 的期望是随机变量 Y 的期望的无偏估计量 (unbiased estimator)。
- 方程2表明：随机变量 \bar{y} 的方差与随机变量 Y 的方差存在以上关系。
- 方程3表明：随机变量 \bar{y} 的方差的无偏估计量可以通过样本数据计算得到，其结果为 $\widehat{Var}(\bar{y}) \equiv \frac{s^2}{n}$ 。



抽样分布：骰子游戏

随机样本（random sampling）是从总体中随机抽取个体的集合。

六点骰子的可能结果为 $\{1, 2, 3, 4, 5, 6\}$ ，如果随机投掷2次，可以得到2次结果的数值加总：

```
set.seed(520)
toll_2 <- sample(1:6, 2, replace = T)
toll_2
```

```
[1] 3 6
```

```
sum(toll_2)
```

```
[1] 9
```



抽样分布：骰子游戏

随机投掷2次的所有可能结果共有
 $6^2 = 36$ 种可能：

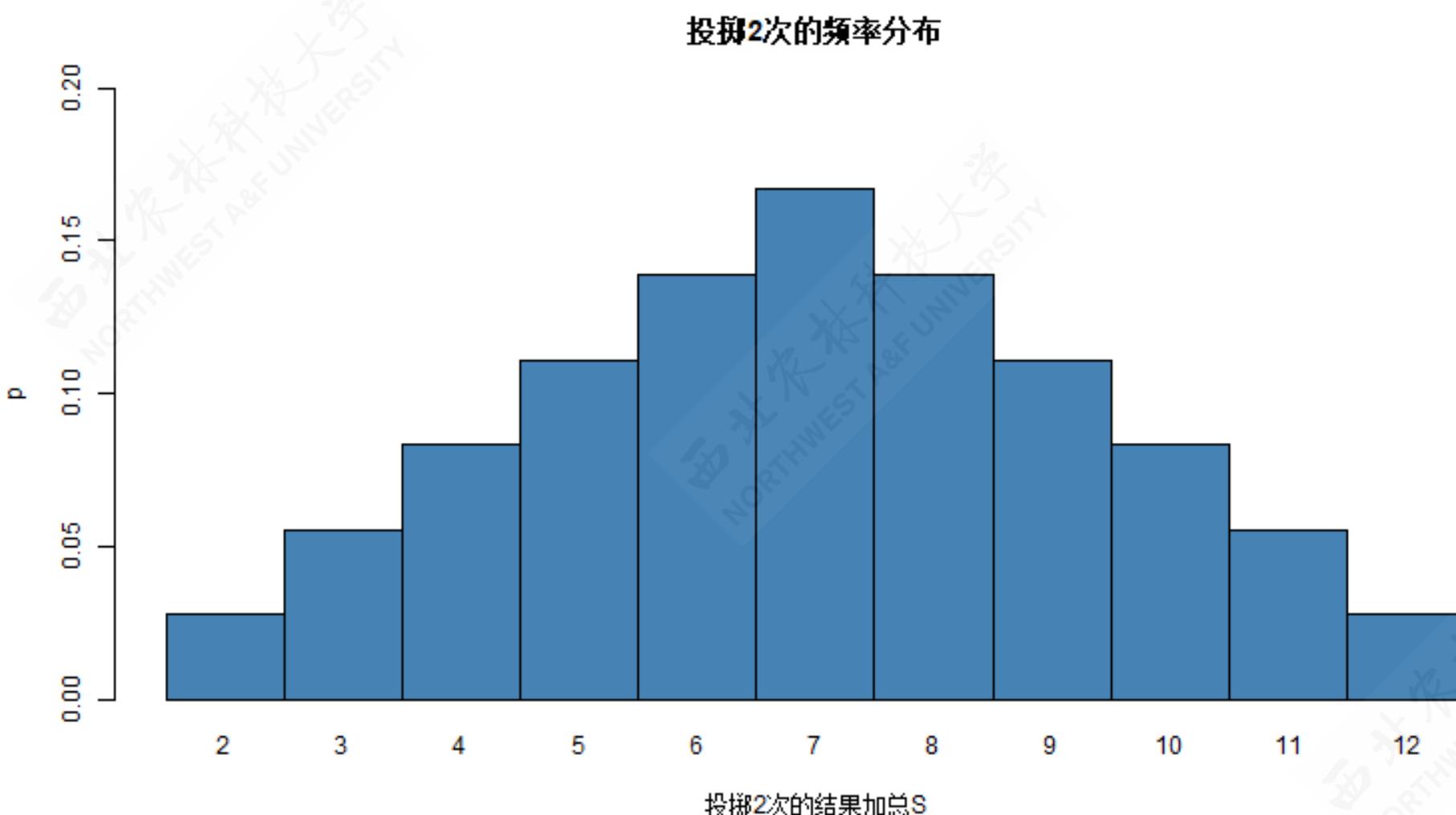
(1, 1)(1, 2)(1, 3)(1, 4)(1, 5)(1, 6)
(2, 1)(2, 2)(2, 3)(2, 4)(2, 5)(2, 6)
(3, 1)(3, 2)(3, 3)(3, 4)(3, 5)(3, 6)
(4, 1)(4, 2)(4, 3)(4, 4)(4, 5)(4, 6)
(5, 1)(5, 2)(5, 3)(5, 4)(5, 5)(5, 6)
(6, 1)(6, 2)(6, 3)(6, 4)(6, 5)(6, 6)

以上的全部组合，共有11种加总结果
 $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ ，
每种加总结果的概率分别是：

$$P(S) = \begin{cases} 1/36, & S = 2 \\ 2/36, & S = 3 \\ 3/36, & S = 4 \\ 4/36, & S = 5 \\ 5/36, & S = 6 \\ 6/36, & S = 7 \\ 5/36, & S = 8 \\ 4/36, & S = 9 \\ 3/36, & S = 10 \\ 2/36, & S = 11 \\ 1/36, & S = 12 \end{cases}$$



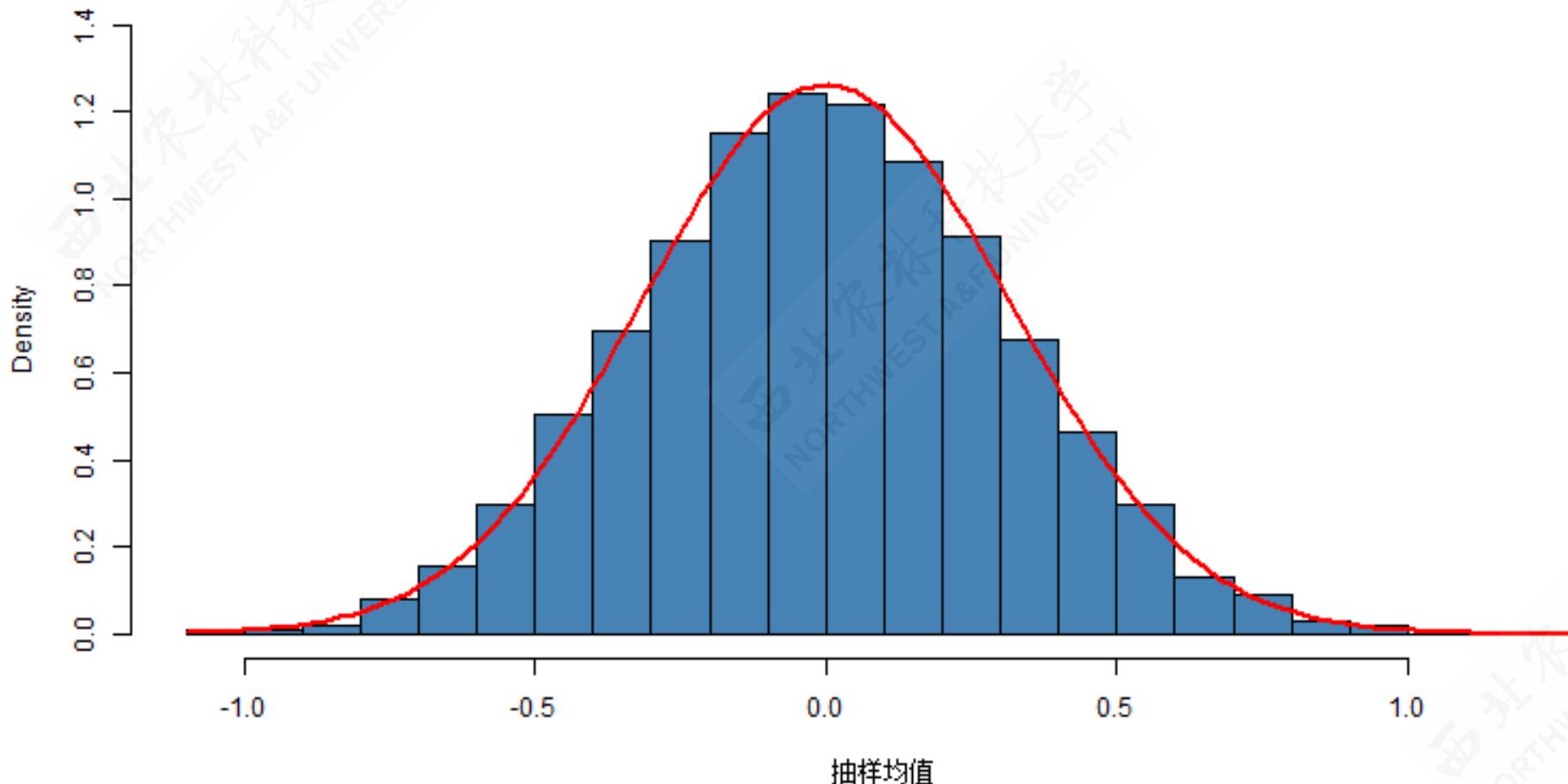
抽样分布：骰子游戏





抽样分布：骰子游戏

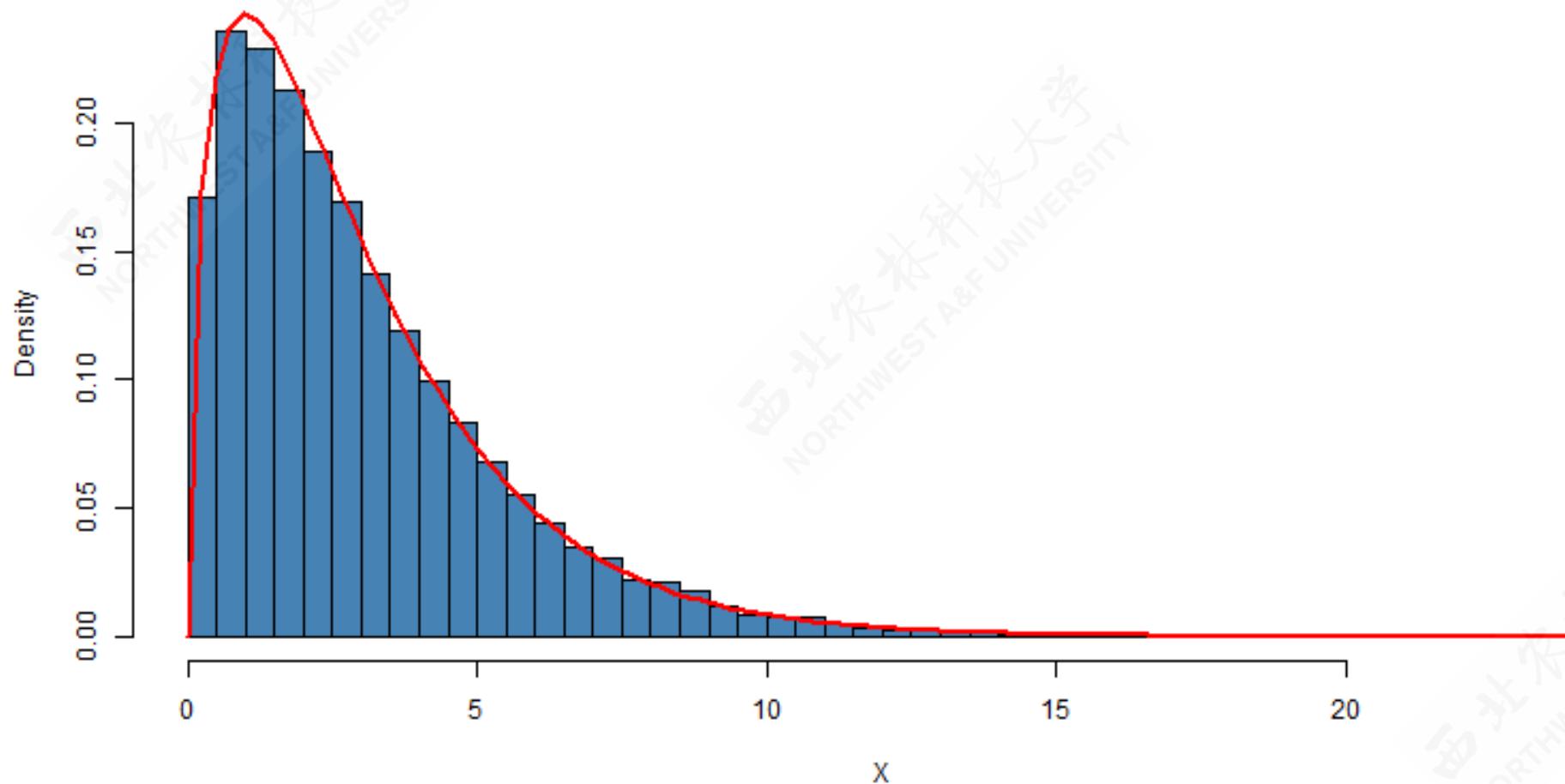
10000份随机样本（每份样本n=10）均值的直方图





抽样分布：骰子游戏

10000份随机样本（每份样本n=3）平方和的直方图





抽样误差：均值和方差

假定随机样本 y_1, \dots, y_n 是独立随机抽取自正态分布总体 $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, 那么前述随机样本数据的均值 \bar{y} 将服从如下正态分布:

$$\bar{y} \sim \mathcal{N}(\mu_Y, \sigma_Y^2/n) \quad (2.4)$$

其中:

$$E(\bar{y}) \equiv \mu_{\bar{y}} = E\left(\frac{1}{n} \sum_{i=1}^n y_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n y_i\right) = \frac{1}{n} \sum_{i=1}^n E(y_i) = \frac{1}{n} \cdot n \cdot \mu_Y = \mu_Y$$

$$\begin{aligned} \text{Var}(\bar{y}) \equiv \sigma_{\bar{y}}^2 &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n y_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(y_i) + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \text{cov}(y_i, y_j) = \frac{\sigma_Y^2}{n} \end{aligned}$$



抽样误差：中心极限定理

然而，实际中我们往往并不知道总体方差 σ_Y^2 。此时，上述方差公式是不能够计算的。

有限总体中心极限定理 (The finite population Central Limit Theorem) 对于随机变量 \bar{y} 的意义在于：我们可以用样本方差 s_y^2 来近似替代总体方差 σ_Y^2 。也即：

$$\text{Var}(\bar{y}) \equiv \sigma_{\bar{y}}^2 = \frac{\sigma_Y^2}{n}$$

$$\widehat{\text{Var}}(\bar{y}) \equiv \hat{\sigma}_{\bar{y}} = \frac{s_y^2}{n}$$



抽样误差：中心极限定理

如果样本容量 n 很小，随机变量 \bar{y} 的可能分布会是多种多样的。有限总体中心极限定理表明，随着样本容量 n 的不断增大，随机变量 \bar{y} 的分布会越来越稳定，并趋向于正态分布（normal distribution），从而有：

$$\begin{aligned}\bar{y} &\sim \mathcal{N}(\mu_{\bar{y}}, \sigma_{\bar{y}}^2) \\ \frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} &= \frac{\bar{y} - \mu_{\bar{y}}}{\sqrt{Var(\bar{y})}} \sim \mathcal{Z}(0, 1)\end{aligned}$$



抽样误差：置信区间

如果随机变量 \bar{y} 的总体方差 $Var(\bar{y})$ 未知，则无法使用上述正态分布 \mathcal{N} 或者标准正态 Z 分布，进行有关置信区间的样本推断。幸运的是，我们可以构造出如下服从 t 分布的随机变量：

$$\frac{\bar{y} - \mu_{\bar{y}}}{\hat{\sigma}_{\bar{y}}} = \frac{\bar{y} - \mu_{\bar{y}}}{\sqrt{\widehat{Var}(\bar{y})}} = \frac{\bar{y} - \mu_{\bar{y}}}{s_y / \sqrt{n}} \sim t(n-1)$$

因此可以进一步得到参数 $\mu_{\bar{y}}$ 的 $1 - \alpha$ 置信区间：

$$\bar{y} - t_{1-\alpha/2} \sqrt{\widehat{Var}(\bar{y})} \leq \mu_{\bar{y}} \leq \bar{y} + t_{1-\alpha/2} \sqrt{\widehat{Var}(\bar{y})}$$

对于有放回的简单随机抽样，参数 $\mu_{\bar{y}}$ 的 $1 - \alpha$ 置信区间具体为：

$$\bar{y} - t_{\alpha/2} \sqrt{\left(\frac{N-n}{N} \right) \left(\frac{s^2}{n} \right)} \leq \mu_{\bar{y}} \leq \bar{y} + t_{\alpha/2} \sqrt{\left(\frac{N-n}{N} \right) \left(\frac{s^2}{n} \right)}$$



抽样误差：简单随机抽样

对于无放回的简单随机抽样方案，采用无偏估计法（unbiased estimator）下的均值 \bar{y}_{st} 和方差 $\widehat{\text{var}}(\bar{y}_{st})$ 分别为：

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}$$
$$\widehat{\text{Var}}(\hat{\mu}) = \frac{N-n}{N} \cdot \frac{s_y^2}{n}$$

上述方差公式中， $s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ 。而 $\frac{N-n}{N}$ 又被称为有限总体校正比值 (finite population correction)：

- 如果采用有放回的简单随机抽样，则上述方差公式需要去掉有限总体校正比值。
- 如果采用无放回的简单随机抽样，但是n相对于N非常小，则上述方差公式中有限总体校正比值会接近于1，因此也可以忽略。



(示例) 田野甲虫数量案例

案例说明：为估计出一片农地中甲虫的总数。研究人员将农地细分为100个大小相等的区块。

研究者决定采用简单随机抽样方案，随机抽选了其中的8个区块（编号见列field），并分别统计出其中的甲虫数量（见列beetles）。最终抽样统计表见右：

简单随机抽样结果

	field	beetles
	41	234
	42	256
	18	128
	13	245
	80	211
	68	240
	25	202
	100	267



(示例) 简单随机抽样下估计期望和方差：计算结果

根据案例，容易计算得到：全部区块数量 $N = 100$ ；抽选区块数量 $n = 8$ 。抽选区块下甲虫数量的样本方差为 $s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = 1932.70$ 。

因此根据简单随机抽样无偏估计法下的计算公式，分别可以计算得到估计的均值 $\hat{\mu}$ 和方差 $\widehat{\text{Var}}(\hat{\mu})$ 分别为：

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y} = 222.88$$

$$\begin{aligned}\widehat{\text{Var}}(\hat{\mu}) &= \frac{N-n}{N} \cdot \frac{s_y^2}{n} \\ &= \frac{100-8}{100} \cdot \frac{1932.70}{8} = 222.2601\end{aligned}$$



(示例) 简单随机抽样下估计期望和方差：计算结果

根据上述计算，给定 $\alpha = 0.05$ 下，平均每个区块甲虫数 $\mu_{\bar{y}}$ 的置信区间计算结果为：

$$\begin{aligned}\hat{\mu} - t_{1-\alpha/2} \sqrt{\widehat{Var}(\hat{\mu})} &\leq \mu_{\bar{y}} \leq \hat{\mu} + t_{1-\alpha/2} \sqrt{\widehat{Var}(\hat{\mu})} \\ 222.88 - 2.36 \times \sqrt{222.2601} &\leq \mu_{\bar{y}} \leq 222.88 + 2.36 \times \sqrt{222.2601} \\ 187.62 &\leq \mu_{\bar{y}} \leq 258.13\end{aligned}$$

思考提问：全部地块的甲虫数量和置信区间是多少？

说明：此时，t查表值为 $t_{1-\alpha/2}(n-1) = t_{0.975}(8-1) = 2.36$ 。



必要样本数

不管采用哪种抽样方法，在哪一层抽样，在哪个阶段抽样，到底要抽多少样本合适啊？

假定 $\hat{\sigma}$ 是参数 σ 的无偏、正态估计量。则有

$$\frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \sim \mathcal{Z}(0, 1)$$

$$P\left(\frac{|\hat{\theta} - \theta|}{\sqrt{\text{Var}(\hat{\theta})}} > \mathcal{Z}_{1-\alpha/2}\right) = \alpha$$

$$P\left(|\hat{\theta} - \theta| > \mathcal{Z}_{1-\alpha/2} \cdot \sqrt{\text{Var}(\hat{\theta})}\right) = \alpha$$



必要样本数

令 $d = |\bar{y} - \mu|$ 为抽样极限误差，则简单随机抽样（不放回）方案下必要样本数的计算公式为：

$$P\left(|\bar{y} - \mu_{\bar{y}}| > Z_{1-\alpha/2} \cdot \sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}}\right) = \alpha$$
$$Z_{1-\alpha/2} \sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}} \equiv d$$
$$n = \frac{1}{\frac{d^2}{Z_{1-\alpha/2}^2 \cdot \sigma^2} + \frac{1}{N}}$$



(示例) 简单随机抽样方案下必要样本数的计算

在前述甲虫数量案例中, 给定 $\alpha = 0.05$ 下, 且抽样极限误差不超过 1000 只, 请计算简单随机抽样方案下的必要样本数?

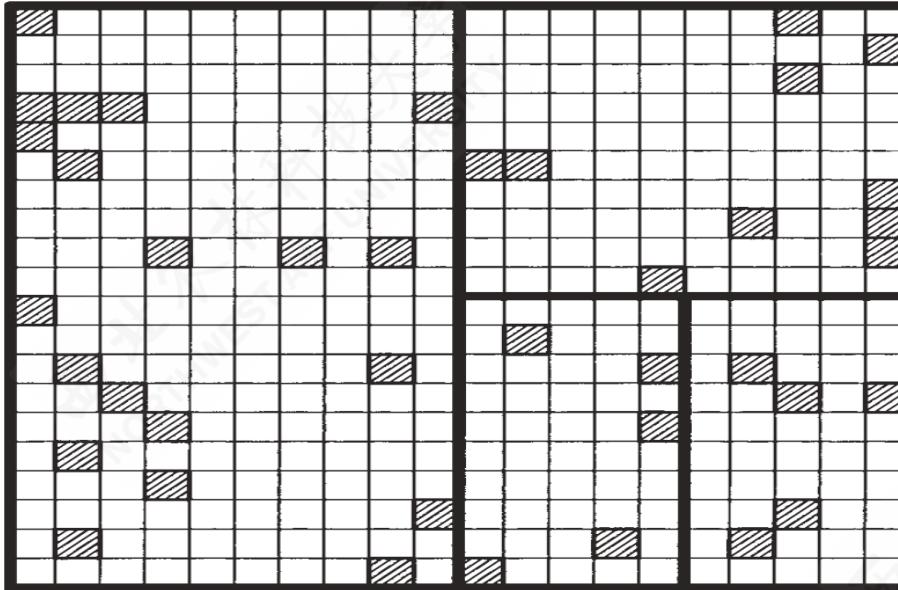
解答: 根据案例已知 $N = 100$, 抽样极限误差为 $d = 1000$, 给定 $\alpha = 0.05$ 下 $Z_{1-\alpha/2} = Z_{0.975} = 1.96$ 。

因为我们不知道总体方差 σ_y^2 , 但是可以使用样本方差 $s_y^2 = 1932.70$ 进行替代

$$\begin{aligned} n &= \frac{1}{\frac{d^2}{N^2 \cdot Z_{1-\alpha/2}^2 \cdot \sigma^2} + \frac{1}{N}} = \frac{1}{\frac{d^2}{N^2 \cdot Z_{1-\alpha/2}^2 \cdot s_y^2} + \frac{1}{N}} \\ &= \frac{1}{\frac{(1000)^2}{(100)^2 \cdot (1.96)^2 \cdot 1932.70} + \frac{1}{100}} = 42.61 \doteq 43 \end{aligned}$$



抽样误差：分层抽样



- 分层数量: $L = 4$; 各个分层的单位数:

$N_1 = 200, N_2 = 100, N_3 = N_4 = 50$;
全体单位总数

$$N = \sum_{h=1}^L N_h = 200 + 100 + 50 + 50 = 400$$

◦

- 各个分层的抽样单位数:

$n_1 = 20, n_2 = 10, n_3 = n_4 = 5$; 全部抽样总数

$$n = \sum_{h=1}^h n_L = 20 + 10 + 5 + 5 = 40$$

◦



抽样误差：分层抽样

分层抽样的均值 $\hat{\mu}_{st}$ 和方差 $\widehat{Var}(\hat{\mu}_{st})$ 分别为：

$$\hat{\mu}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h$$
$$\widehat{Var}(\hat{\mu}_{st}) = \sum_{h=1}^L \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h}$$

其中：

- L 分层数量； N_h 表示第 h 个分层的所有单位数，其中 $h \in \{1, 2, \dots, L\}$ ；
 $N = N_1 + N_2 + \dots + N_L$ 为所有单位数。 n_h 表示第 h 个分层的抽样数；
 $n = n_1 + n_2 + \dots + n_L$ 为所有抽样单位总数。
- 各个分层的样本方差为： $s_h^2 = \frac{1}{n_h-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$ ； \bar{y}_h 表示各个分层的样本均值。



(示例)家庭观看电视时长案例

案例说明：一家广告公司为了有针对性地在某个县投放电视广告，公司决定进行抽样调查，以估计该县家庭每周观看电视的平均小时数。该县有两个镇区A和镇区B，还有农村区域C。A区建在一家工厂周围，大多数家庭都有带学龄儿童的工厂工人。B区主要是退休人员，C区主要是农民。A区有155户，B区有62户，C区有93户。公司决定从A区抽选20户，B区抽选8户，C区抽选12户。具体抽样结果如下：

Stratification sampling		n	h	N	h
Town A	35, 43, 36, 39, 28, 28, 29, 25, 38, 27, 26, 32, 29, 40, 35, 41, 37, 31, 45, 34	20	155		
Town B	27, 15, 4, 41, 49, 25, 10, 30		8	62	
Rural Area C	8, 14, 12, 15, 30, 32, 21, 20, 34, 7, 11, 24		12	93	



(示例) 分层抽样下的抽样误差：计算结果

各分层的计算表如下：

Stratification	n_h	N_h	mean_h	sd_h	var_p1	var_p2	var_p3	var_
Town A	20	155	33.90	5.95	0.25	0.87	1.77	0.
Town B	8	62	25.13	15.25	0.04	0.87	29.05	1.
Rural Area C	12	93	19.00	9.36	0.09	0.87	7.30	0.

根据该分层抽样，估计的总体均值（该县住户平均收看电视时间）结果为：

$$\begin{aligned}\hat{\mu}_{st} &= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3) \\ &= \frac{1}{155 + 62 + 93}[(155 \times 33.9) + (62 \times 25.12) + (93 \times 19.0)] \\ &= 27.7\end{aligned}$$



(示例) 分层抽样下的抽样误差：计算结果

各分层的计算表如下：

Stratification	n_h	N_h	mean_h	sd_h	var_p1	var_p2	var_p3	var_
Town A	20	155	33.90	5.95	0.25	0.87	1.77	0.
Town B	8	62	25.13	15.25	0.04	0.87	29.05	1.
Rural Area C	12	93	19.00	9.36	0.09	0.87	7.30	0.

根据该分层抽样，上述关于的总体均值估计（住户平均收看电视时间）的方差为：

$$\widehat{Var}(\hat{\mu}_{st}) = \sum_{h=1}^3 \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h} = \frac{1}{(310)^2} \left[\left((155)^2 \cdot \frac{(155 - 20)}{155} \cdot \frac{(5.95)^2}{20} \right) + \left((62)^2 \cdot \frac{(62 - 8)}{62} \cdot \frac{(15.25)^2}{8} \right) + \left((93)^2 \cdot \frac{(93 - 12)}{93} \cdot \frac{(9.36)^2}{12} \right) \right] = 1.97$$



(示例) 分层抽样下的抽样误差：计算结果

各分层的计算表如下：

Stratification	n_h	N_h	mean_h	sd_h	var_p1	var_p2	var_p3	var_
Town A	20	155	33.90	5.95	0.25	0.87	1.77	0.
Town B	8	62	25.13	15.25	0.04	0.87	29.05	1.
Rural Area C	12	93	19.00	9.36	0.09	0.87	7.30	0.

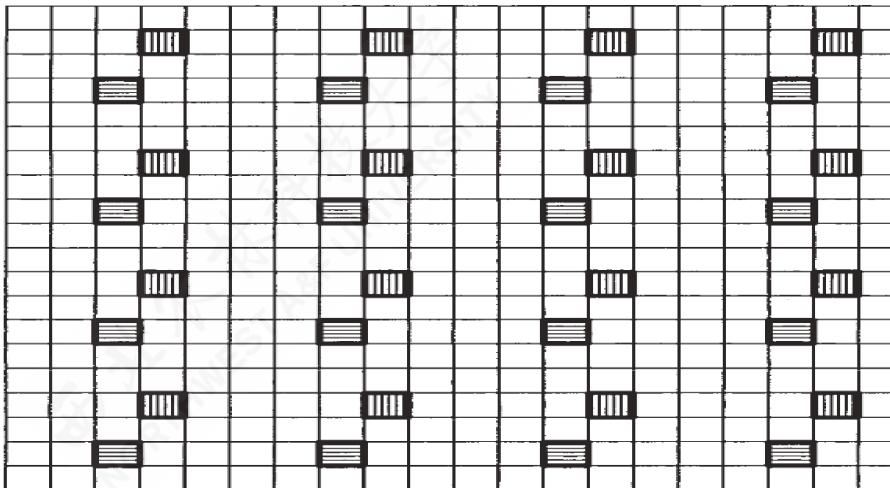
根据该分层抽样，上述关于的总体均值估计（住户平均收看电视时间）的95%置信区间为（t查表值为 $t_{1-0.05/2}(39) = 2.02$ ）*：

$$\begin{aligned}\hat{\mu}_{st} &\pm t_{1-\alpha/2}(df) \cdot \sqrt{\widehat{Var}(\hat{\mu}_{st})} \\ &= 27.7 \pm 2.02 \times \sqrt{1.97} = 27.7 \pm 2.84\end{aligned}$$

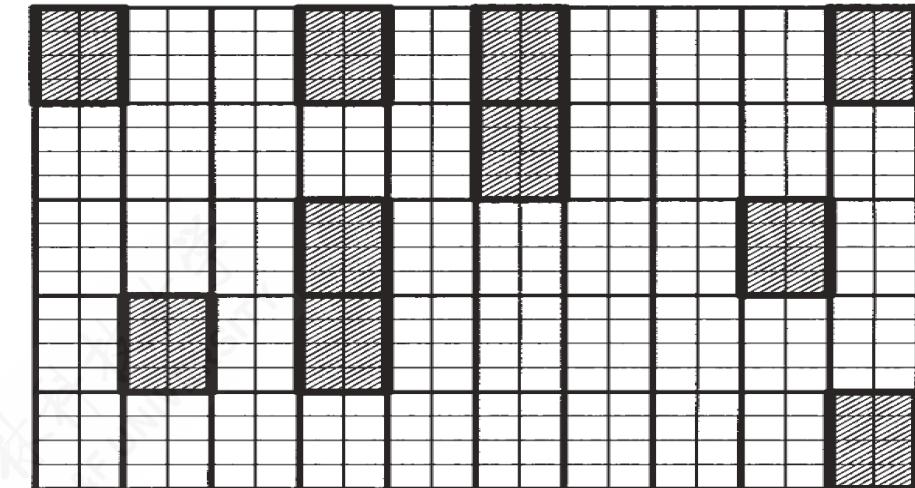
说明： * 如果不是等比例分层抽取，那么自由度的确定需要用到一个计算公式



抽样误差：系统抽样和整群抽样的关系



系统抽样示例



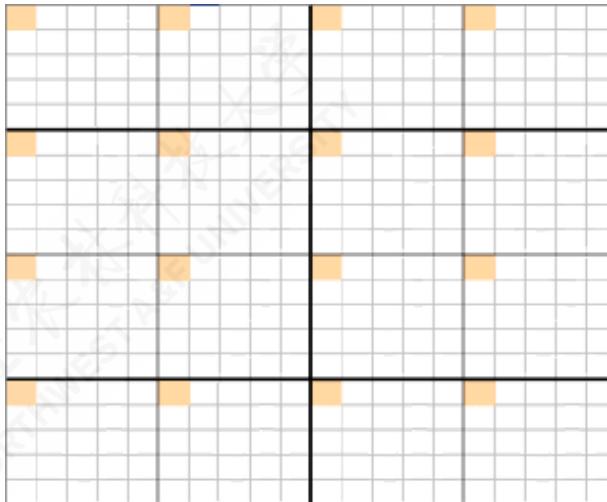
整群抽样示例

从表面上看，系统抽样（systematic sampling）和整群抽样（cluster sampling）非常不同。实际上，这两种方式具有相同的抽样结构：

- 利用主要抽样单位（PSU）划分群组，而每个主要抽样单位又是由次要抽样单位（SSU）组成。
- 如果主要抽样单位（PSU）被随机抽中，则其所有次要单位（SSU）的 y 值将
都会被抽中

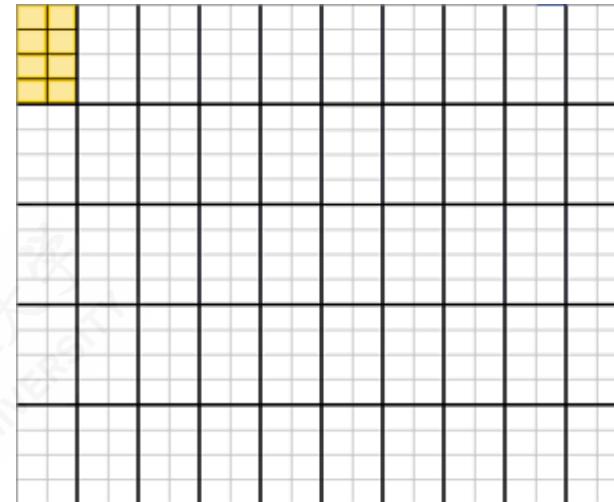


抽样误差：系统抽样和整群抽样的关系



系统抽样示例

- 该案例共有25个主要抽样单位(PSU)：每个 5×5 中型方框都有25个小格。
- 着色色区块是一次典型的随机抽取的系统抽样结果，共抽取样本数($n=16$)。



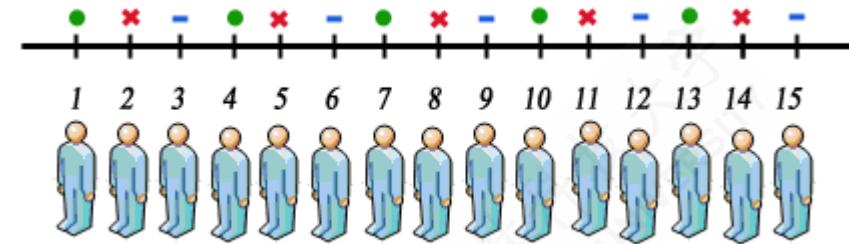
整群抽样示例

- 该案例共有50个主要抽样单位(PSU)：共有50个 2×4 型方框。
- 着色色区块是一次典型的随机抽取的整群抽样结果，共抽取样本数($n=8$)。



(示例) 系统抽样和整群抽样的关系

案例说明：下面展示的是“三采一”的系统采样：我们从三个主要抽样单元(PSU)中随机选择一个，然后再连续地每隔三个选择一个。



从主要抽样单位PSU $\{1, 2, 3\}$ 中随机选择一个值。例如，如果选择2，那么我们将选择上图中所有的红叉个体 $*$ 的 $\{2, 5, 8, 11, 14\}$ 。

- 抽样得到的样本数据 $\{2, 5, 8, 11, 14\}$ ，只是我们随机选定了1个主要抽样单位(PSU)红叉 $*$ ，因而所有具有该主要抽样单位的全部个体都被抽中（全部红叉）。
- 实际上，只抽选1个主要抽样单位(PSU)的情况并不少见，例如以上“三采一”的系统样本。我们只采样3个主要抽样单位（分别是绿点●、红叉 $*$ 、蓝



抽样误差：系统抽样和整群抽样（记号表达）

我们约定系统抽样和整群抽样的记号表达体系如下：

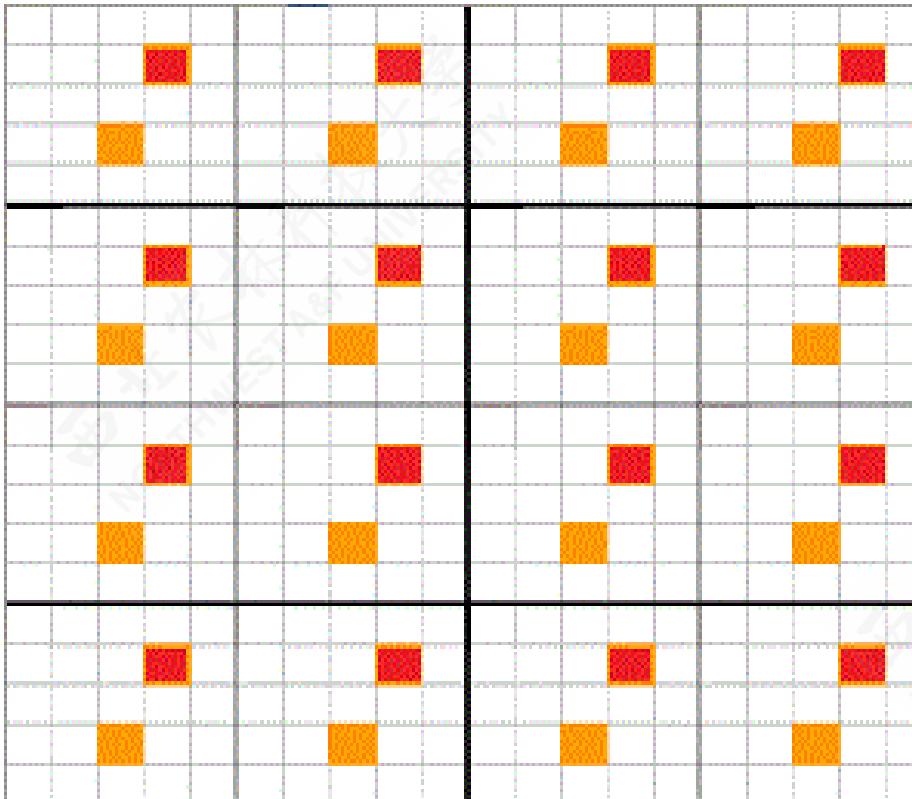
- N 表示总体中的主要抽样单元 (PSU) 的数量； n 表示样本中的主要抽样单元 (PSU) 的数量； M_i 表示第 i 个主要抽样单元 (PSU) 中次要抽样单元 (SSU) 的数量；
 $M = \sum_{i=1}^N M_i$ 表示总体中的所有次要抽样单元 (SSU) 的数量；
- y_{ij} 表示第 i 个主要抽样单元 (PSU) 中第 j 次要抽样单元 (SSU) 的个体的变量值。
 $y_i = \sum_{j=1}^{M_i} y_{ij}$ 表示第 i 个主要抽样单元 (PSU) 下所有个体的变量值之和。
- 主要抽样单元 (PSU) 的均值记为 μ_1 ，次要抽样单元 (SSU) 的均值记为 μ ，二者的计算公式分别为：

$$\mu_1 = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij} = \frac{1}{N} \sum_{i=1}^N y_i$$

$$\mu = \frac{1}{M} \sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij} = \frac{1}{M} \sum_{i=1}^N y_i$$



(示例) 系统抽样的记号表达

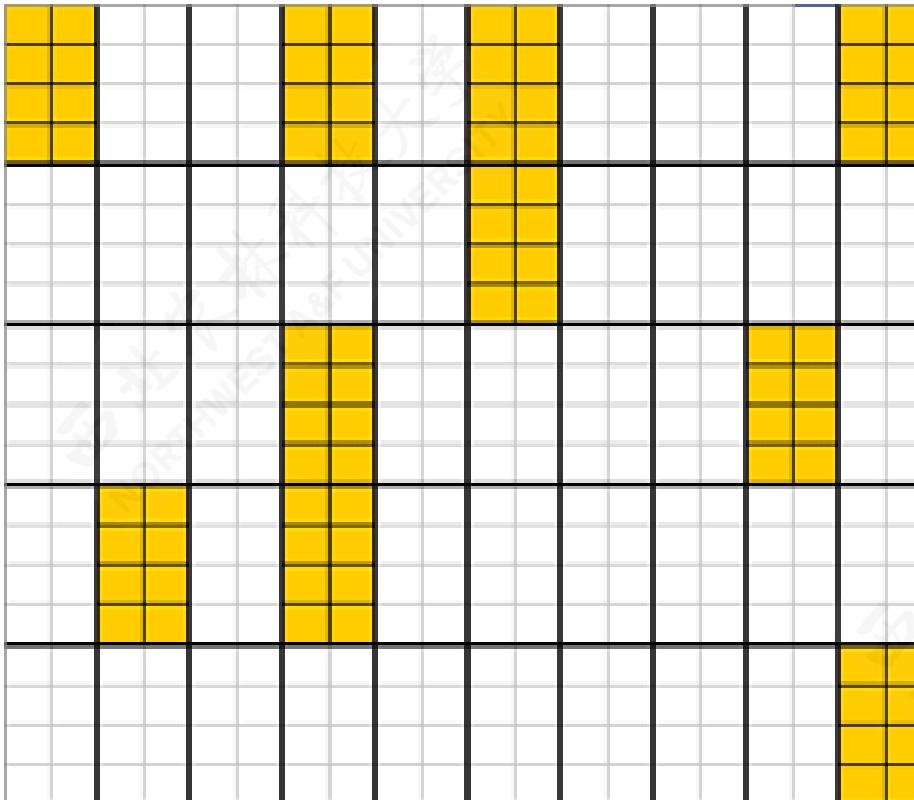


系统抽样示例

- 总体的主要抽样单元（PSU）的数量 $N = 25$: 每个 5×5 中型方框的全部小格，共25个。
- 样本中的主要抽样单元（PSU）的数量 $n = 2$: 每个 5×5 中型方框的都抽中了2个着色小格。
- 每个主要抽样单元（PSU）中次要抽样单元（SSU）的数量 $M_i = 16$: 所有 5×5 中型方框，共有16个。



(示例) 整群抽样的记号表达



整群抽样示例

- 总体的主要抽样单元（PSU）的数量
 $N = 50$: 每个 2×4 中型方框，共50个。
- 样本中的主要抽样单元（PSU）的数量 $n = 10$: 随机抽中的 2×4 中型方框，共抽中10个 2×4 中型着色方框。
- 每个主要抽样单元（PSU）中次要抽样单元（SSU）的数量 $M_i = 8$: 每个 2×4 中型方框中的8个小格。



抽样误差：系统抽样的抽样误差计算

在“1-in-n”系统抽样方案下，估计次要抽样单元（SSU）的均值 $\hat{\mu}_{sy}$ 和方差 $\widehat{Var}(\hat{\mu}_{sy})$ 分别为：

$$\hat{\mu}_{sy} = \bar{y}_{sy} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i$$

$$\begin{aligned}\widehat{Var}(\hat{\mu}_{sy}) &= \frac{M - n \cdot \bar{M}}{M \cdot n} \cdot \frac{1}{(n-1)} \cdot \sum_{i=1}^n (\bar{y}_i - \hat{\mu})^2 \\ &= \frac{M - n \cdot \bar{M}}{M \cdot n} \cdot s_{\bar{y}_i}^2\end{aligned}$$

其中：

- $\bar{y}_i = \frac{y_i}{M_i} = \frac{\sum_{j=1}^{M_i} y_{ij}}{M_i}; i \in 1, 2, \dots, n$
- $\bar{M} = M_1 = M_2 = \dots = M_n$



(示例) 渡船汽车载客量案例

案例说明：载有汽车横渡海湾的渡轮是按载客量而不是按人收取费用的。轮渡公司希望估算8月份每辆车的平均载人数。该公司知道去年有400辆车乘坐轮渡（见右表）。

1	2	3	4	5	6	7	8	10
12	22	59	102	108	66	24	6	1

公司想对其中80辆车进行采样，为了便于估计系统样本的方差，研究人员决定选择使用系统抽样方法，反复抽

id	persons
1	5
2	6
3	4
4	8
5	5
6	6
7	5
8	2

Showing 1 to 8 of 400 entries



(示例) 系统抽样下估计期望和方差：抽样结果

公司决定采用 $1 - in - 50(400/8)$ 的系统抽样方案，也即：

- 从1到50的序号中，不重复随机选择10个序号：8、16、40、6、2、26、37、14、47、46；
- 然后分别以这10个序号作为起始点，每隔50个抽取1个单位，每份样本都会抽取得到8个单位；
- 最终共获得10份系统抽样样本（每份样本含8个个体）。抽样结果如下（括号内为车内人数）：

select	out	mean
sample_1	8(2), 58(3), 108(2), 158(3), 208(3), 258(6), 308(4), 358(1)	3.00
sample_2	16(4), 66(5), 116(5), 166(6), 216(2), 266(4), 316(4), 366(6)	4.50
sample_3	40(5), 90(5), 140(7), 190(7), 240(5), 290(5), 340(4), 390(6)	5.50
sample_4	6(6), 56(3), 106(7), 156(4), 206(6), 256(6), 306(3), 356(3)	4.75
sample_5	2(6), 52(6), 102(6), 152(4), 202(4), 252(5), 302(4), 352(3)	4.75



(示例) 系统抽样下估计期望和方差：计算结果

根据案例，容易计算得到：主要抽样单位（PSU）数量 $N = 50$ ；样本中的主要抽样单位（PSU）数量 $n = 10$ ；第 i 个主要抽样单位下的次要抽样单位的数量 $M_i = 8(i \in 1, 2, \dots, 50)$ ；总体的全部次要抽样单位（SSU）数量

$$M = \sum_{i=1}^{50} M_i = 8 \times 50 = 400$$

$$\hat{\mu}_{sy} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i = 4.62$$

$$\begin{aligned}\widehat{Var}(\hat{\mu}_{sy}) &= \frac{M - n \cdot \bar{M}}{M \cdot n} \cdot \frac{1}{(n-1)} \cdot \sum_{i=1}^n (\bar{y}_i - \hat{\mu})^2 = \frac{M - n \cdot \bar{M}}{M \cdot n} \cdot s_{\bar{y}_i}^2 \\ &= \frac{400 - 10 \times 8}{400 \times 10} \cdot 0.4931 = 0.0394\end{aligned}$$

- 其中： $\bar{y}_i = \frac{y_i}{M_i} = \frac{\sum_{j=1}^{M_i} y_{ij}}{M_i}; i \in 1, 2, \dots, n;$ 以及 $\bar{M} = M_1 = M_2 = \dots = M_n$



抽样误差：整群抽样的抽样误差计算方法I

为了次要抽样单元（SSU）均值和方差，我们可以采用无偏估计法（unbiased estimator）：

$$\hat{\mu} = \frac{N}{M} \cdot \frac{\sum_{i=1}^n y_i}{n}$$
$$\widehat{Var}(\hat{\mu}) = \frac{N(N-n)}{M^2} \cdot \frac{s_u^2}{n}$$

- $s_u^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$
- N 表示总体中的主要抽样单元（PSU）的数量； n 表示样本中的主要抽样单元（PSU）的数量； M_i 表示第 i 个主要抽样单元（PSU）中次要抽样单元（SSU）的数量； M 表示总体中的所有次要抽样单元（SSU）的数量；
- y_{ij} 表示第 i 个主要抽样单元（PSU）中第 j 次要抽样单元（SSU）的个体的变量值。 $y_i = \sum_{j=1}^{M_i} y_{ij}$ 表示第 i 个主要抽样单元（PSU）下所有个体的变量值之和。



抽样误差：整群抽样的抽样误差计算方法2

此外，当群组变量值总和与群组单位数呈正相关关系时，使用比率估计法（ratio estimator）比使用无偏估计更好。此时，估计的次要抽样单元（SSU）均值 $\hat{\mu}_r$ 和方差 $\widehat{Var}(\hat{\mu}_r)$ 分别为：

$$\hat{\mu}_r = r = \frac{\sum_{i=1}^n y_i}{M} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i}$$

$$\widehat{Var}(\hat{\mu}_r) = \frac{N(N-n)}{n(n-1)} \cdot \frac{1}{M^2} \sum_{i=1}^n (y_i - rM_i)^2$$

- N 表示总体中的主要抽样单元（PSU）的数量； n 表示样本中的主要抽样单元（PSU）的数量； M_i 表示第 i 个主要抽样单元（PSU）中次要抽样单元（SSU）的数量； M 表示总体中的所有次要抽样单元（SSU）的数量；
- y_{ij} 表示第 i 个主要抽样单元（PSU）中第 j 次要抽样单元（SSU）的个体的变量值。 $y_i = \sum_{j=1}^{M_i} y_{ij}$ 表示第 i 个主要抽样单元（PSU）下所有个体的变量值之和。



(示例)家庭休假支出案例

案例说明：社会学家想要估计某个城市中每个家庭的平均年休假预算。据统计，这个城市有3100户。

社会学家将整个城市划分为400个街区，并将其视为400个集群。然后，他随机抽样了24个集群，采访了该集群中的每个家庭。

整群抽样的结果见右边数据表：

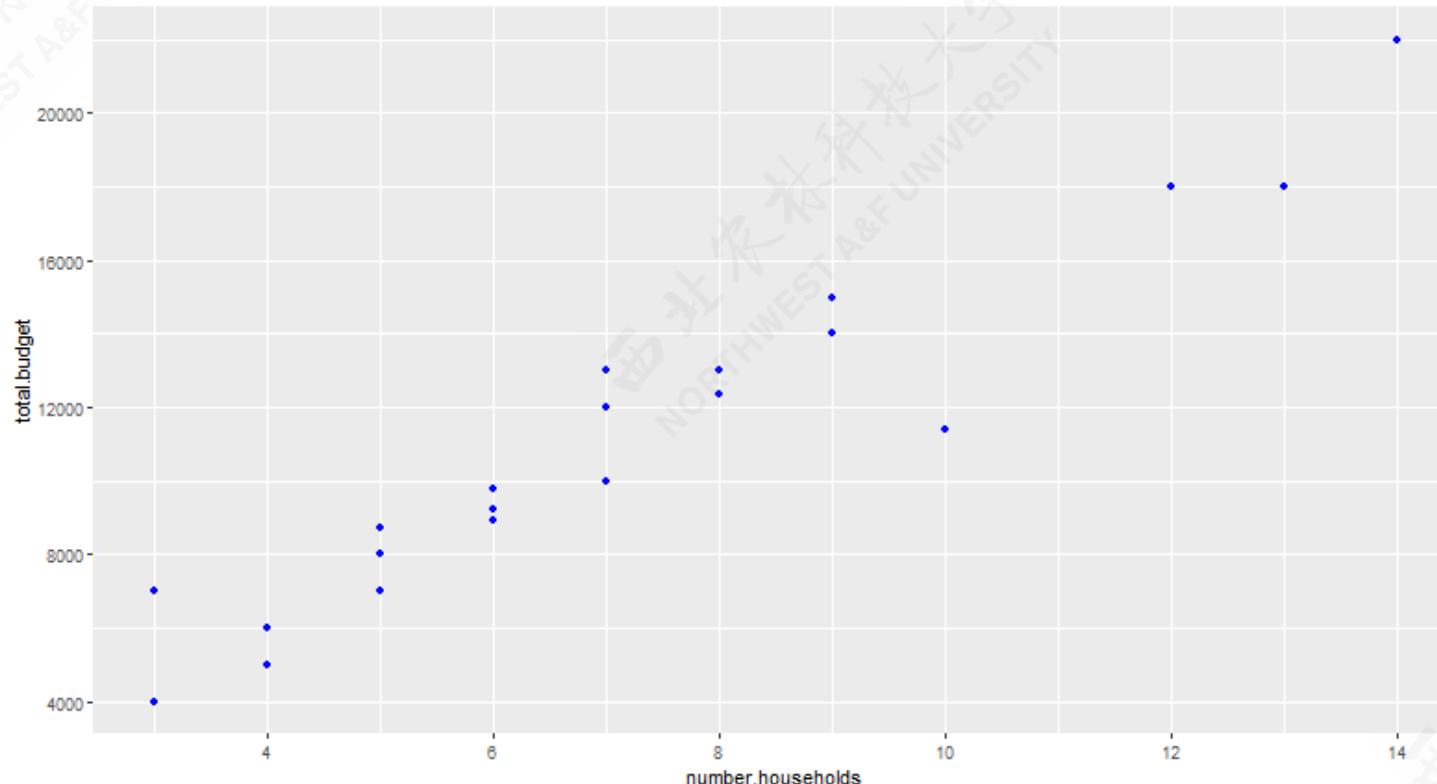
cluster	number.households	total.budget
389	7	12000
202	9	15000
39	5	8000
286	8	13000
6	12	18000
180	5	7000
143	4	6000
280	8	13000

Showing 1 to 8 of 25 entries



(示例) 整群抽样下的抽样误差：相关性分析

初步分析抽样的各群组我们可以发现，主要抽样单元（PSU）的变量总值 y_i （各群组内全部家庭的旅游支出总和）与主要抽样单元（PSU）的单位规模 M_i （各群组的家庭数）存在高度正相关关系。





(示例) 整群抽样下的抽样误差：回归分析

利用R软件进行回归分析，可以进一步发现二者呈现显著线性关系。

$$\begin{aligned}total.\ budget &= + 647.98 + 1441.94number.\ households_i + e_i \\(s) &705.8674 \quad 92.5852 \\(t) &+ 0.92 \quad + 15.57 \\(p) &0.3686 \quad 0.0000\end{aligned}$$



(示例) 整群抽样下的抽样误差：比率估计法

根据整群抽样数据，我们可以计算得到次要抽样单元（SSU）的均值为：

$$\hat{\mu}_r = r = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} = \frac{259240}{169} = 1533.96$$

$$\widehat{Var}(\hat{\mu}_r) = \frac{N(N-n)}{n \cdot M^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - rM_i)^2$$

因为已知： $M = 3100$; $N = 400$; $n = 24$, 次要抽样单元（SSU）的方差计算结果为：

$$\begin{aligned}\widehat{Var}(\hat{\mu}_r) &= \frac{N(N-n)}{n \cdot M^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - rM_i)^2 \\ &= \frac{400 \times (400 - 24)}{24 \times 9610000} \times \frac{1}{24 - 1} \times 40387519.04 \\ &= \frac{150400}{230640000} \times \frac{1}{23} \times 40387519.04 = 1145.07\end{aligned}$$



(示例) 整群抽样下的抽样误差：比率估计法

前述比率估计法需要用到的计算表如下所示：

cluster	number.households	total.budget	r_Mi	part_sqr
389	7	12000	10,737.75	1,593,271.33
202	9	15000	13,805.68	1,426,399.13
39	5	8000	7,669.82	109,017.19
286	8	13000	12,271.72	530,397.62
6	12	18000	18,407.57	166,116.54
180	5	7000	7,669.82	448,662.16
143	4	6000	6,135.86	18,457.39
280	8	13000	12,271.72	530,397.62
126	14	22000	21,475.50	275,097.15

Showing 1 to 9 of 25 entries

Previous

1

2

3

Next



(示例) 整群抽样下的抽样误差：无偏估计法

作为对比，下面我们再采用无偏估计法公式进行计算。

容易计算： $M = 3100$; $N = 400$; $n = 24$ ，以及

$s_u^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = 20208762.32$ 。因此，次要抽样单元（SSU）的均值和方差计算结果分别为：

$$\hat{\mu} = \frac{N}{M} \cdot \frac{\sum_{i=1}^n y_i}{n} = \frac{400}{3100} \cdot \frac{259240}{24} = 1393.81$$

$$\begin{aligned}\widehat{Var}(\hat{\mu}) &= \frac{N(N-n)}{M^2 \cdot n} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{400(400-24)}{(3100)^2 \cdot 24} \cdot s_u^2 = \frac{400(400-24)}{(3100)^2 \cdot 24} \times 20208762.32 = 13178.1\end{aligned}$$



抽样误差：整群抽样两个误差估计方法的比较

- 当群组变量值总和与群组单位数大小成正比时，使用比率估计比使用无偏估计更好。因为无偏估计法的方差会非常大，估计结果非常不满意。
我们可以简单使用的随机抽样的公式来计算方差吗？——抱歉不行！
- 如果采用简单随机抽样，那么就应该相对应使用简单随机抽样公式计算方差，而且必须通过简单随机抽样收集数据。注意：如果不按照抽样方案计算方差，这会是一个很大的错误！



抽样误差：整群抽样的抽样误差计算方法3

有时候，群组被抽中的概率 p_i 就等于群组单位数占总体单位数的比率，也即 $p_i = M_i/M$ 。我们一般称的这种情形为主要抽样单元（PSU）满足比例概率条件（probabilities proportional to size, pps）。那么，在满足pps条件下进行的整群抽样，估计次要抽样单元（SSU）的均值 $\hat{\mu}_p$ 和方差 $\widehat{Var}(\hat{\mu}_p)$ 分别为：

$$\hat{\mu}_p = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i}{M_i} \right)$$

$$\widehat{Var}(\hat{\mu}_p) = \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_p)^2$$

- $\bar{y}_i = \frac{y_i}{M_i}$ 表示第 i 个群组的抽样均值。 n 表示样本中的主要抽样单元（PSU）的数量；
 M_i 表示第 i 个主要抽样单元（PSU）中次要抽样单元（SSU）的数量； M 表示总体中的所有次要抽样单元（SSU）的数量。



(示例) 请求计算机援助案例

案例说明：一家大型公司共有10个部门，每个部门的员工人数各不相同（见下左表）。IT部门主管计划对该公司的3个部门进行随机抽样，以估计该公司平均每个部门的计算机帮助请求数。然后，他采用可重复抽样的比例概率整群抽样法(pps)，随机抽取了三个部门的样本数据（见下右表）：

cluster	employees	requires
1	1000	643
2	650	427
3	2100	1266
4	860	544
5	2840	1938
6	1910	1308

Showing 1 to 6 of 11 entries

Previous

cluster	employees	requires
2	650	427
8	3200	1933
10	1200	770



(示例) 整群抽样下的抽样误差：计算结果

cluster	employees	requires	ratio	minus	sqr
2	650	427	0.6569	0.0227	0.000516
8	3200	1933	0.6041	-0.0302	0.000909
10	1200	770	0.6417	0.0074	0.000055
Total	5050	3130	1.9027	0.0000	0.001480

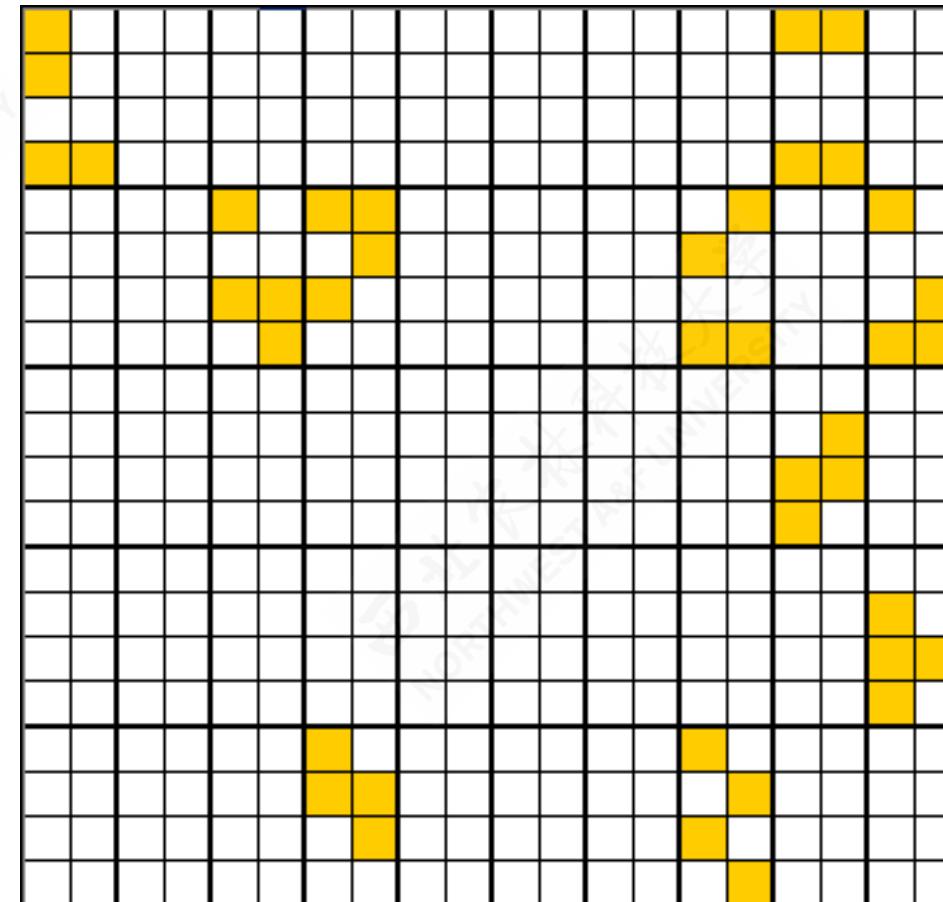
依据抽样数据，我们知道 $n = 3$ ，容易计算得到 $\bar{y}_i = \frac{y_i}{M_i}$ ，并进一步计算得到估计的均值 $\hat{\mu}_p$ ，然后再计算出方差 $\widehat{Var}(\hat{\mu}_p)$:

$$\hat{\mu}_p = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{M_i} = mean(\bar{y}_i) = \bar{y}_i = 0.6342$$

$$\widehat{Var}(\hat{\mu}_p) = \frac{1}{n(n-1)} \cdot \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_p)^2 = \frac{1}{n} \cdot s^2(\bar{y}_i) = 0.000247$$



抽样误差：多阶段抽样

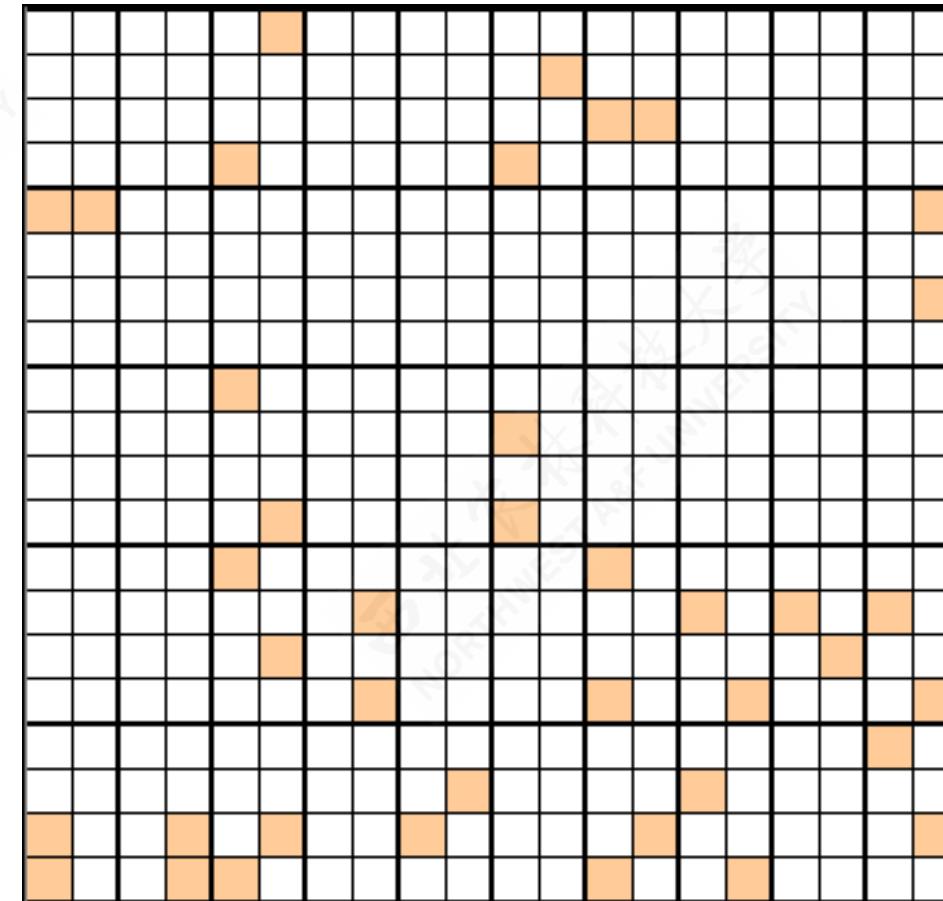


两阶段抽样示例1：10个PSU，4个SSU/PSU

西北农林科技大学
NORTHWEST A&F UNIVERSITY



抽样误差：多阶段抽样



两阶段抽样示例2：20个PSU，2个SSU/PSU

西北农林科技大学
NORTHWEST A&F UNIVERSITY



抽样误差：多阶段抽样的符号约定

多阶段抽样下，一些重要的符号约定如下：

- N 表示总体中的全部群组数量； n 表示随机抽样后抽选得到的群组数量； M_i 表示总体中，第 i 个群组中的单位数量； m_i 表示随机抽中的第 i 个群组中的单位数量； $M = \sum_{i=1}^N M_i$ 表示总体中的所有单位数量；
- y_{ij} 表示随机抽中的第 i 个群组中的第 j 个单位的变量值； $\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$ 表示被抽中的第 i 个群组的样本均值。 $\hat{y}_i = M_i \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i} = M_i \bar{y}_i$ 表示对总体中第 i 个群组的变量总值的估计值。



抽样误差：多阶段抽样下的抽样误差计算方法(

多阶段抽样下，次要抽样单元（SSU）均值 $\hat{\mu}$ 和方差 $\widehat{Var}(\hat{\mu})$ 的无偏估计法 (unbiased estimator) 计算公式分别为：

$$\hat{\mu} = \frac{N}{M} \cdot \frac{\sum_{i=1}^n \hat{y}_i}{n} = \frac{N}{M} \cdot \frac{\sum_{i=1}^n M_i \bar{y}_i}{n}$$

$$\widehat{Var}(\hat{\mu}) = \frac{N(N-n)}{M^2} \cdot \frac{s_u^2}{n} + \frac{N}{nM^2} \sum_{i=1}^n M_i (M_i - m_i) \frac{s_i^2}{m_i}$$

两个样本方差，其中 s_u^2 表示主要抽样单位（PSU）的样本方差；而 s_i^2 表示被抽中的第 i 个群组的样本方差。



抽样误差：多阶段抽样下的抽样误差计算方法2

对于两阶段抽样方案：第一阶段和第二阶段都采用简单随机抽样。

- 如果总体的次要抽样单元（SSU）总数 M 不可知，则不能使用前述的无偏估计法。
- 此外，如果群组的变量加总值（sum value）与群组的个体数量（element size）与呈比率关系，则应该采用下述比率估计法。

对于这样的多阶段抽样方案，次要抽样单元（SSU）均值 $\hat{\mu}_r$ 和方差 $\widehat{Var}(\hat{\mu}_r)$ 的比率估计法（ratio estimator）计算公式分别为：

$$\hat{\mu}_r = \hat{r} = \frac{\sum_{i=1}^n \hat{y}_i}{\sum_{i=1}^n M_i} = \frac{\sum_{i=1}^n M_i \bar{y}_i}{\sum_{i=1}^n M_i}$$

$$\widehat{Var}(\hat{\mu}_r) = \frac{N(N-n)}{nM^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - M_i \hat{r})^2 + \frac{N}{nM^2} \sum_{i=1}^n M_i (M_i - m_i) \frac{s_i^2}{m_i}$$



(示例) 连锁餐厅满意度案例

案例说明：一家餐饮连锁店想估计员工对工作的平均满意度（里克特量表1-7分制）。该连锁店共有120家餐厅，连锁店的全体员工总数为6860人。研究人员决定使用两阶段随机抽样方案，第一阶段采用简单随机抽样来采样10家餐厅（被抽中的序号见列 `id`，餐厅总员工数见列 `tot_worker`）。然后，第二阶段也使用简单随机抽样对这些餐厅中约20%的员工（被抽中的员工数量见列 `sel_worker`）进行抽样和工作满意度采访（见列 `satisfaction`）。最终抽样数据结果如下：

<code>id</code>	<code>tot_worker</code>	<code>sel_worker</code>	<code>satisfaction</code>
41	54	11	5, 7, 4, 7, 6, 7, 6, 5, 3, 4, 7
119	48	10	6, 3, 7, 3, 6, 3, 5, 6, 7, 7
42	68	14	5, 5, 7, 6, 4, 3, 5, 5, 6, 3, 6, 4, 7, 4
18	70	14	3, 3, 3, 4, 7, 5, 7, 4, 6, 7, 3, 3, 3, 3
13	52	11	6, 5, 5, 3, 6, 3, 3, 5, 7, 7, 3

Showing 1 to 5 of 10 entries

Previous 1 2 Next



(示例) 多阶段抽样的抽样误差：基本计算量

根据案例数据，容易得到：

- 所有餐厅数量为 $N = 120$ 。
- 简单随机抽样选中的餐厅数量为 $n = 10$ 。
- 第 i 个餐厅的总员工数为 $M_i = \text{total_worker}$ 列。
- 随机抽中的第 i 个餐厅中的被抽中的员工数量为 $m_i = \text{sel_worker}$ 列。
- 连锁店全体员工的总人数为 $M = \sum_{i=1}^N M_i = 6860$ 。
- 随机抽中的第 i 个餐厅中的平均工作满意度评分为 $\bar{y}_i = \text{mean}$ 列，工作满意度的样本方差 $s_i^2 = \text{variance}$ 列。
- 估计得到的第 i 个酒店的加总满意度评分为 $\hat{y}_i = \text{y_hat}$ 列，10家被抽中酒店估计的平均加总满意度评分为 $\bar{\hat{y}} = \frac{\sum_{i=1}^n \hat{y}_i}{n} = 280.29$ 。



(示例) 多阶段抽样的抽样误差：无偏估计法的计算量1

我们可以根据无偏估计法的相关理论公式，得到如下的计算表：

id	tot_worker	sel_worker	satisfaction	mean	variance	y_hat	sum_
41	54	11	5, 7, 4, 7, 6, 7, 6, 5, 3, 4, 7	5.55	2.07		
119	48	10	6, 3, 7, 3, 6, 3, 5, 6, 7, 7	5.30	2.90		
42	68	14	5, 5, 7, 6, 4, 3, 5, 5, 6, 3, 6, 4, 7, 4	5.00	1.69		
18	70	14	3, 3, 3, 4, 7, 5, 7, 4, 6, 7, 3, 3, 3, 3	4.36	2.86		
13	52	11	6, 5, 5, 3, 6, 3, 3, 5, 7, 7, 3	4.82	2.56		
80	62	13	5, 5, 4, 3, 6, 5, 7, 7, 3, 5, 3, 3, 6	4.77	2.19		
68	41	9	6, 4, 3, 6, 6, 3, 3, 5, 7	4.78	2.44		
25	53	11	3, 4, 5, 7, 4, 7, 7, 7, 4, 7, 5	5.45	2.47		

Showing 1 to 8 of 10 entries

Previous

1

2

Next



(示例) 多阶段抽样的抽样误差：无偏估计法的计算量2

从而容易计算得到到如下两个无偏估计法需要用到的样本方差：

$$s_u^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\hat{y}_i - \frac{\sum_{i=1}^n \hat{y}_i}{n} \right)^2 = \frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 = 1591.18$$

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2 = \text{variance}$$

| 上述 s_i^2 计算结果见前页 ppt 的计算表中的 variance 列。



(示例) 多阶段抽样的抽样误差：无偏估计法的结果

因此，采用无偏估计法估计得到的酒店满意度的均值和方差分别为：

$$\hat{\mu} = \frac{N}{M} \cdot \frac{\sum_{i=1}^n \hat{y}_i}{n} = \frac{N}{M} \cdot \bar{\hat{y}} = \frac{120}{6860} \times 280.29 = 4.90$$

$$\begin{aligned}\widehat{Var}(\hat{\mu}) &= \frac{N(N-n)}{M^2} \cdot \frac{s_u^2}{n} + \frac{N}{nM^2} \sum_{i=1}^n M_i (M_i - m_i) \frac{s_i^2}{m_i} \\ &= \frac{120(120-10)}{10 \times 6860^2} \times 1591.18 + \frac{120}{10 \times 6860^2} \times 4615.55 \\ &= 0.0458\end{aligned}$$

上述求和项内部个值计算结果见前页ppt的计算表，其中：

- $M_i (M_i - m_i) \frac{s_i^2}{m_i}$ 见计算表中的 sum_right 列；



(示例) 多阶段抽样的抽样误差：比率估计法的计算量1

我们可以根据比率估计法的相关理论公式，得到如下的计算表：

id	tot_worker	sel_worker	satisfaction	mean	variance	y_hat	sum
41	54	11	5, 7, 4, 7, 6, 7, 6, 5, 3, 4, 7	5.55	2.07		
119	48	10	6, 3, 7, 3, 6, 3, 5, 6, 7, 7	5.30	2.90		
42	68	14	5, 5, 7, 6, 4, 3, 5, 5, 6, 3, 6, 4, 7, 4	5.00	1.69		
18	70	14	3, 3, 3, 4, 7, 5, 7, 4, 6, 7, 3, 3, 3, 3	4.36	2.86		
13	52	11	6, 5, 5, 3, 6, 3, 3, 5, 7, 7, 3	4.82	2.56		
80	62	13	5, 5, 4, 3, 6, 5, 7, 7, 3, 5, 3, 3, 6	4.77	2.19		
68	41	9	6, 4, 3, 6, 6, 3, 3, 5, 7	4.78	2.44		
25	53	11	3, 4, 5, 7, 4, 7, 7, 7, 4, 7, 5	5.45	2.47		

Showing 1 to 8 of 10 entries

Previous

1

2

Next



(示例) 多阶段抽样的抽样误差：比率估计法的计算量2

容易计算得到到如下比率估计法需要用到的样本方差：

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2 = \text{variance}$$

| 上述 s_i^2 计算结果见前页 ppt 的计算表中的 variance 列。



(示例) 多阶段抽样的抽样误差：比率估计法的结果

因此，采用比率估计法估计得到的酒店满意度的均值和方差分别为：

$$\hat{\mu}_r = \hat{r} = \frac{\sum_{i=1}^n \hat{y}_i}{\sum_{i=1}^n M_i} = \frac{2802.86}{555} = 5.05$$

$$\begin{aligned}\widehat{Var}(\hat{\mu}_r) &= \frac{N(N-n)}{nM^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - M_i \hat{r})^2 + \frac{N}{nM^2} \sum_{i=1}^n M_i (M_i - m_i) \frac{s_i^2}{m_i} \\ &= \frac{120(120-10)}{10 \times 6860^2} \times \frac{1}{10-1} \times 7120.48 + \frac{120}{10 \times 6860^2} \times 4615.55 \\ &= 0.0234\end{aligned}$$

上述两个求和项内部个值计算结果见前页ppt的计算表，其中：

- $(\hat{y}_i - M_i \hat{r})^2$ 见计算表中的 sum1_yi_sqr 列；
- $M_i (M_i - m_i) \frac{s_i^2}{m_i}$ 见计算表中的 sum2_si_sqr 列。



抽样误差：多阶段抽样下的抽样误差计算方法3

对于两阶段抽样方案：第一阶段采用比例概率抽样法（PPS），第二阶段采用简单随机抽样法：

- 那么抽样误差计算应该使用比例概率估计法（pps估计法，具体为 Hansen–Hurwitz estimator）。

对于这样的多阶段抽样方案，次要抽样单元（SSU）均值 $\hat{\mu}_p$ 和方差 $\widehat{Var}(\hat{\mu}_p)$ 的比例概率估计法（pps estimator）计算公式分别为：

$$\hat{\mu}_p = \frac{1}{n} \cdot \sum_{i=1}^n \frac{\hat{y}_i}{M_i} = \frac{1}{n} \cdot \sum_{i=1}^n \frac{\bar{y}_i * M_i}{M_i} = \frac{1}{n} \cdot \sum_{i=1}^n \bar{y}_i$$

$$\widehat{Var}(\hat{\mu}_p) = \frac{1}{n(n-1)} \cdot \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_p)^2$$



(示例) 学生书本支出案例

案例说明：一个学院共有36个专业（major）。研究者想估算出上学期学生在教科书上花费（expenses）的平均金额。由于每个专业的规模差异很大，因此采用的两阶段抽样方案，其中第一阶段采用的是pps抽样，第二阶段是简单随机抽样。最终抽样数据结果如下：

major	tot_students	sel_students	expenses
18	10	4	326, 400, 423, 443
13	20	8	278, 312, 450, 350, 227, 438, 512, 403
16	30	12	512, 256, 332, 402, 512, 309, 411, 610, 422, 630, 550, 470
4	15	6	426, 312, 512, 440, 342, 533



(示例) 多阶段抽样的抽样误差：PPS估计法的计算表

我们可以根据比例概率估计法的相关理论公式，得到如下的计算表：

major	tot_students	sel_students	expenses	mean	variance	y_hat	z
18	10	4		326, 400, 423, 443			
13	20	8		278, 312, 450, 350, 227, 438, 512, 403			
16	30	12		512, 256, 332, 402, 512, 309, 411, 610, 422, 630			
4	15	6		426, 312, 512, 440, 342, 533			

建议使用html浏览本课件，此表可以往右拉动，查看更多计算列。



(示例) 多阶段抽样的抽样误差：PPS估计法的结果

因此，采用比例概率估计法估计得到学生书本支出的均值和方差分别为：

$$\hat{\mu}_p = \frac{1}{n} \cdot \sum_{i=1}^n \frac{\hat{y}_i}{M_i} = \frac{1}{n} \cdot \sum_{i=1}^n \frac{\bar{y}_i * M_i}{M_i} = \frac{1}{n} \cdot \sum_{i=1}^n \bar{y}_i = \bar{\bar{y}}_i = 412.02$$

$$\begin{aligned}\widehat{Var}(\hat{\mu}_p) &= \frac{1}{n(n-1)} \cdot \sum_{i=1}^n (\bar{y}_i - \hat{\mu}_p)^2 = \frac{1}{n} \cdot s_{\bar{y}_i}^2 \\ &= \frac{1}{4} \times 1214.6406 = 303.6602\end{aligned}$$

上述计算结果的中间步骤计算值见前页ppt的计算表。 其中：

- \bar{y}_i 见计算表中的mean列；
- \hat{y}_i 见计算表中的y_hat列。

2.5 数据整理

工作流程

工作记录

数据库化

数据检索

数据安全



数据整理的工作流程

在数据收集过程中，重要的是条分缕析。

- 分类存储
 - 依据数据的载体类型、研究的时间需来进行分类，采用合适存放工具进行存放。
 - 纸版问卷，不能随便堆放，需要按照一定分类标准进行存放，便于后续工作。
- 建立目录
 - 存放的目的不仅仅只为了存储，更重要的是为了便于使用，建立目录就是便于利用的方式之一。
 - 目录是用于检索的，对调查获得数据建立目录，也是为了方便检索。
- 编制索引
 - 对于复杂数据，还需要在目录与存储之间建立关联，这就是索引



收集整理的工作记录

数据收集和整理中不仅需要核实，还需要记录。主要记录：

- 数据来源信息：
 - 如调查项目，调查人，采集人，采集时间，地点，对象。
- 数据载体类型信息：
 - 具体是什么载体？比如，纸张的、数字的。
- 数据描述信息：
 - 有多大规模，什么内容，关联什么主题，等等。
- 数据分类信息：
 - 无论是按照载体形态分类还是按照其他标准分类，一个大型项目需要对原始数据根据数据使用，建立基本分类。
- 数据存储信息：
 - 数据以什么样的载体，什么样的方式存储在什么位置？
 - 与数据安全相关的信息，如存储的版本、份数、时间变化关系等。



数据化检索和数据安全

“老师，能把上星期发给我的课件再发一遍吗？我忘记放到哪了？”

“老师，非常的崩溃！电脑的硬盘坏了，写的东西都没有了！”

上述记录信息要尽可能的保存若干个版本。

- 纸和笔的传统版本。便于在需要的时候翻阅，尤其是使用范围相对较广的数据。
- 数字化可检索的版本。为什么要做数字化的可检索版本？
 - 目录树法（相对简单的数据）
 - 建立专门的数据库（针对异常复杂或庞大的数据）



数据化检索和数据安全

数字化数据有一些需要特别注意的问题：

- 数据存储。随时都有若干个备份！
 - 数字化的数据从最初的纸袋到今天的磁盘、硬盘，有各种介质。由于介质的可靠性不同，数据的安全性也不相同。
 - 美国的“911”事件。美国联储会的主席格林斯潘知道这个消息的第一时间，他担心的不是“911”的伤亡情况，而是美国金融数据的安全。
- 数据安全。安全的风险，要么来自于使用者的误操作，要么来自于内部或者外部的有意攻击。
 - 离线保存的目的不仅仅是为了应对各种预想不到的不测，更重要的是为了防止数据泄露。
 - 斯诺登事件：任何在线数据事实上都是不安全的，都有安全隐患。



数据化检索和数据安全

文本数据的安全：

- 文本数据的安全威胁主要来自于不可抗力的一些因素，比如说自然灾害、风蚀等。
- 当然也来源于人为因素。比如说错误的识别，本来是很重要的数据，却被当作了废纸。

非数字化数据的安全：

- 图片数据的载体形态比较复杂，胶片、图片由于介质存储特征的差异，不可以混合放置而保管。如胶片就需要防潮，通常要使用防潮器皿。
- 实物数据的安全具有独特性，应根据实物特征进行科学整理和安全管理。比如说兵马俑，那就在兵马俑的原址上盖一个博物馆进行整理。

2.6 数据清洗

工作内容

清洗记录

备份安全

清洗举例



数据清洗工作的内容

数据整理主要是分类和梳理，数据清洗主要讨论的则是检错。通过这两部分的工作减少人为的误差，降低调查误差。数据清洗包括四个工作内容：

- 真实性评估：确认数据是真实的，不是道听途说，不是张冠李戴，更不是杜撰臆想。
 - “假新闻”现象就是调查数据在真实性层面出现的问题。
 - 微信群里“令人发指”的各类长辈转发
- 完整性评估：数据应与研究工作的目标要求相符，研究不需要的就不应该出现在数据中，研究需要的在数据中就不应该缺少。
 - 如果需要补值，就应继续补充收集数据。



数据清洗工作的内容

- 可用性评估：数据是不是可以用于数据库化了？如果不能，还需要做怎样的数据加工？
 - 比如对图片数据、音频、视频数据，甚至文本数据是不是还要做数字化工作。
 - 对于痕迹数据，尤其是大数据，如果不是直接采用大数据分析，而是应用于单机分析或服务器分析，是不是还要根据数据量进行抽样。
 - 脱敏化处理。对有可能泄露受访者隐私、泄露传感器使用者隐私的部分，还需要做匿名化工作。
- 错误性评估：评估数据可能的错误来源、可能的错误大小，及其对数据质量的影响。



数据清洗工作的内容

以调查问卷数据的清洗为例：

- 真实性的清洗：要确认数据来自于受访者。
- 完整性的清洗：主要看样本无应答，也就是一整份问卷没有应答。以及选项无应答，也就是应该应答的访题没有应答。
- 可用性的清洗：主要是看编码是否完成，权数是否可行，以及缺失值如何标记和处理。
- 错误性的清洗：主要是清洗调查环节的错误，比如样本错误、应答人错误、应答方式错误。



数据清洗工作的记录

清洗数据工作中的每一项活动都要有记录。记录信息包括：

- 清洗工作的信息记录：
 - 数据清洗每一个步骤的做法、参与人、时间、地点、过程信息。
- 与清洗内容相关联的信息记录：
 - 数据真实性信息。比如是否真实？是否存在编造、作弊嫌疑？哪些部分存在不真实？怎么样不真实？。
 - 数据完整性信息。比如是否完整？是否有缺失？如果有缺失，哪些部分缺失？缺失哪些数据？
 - 数据可用性信息。比如问卷数据是否加权？痕迹数据是否数据化了？大数据如何处理？是运用云计算策略，还是裁剪为单机计算容量？
 - 数据错误性信息。比如问卷数据中的缺失，文献数据中的差错等。



数据清洗记录的备份与安全

数据清洗的记录信息应尽可能地保留若干不同的版本。一般包括纸笔版本和数字化版本。纸笔版本便于随时翻阅，数字化版本，便于交流，也便于检索。

- 笔记的清洗。不管是哪一类的笔记，所有的笔记都有私用和公用之别，通常人们做笔记都是做给自己看的（私用笔记）。
 - 你把自己的笔记给别人看，别人能看懂吗？
 - 在正式使用之前，需要把笔记数据通过清洗，变成任何使用者都可读的笔记（公用笔记），
 - 这就是格式化问题，就是把你个人的笔记清洗为数据笔记。



数据清洗记录的备份与安全

- 对音频要抄录：
 - 把语音文档，不管是磁带录音，还是数字录音，抄录为文字，表述为文字或者文字加图片这样的格式。
 - 数字录音还有一个格式清洗问题，不同数字设备的录音，可能会采用不同的格式。
 - 比如olympus的早期设备，采用的就是它自己的格式，DSS格式；如果不是采用它自己的软件就读不出来，最好呢，是转化为通用的格式，比如mp3格式。
- 对视频清洗编码：
 - 如果是非数字录像，最好先转化为数字格式
 - 如果已经是数字录像，对视频清洗编码需要给出时间记录码。



数据清洗工作的几点忠告

哦，已经数字化了，可以扔了，那个没用了，可以扔了。

- 不要轻易地丢弃任何一段看起来没有用处的信息，信息载体。
- 清洗不是扔东西，是清洗数据，让数据清晰化。
- 清洗的目的就是将特异性的数据，转化为公共性的数据、分析研究者都可以读的数据。
- 在清洗的过程中，千万要保留原始观察记录。
 - 一般而言，原始问卷至少要保留十年以上，访谈记录和观察笔记一般要求永久保留。



数据清洗例举1：观测性数据

以观察性研究中数据的清洗为例：

- 观察性数据有一个特点就是差异性，对同一个场景、同一个事件，不同的人去观察，看到的并非完全一致。
- 每个人的观察记录，都有自己的习惯，有的习惯于采用速写和密写，比如说有些人为了防止别人看他的笔记，长采用密写的方式。即使是结构式的观察，不同的观察者也会有特异性。
- 观察性数据的清洗就需要把各类个性化的个人观察数据转变为标准化的观察记录。



数据清洗例举2：文献数据

以文献数据的清洗为例：

- 笔记的清洗。比如说：研究用的素材如文献的阅读、标注与笔记、摘录，如果希望未来继续使用，那就需要格式化清洗，把素材转化为数据。如果有必要，还可以为下一步的数据库化做准备，比如编码。
- 文献的清洗。对阅读过的文献，如果已经获得了数字版本，就需要与数字版本关联的编目信息、阅读信息关联起来整理，结合后边讨论的数据库化工作，把它们转化为个人档案馆。如果没有数字化的版本，则需要将文献信息与阅读笔记信息关联，结合后边讨论的数据库化工作，把它们变成个人的档案阅读目录数据馆。



数据清洗例举3：痕迹数据

对痕迹数据的“四性”评估和清洗，一般是直接依据数据的来源来确认的。比如，来自于网络爬取的数据，和来自于数据拥有者机构提供的数据，其它的平行数据等等。

一般而言，如果数据来源的渠道没有问题，数据的四性就不会有太大的问题。
清洗痕迹数据最重要的一项工作，就是把非格式化数据 清洗为格式化数据
(Why？至少目前的分析工具还不支持直接分析非格式化的数据)

数据格式化：把混杂在一堆数据中的各类数据清洗出来，分门别类。比如说日志数据中的用户行为数据，以淘宝数据为例，订单数据、发单数据、物流数据等等，分门别类整理出来。

数据结构化：把各类数据和变量进行多维度关联。比如把以上日志数据中的各个子集关联到用户之下，形成类似于问卷调查数据的每个样本数据。



数据清洗例举4：大数据

如果痕迹数据是大数据，情况就有些不同了。

在清洗数据之前，需要把清洗策略测试一遍，然后就可以直接采用大数据的清洗模式了。

- 从大数据中抽取数据，或者是从网页上爬取数据，在处理中尽管不一定会用到云计算，在处理逻辑上还是一致的。
- 大数据的清洗，目前运用比较普遍的是Hadoop框架下的Map Reduce。



数据清洗例举4：大数据（以阿里巴巴案例）

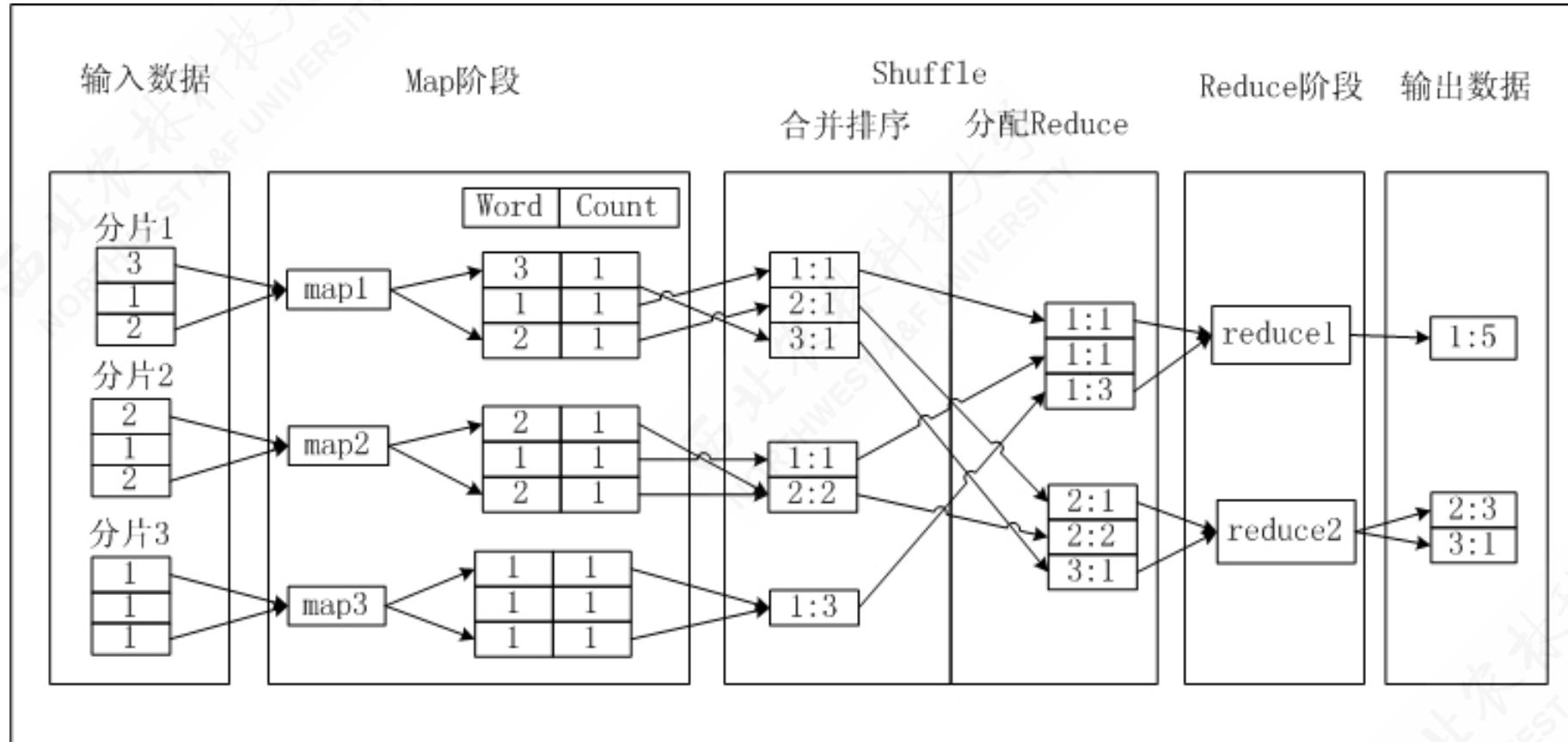
阿里巴巴有淘宝、天猫、一淘等等业务，这些业务每时每刻都在产生数据，这些数据涉及到信用、金融、物流、管理等等业务操作。所有这些操作的数据都会汇集到数据交换平台，由此构成了阿里巴巴的数据动态。

2014年的双十一期间，6个小时之内的处理量就已经达到了100个PB。在产生的这些数据中，既有结构化的数据，也有非结构化的数据，进出数据平台的数据不是个，不是匹，而是流。这些数据流，通过数据处理就变成了中间层的数据，可以运用和应用于服务，中间服务，既可以对内，又可以对外。



数据清洗例举4：大数据（以阿里巴巴案例）

问题是，这些数据是怎么处理的呢？数据清洗关心的正是这个问题。





CGSS数据的清洗：介绍

中国综合社会调查（China General Society Survey, CGSS）：

- CGSS2013, CGSS2015
- 属于混合截面数据：也即不同年份的观测单位不是固定的。
- CGSS2015一共有样本数10968，变量总数有1398



CGSS数据的清洗：数据视图

随机抽取300份样本的前20个变量。CGSS2015数据（样本数=10968）数据如下：

id	s1	s41	s42	s43	s44	s45	token	a0	a1011	a11
1	1	4	7	17	30	55	11001466204877	2015-10-04 16:27:00	配偶	
158	1	4	7	22	38	68	11051921620499	2015-08-31 16:57:00	丈夫	
224	1	4	7	18	31	57	11077761043328	2015-10-04 15:41:00	丈夫	



CGSS数据的清洗：变量视图(全景)

CGSS2015变量体系 (变量数=1398)

题号	变量名	题项	变量类型	是否标签
1	id	问卷编号	double	false
2	s1	样本类型	double	true
3	s41	采访地点-省/自治区/直辖市编码	double	true
4	s42	采访地点-地级市编码	double	false
5	s43	采访地点-县/区编码	double	false

Showing 1 to 5 of 1,398 entries

Previous

1

2

3

4

5

...

280

Next



CGSS数据的清洗：变量视图(局部)

含有“收入”的变量 (变量数=20)

题号	变量名	题项	标签
222	a8a	您个人去年全年的总收入	c(无法回答 = -8, 拒绝回答 = -3, 不知道 = -2, 不适用 = -1, 个人全年总收入高于百万位数 = 9999996)
223	a8b	您个人去年全年的职业/劳动收入	c(无法回答 = -8, 拒绝回答 = -3, 不知道 = -2, 不适用 = -1, 个人全年职业劳动收入高于百万位数 = 9999996)

Showing 1 to 2 of 20 entries

Previous 1 2 3 4 5 ... 10 Next



CGSS数据的清洗：缺失值1

挑选出如下几个变量来观测：

题号	变量名	题项	标签
320	a5606	您在最近三个月内采取过以下哪些方式寻找工作-为自己经营做准备	c(无法回答 = -8, 拒绝回答 = -3, 不知道 = -2, 不适用 = -1, 否 = 0, 是 = 1)
457	b8b	您认为每月户平均收入高于多少元就属于富裕户了	c(拒绝回答 = -3, 不知道 = -2)
478	b1011	在不直接涉及金钱利益的一般社会交往/接触中的信任度-一起参加宗教活动的人士	c(无法回答 = -8, 拒绝回答 = -3, 不知道 = -2, 不适用 = -1, 绝大多数不可信 = 1, 多数不可信 = 2, 可信者与不可信者各半 = 3, 多数可信 = 4, 绝大多数可信 = 5)



CGSS数据的清洗：缺失值2

挑选出如下几个变量来观测。CGSS2015回答情况一瞥(随机40个样本)：

id	b8b	b1011	a5606
17161	5000	-8	
6815	20000	3	0
3910	10000	3	
16452	-2	1	
13069	30000	-8	
12047	10000	-8	0
15015	10000	-8	0
3389	20000	-8	0

Showing 1 to 8 of 40 entries

Previous 1 2 3 4 5 Next



CGSS数据的清洗：变量处理

变量重新命名前

id	a36	a10	a8a
225	4	1	13000
435	3	2	6000
884	5	1	10000
887	3	1	3000
1042	4	1	38000
1500	4	1	18000
1577	2	1	24000
2111	5	2	15000

Showing 1 to 8 of 100 entries

Previous 1 2 3 4 5 ... 13 Next

变量重新命名后

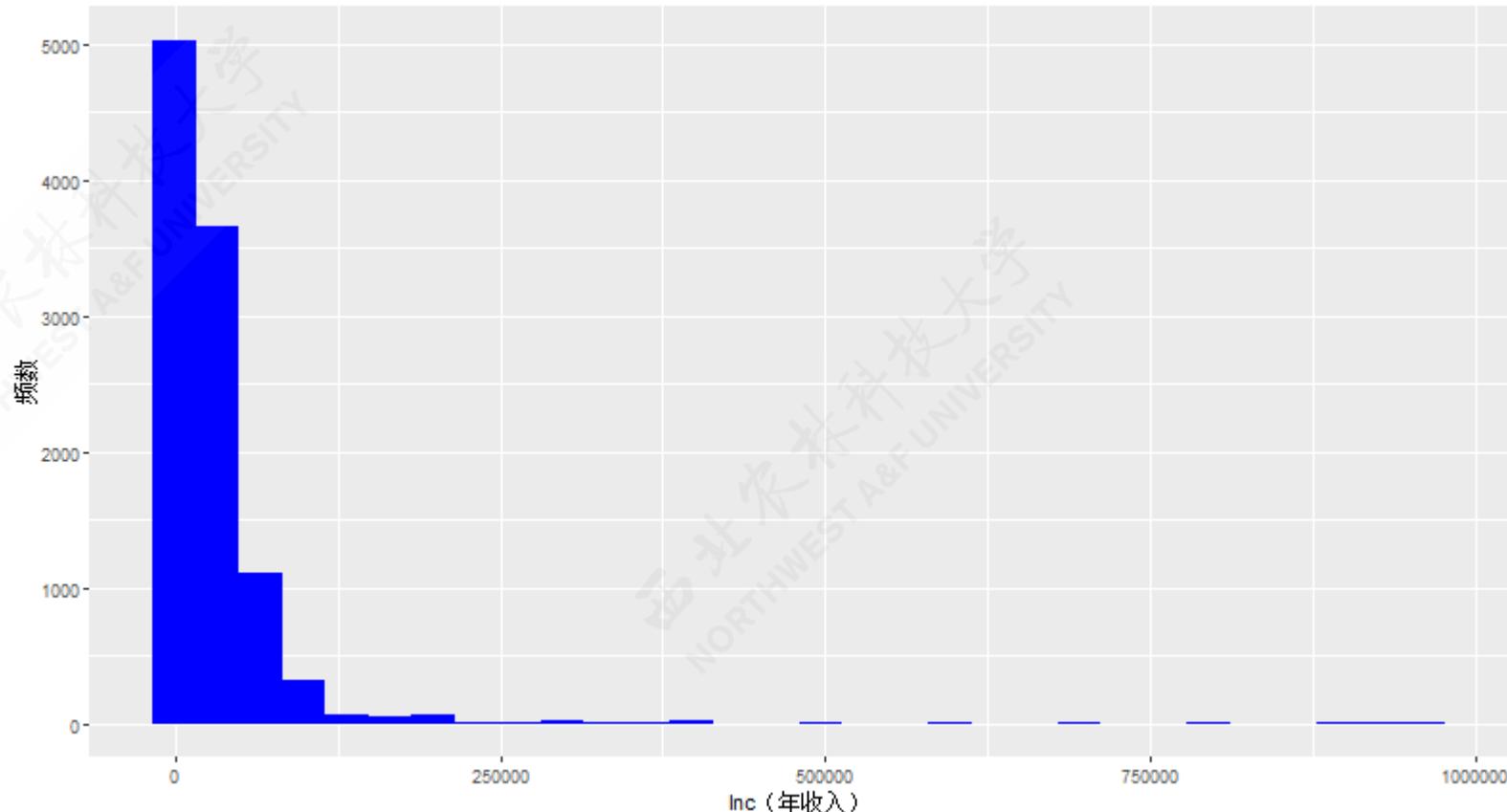
id	Happ	Pol	Inc	Lninc
225	4	1	13000	9.47
435	3	2	6000	8.7
884	5	1	10000	9.21
887	3	1	3000	8.01
1042	4	1	38000	10.55
1500	4	1	18000	9.8
1577	2	1	24000	10.09
2111	5	2	15000	9.62

Showing 1 to 8 of 100 entries

Previous 1 2 3 4 5 ... 13 Next



CGSS数据的清洗：异常值处理前



年收入的直方图



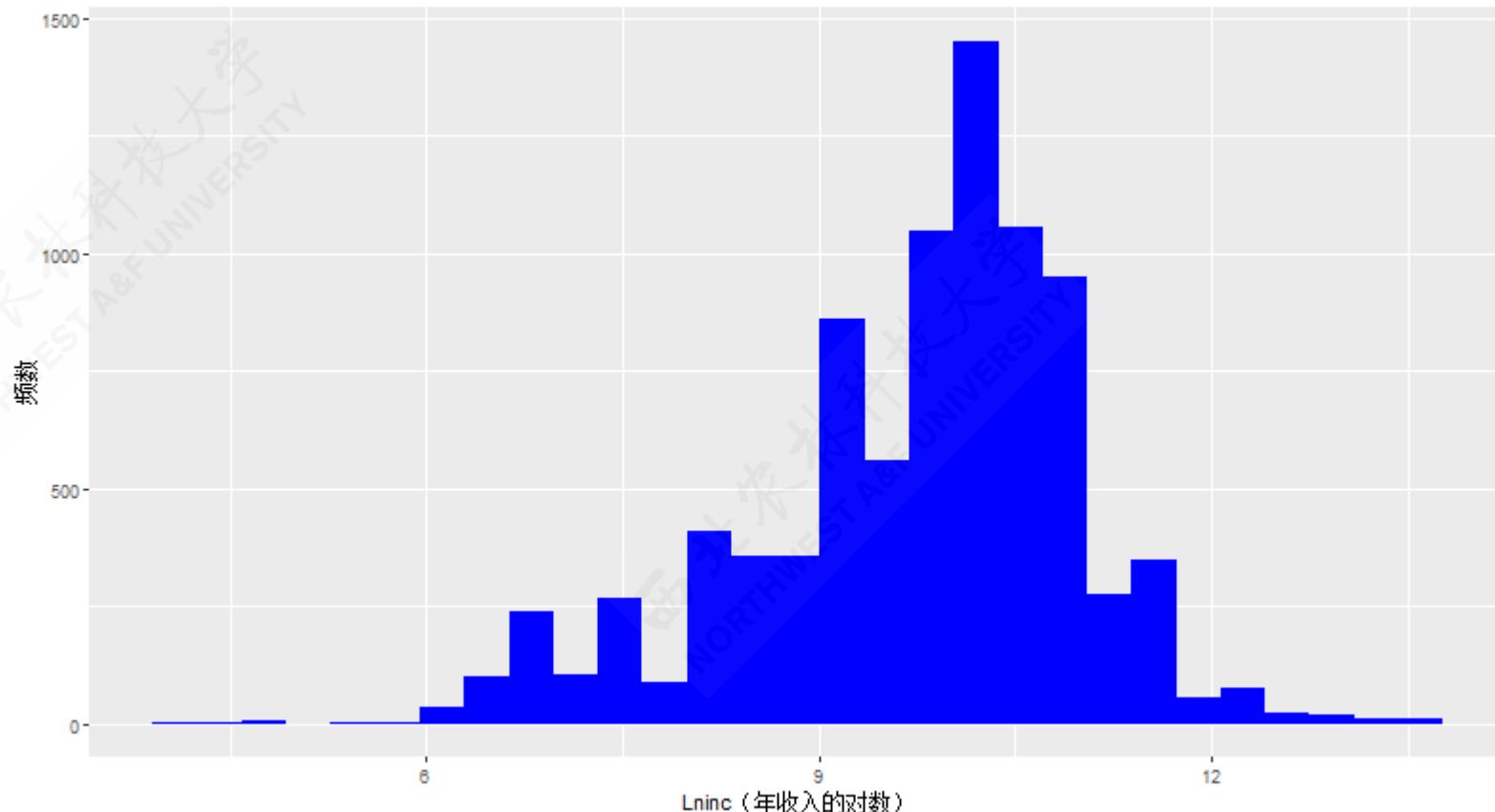
CGSS数据的清洗：异常值处理办法

异常值的处理办法：

- 截尾：比如截掉大于0.99分位数的观测值。
- 数据的转换：对于右偏分布较严重的变量，即右侧异常值较多，自然对数($\ln()$)可以使其实现更加对称。比如，年收入及其对数的分布。



CGSS数据的清洗：异常值对数化处理后



年收入对数的直方图

2.7 数据的数据库化

数据库化的类型

一手资料的数据库化

二手资料的数据库化



为什么需要把数据进行数据库化？

数据不仅整理好了，也清理好了，是不是就可以分析研究了呢？

- 采用手工计算的情形几乎已经消失了。数据的数量与复杂程度，已经超出了人们运用大脑、纸和笔，直接处理的程度。
- 调查数据的分析与研究，从计算机应用普及以来，就已经主要依靠计算机了。运用计算机是最有效和最快捷的方式。
- 运用计算机就需要满足计算机对数据的要求，那就是数据库。
 - 清理整理好的数据，要变成计算机可以读取并进行运算的数据格式，通常这一类的格式都是数据库格式。计算机应用程序不同对数据库的格式要求也不相同。

数据库化的目的就是为了便于分析和使用。基本的要求是通过数据库化，让调查数据格式化、结构化，符合统计分析、计算的要求。



数据库化的类型

数据的数据库化，就是把得到的变量、变量属性或者标签输入计算机，变成结构化的数据矩阵。从数据库化的目标来分，主要有如下两类：

- 计算机网络系统的数据库化，主要是用于存储数据，有各种类型的数据库应用程序。
 - 常见的结构化数据库SQL数据库有多种，比如开源的免费的MySQL。
 - 分析计算用的数据库化，主要是通过建立数据库，用于统计分析软件的计算。
 - 我们这里所学的就是这一类数据库化。

我们主要学习常用的运用于计算机单机统计计算与分析用的数据库化。

大数据的数据库化有不一样的特点和需求。



数据的数据库化示例

SPSS是社会科学统计计算运用比较多的一个大型统计计算软件

SPSS数据库的数据视图：

- 每一行代表样本，
- 每一列代表变量
- 中间单元格表示数据取值

SPSS数据库的变量视图

- 每一行代表一个变量
- 每一列表示变量的性质和特征



4. 调查数据的数据库化（主要步骤）

问卷调查的数据，在完成了问卷的审核、归档、清理以后，在用于分析软件的分析之前，就需要把它转化为数据表示的数据库。通常有三个步骤：

- 第一步，编码。
 - 在清理工作中，这项工作应该已经完成了，不过在数据入库之前还需要审核。
- 第二步，数据录入与转化。
 - 如果是纸版问卷调查，这个时候就需要录入数据。建议采用专门的录入软件进行录入，尽量避免录入中出现的差错，进而降低调查误差。
 - 如果是计算机辅助调查，这个时候就需要转化数据。无论是内容转化还是格式转化，也建议尽量采用可靠的工具，避免出现差错。
- 第三步，对录入完成和转化完成的数据，做基本的检验和清理。
 - 最容易出现的差错就是错行、错列造成数据的混乱。



4. 调查数据的数据库化（编码）

编码：就是把调查问卷的每一道访题用符号或者数字组合代码换，包括对每一道访题的选项或应答赋值。

- 每一道访题的编码就是数据库表中的变量。
- 应答赋值就是数据库表中的变量值，这个只有各种属性，就是数据库表变量视图中的各种标签，又称为变量标签。



4. 调查数据的数据库化（编码示例）

我们来看例子，这是Self PS中的一道访题，

【问题】1.5，请问您希望孩子念书，最高念到哪一个程度？（共7个选项）。

【选项】A. 小学；B. 初中；C. 高中；D. 专科、职高、技校、大专；E. 大学本科；F. 本科以上；G. 不必念书

对这套访题我们可以这样编码，访题可以编为B15。（为什么这么编？）

对选项的编码，就以选项的编号做编码。



4. 调查数据的数据库化（编码）

问卷调查数据的编码，一般有三种方法：

- 第一：原始编码，就是直接运用问卷的编码。
 - 通常这种方法仅仅用在访题数量极少，应答非常简单的情况下。
- 第二：先编码，在调查开始之前，编码工作就已经做好了。
 - 通常这种方法会用在基本上都是封闭访题的情况下。
- 第三，后编码，就是在问卷调查完成以后再做编码。
 - 只要是有开放访题，一般都会采用这种编码方式。

编码部相当于问卷数据的一个索引，把变量、变量值，变量标签关联起来，类似于一本问卷数据字典。



4. 调查数据的数据库化（录入）

- 对于简单的问卷调查，可以运用常用的办公软件或统计分析工具来做录入，
 - MS Office Excel
 - Mac Numbers
 - SPSS
 - Stata、statistica、R…
- 对于相对庞大复杂的问卷调查，需要使用专门的数据录入软件。
 - 商业收费的SPSS [Data Entry](#) 模块
 - 免费的[EpiData](#)

专用录入软件的能提高录入效率，并减少录入误差：

- 可把纸版问卷计算机界面化，把纸版问卷完整呈现在计算机屏幕上。
- 可通过对跳转、阈值、变量类型等的控制，尽量减少录入所带来的误差。
- 在录入完成以后，还可以直接把录好的数据导出为数据库表文件。



4. 调查数据的数据库化（检验和清洗）

针对的已经数据库化的数据，通常需要运用统计分析方法进行检验和清洗：

- 第一，录入错误清理。可以把双录入的数据输出为一个清理数据库，核对录入中出现的冲突数据。
- 第二，编码清理。对不在编码值范围的变量值进行清理。
 - 假设性别属性值的编码原本只有0和1，如果在数据表中出现了其它值，那就一定是哪里有错误了，就需要清理并且改正错误。
- 第三，逻辑清理。主要是针对基本事实逻辑的清理。
 - 比如样本为男性，在是否怀孕的访题下，变量值说明他有怀孕记录，这就是逻辑错误



4. 调查数据的数据库化（检验和清洗）

数据库检验和清洗还需要注意如下问题：

- 离群值：偏离了日常理解的范围，但实际上可能是有效值的一部分。
 - 男性怀孕令人奇怪，女性怀孕就没有什么让人奇怪的了，对不对？
 - 女性16岁-49岁之间怀孕都是正常的。如果数据显示有一位七十岁的老奶奶怀孕了， 有没有可能呢？
- 极大值和极小值， 都是需要再次确认的变量值。
- 无应答的处理，通过分析已经应答的数值，确定对无应答的处理方式，比如差值。
- 变量的再编码，在数据的清理中也可以产生衍生变量。
 - 比如受教育程度或者年龄的重新分组
 - 比如依据受教育程度和收入来建构社会经济地位



4. 调查数据的数据库化（清单）

正常的完成了数据库化的问卷数据，至少应该包括以下的文件：

1. 调查问卷（已经有了）
2. 调查问卷的数据库编码手册（已经有了）
3. 两个数据库，一个是完成问卷的数据库，一个是未完成问卷的数据库。
4. 样本数据库，通常抽样完成以后，一定有一个数据库。这个数据库包括了用于抽样的变量、抽样单位、分层变量、权重变量等，这些应该是分析研究之前已经有的数据。
5. 抽样报告、实施报告，这两份报告用于判断数据质量，制订分析策略。
6. 完成的、未完成问卷数量的统计表。通常用表格方式展示出来。
7. 数据清理报告，对变量的可分析性要进行说明。



B. 访谈调查数据的数据库化（主要步骤）

对访谈调查的数据，在完成了访谈笔记的整理、格式化、归档、清理之后，在用于分析之前也需要把相关的信息录到数据库中。虽然不一定可以像问卷调查数据那样完全的数据库化，至少访谈记录与整理信息应该数据库化。

- 第一步，编码。
 - 记录信息的编码（重点工作）
 - 记录内容的编码（如果要进行文本分析，则需要此步骤）
- 第二步，录入。
 - 录入访谈记录信息，便于检索，也便于查找。
 - 如果要做内容分析，访谈内容就需要全部地录入。
- 第三步，清理。一般需要逐行核查。内容数据是没有办法采用统计分析方法的进行核查。



B. 访谈调查数据的数据库化（编码）

访谈数据的编码有两类：

- 访谈记录信息的编码。基本变量有记录编号、访谈时间、地点、人物、主题、位置图。如果有日志信息，也需要把日志信息加入其中。
- 访谈记录的编码。如果希望编码的程度可以直接应用到内容分析软件的分析，那么就需要学习专门的课程，不同的分析软件对编码的要求是不一样的。



B. 访谈调查数据的数据库化（录入）

访谈数据的录入工具：

- 要是涉及到数字数据的，就可以使用Excel、SPSS、Stata、Statistica、r等等
- 对文本数据，就可以使用Word，当然也可以使用Numbers和Pages。
- 对访谈内容，还可以采用内容分析软件，比如Nvivo、Aquad、ATLAS.ti和Qualrus。

访谈数据录入的几个要点：

- 录入策略问题。对于访谈记录信息的录入，尽量采用标准化的格式，目的是便于交换、便于交流。
- 文本格式问题。一般可以先转录为纯文本格式，注意纯文本格式有一个编码问题，最好采用通用的编码，比如Unicode。



B. 访谈调查数据的数据库化（清单）

访谈数据的数据库化产出也有一份清单，至少要有以下的数据文件：

1. 调查提纲或者访谈提纲，或者访谈设计。
2. 访谈记录的整理、清理的数据库
3. 访谈内容的数据库
4. 访谈记录的数字化，也就是数字化的过程及报告
5. 最后还有清理报告



C. 观察数据的数据库化（主要步骤）

观察数据怎么数据库化呢？主要也是三个步骤：

- 第一步，编码。
 - 观察调查数据的编码与其他编码不一样的地方在于观察记录信息比访谈记录信息要丰富得多。当然对观察记录的内容，如果希望用作分析素材，也需要编码。
- 第二步，录入。
 - 在大多数情况下，主要录入观察记录信息，同样，如果要把观察记录的内容作为统计分析的素材，那么也需要把它录到数据库中。
- 第三步，清理。
 - 同样在录入完成之后，要对已经录入的数据进行核查，如果有观察记录的内容，就需要对已经数据库化的内容做仔细的核查，确保内容准确。



C. 观察数据的数据库化（编码）

观察数据的编码主要包括两个方面：

- 观察记录信息的编码。基本变量包括记录编号、观察的时间、地点、事件、主题，还有观察媒体（望远镜/摄像机/眼睛）。如果有日志信息，也可以把日志信息列入其中。
- 观察记录内容的编码。即使观察记录的内容不会作为统计分析的素材，最好还是录入为数据化的文本文件，便于交流。



C. 观察数据的数据库化（录入）

观察记录的录入：

- 文本数据、数字数据的录入。采用word或pages录入。
- 图片数据的录入。可以采用类似于Adobe的Lightroom之类的数据库。可以先扫描，再录入记录信息。
- 视频数据的录入，则可以运用类似于Adobe Premier之类的编辑库。
- 音频数据的录入，也可以寻找适用的音频数据库。



C. 观察数据的数据库化（清单）

一份完整的数据库化的观察数据的数据库，至少要提供以下的数据文件：

1. 观察提纲或者观察设计；
2. 观察记录的整理、清理数据库；
3. 观察内容数据库；
4. 观察记录数据数字化、数据库化过程的数据；
5. 清理报告。



D. 文献数据的数据库化

文献数据一般情况下原本就来源于数据库。因此，运用原来数据库的数据，是文献数据库的特点。文献数据的数据库化包括三个步骤：

1. 编码。指的是文献信息的编码，而不是文献内容的编码，文献信息就是编目信息，文献内容就是文献记载的内容。
2. 录入。就是把原来数据库的文献编目信息和文献内容抄录到研究用的文献数据库中去。
3. 清理。就是在数据录入完成以后，对录入的数据进行核查、清理，包括完整性检查。



- 文献记录信息的录入和管理。
 - 基本变量主要有作者、篇名、时间、载体、存放、DOI，或者ISBN，或者ISSN等。
 - 文献记录的编码可以直接运用文献记录的原始编码，一些数据库化的数据，比如jasdo，还支持编码的数据直接导出。
 - 专门的信息录入和文献管理软件：Endnote和papers。
- 文献内容信息的录入和管理。
 - 主要管理的是文献内容、阅读笔记、思路图谱、总结要点等
 - 专门的内容录入和关系管理软件：onenote、Mindmanager、印象笔记等。



6. 痕迹数据的数据库化（简要）

痕迹数据的数据库化，无论是Map-Reduce的产出，还是网页爬取的数据的整理、清理时的产出，都是基于变量的数据，还没有把变量数据串起来，变成基于样本的数据。

样本在变量上的变异是分析工作的基础，数据库化需要做的工作就是把变量数据串起来，变成类似于样本数据的数据。串起来的方法很多，技术性也很强，基本上依靠脚本来完成。

如果从大数据中抽取数据，由于无需数据录入，故数据库化只有两个步骤可做：

1. 编码。通常原有的数据就已经有编码了，这个手续要做的就是要么确认使用原来的编码，要么呢，因为特殊的原因，需要重新编码，何去何从，完全取决于计算的需要。
2. 清理。与其他调查数据的清理不同，这里主要是在确认编码以后，确认数据的可计算性，也就是格式化、结构化在转化中没有发生问题，以及是否可以直接运用于分布式并行计算或者单机计算。



二手数据的数据库化（实例分享）

研究议题：旱区农业科技资源配置情况研究。具体研究内容如下：

- 2.1 科技装备
 - 2.1.1 农业机械动力
 - 2.1.2 农用拖拉机
 - 2.1.3 农用灌溉机械
 - 2.1.4 农用收获机械
 - 2.1.5 农业化学要素
- 2.2 科技投入
 - 2.2.1 公共财政投入
 - 2.2.2 RD研发投入
- 2.3 科技计划
 - 2.3.1 重大基础类科技计划
 - 2.3.2 国家自然科学基金
 - 2.3.3 农业综合开发投入
- 2.4 科技条件
 - 2.4.1 国家工程技术研究中心
 - 2.4.2 国家重点实验室
- 2.5 科技服务
 - 2.5.1 国家农业科技园区
 - 2.5.2 技术示范转移机构



资料和数据

研究对象：旱区16个省份——北京、天津、河北、山西、内蒙古、辽宁、吉林、黑龙江、山东、河南、西藏、陕西、甘肃、青海、宁夏、新疆

文本资料：政府公开资料、公共信息、图书、文献...

数据资料：统计年鉴、网页数据、商业数据库信息...



数据整理

文件夹管理：

- 1 文献资料文件 (material)：收集到的各种相关资料
(.xlsx、.word、.pdf、.html、.png等)
- 2 粗制的数据文件 (raw data)：摘录、数值化 (.xlsx)
- 3 提取的数据文件 (extract data)：整合、合并 (.xlsx)
- 4 加工的数据文件 (process data)：更新、维护 (.xlsx)
- 5 分析的数据文件 (analysis data)：调用、子集化 (.xlsx)



0. 文献资料

- 囊括了研究涉及的全部材料
- 分门别类在各个文件夹下
- 形成目录树
- 文件以原始状态存放
- 格式各种各样

课题研究 > 6 参加各类课题 > 2012-旱区农业技术报告 > 数据		
名称	修改日期	类型
00 科技统计年鉴	2019/8/14, 星期...	文件夹
00 中国机械工业统计年鉴	2019/9/7, 星期六...	文件夹
00 中国农村统计年鉴	2019/8/23, 星期...	文件夹
00 中国统计年鉴	2019/8/23, 星期...	文件夹
0.0 all data	2019/7/15, 星期...	文件夹
01 农业农村统计年鉴	2019/8/1, 星期四...	文件夹
2.1 科技投入	2019/8/1, 星期四...	文件夹
2.2 科技条件	2017/9/27, 星期...	文件夹
2.3 科技服务机构和组织	2019/7/15, 星期...	文件夹
973&863	2019/7/15, 星期...	文件夹
创新人才推进计划	2019/7/15, 星期...	文件夹
第一章 政策	2019/7/15, 星期...	文件夹
高技术产业和企业	2019/7/15, 星期...	文件夹
国家产业技术创新战略联盟 (2012-2...	2019/7/15, 星期...	文件夹
国家高新区	2019/8/1, 星期四...	文件夹
国家工程技术研究中心	2019/7/15, 星期...	文件夹
国家工程研究中心 (发改委)	2019/7/15, 星期...	文件夹
国家级科技企业孵化器	2019/7/15, 星期...	文件夹
国家级示范生产力促进中心	2019/7/15, 星期...	文件夹
国家技术转移示范机构	2019/7/15, 星期...	文件夹
国家农业科技园区	2019/8/1, 星期四...	文件夹
国家重大科学仪器设备开发专项	2019/7/15, 星期...	文件夹
国家重点实验室	2019/8/1, 星期四...	文件夹
科技部火炬中心	2019/7/15, 星期...	文件夹
科技基础性工作专项	2019/7/15, 星期...	文件夹
科技特派员	2019/7/15, 星期...	文件夹
农业综合开发	2019/7/15, 星期...	文件夹
其他专项	2019/7/15, 星期...	文件夹
全国科技经费投入公报	2019/7/15, 星期...	文件夹
政策引导类	2019/7/15, 星期...	文件夹
重大科技创新基地	2019/7/15, 星期...	文件夹

文献资料文件夹



0. 文献资料I-1

- 历年的《中国科技统计年鉴》
- 数据来源：[人大经济论坛](#)；[中国知网-统计年鉴数据库](#)
- 部分年鉴数值化 (.xls)
- 部分年鉴仅是数字化 (.caj)
- 每本年鉴都有目录
- 年鉴中仅部分内容跟研究相关

课题研究 > 6 参加各类课题 > 2012- 旱区农业技术报告 > 数据 > 00 科技统计年鉴			
名称	修改日期	类型	大小
excel4 高等学校	2019/7/15, 星期...	文件夹	
excel5 高技术产业	2019/7/15, 星期...	文件夹	
excel6 国家科技计划	2019/7/15, 星期...	文件夹	
excel7 科技活动成果	2019/7/15, 星期...	文件夹	
excel9 国际比较	2019/7/15, 星期...	文件夹	
中国高技术产业统计年鉴2017	2018/8/29, 星期...	文件夹	
中国高技术产业统计年鉴20172	2018/8/29, 星期...	文件夹	
中国科技统计年鉴2009	2019/7/15, 星期...	文件夹	
中国科技统计年鉴2010	2019/7/15, 星期...	文件夹	
中国科技统计年鉴2011	2019/8/14, 星期...	文件夹	
中国科技统计年鉴2012	2019/8/14, 星期...	文件夹	
中国科技统计年鉴2013	2019/8/14, 星期...	文件夹	
中国科技统计年鉴2014	2019/8/14, 星期...	文件夹	
中国科技统计年鉴2015	2019/8/14, 星期...	文件夹	
中国科技统计年鉴2015 (2)	2019/8/14, 星期...	文件夹	
中国科技统计年鉴2016	2018/8/22, 星期...	文件夹	
中国科技统计年鉴2017	2019/7/15, 星期...	文件夹	
中国科技统计年鉴2018	2019/8/1, 星期四...	文件夹	
2016 6-01.xls	2016/10/30, 星期...	Microsoft Excel ...	40 KB
2017年中国科技统计年鉴.zip	2018/8/21, 星期...	ZIP 压缩文件	2,105 KB
中国高技术产业统计年鉴2017.rar	2018/8/29, 星期...	RAR 压缩文件	2,823 KB
中国高技术产业统计年鉴20172.rar	2018/8/29, 星期...	RAR 压缩文件	49,394 KB
中国科技统计年鉴2012 (在线).caa	2014/6/16, 星期...	CAA 文件	1 KB
中国科技统计年鉴2013 (在线).caa	2014/6/16, 星期...	CAA 文件	1 KB
中国科技统计年鉴2014 (在线).caa	2015/10/13, 星期...	CAA 文件	1 KB
中国科技统计年鉴2014.rar	2015/7/8, 星期三...	RAR 压缩文件	1,312 KB
中国科技统计年鉴2015.zip	2016/10/27, 星期...	ZIP 压缩文件	45,793 KB
中国科技统计年鉴2016.zip	2018/8/22, 星期...	ZIP 压缩文件	1,725 KB
中国科技统计年鉴2016readonline.caa	2017/9/27, 星期...	CAA 文件	1 KB
中国科技统计年鉴2018.zip	2019/7/31, 星期...	ZIP 压缩文件	2,732 KB

子文件夹



0. 文献资料I-2

- 历年《国家工程技术研究中心》资料
- 数据来源：[科技部网站](#)
- 部分资料以年度报告呈现 (.pdf)
- 部分资料以公开网页呈现
(.html、.doc)
- 资料发布时间不确定
- 资料非标准化，需手工收集整理

课题研究 > 6 参加各类课题 > 2012- 旱区农业技术报告 > 数据 > 国家工程技术研究中心			
名称	修改日期	类型	28 个项目
年度报告 2012年	2019/7/1...	文件夹	
年度报告 2013年	2019/7/1...	文件夹	
年度报告 2014年	2019/7/1...	文件夹	
[国家工程技术研究中心] 2014 科技部关于2014年度国家工程技术研究...	2015/7/1...	Microsoft Word ...	
[国家工程技术研究中心] 2014 科技部关于2014年度国家工程技术研究...	2015/7/1...	PDF Document	
[国家工程技术研究中心信息网]2012年国家工程技术研究中心名单.doc	2014/6/1...	Microsoft Word ...	
[国家工程技术研究中心信息网]2012年国家工程技术研究中心组建项目...	2014/6/1...	Microsoft Word ...	
[国家工程技术研究中心信息网]2012农业领域发展情况.pdf	2014/6/1...	PDF Document	
[国家工程技术研究中心信息网]附件1：2012年国家工程技术研究中心第...	2014/6/1...	Microsoft Word ...	
[国家工程技术研究中心信息网]附件2：2012年提出警告或撤销称号国家...	2014/6/1...	Microsoft Word ...	
[国家工程技术研究中心信息网]科技部关于国家工程技术研究中心第四次...	2014/6/1...	PDF Document	
[科技部]2013年国家工程技术研究中心组建计划表.doc	2014/6/1...	Microsoft Word ...	
0资料来源.txt	2014/6/1...	文本文档	
2014年度报告_国家工程技术研究中心.pdf	2016/10/...	PDF Document	
2015年度报告_国家工程技术研究中心.pdf	2016/11/...	PDF Document	
2016年报告_国家工程技术研究中心2014年度报告.pdf	2016/10/...	PDF Document	
2016年度报告_国家工程技术研究中心.pdf	2018/8/2...	PDF Document	
分析_计划组建名单.xlsx	2017/11/...	Microsoft Excel ...	
国家工程技术研究中心名单.xls	2016/11/...	Microsoft Excel ...	
计划组建名单.xlsx	2015/10/...	Microsoft Excel ...	
科技部关于2012年度国家工程技术研究中心立项的通知.pdf	2014/6/1...	PDF Document	
科技部关于2013年度国家工程技术研究中心立项的通知.pdf	2014/6/1...	PDF Document	
评估2012 附件1：国家工程技术研究中心第四次运行评估结果（优秀、...	2014/6/2...	Microsoft Word ...	
评估2012 附件2：提出警告或撤销称号国家工程技术研究中心名单.doc	2014/6/2...	Microsoft Word ...	
评估2012 国家工程研究中心第四次评估.xlsx	2014/6/2...	Microsoft Excel ...	
评估2012 科技部关于国家工程技术研究中心第四次运行评估结果的通知...	2014/6/2...	PDF Document	
新建 Microsoft Word 文档.docx	2018/8/2...	Microsoft Word ...	
新建文本文档.txt	2018/8/2...	文本文档	

文献《国家工程技术研究中心》



0. 文献资料I-I-I

- 《中国科技统计年鉴2018》
- 数据来源：[人大经济论坛](#)
- 该年鉴已数值化 (.xls)
- 年鉴统计资料依次以.xls格式呈现
- 具体文件含义可以查看目录
- 年鉴中仅部分.xls跟研究相关，需要提取出来

名称	修改日期	类型	大小
01-25.xls	2019/1/16, 星期...	Microsoft Excel ...	42 KB
01-26.xls	2019/1/16, 星期...	Microsoft Excel ...	50 KB
01-27.xls	2019/1/16, 星期...	Microsoft Excel ...	58 KB
02-01.xls	2019/1/16, 星期...	Microsoft Excel ...	46 KB
02-02.xls	2019/1/16, 星期...	Microsoft Excel ...	42 KB
02-03.xls	2019/1/16, 星期...	Microsoft Excel ...	42 KB
02-04.xls	2019/1/16, 星期...	Microsoft Excel ...	41 KB
02-05.xls	2019/1/16, 星期...	Microsoft Excel ...	42 KB
02-06.xls	2019/1/16, 星期...	Microsoft Excel ...	43 KB
02-07.xls	2019/1/16, 星期...	Microsoft Excel ...	42 KB
02-08.xls	2019/1/16, 星期...	Microsoft Excel ...	46 KB
02-09.xls	2019/1/16, 星期...	Microsoft Excel ...	47 KB
02-10.xls	2019/1/16, 星期...	Microsoft Excel ...	47 KB
02-11.xls	2019/1/16, 星期...	Microsoft Excel ...	42 KB
02-12.xls	2019/1/16, 星期...	Microsoft Excel ...	43 KB

文献《中国科技统计年鉴》



0. 文献资料 I-I-I-I

- 《中国科技统计年鉴2018》
- 数据来源：[人大经济论坛](#)
- “表1-7 2017年中国RD支出类型”
- 原始表格有各种“烦人状况”！
 - 看行：空行？字符有空格？意外字符？
 - 看列：列变量？中英文？跨多行？
 - 看单元格：数值（number）还是文字（character）？

地 区	Region	R&D经费 内部支出			
		Total	基础研究 Basic Research	应用研究 Applied Research	试验发展 Experimental Development
全 国	National Total	176061295	9754893	18492095	147814307
东 部 地 区	Eastern Region	118848464	6435902	11611479	100801082
中 部 地 区	Middle Region	28201677	1090689	2742737	24368251
西 部 地 区	Western Region	21966359	1526173	2977123	17463063
东 东北 地 区	Northeast Region	7044796	702129	1160755	5181911
北 京	Beijing	15796512	2323632	3616704	9856177
天 津	Tianjin	4587227	336505	689647	3561075
河 北	Hebei	4520312	105087	379702	4035523
山 西	Shanxi	1482347	83269	260841	1138237
内 蒙 古	Inner Mongolia	1323278	37402	104900	1180976
辽 宁	Liaoning	4298825	305773	664538	3328515

文献-中国科技统计年鉴2018

-表I-7RD支出类型



A. 粗制数据 (raw data) |

- 《中国科技统计年鉴2010–2018》
- 数据来源：[人大经济论坛](#)
- 各年年鉴整合
- 不按年份，而按内容来管理文件夹
- 文件夹命名坚持用英文！

github > tech-report > data-raw				
□ 名称	修改日期	类型	大小	
agri-development	2019/8/18, 星期...	文件夹		
nation-yearbook	2019/8/23, 星期...	文件夹		
rural-yearbook	2019/8/26, 星期...	文件夹		
<input checked="" type="checkbox"/> tech-yearbook	2019/8/14, 星期...	文件夹		
raw-NERC-report-2016-lists.xlsx	2018/8/28, 星期...	Microsoft Excel ...	37 KB	
raw-pannel-NFSC-year2018.xlsx	2019/7/31, 星期...	Microsoft Excel ...	16 KB	
report-NERC-yearbook-2016.pdf	2018/8/28, 星期...	PDF Document	5,191 KB	
report-NFSC-yearbook-2012.pdf	2014/6/28, 星期...	PDF Document	580 KB	
report-NFSC-yearbook-2013.pdf	2014/6/28, 星期...	PDF Document	861 KB	
report-NFSC-yearbook-2017.pdf	2018/8/22, 星期...	PDF Document	750 KB	
report-NFSC-yearbook-2018.pdf	2019/7/31, 星期...	PDF Document	712 KB	
report-NFSC-yearbook-2018.xlsx	2019/7/31, 星期...	Microsoft Excel ...	1,573 KB	
report-NFSC-yearbook-2018-extra.pdf	2019/7/31, 星期...	PDF Document	713 KB	
report-SKL-firm-yearbook-2016.pdf	2018/8/28, 星期...	PDF Document	2,330 KB	
report-SKL-province-yearbook-2016.pdf	2018/8/28, 星期...	PDF Document	1,740 KB	
report-SKL-state-yearbook-2016.pdf	2018/8/28, 星期...	PDF Document	5,669 KB	

github > tech-report > data-raw > tech-yearbook > part01-over				
□ 名称	修改日期	类型	大小	
01-labor-hour	2019/8/14, 星期...	文件夹		
02-spend-intense	2019/8/14, 星期...	文件夹		
<input checked="" type="checkbox"/> 03-spend-inner	2019/8/16, 星期...	文件夹		
04-spend-outer	2019/8/14, 星期...	文件夹		
05-public-professionals	2019/8/26, 星期...	文件夹		

重新整理后的科技统计年鉴文件夹



A. 粗制数据 (raw data) 2

- 《中国科技统计年鉴2010–2018》
- 数据来源：[人大经济论坛](#)
- 表 “中国RD支出类型” (.xls)
- 取你所需！
 - 每年的表格来自每年的年鉴
 - 每年的表格单独命名
 - 文件命名要有规则
 - 确保每个文件的行列数据保持一致！

github > tech-report > data-raw > tech-yearbook > part01-over > 03-spend-inner			
□	名称	修改日期	类型
▼ Microsoft Excel 97-2003 工作表 (24)			
	2010.xls	2019/8/14, ...	Microsoft Excel...
	2011.xls	2012/12/3, ...	Microsoft Excel...
	2012.xls	2013/10/31, ...	Microsoft Excel...
	2013.xls	2019/8/16, ...	Microsoft Excel...
	2014.xls	2019/8/14, ...	Microsoft Excel...
	2015.xls	2016/11/21, ...	Microsoft Excel...
	2016.xls	2017/12/11, ...	Microsoft Excel...
	2017.xls	2019/1/16, ...	Microsoft Excel...

历年的RD支出类型 (2010-2017)





B. 精制数据 (extract data) |

- 《中国科技统计年鉴2010-2018》
- 表 RD支出类型 (2010-2017)
- 数据来源: [人大经济论坛](#)
- 依次读取整合每一年的表 中国RD支出类型.xlsx
 - 统一变量命名
 - 分别写入年份信息
 - 行合并年度文件数据
 - 确保数据是正确读取的!

名称	修改日期	类型
part01-01-machine-2010t2017.xlsx	2019/9/10, ...	Microsoft Excel 工作表
part01-01-machine-2010t2017-0.xlsx	2019/8/27, ...	Microsoft Excel 工作表
part01-02-fertilizer-2010t2017.xlsx	2019/8/27, ...	Microsoft Excel 工作表
part01-03-plastic-2010t2017.xlsx	2019/8/27, ...	Microsoft Excel 工作表
part01-04-peptide-2010t2017.xlsx	2019/8/27, ...	Microsoft Excel 工作表
part02-01-finance-public-budget.xlsx	2019/8/23, ...	Microsoft Excel 工作表
<input checked="" type="checkbox"/> part02-03-spend-inner-1activity.xlsx	2019/8/16, ...	Microsoft Excel 工作表
part02-03-spend-inner-2purpose.xlsx	2019/8/16, ...	Microsoft Excel 工作表
part02-03-spend-inner-3source.xlsx	2019/8/16, ...	Microsoft Excel 工作表
part02-05-professionals.xlsx	2019/8/26, ...	Microsoft Excel 工作表
part08-03-output-pull-amount-2017.xlsx	2019/9/23, ...	Microsoft Excel 工作表
part08-03-output-pull-funds-2017.xlsx	2019/9/23, ...	Microsoft Excel 工作表
part08-output-techmarket-2017.xlsx	2019/9/23, ...	Microsoft Excel 工作表
SKL-firm-2016.xlsx	2018/8/28, ...	Microsoft Excel 工作表
SKL-province-2016.xlsx	2018/8/28, ...	Microsoft Excel 工作表
SKL-state-2016.xlsx	2019/8/18, ...	Microsoft Excel 工作表
techyearbook-part05-industry-01-RD-activity.xlsx	2019/9/24, ...	Microsoft Excel 工作表
techyearbook-part05-industry-02-new-product.xlsx	2019/9/24, ...	Microsoft Excel 工作表
techyearbook-part05-industry-03-patent.xlsx	2019/9/24, ...	Microsoft Excel 工作表
techyearbook-part05-industry-04-tech-renew.xlsx	2019/9/24, ...	Microsoft Excel 工作表
techyearbook-part05-industry-05-investment.xlsx	2019/9/24, ...	Microsoft Excel 工作表
techyearbook-part05-industry-06-trade.xlsx	2019/9/24, ...	Microsoft Excel 工作表

提取整合后的RD支出类型 (2010-2017)



B. 精制数据 (extract data) 2

- 《中国科技统计年鉴2010-2018》
- 表 RD支出类型 (2010-2017)
- 数据来源: [人大经济论坛](#)
- 基本保持原来的数据形态:
 - 看行(257行): 无空行、地区字符正确标准
 - 看列: 列变量统一命名
 - 看单元格: 全部是数值(number)

A	B	C	D	E	F	G	H
1	province	v4_zh_nbzc_hj	v4_zh_nbzc_jcyj	v4_zh_nbzc_yyyj	v4_zh_nbzc_syfz	year	
20	湖南	1865583.7	69103.9	249105.3	1547372.5	2010	
21	广东	8087477.6	167217.5	373206.6	7547055.5	2010	
22	广西	628696.2	36005.4	95587.8	497106	2010	
23	海南	70203.5	10680.7	24800.9	34723.9	2010	
24	重庆	1002663.3	64983.6	131762.9	805916.7	2010	
25	四川	2642695.3	154231.8	829966.4	1658499	2010	
26	贵州	299664.6	21804.8	36001.6	241860.2	2010	
27	云南	441671.8	55210.9	110390.5	276073.4	2010	
28	西藏	14598.5	1977	5331.2	7290.3	2010	
29	陕西	2175042.2	101416.3	383491.6	1690132.3	2010	
30	甘肃	419384.6	56508.5	88006.4	274870.7	2010	
31	青海	99437.9	9795.2	15588.7	74055.1	2010	
32	宁夏	115101.3	9861.3	10336.2	94903.8	2010	
33	新疆	266545.4	14325.5	63005.1	189215.8	2010	
34	全国	86870092.6	4118142.5	10283899.3	72468050.8	2011	
35	北京	9366438.8	1085319.2	2280063.7	6001058.9	2011	
36	天津	2977580.2	130039.8	372976.7	2474565.6	2011	
37	河北	2013376.9	63349.4	257791.5	1692230	2011	
38	山西	1133926.3	27462.7	171676.8	934787.7	2011	
39	内蒙古	851685.3	16306	61627.5	773749.9	2011	
40	辽宁	3638347.6	116267.6	527147.1	2994934.9	2011	
41	吉林	891337.3	89293.4	268496.8	533549.1	2011	
42	黑龙江	1287788.1	125096.4	190438.9	972252.9	2011	
43	上海	5977130.7	377819	924251.6	4675064.2	2011	
44	江苏	10655109.1	234922.1	567978.7	9852207.4	2011	
45	浙江	5980824.4	136506.3	318300.3	5196021.8	2011	

提取整合后的RD支出类型 (2010-2017)



C. 加工数据 (process data) |

- 《中国科技统计年鉴2010-2018》
- 表 RD支出类型 (2010-2017)
- 数据来源: 人大经济论坛
- 需要继续对数据形态加工变形
- 目标是标准化的数据集！！？

github > tech-report > data-proc

名称	修改日期	类型	大小
basic-province.xlsx	2019/6/23...	Microsoft Excel 工...	9 KB
basic-telephone-code.xlsx	2018/8/28...	Microsoft Excel 工...	18 KB
basic-telephone-code-2018.xlsx	2018/8/28...	Microsoft Excel 工...	14 KB
basic-vars.xlsx	2019/8/3, ...	Microsoft Excel 工...	34 KB
basic-vars-2019-8-3.xlsx	2019/9/24...	Microsoft Excel 工...	50 KB
long-format-all-end2015.xlsx	2019/6/23...	Microsoft Excel 工...	436 KB
new-data.xlsx	2019/6/23...	Microsoft Excel 工...	193 KB
part01-02-fpp-2010t2017.xlsx	2019/9/2, ...	Microsoft Excel 工...	17 KB
part01-03-spend-inner-1activity-2010t2017.xlsx	2019/8/16...	Microsoft Excel 工...	31 KB
part01-03-spend-inner-2purpose-2010t2017.xlsx	2019/8/16...	Microsoft Excel 工...	31 KB
part01-03-spend-inner-3source-2010t2017.xlsx	2019/8/16...	Microsoft Excel 工...	30 KB
part01-05-professionals-2010t2017.xlsx	2019/8/26...	Microsoft Excel 工...	45 KB
part01-07-finance-public-budget2010t2017.xlsx	2019/8/23...	Microsoft Excel 工...	30 KB
part02-02-NFSC-2012t2017.xlsx	2019/7/31...	Microsoft Excel 工...	61 KB
part08-output-techmarket-end2017.xlsx	2019/9/23...	Microsoft Excel 工...	32 KB
techyearbook-part05-industry-01-RD-activity.xlsx	2019/9/24...	Microsoft Excel 工...	14 KB
techyearbook-part05-industry-02-new-product.xlsx	2019/9/24...	Microsoft Excel 工...	13 KB
techyearbook-part05-industry-04-tech-renew.xlsx	2019/9/24...	Microsoft Excel 工...	13 KB
techyearbook-part05-industry-05-investment.xlsx	2019/9/24...	Microsoft Excel 工...	14 KB
techyearbook-part05-industry-06-trade.xlsx	2019/9/24...	Microsoft Excel 工...	12 KB
update-ACEP-end2016.xlsx	2019/6/23...	Microsoft Excel 工...	198 KB
update-hightech-2016.xlsx	2019/6/23...	Microsoft Excel 工...	17 KB
update-part01-RD-year2016.xlsx	2019/7/31...	Microsoft Excel 工...	8 KB
update-part01-RD-year2017.xlsx	2019/7/31...	Microsoft Excel 工...	8 KB
update-part02-02-NFSC-year2018.xlsx	2019/8/18...	Microsoft Excel 工...	13 KB
update-part03-03-techmarket-year2016.xlsx	2019/6/23...	Microsoft Excel 工...	11 KB

已选择 4 个项

加工变形后的RD支出类型

(2010-2017)



C. 加工数据 (process data) 2

- 《中国科技统计年鉴2010–2018》
- 表 RD支出类型 (2010–2017)
- 数据来源: [人大经济论坛](#)
- 这是一份标准化的数据集！！！
 - 看行(1025行): 按年度(year)、按省份(province)
 - 看列: 4个变量被折叠对方为1列(variables)!
 - 看单元格: 全部数值被折叠对方为1列(value)!

	A	B	C	D	E	F	G	H	I	J
1	province	year	variables	value						
20	湖南	2010	v4_zh_nbzc_hj	1865584						
21	广东	2010	v4_zh_nbzc_hj	8087478						
22	广西	2010	v4_zh_nbzc_hj	628696.2						
23	海南	2010	v4_zh_nbzc_hj	70203.5						
24	重庆	2010	v4_zh_nbzc_hj	1002663						
25	四川	2010	v4_zh_nbzc_hj	2642695						
26	贵州	2010	v4_zh_nbzc_hj	299664.6						
27	云南	2010	v4_zh_nbzc_hj	441671.8						
28	西藏	2010	v4_zh_nbzc_hj	14598.5						
29	陕西	2010	v4_zh_nbzc_hj	2175042						
30	甘肃	2010	v4_zh_nbzc_hj	419384.6						
31	青海	2010	v4_zh_nbzc_hj	99437.9						
32	宁夏	2010	v4_zh_nbzc_hj	115101.3						
33	新疆	2010	v4_zh_nbzc_hj	266545.4						
34	全国	2011	v4_zh_nbzc_hj	86870093						
35	北京	2011	v4_zh_nbzc_hj	9366439						
36	天津	2011	v4_zh_nbzc_hj	2977580						
37	河北	2011	v4_zh_nbzc_hj	2013377						
38	山西	2011	v4_zh_nbzc_hj	1133926						
39	内蒙古	2011	v4_zh_nbzc_hj	851685.3						
40	辽宁	2011	v4_zh_nbzc_hj	3638348						
41	吉林	2011	v4_zh_nbzc_hj	891337.3						
42	黑龙江	2011	v4_zh_nbzc_hj	1287788						
43	上海	2011	v4_zh_nbzc_hj	5977131						
44	江苏	2011	v4_zh_nbzc_hj	10655109						

加工变形后的RD支出类型 (2010-2017)



D. 分析数据 (analysis data) |

- 《中国科技统计年鉴2010–2018》
- 完整的RD数据集(part01-over-2010t2017.xlsx)
- 数据来源：人大经济论坛
- 每一个数据子集被加工完成后，需要继续进行整合
- 目标是一个标准化的完整数据集！！？

github > tech-report > data-analysis				
名称	修改日期	类型	大小	
part01-07-finance-public-budget2010t2017.xlsx	2019/8/23...	Microsoft Excel ...	30 KB	
part01-over-2010t2017.xlsx	2019/8/16...	Microsoft Excel ...	82 KB	
part01-RD.xlsx	2019/7/31...	Microsoft Excel ...	22 KB	
part02-02-NFSC-renew2018.xlsx	2019/8/18...	Microsoft Excel ...	51 KB	

part01-RD.xlsx
Microsoft Excel 工作表



聚合各个子数据集为一个完整RD数据集 (2010-2017)



D. 分析数据 (analysis data) 2

- 《中国科技统计年鉴2010–2018》
- 完整的RD数据集(part01-over-2010t2017.xlsx)
- 数据来源：人大经济论坛
- 这是一份完整的、标准化的数据集！！！
 - 看行(3329行)：按年度(year)、按省份(province)
 - 看列：全部变量被折叠对方为1列(variables)！
 - 看单元格：全部数值被折叠对方为1列(value)！

	A	B	C	D	E	F	G	H
1	province	year	variables	value				
1016	重庆	2017	v4_zh_nbzc_syfz	3136824				
1017	四川	2017	v4_zh_nbzc_syfz	5136224				
1018	贵州	2017	v4_zh_nbzc_syfz	735063.2				
1019	云南	2017	v4_zh_nbzc_syfz	1220862				
1020	西藏	2017	v4_zh_nbzc_syfz	10566.1				
1021	陕西	2017	v4_zh_nbzc_syfz	3519604				
1022	甘肃	2017	v4_zh_nbzc_syfz	607087				
1023	青海	2017	v4_zh_nbzc_syfz	127149.4				
1024	宁夏	2017	v4_zh_nbzc_syfz	305381.9				
1025	新疆	2017	v4_zh_nbzc_syfz	428132.1				
1026	全国	2010	v4_zh_nbzc_rczc	59252528				
1027	北京	2010	v4_zh_nbzc_rczc	6559340				
1028	天津	2010	v4_zh_nbzc_rczc	1853558				
1029	河北	2010	v4_zh_nbzc_rczc	1251688				
1030	山西	2010	v4_zh_nbzc_rczc	744856.2				
1031	内蒙古	2010	v4_zh_nbzc_rczc	541394				
1032	辽宁	2010	v4_zh_nbzc_rczc	2512699				
1033	吉林	2010	v4_zh_nbzc_rczc	668321.7				
1034	黑龙江	2010	v4_zh_nbzc_rczc	1041582				
1035	上海	2010	v4_zh_nbzc_rczc	4204858				
1036	江苏	2010	v4_zh_nbzc_rczc	7292419				
1037	浙江	2010	v4_zh_nbzc_rczc	4323623				
1038	安徽	2010	v4_zh_nbzc_rczc	1340932				
1039	福建	2010	v4_zh_nbzc_rczc	1377018				
1040	江西	2010	v4_zh_nbzc_rczc	697913.9				
1041	山东	2010	v4_zh_nbzc_rczc	5807270				

聚合各个子数据集为一个完整RD数据集（2010–2017）



数据和变量关联与管理

	A	C	D	E	F	G	H	I	J	K	L	M	O
1	variables	short_chn	short_en	unit	block1	block2	block3	block4	chn_blk	chn_block	chn_block3	chn_block4	flag
182	v4_ztr_jf_RD			亿元	v4	ztr	jf	RD	科技	总投入	经费	R&D经费	v2018.6
183	v4_ztr_qd_RD			%	v4	ztr	qd	RD	科技	总投入	强度	R&D强度	v2018.6
184	v4_zh_nbzc_hj	合计	total	万元	v4	zh	nbzc	hj	科技	综合	内部支出	合计	v2019.8
185	v4_zh_nbzc_jcyj	基础研究	basic	万元	v4	zh	nbzc	jcyj	科技	综合	内部支出	基础研究	v2019.8
186	v4_zh_nbzc_yyyj	应用研究	apply	万元	v4	zh	nbzc	yyyj	科技	综合	内部支出	应用研究	v2019.8
187	v4_zh_nbzc_syfz	试验发展	test	万元	v4	zh	nbzc	syfz	科技	综合	内部支出	试验发展	v2019.8
188	v4_zh_qd_RD			%	v4	zh	qd	RD	科技	综合	强度	R&D强度	v2019.8
189	v4_zh_nbzc_rczc			万元	v4	zh	nbzc	rczc	科技	综合	内部支出	日常性支出	v2019.8
190	v4_zh_nbzc_rylwf			万元	v4	zh	nbzc	rylwf	科技	综合	内部支出	人员劳务费	v2019.8
191	v4_zh_nbzc_zcxzc			万元	v4	zh	nbzc	zcxzc	科技	综合	内部支出	资产性支出	v2019.8
192	v4_zh_nbzc_yqsb			万元	v4	zh	nbzc	yqsb	科技	综合	内部支出	仪器和设备	v2019.8
193	v4_zh_wbzc_hj		total	万元	v4	zh	wbzc	hj	科技	综合	外部支出	合计	v2019.8
194	v4_zh_wbzc_jnjg			万元	v4	zh	wbzc	jnjg	科技	综合	外部支出	对境内研究机构支出	v2019.8
195	v4_zh_wbzc_jngx			万元	v4	zh	wbzc	jngx	科技	综合	外部支出	对境内高等学校支出	v2019.8
196	v4_zh_wbzc_jnqy			万元	v4	zh	wbzc	jnqy	科技	综合	外部支出	对境内企业支出	v2019.8
197	v4_zh_wbzc_jwjg			万元	v4	zh	wbzc	jwjg	科技	综合	外部支出	对境外机构支出	v2019.8
198	v4_zh_nbzc_zfzj			万元	v4	zh	nbzc	zfzj	科技	综合	内部支出	政府资金	v2019.8
199	v4_zh_nbzc_qyzj			万元	v4	zh	nbzc	qyzj	科技	综合	内部支出	企业资金	v2019.8

变量命名是一门学问！



数据和变量关联与管理

- 原始文件没有变量?
- 变量形式与其含义?
 - 唯一识别变量名(variable): v4_zh_nbzc_hj、v4_zh_nbzc_jcyj、v4_zh_nbzc_yyyj、v4_zh_nbzc_syfz
 - 中文变量名(short_chn): 合计、基础研究、应用研究、试验发展
 - 英文变量名(short_eng): total、basic、apply、test
- 变量命名如何动态调整?
 - 备注变量系统的版本号(flag): v2018.6、v2019.8、v2019.9

2.8 数据质量

数据质量内涵与评估

影响数据质量的主要原因



数据质量（评判原则）

没有质量的数据，就是垃圾。“垃圾进，垃圾出”。

数据的质量会受哪些因素的影响？如何评估数据质量？

数据质量评估是一项专门的技术，对不同来源的数据，有不同的评估方法。判断数据质量的基本原则有三项：

1. 真实性。真实性指的就是数据确实来源于调查，与数据产生有关的过程真实存在，调查对象真实存在；访问、观察真实存在；应答、场景、文献真实存在。
2. 准确性。数据的调查人员准确按照研究设计在执行，准确地处理了调查对象和调查对象的反馈，或者是，准确地转录了原始数据。
3. 时效性。对于有时效要求的数据，还要考虑调查的实施过程是不是符合规定的时间要求。如果上述三项原则都能得到满足，就可以进一步考察数据的基本质量，那就是符合性。



数据质量（评判维度）

对于数据质量的评估，总体上有两个维度：

1. 正向评估，是与标准要求的距离到底有多远，也就是符合性问题。
2. 反向评估，就是误差的大小。



数据质量（误差分类1）

事实上，数据收集、整理、清理的每一个环节都有可能产生误差。

1. 覆盖性误差。就是涉及到调查对象的备选机会而可能产生的误差。抽样问卷调查、访谈调查、观察调查、文献调查都有可能产生覆盖性的问题。
2. 测量性误差。就是调查数据中可能产生的误差。调查大都涉及到测量的信度和效度。只要信效度有问题，那么测量性误差就可能存在。
3. 应答性误差。观察调查、文献调查看起来没有应答类型的问题，实际上不是。只要是访问员提出的要求都存在应答类型的问题。只是不同类型的调查，应答性误差的表现形式、计算方法不同。因此，应答性误差也是调查数据中可能存在的误差。
4. 抽样性误差，仅出现在抽样问卷调查中的一类误差。



数据质量（误差分类2）

以上误差，如果依据在调查活动中的可改进性来看，又可以被归纳为两类误差：

1. 随机误差，就是在调查活动中随机产生的误差。

- 比如访问员的不规范行为产生的误差。通过规范访问员行为就可以减少这一类型的误差。在表现形式上，这类误差会增大变量测量的方差。

2. 系统误差，是由设计因素影响所产生的误差。

- 比如测量工具带来的误差，由于测量工具有问题，导致凡是采用这个测量工具的调查都会产生同一类、甚至同样程度的误差。在表现形式上，这类误差会增大测量的偏移量，就是bias。

调查总误差：所有这些由数据收集、整理、清洗活动产生的误差的综合，被称为调查总误差。通常用均方误（MSE）来表示。



覆盖性误差（概念）

覆盖性误差，又称为抽样框误差，指的就是目标总体与抽样框总体不一致所导致的调查对象错位所产生的误差。

覆盖性误差存在于所有通过调查方法获取数据的研究活动中。

- 目标总体就是调查对象总体，有明确的调查对象所指。
- 抽样框总体，简称框总体，是用于抽样的所有调查对象的集合。
- 样本总体，是被抽中的，且被作为调查对象的集合。

文献调查中，已知需要查阅的涉及某件事的所有文献，在查阅之前，却打算把查阅文献的范围扩大或者缩小，这就产生了覆盖性误差。



覆盖性误差（来源）

1. 丢失或者重叠目标总体要素。

- 框总体小于或者看起来大于目标总体，进而让部分要素失去或者获得了多次被抽中的机会，这里既有覆盖过度的现象，也有覆盖不足的现象。
- 比如在“北京大学本科生入学机会地区不平等”的调查中，如果以已经入学的学生为总体，丢掉了某个院系，或者既用院系、又用地区做抽样框，就会产生丢失，或者重叠问题。

2. 在抽样框总体中，包含着非目标总体要素。

- 这会使得况总体看起来会大于目标总体，进而让目标总体的备选概率小于理论概率
- 比如“北京大学本科生入学机会的地区不平等”研究，把北京大学的保安纳入到了抽样框，就会让目标总体学生的备选概率降低。

3. 不正确的辅助信息。



覆盖性误差（影响）

那么覆盖性误差对调查误差到底会有怎样的影响呢？

- 如果是抽样问卷调查，那么就会通过影响等概率，进而影响到代表性，影响了代表性，就影响到数据质量。
- 在非抽样问卷调查中，虽然不存在影响等概率的问题，但覆盖性问题依然存在，只是表现形式不同而已。如果覆盖过度，虽然不会对调查数据质量造成可计算的影响，却可能会干扰研究判断，比如冲淡了真正对象的变异性或者影响。如果覆盖不足呢，则有可能对研究判断造成致命的影响。
 - 在文献调查中，缺失了最关键的文献就有可能会认为没有这类文献，进而出现错误判断。
 - 在访谈调查中，如果没有访问到事件的当事人，就有可能出现关键信息不全甚至缺失，进而也导致错误的判断。
 - 在观察调查中，漏掉了关键的场景，比如研究庙会的，却没有去观察某个庙会，就无法对场景的现象做正确的判断。



测量性误差

测量性误差，指来源于测量工具的误差，和运用测量工具的误差。

在测量长度的时候你拿着尺子来量，尺子很准，很可靠，不过呢你的眼神不好，测量过程就有可能带来误差。

如果工具不好，即使你非常认真，也会产生误差。如果工具很好，没有用好，也不行，也会产生误差。

两个来源的误差都会反映在测量的质量参数上来，这就是信度和效度。

- 信度测量：前后测信度、折半信度、复本信度、一致性信度。
- 效度测量：表面效度、准则效度（校标效度）、建构效度、内在效度。

信度和效度的测量是针对结构式测量的。事实上无结构式调查中也同样存在信度和效度问题。只是因为没有结构，对信度和效度的测量比较困难而已。



测量质量的检验I（信度）：概念

信度 (reliability)：是指测量工具的可靠性，也即使用同一个测量工具、重复测量同一个对象，得到相同结果的概率。

- 得到相同结果的概率越高，测量工具的信度也就越高。
- 信度对测量而言，就是测量工具的稳定性。
- “重测信度”，就是看前后之间有没有差异，前后之间的差异越小，信度就越高。



测量质量的检验I(信度)：实践类型

假设我们在做调查，用问卷在做调查，用访题在做测量。

- 垂直重复信度：又叫前-后测信度，在实践上一前一后测试两次。适合变量不随时间变化的测量。
- 水平重复信度：又称为复本信度，或等值信度，也是水平的重复测量。要求测量对象具有等价性。

假设我们有一组访题，一般是5-6道，或者6-7道访题。针对主观变量又如何检验测量的信度呢？



测量质量的检验I(信度)：计算办法A

折半信度法：如果访题的一致性很好，奇数题得分与偶数题得分之间的相关系数也应该很高，如果访题之间的一致性有问题，相关系数也不会高就说明访题的稳定性不高。

- 把访题编号，编成奇偶数。
- 对同一组对象，用奇数题和偶数题分别进行一次测量
- 计算奇数题偶数题得分的相关系数。
- 再用Spearman Brown公式计算信度。



测量质量的检验I(信度)：计算办法B

克隆巴赫系数法，一般记为“Cronbach α ”，主要运用了内部方差原理，也就是如果访题的内部方差越大，则测量的一致性也就越差。

一般表达式为：

$$\alpha = \frac{K}{K-1} \left(1 - \frac{\sum_{i=1}^K \sigma_{Y_i}^2}{\sigma_{X_i}^2} \right)$$

或者也可以表达为：

$$\alpha = \frac{K\bar{c}}{(\bar{v} + (K-1)\bar{c})}$$

其中， \bar{v} 表示每个题项之间的平均方差， \bar{c} 表示不同被测者之间在不同题项上的平均协方差。

- $\alpha \in [0.9, 1)$ 表明测量可靠性极高
- $\alpha \in [0.8, 0.9)$ 表明测量可靠性较好
- $\alpha \in [0.7, 0.8)$ 表明测量可靠性能够接受



测量质量的检验2(效度)：概念和类型

效度 (validity)：指的是测量工具是否正确和有效。

- 预测效度：一个试点的测量结果与另一个试点测量结果之间的相关程度，相关程度越高 预测效度也就越高。

一模、二模的成绩能在多大程度上预测高考成绩就是模考的预测效度。

- 同时效度：指的是测量结果与既有的有效测量之间的相关程度，相关程度越高同时效度也就越高

笔试与实际能力之间的关系既涉及到预测效度，也涉及到同时效度



测量质量的检验2(效度)：概念和类型

- 结构效度：指一组题在多大程度上可以测量到理论上期望的特征，或者说在多大程度上能测量到事物之间的关系模式。

一组题，能在多大程度上发现婚姻满意度与夫妻之间相互忠诚之间的关系模式。

- 内容效度：直接测量变量的属性，是指测量在多大的意义上包含了概念的含义。

身高和体重，用什么测？



测量质量的检验2（效度）：示例

效度的检验不像信度的检验，总是需要用到统计检验，大多数的情况下都是主观判断。也有复杂的效度测量，难度超出了课程的要求。

北京大学本科生入学机会地区不平等的研究案例：

- 把地区之间的差距操作化为地区之间的人均GDP，虽然测量起来比较容易可是测量的并不是每个地区人们可以用于教育的资源。
- 测量人均GDP倒是很稳定，信度很好，却没有很准确地测量到我们希望测量的、可以用于教育的资源。
- 如果追求数度，测量每一个毕业生家庭可以用于教育的资源，虽然测量到了要测量的内容，可是测量起来却很困难。



应答性误差（概念）

应答性误差，是指访员发出了调查请求，调查对象却没有做出回应或者做出应答，由此带来的直接后果就是调查数据的缺失，从而引起数据误差。

在不同的调查中，应答性误差的表现形式并不一样，

数据缺失如果是样本层面的、对象群体层面的、场景层面的、文献类别层面的，那么应答性误差就可以被理解为广义覆盖性误差中的一种。

即使我们获得了等概率样本，或者必须调查的对象列表，在调查中，调查对象拒访、场景不可及、文献不可及等等情况总是会有的。

即使接受了访问，场景也可及，文献也找到了，可是某几道访题受访者不作答，或者不知道如何作答；或者没有遇到具体的场景。

希望看婚庆，但没有遇到有人结婚；或者文献中的某几页缺失了。

如此，就相当于覆盖不足，或者数据缺失。



应答性误差（概念）

无应答从类型上看主要有两种。

1. 对象无应答：

- 在抽样问卷调查中，常常被称之为样本无应答，或者单元无应答，英文是unit nonresponse；
- 在非抽样问卷调查中，对象无应答被称之为“失访”，就是没有接触到、观察到或者访问到设计中需要调查的对象、文献、痕迹。

2. 某些议题没有得到应答：

- 在抽样问卷调查中，如果部分访题没有得到应答，就会被称之为选项无应答，又称之为项目无应答，英文叫item nonresponse。
- 在非抽样问卷调查中，指一个或者具体几个议题，没有“访到”，自己忘记了、遗漏了，或者缺失了。



应答性误差（应答率）

在抽样问卷调查中，应答率是评估数据质量的基本参数之一。应答率等于应答样本数除上样本总数，再乘上百分之一百。

从分子角度来看：

- 一种情形是完全应答，完成了所有应回答的访题。
- 另一种情形是如果只是部分地应答了，没有完成所有应该回答的访题呢，那么到底完成了多少算是应答了呢？通常会根据访题的数量算出一个百分数，也就是完成了百分之多少访题的应答率是多少。

从分母角度来看：

- 无效的样本，比如不符合样本约束条件的对象；
- 未接触到的样本，也不知道是不是符合样本的约束条件；
- 接触到了，却完全无应答的样本；
- 即使没有接触到，却被认为是有效的样本；



应答性误差（影响）

应答率对数据质量有什么影响呢？

假设应答率为 p , 无应答率其实就是 $1 - p$, 由于无应答既可能是随机现象，也可能是系统现象。

- 随机现象，比如某个访题遗漏了，某个样本遗漏了；
- 系统现象，比如高收入的人群完全接触不到。

因此，无应答对样本估计值的影响主要来自于满足约束条件的样本的无应答，对代表性的影响。

- 高收入人群完全访问不到就会造成这一部分人群没有样本，进而影响到让样本满足等概率性。



抽样性误差（内涵）

在抽样调查中，覆盖性误差、测量性误差、应答性误差，三类误差都是可计算的。

抽样调查中抽样性误差的来源

- 主要来自于制作抽样框时候形成的误差，比如对样本的覆盖性。换句话说，在抽样调查中，覆盖性误差其实是抽样性误差的一部分。
- 还有在抽样过程中形成的误差，比如分层、多阶段，尤其是在末端抽样中，采用的方法、抽样的人都有可能形成误差。

在文献调查中，因为使用二手文献、因为选择版本等所带来的误差；

在观察调查中，因为选择场景所带来的误差。

在访谈调查中，因为访谈对象变动所带来的误差。



抽样性误差（计算）

抽样误差的计算也是针对具体变量的。

抽样的目的是为了获得有代表性的样本；获得有代表性的样本是为了用样本推论总体，误差尽可能的小；而推论是针对具体变量的推论；可是任何一项调查，误差总是要体现在这个变量上的，没有变量，哪来的误差呢？

1. 均值的变异系数。等于样本均值除以标准误，也即 $\frac{\bar{X}}{\sigma}$ 。如果是比例值，则为 $\frac{p}{\sqrt{p(1-p)}}$ 。经验上，如果一项调查样本均值的变异系数小于50%，就认为质量是可以接受的。
2. 样本均值的相对方差。等于样本方差除上均值的平方，也即 $\frac{\bar{X}^2}{\sigma^2}$ 。如果是比例值，则为 $\frac{p}{p(1-p)}$ 。

本章结束

