

第一周作业

20307100013 蒋翌坤

1. 请从 Lohr, Sampling Design and Analysis, 3rd ed., 2022 § 1.9 的第 1~32 个案例中选择 1 个, 根据你的理解对此案例作简要描述, 可供参考的描述要点包括:
 - (1) 该案例希望通过调查达到哪些目的、了解哪些信息?
 - (2) 调查的目标总体、个体、抽样单元、抽样框、抽样总体各是什么?
 - (3) 列举一、二个需要调查的变量、需要推测的总体参数。
 - (4) 抽样及调查过程中可能存在哪些样本选择偏差、测量误差? 可能是哪些原因导致的? 哪些偏差、误差事先可以想办法防范或减小它们的影响?

选取案例 Exercise 9:

Entries in the online encyclopedia Wikipedia can be written or edited by anyone with internet access. This has given rise to concern about the accuracy of the information. Giles (2005) reported on a Nature study assessing the accuracy of Wikipedia science articles. Fifty subjects were chosen "on a broad range of scientific disciplines." For each subject, the entries from Wikipedia and Encyclopaedia Britannica were sent to a relevant expert; 42 sets of usable reviews were returned. The editors of Nature then tallied the number of errors reported for each encyclopedia. (Giles, J. (2005). Internet encyclopaedias go head to head. Nature 438, 900–901.)

Wikipedia 使人们能够通过互联网快速、便捷地获取信息。但是由于 Wikipedia 可以被互联网上的任何人编辑, 信息可能不准确。案例希望通过调查 Wikipedia 上关于科学的文章或词条, 通过相关专家的评审, 来判断 Wikipedia 上信息的准确性。

该调查的目标总体为 Wikipedia 上的所有科学文章或词条及专家的评审; 抽样总体为 Wikipedia 上的所有科学文章或词条及专家的评审; 个体为 Wikipedia 上的一篇科学文章或词条及专家的评审、抽样单元为个体、抽样框为 Wikipedia 上的一篇科学文章或词条、对应 Encyclopaedia Britannica 上的文章或词条及专家的评审。

该调查需要调查的变量可能有: Wikipedia 上的一篇科学文章或词条存在科学性错误的个数、Wikipedia 上的一篇科学文章或词条的信息准确性。需要推测的总体参数可能有: Wikipedia 上的一篇科学文章或词条的信息准确性均值及方差。

抽样及调查过程中可能存在以下样本选择偏差。调查员主观判断、经验抽样: 调查员选择和自己学科相关或者热门学科的 Wikipedia 文章或词条, 造成样本选择偏差; 按方便程度抽样: 有些小众学科可能无法找到相关专家进行评审, 因此抽样时只找调查员可以接触到的专家所处学科对应的 Wikipedia 文章或词条。抽样及调查过程中还可能存在以下测量误差: 调查员在进行调查时定义的准确性有歧义, 造成专家的误解或难以精确回答; 专家可能会对文章的准确性进行主观判断, 文章中表示相同的含义但不同的措辞可能造成专家对准确性的误判。

2. 请从报纸、杂志、书籍、网络等处或者自己的实习、科研工作中，找一个抽样调查的实际案例，如同第1题那样作个简要介绍。请注明案例的出处。

实际案例：美国 Current Population Survey（这是我在宏观经济学课程中了解到的一个抽样调查案例）。

出处：

https://en.wikipedia.org/wiki/Current_Population_Survey

<https://www.census.gov/programs-surveys/cps.html>

就业情况是衡量一个国家经济情况的重要指标，失业率和通货膨胀紧密相关。美国人口统计局每月会举行 Current Population Survey (CPS)，样本量大约为 60,000 户美国家庭。CPS 询问被调查者的有关工作、就业情况，希望通过样本数据来获取全美国的劳动就业情况。

该调查的目标总体为全美国 16 岁及以上的居民；个体为一位 16 岁及以上的美国居民；抽样单元分级，首先将美国分成 754 个地理区域（先以州为单位然后再划分小区域），再从这些区域中选取一定数量的家庭作为样本；抽样框、抽样总体均为目标总体。

该调查需要调查的变量可能有：被调查者是否有工作、如果被调查者有工作，周工作时长、薪酬等。需要推测的总体参数可能有：美国居民的就业、失业率、平均工资等。

抽样及调查过程中可能存在以下样本选择偏差。被调查者可以不做应答或拒绝被调查，这些被调查者可能有特质在样本数据中被忽略；被调查者全是以自愿原则参加调查的，这也会造成样本选择偏差。美国人口统计局无法避免这些样本选择偏差，尤其是无应答所造成的偏差，因此，美国人口统计局通过统计手段以减少这些偏差。抽样及调查过程中还可能存在以下测量误差：问卷的表述不清楚造成歧义；被调查者随意回答等。但是，美国人口统计局通过详细的问题解释并提供相关文档、控制问卷题目数量等措施以减少这些偏差。