

doi:10.13582/j.cnki.1674-5884.2016.09.025

统计学课程教学的思考

曾惠芳^a,熊培银^b

(湖南科技大学 a. 商学院; b. 信息与电气工程学院, 湖南 湘潭 411201)

摘要: 统计学是研究数据背后规律的一门科学。为使学生理解抽象的统计概念,培养统计思维,提高统计学修养,围绕统计学的概念,以生动形象的例子阐述数据的获得,数据背后规律的刻画,以及统计推断的思想。

关键词: 统计学; 规律性; 随机性; 统计思维

中图分类号: F224.9, O212

文献标志码: A

文章编号: 1674-5884(2016)09-0078-03

统计学是研究如何获取数据、分析数据、解释数据,从数据中提取信息、寻找规律性的科学。当今,对统计学的理解、研究和实际应用已经扩展到整个自然科学、社会科学、工程技术、管理、经济、艺术和文学等领域。统计学家威尔斯(H. G. Wells)曾指出:“就像读和写的能力一样,将来有一天统计的思维方法会成为公民的必备能力。”^[1]所以,学好统计学对学生来说具有深远的意义。统计学教学中应注重抽象概念与生动例子的结合,提高学生的学习兴趣。同时,培养学生的统计修养,使其学会用统计思维观察和认识社会现象^[2]。

1 什么是数据,如何获得数据

统计学可以应用于各领域。按萨维奇(L. J. Savage)的说法:“统计学基本上是寄生的,靠研究其他领域内的工作而生存。这不是对统计学表示轻视,这是因为对很多寄主来说,如果没有寄生虫就会死。对有的动物来说,如果没有寄生虫就不能消化它们的食物。因此,人类奋斗的很多领域,如果没有统计学,虽然不会死亡,但一定会变得很弱。”

统计学是一门非常实用的科学,它的任务就是处理数据。那么什么是数据?拿掷骰子来说,掷骰子会得到什么值,是个随机变量;而每次取得1~6点中任意点数的概率在理论上都是1/6(如果骰子均匀)。而在实际掷骰子过程中,如果掷100次,会得到100个由1~6点组成的数字串;再掷100次,又得到一个数字串,和前一次的结果很可能不一样。这些试验结果就是数据。所以说,数据是关于变量的观测值。

社会经济统计数据资料的来源主要有两种渠道:一种是通过直接的调查获得原始数据,这是统计数据的直接来源,一般称之为第一手或直接的统计数据;另一种是别人调查的数据,并将这些数据进行加工和汇总后公布的数据,通常称之为第二手或间接的统计数据。一切间接的统计数据都是从直接的、第一手数据过渡而来的。

我们处于一个大数据时代,每天都可以获得大量的信息。应该如何去认识这些信息?如何在具有欺骗性的数据海洋中找到可靠有用的资料?怎样凭借双眼就能识破虚假的统计资料并揭穿它?对此,可以从五个方面对数据进行探索检验分析:1)数据是如何收集、记录的?2)数据中含有测量误差和记录误差吗?有关测量值的概念和定义明确吗?观察值之间有什么区别吗?3)数据是真实的吗?是

收稿日期:20150722

基金项目:国家自然科学基金项目(41301421);湖南科技大学教学研究与改革项目(G31533)

作者简介:曾惠芳(1981-),女,湖南邵阳人,副教授,主要从事贝叶斯统计研究。

所调查的原样,还是以任何方式经过人工伪造、编纂或修改过的?是否由观察者自行决定删除了任何观察值?数据中是否存在任何或许会过度影响统计推断的异常值?4)提供信息的观察数据来自什么样的总体?作为抽样调查总体中所选定部分是否存在没有回答的(部分或全部)?数据信息是来自单一总体,还是混合总体?与抽出样本单位的识别和分类有关的因素都记录下来了吗?5)对所要调查研究的课题或是观察数据的性质是否存在任何先验信息?

任何对数据的探索和检验分析都是为了了解数据的性质,剔除测量误差、记录误差和异常值,检验先验信息的有效性,检测数据的真伪。数据的初始研究也用于检验一个指定的模型的有效性或是对进一步的数据分析选择一个更合适的随机概率模型或随机概率模型族^[3]。

2 什么是统计规律,如何刻画这种规律

统计学的重要作用是利用已有的数据来预测未来。利用已有的数据来预测未来,必须要找到一个恰当的统计模型,确切地说,就是一条分布曲线去拟合已有的数据,然后用拟合的分布去预测在未来的观察中各种值出现的可能性大小^[4]。

例 1:19 世纪末,骑兵部队是多数军队的一个组成部分,在骑兵部队内间或就有人被马踢死。俄国统计学家博特基维茨对普鲁士军队的 10 个骑兵部队在 1875 ~ 1894 年这 20 年间的此类死亡情况进行了记录,并对此进行了统计分析。表 1 给出了这 200 支骑兵部队一年中被马踢死的人数信息。

表 1 200 支骑兵部队一年中被马踢死的人数统计

每个部队每年的死亡人数	部队数
0	109
1	65
2	22
3	3
4	1
大于 4	0

在概率论中,大家对泊松分布产生的一般条件已有所了解,容易想到,200 支骑兵部队一年中被马踢死的人数,可以用一个泊松随机变量来近似描述。也就是说,可以假设 200 支骑兵部队一年中被马踢死的人数的分布 X 近似泊松分布。

现在的问题是:上面的数据能否证实 X 具有泊松分布的假设是正确的?卡方检验法是在总体 X 的分布未知时,根据来自总体的样本,检验关于总体分布的假设的一种检验方法。

提出假设 H_0 : X 服从参数为 λ 的泊松分布。根据观察结果,得参数 λ 的极大似然估计为

$$\lambda = \bar{X} = 0.61。$$

按参数 λ 为 0.61 的泊松分布,计算事件 $p_i = \Pr(X = i)$ 的概率, p_i 的估计为

$$\hat{p}_i = \frac{0.61^i}{i!}e^{-0.61}, i = 0, 1, 2, 3, 4, \cdots$$

将 $n\hat{p}_i < 5$ 的组予以合并,即将死亡人数是 3, 4 和大于 4 的组归并为一组。因 H_0 所假设的理论分布中有一个未知参数,故自由度为 $4 - 1 - 1 = 2$ 。按 $\alpha = 0.05$, 自由度为 $4 - 1 - 1 = 2$ 查卡方分布表得 $\chi_{0.05}^2(2) = 5.991$ 。由于卡方统计量实际观察值为 $0.32 < 5.991$ 。未落入否定域,故认为每年部队被踢死的人数 X 服从参数为 0.61 的泊松分布。

3 如何根据数据来推断统计规律

统计学的作用,一方面是对数据进行概要的描述,另一方面是基于数据做出推断,包括评价推断的有效性。所谓统计推断,是基于一个指定的随机概率模型来估计未知参数,进行相应的假设检验,预测未来的观测值,以及做出决策等的统计方法。统计分析的目的是“从观测得到的数据中提取有效的信息”。所记录的数据中有时有某种缺陷,如存在记录误差和异常值,有时甚至可能是伪造的,一个统计学者首先应做的是详细考察或交叉检验数据,以便发现可能有的缺陷并了解数据的特征。下一步则是利用先验信息和交叉核实技术,对数据提出一个合适的随机概率模型。基于被选择的模型进行数据推断分析,包括未知参数的估计,假设检验,对未来观测值的预报以及做出决策^[5]。

例2:一个鱼塘的主人想知道池塘中有多少鱼,他猜想数量在3 000条左右,但其实可能仅有它的一半,又或者是它的两倍。逐一数鱼是不可能的,那么他该怎么办?

假设池塘中鱼种的平均寿命大约为3年,所以花一个月左右的时间获取鱼的数量过程,其出生与死亡数对总数量不会产生大的影响。他先捕到400条鱼,贴上标签,再放回池塘。然后再从池塘捕300条鱼,其中60条贴有标签。利用样本比例估计总体比例,可得到鱼塘中贴有标签的鱼的比的最佳估计为

$$p = \frac{60}{300} = 0.2。$$

若鱼塘中鱼的数量为 N ,且有400条鱼被贴了标签,则总体比例真实值为 $\frac{400}{N}$ 。实际上,样本比例 p 是一个统计量,可以看作是一个服从两点分布的随机变量 X 的和,即

$$p = \frac{1}{300} \sum_{i=1}^{300} x_i。$$

其中,若捕到的鱼贴了标签 $x_i = 1$,若没有贴标签 $x_i = 0$ 。因此, p 也可以看作是一个样本均值统计量,根据中心极限定理,样本比例 p 服从均值为 π ,方差为 $\frac{\pi(1-\pi)}{300}$ 。根据正态分布的特征,可得

$$\Pr\left(1.96 \leq \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{300}}} \leq 1.96\right) = 0.95。$$

为了简便起见, $\sqrt{\frac{\pi(1-\pi)}{300}}$ 中的 π 用样本比例 p 代替。由此我们可以得到总体比例 π 的95%的置信区间为 $p \pm 1.96 \sqrt{\frac{p(1-p)}{300}}$ 。所以总体比例 $\pi = \frac{400}{N}$ 的95%的置信区间为 $0.2 \pm 1.96 \sqrt{\frac{0.2(1-0.2)}{300}}$ 。进一步,可以计算在95%的概率下,鱼塘中鱼的数量范围是 $[1625, 2601]$,或者说,我们有95%的把握,鱼塘中鱼的数量是在1625~2601之间变化。可能性最大的鱼塘中的数量 $N = 2\ 000$ 。

4 结语

统计学是一门十分实用的科学,它并不是枯燥的数据和公式的堆砌。为了提高学生的学习兴趣,调动学生学习的积极性与主动性,在教学的过程中,应强调统计学的实用性,通过生动形象的例子和故事来表达统计思想。同时,应注重实践教学,让学生学会关注社会问题,并通过统计调查和统计分析,提高学生的统计应用能力。

参考文献:

- [1] C R 劳. 统计与真理——怎样运用偶然性[M]. 北京:科学出版社,2004.
- [2] 孙晓祥,惠英杰. 数理统计课程教学中培养学生统计思维的几点认识[J]. 吉林农业科技学院学报,2013,22(1):80-82.
- [3] 刘旭华,田英,陈薇. 对研究生数理统计课程教学的思考与探索[J]. 高等农业教育,2010(7):76-78.
- [4] 云霞. 统计思维是现代人的标签[J]. 调研世界,2012(12):58-59.
- [5] 刘娟. 概率论与数理统计案例教学探讨[J]. 当代教育理论与实践,2014,6(12):36-37.

(责任校对 莫秀珍)