计算机时代统计推断

张晨阳

2018-08-02

献给……

呃，爱谁谁吧

目录

第一章　算法和推断 Algorithms and Inference **1**

1.1　例子：回归 A Regression Example 4

第二章　白苹风末 **7**

2.1　张老爷子 7

2.2　彭大将领 7

附录 **9**

附录 A　余音绕梁 **9**

表格

1.1 雷猴啊，iris 数据！ 6

插图

1.1 雷猴啊，散点图！ 6

前言

你好，世界。我写了一本书。这本书是这样的，第 一 章介绍了啥啥，第 二 章说了啥啥，然后是啥啥……

我用了两个 R 包编译这本书，分别是 **knitr** (Xie, 2015) 和 **bookdown** (Xie, 2018)。以下是我的 R 进程信息：

```
sessionInfo()
```

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS High Sierra 10.13
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRb
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/lib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets
## [6] methods   base
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.17    bookdown_0.7    digest_0.6.15
##  [4] rprojroot_1.3-2 backports_1.1.2 magrittr_1.5
```

```
##  [7] evaluate_0.10.1 highr_0.7       stringi_1.2.3
## [10] rstudioapi_0.7  rmarkdown_1.10  tools_3.5.0
## [13] stringr_1.3.1   xfun_0.3        yaml_2.1.19
## [16] compiler_3.5.0  htmltools_0.3.6 knitr_1.20
```

致谢

非常感谢谁谁以及谁谁对我的帮助。艾玛，要不是他们神一样的队友，我两年前就写完这本书了。

张三
于 A 村某角落

作者简介

上不了厅堂，下得了厨房。敲得了代码，逮得住蟑螂。

第一章　算法和推断 Algorithms and Inference

Statistics is the science of learning from experience, particularly experience that arrives a little bit at a time: the successes and failures of a new experimental drug, the uncertain measurements of an asteroid's path to-ward Earth. It may seem surprising that any one theory can cover such an amorphous target as "learning from experience." In fact, there are *two* main statistical theories. Bayesianism and frequentism, whose connections and disagreements animate many of the succeeding chapters.

统计学是一个从经验中学习的学科，尤其是每次得到一点点的经验：一种新型试验药的成功与失败，小行星驶向地球路径的不确定性测量。任何一个理论都能适用这样一个模糊的目标："从经验中学习"，这似乎令人惊讶。事实上，有两种主要的统计理论。贝叶斯主义和频率派，它们的联系和分歧推动了许多后续章节。

First, however, we want to discuss a less philosophical, more operational division of labor that applies to both theories: between the *algorithmic* and *inferential* aspects of statistical method, averaging. Suppose we have observed numbers $x_1, x_2, \ldots ...x_n$applying to some phenomenon of interest, perhaps the automobile accident rates in the n = 50 states. The *mean*

$$\overline{x} = \sum_{i=1}^{n} x_i/n$$

然而，首先，我们要讨论适用于两种理论的较少哲学，更具操作性的分工：统计方法的算法和推断方面之间的平均。假设我们观察到某些我们感兴趣的数：$x_1, x_2, \ldots...x_n$，就当是 50 个州汽车事故率，它们的均值是：

$$\overline{x} = \sum_{i=1}^{n} x_i/n$$

Summarizse the result in a single number.

将结果汇总到一个数值中。

How accurate is that number? The textbook answer is given in terms of the *standard error*,

那么得到数值的准确度是多少呢？教科书给出的答案是**标准误**，

$$\hat{se} = \left[\sum_{i=1}^{n} (x_i - \overline{x})^2 / (n(n-1))\right]^{1/2}$$

Here *averaging(1)* is the algorithm, while the standard error provides an inference of algorithm's accuracy . It is a crucial, aspect of the statistical theory that same data that supplied an estimate can also assess its accuracy.[1]

这里平均（1）是算法，而标准误差提供算法精度的推断。统计理论的一个重要方面是，用于估计的数据也可以评估估计的准确性。

Of course, $\hat{se}(3)$ is itself an algorithm, which could be (and is) subject to further inferential analysis concerning *its* accuracy. The point is that the algorithm come first and the inference follows at the second level of statistical consideration. In practice this means that algorithm invention is a more free-wheeling and adventurous enterprise, with inference playing catch-up as it strives to assess the accuracy, good or bad, of some hot new algorithmic methodology.

[1] "Inference" concerns more than accuracy: speaking broardly, algorithms say what the statistician does while inference says why he or she does it. - "推断" 不仅仅关注准确性：从广义上讲，算法说明统计学家所做的事情，而推理则说明为什么他或她这样做。

当然，$\hat{se}$ 本身就是一种算法，它可以 (而且是) 对其准确性进行进一步的推论分析。关键是算法首先出现，推断遵循统计考虑的第二个层次。在实践中，这意味着算法发明是一个更加自由和冒险的事业，当它努力评估一些热门的新算法方法的准确性，无论是好还是坏，推断都在追赶。

If the inference/algorithm race is a tortoise-and-hare affair, then modern electronic computation has bred a bionic hare. There are two effects at work here: computer-based technology allows scientists to collect enormous data sets, orders of magnitude larger than those that classic statistical theory was designed to deal with; huge data demands new methodology, and the demand is being met by a burst of innovative computer-based statistical algorithms. When one reads of "big data" in the news, it is usually these algorithms playing the starring roles.

如果推理/算法竞赛是一场龟兔赛跑，那么现代电子计算已经培育出了一只仿生兔子。这里有两个效应: 基于计算机的技术允许科学家收集大量的数据集，这些数据集的数量级要比经典统计学理论所要处理的数据大得多; 巨大的数据需要新的方法，而需求正被大量创新的基于计算机的统计算法所满足。当人们在新闻中读到 "大数据" 时，通常是这些算法扮演主角。

Our book's title, Computer Age Statistical Inference, emphasizes the tortoise's side of the story. The past few decades have been a golden age of statistical methodology. It hasn't been, quite, a golden age for statistical inference, but it has not been a dark age either. The efflorescence of ambitious new algorithms has forced an evolution (though not a revolution) in inference, the theories by which statisticians choose among competing methods. The book traces the interplay between methodology and inference as it has developed since the 1950s, the beginning of our discipline's computer age. As a preview, we end this chapter with two examples illustrating the transition from classic to computer-age practice.

我们这本书的标题，计算机时代的统计推断，强调了乌龟的一面。过去几十年是统计方法的黄金时代。这并不是统计推断的黄金时代，但也不是黑暗时代。雄心勃勃的新算法的繁荣期迫使推断 (尽管不是革命) 发生了变化。推断是统计学家在相互竞争的方法中选择的理论。这本书追溯

了方法论和推理之间的相互作用，从 20 世纪 50 年代开始，也就是我们学科计算机时代的开始，这本书就开始发展。作为一个预览，我们结束这一章的两个例子说明从经典到计算机时代的实践的转变。

1.1 例子：回归 A Regression Example

Figure 1.1 concerns a study of kidney function. Data points (x_i, y_i)have been observed for n = 157 healthy volunteers, with x_i the ith volunteer's age in years, and y_i acomposite measure "tot" of overall function. Kidney function generally declines with age, as evident in the downward scatter of the points. The rate of decline is an important question in kidney transplantation: in the past, potential donors past age 60 were prohibited, though, given a shortage of donors, this is no longer enforced.

图 1.1 是关于肾功能的研究。观测源自 $n = 157$ 的健康志愿者的数据点 (x_i, y_i) ，其中 x_i 第 i 个志愿者的年龄，y_i 是一个整体功能的对应值"tot"。在下方的散点图可以看出，肾功能通常随着年龄的增长而下降。下降的速度是肾脏移植的一个重要问题: 在过去，60 岁以上的潜在捐赠者是被禁止的，但是，由于缺乏捐赠者，这一规定不再被执行。

The solid line in Figure 1.1 is a linear regression

图 1.1 中的实线是一个线性回归，

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

fit to the data by least squares, that is by minimizing the sum of squared deviationston

以最小二乘法拟合数据，即在所有选择的 $(\beta_0\beta_1)$ 中最小化方差之和。

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

over all choices of (β_0, β_1). The least squares algorithm, which dates back to Gauss and Legendre in the early 1800s, gives $\hat{\beta}_0 = 2.86$ and $\hat{\beta}1 = -0.079$ as the least squares estimates. We can read off of the fitted

line an estimated value of kidney fitness for any chosen age. The top line of Table 1.1 shows estimate 1.29 at age 20, down to -3.43 at age 80.

可以追溯到 19 世纪早期的高斯和勒让德雷算法的最小二乘算法，它给出了 $\hat{\beta}_0 = 2.86$ 和 $\hat{\beta}_1 = -0.079$ 为最小二乘的估计值。我们可以从拟合线中读出任何选定年龄的肾脏适合度的估计值。表 1.1 的第一行显示了 20 岁时的估计值为 1.29,80 岁时的估计值为-3.43。

How accurate are these estimates? This is where inference comes in: an extended version of formula (1.2), also going back to the 1800s, provides the standard errors, shown in line 2 of the table. The vertical bars in Figure 1.1 are ± two standard errors, giving them about 95% chance of containing the true expected value of tot at each age.

这些估计有多精准？这就是推断的作用：公式（1.2）的扩展，也可以追溯到 19 世纪，提供了标准误，如表 2 第 2 行所示。图 1.1 中的竖线是 ± 两个标准误差，使得它们在每个年龄段包含 95 % 的真实预期值的可能性。

现在我们可以试试 **bookdown** 的一些初级功能了，例如图表。图 1.1 是一幅无趣的散点图，表 1.1 是一份枯燥的数据。

```
par(mar = c(4, 4, 1, .1))
plot(cars, pch = 19)
```

```
knitr::kable(
  head(iris), caption = ' 雷猴啊，iris 数据！ ',
  booktabs = TRUE
)
```

就这样，你可以一直编下去，直到编不下去。

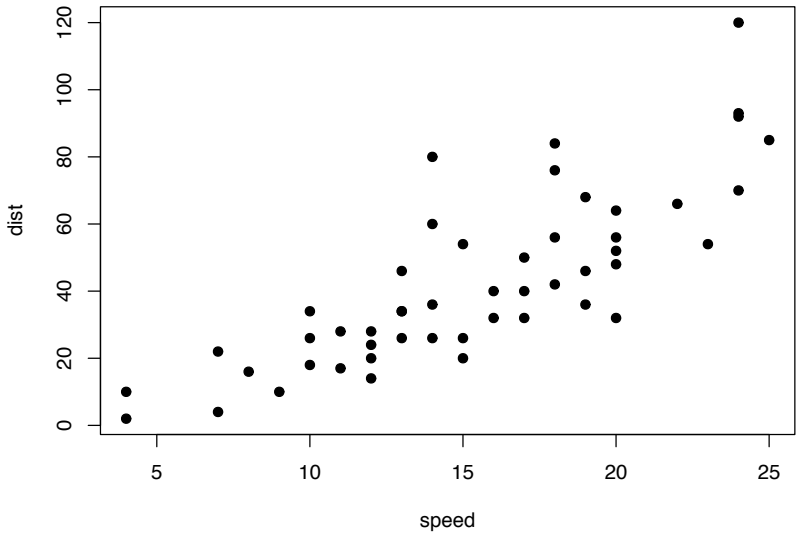

图 1.1: 雷猴啊，散点图!

表 1.1: 雷猴啊，iris 数据!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

第二章　白苹风末

瞎扯几句。

2.1　张老爷子

话说张老爷子写了一首诗：

> 姑苏开遍碧桃时，邂逅河阳女画师。
> 红豆江南留梦影，白苹风末唱秋词。

2.2　彭大将领

貌似大家都喜欢用白萍风这个意境。又如彭玉麟的对联：

> 凭栏看云影波光，最好是红蓼花疏、白苹秋老；
> 把酒对琼楼玉宇，莫孤负天心月到、水面风来。

嘿，玛尼玛尼哄。

附录 A　余音绕梁

呐，到这里朕的书差不多写完了，但还有几句话要交待，所以开个附录，再啰嗦几句，各位客官稍安勿躁、扶稳坐好。

参考文献

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.7.

索引

bookdown, ix

knitr, ix