《生物统计学》课程讲义

王超

目录

1 概率与概率分布

1.1 由二项分布推导正态分布

正态分布是自然科学与行为科学中的定量现象的一个方便模型。各种各样的心理学测试分数和物理现象比如光子计数都被发现近似地服从正态分布。尽管这些现象的根本原因经常是未知的，理论上可以证明如果把许多小作用加起来看做一个变量，那么这个变量服从正态分布（在 R.N.Bracewell 的 Fourier transform and its application 中可以找到一种简单的证明）。正态分

布出现在许多区域统计：例如，采样分布均值是近似地正态的，即使被采样的样本的原始群体分布并不服从正态分布。另外，正态分布信息熵在所有的已知均值及方差的分布中最大，这使得它作为一种均值以及方差已知的分布的自然选择。正态分布是在统计以及许多统计测试中最广泛应用的一类分布。在概率论，正态分布是几种连续以及离散分布的极限分布。

正态分布最早是棣莫弗在 1718 年著作的书籍的（Doctrine of Change），及 1734 年发表的一篇关于二项分布文章中提出的，当二项随机变量的位置参数 n 很大及形状参数 p 为 1/2 时，则所推导出二项分布的近似分布函数就是正态分布。拉普拉斯在 1812 年发表的《分析概率论》（Theorie Analytique des Probabilites）中对棣莫佛的结论作了扩展到二项分布的位置参数为 n 及形状参数为 1>p>0 时。现在这一结论通常被称为棣莫佛－拉普拉斯定理。拉普拉斯在误差分析试验中使用了正态分布。勒让德于 1805 年引入最小二乘法这一重要方法；而高斯则宣称他早在 1794 年就使用了该方法，并通过假设误差服从正态分布给出了严格的证明。将正态分布称作 “钟形曲线” 的习惯可以追溯到 Jouffret 他在 1872 年首次提出这个术语（Bell curve）用来指代二元正态分布。正态分布这个名字还被查尔斯 • 皮尔士、法兰西斯 • 高尔顿、威尔赫姆 • 莱克希斯在 1875 分别独立地使用。这个术语是不幸的，因为它反映和鼓励了一种谬误，即很多概率分布都是正态的。这个分布被称为 “正态” 或者 “高斯” 正好是史蒂格勒名字由来法则的一个例子，这个法则说 “没有科学发现是以它最初的发现者命名的”。

当二项随机变数的位置参数 n 很大及形状参数 p 为 1/2 时，则所推导出二项分布的近似分布函数就是正态分布。当然这个其实就是个极限问题，有兴趣之后我们可以具体讨论。但是这个结果确实是我们直观上可以相像的，当然你还是无法想像，那我们来看看这个计算机的模拟试验。

我们的 R3,R4,R5 分别是从 N=100，1000，10000 次二项分布中生成的，清晰的看到随着 N 的增加，这个分布越来越接近我们这个具有代表性的的这个正态分布了。

事实上，这个东西的严格的讲还有特别厉害的名字，中心极限定理，中心极限定理有着有趣的历史。这个定理的第一版被 法国 数学家 棣莫弗发现，他在 1733 年发表的卓越论文中使用 正态分布去估计大量抛掷硬币出现正面次数的分布。这个超越时代的成果险些被历史遗忘，所幸著名法国数学家拉普拉斯在 1812 年发表的巨著 Théorie Analytique des Probabilités 中拯

救了这个默默无名的理论。 拉普拉斯扩展了 棣莫弗的理论，指出二项分布可用正态分布逼近。但同 棣莫弗一样， 拉普拉斯的发现在当时并未引起很大反响。直到十九世纪末中心极限定理的重要性才被世人所知。1901 年，俄国数学家 里雅普诺夫用更普通的随机变量定义中心极限定理并在数学上进行了精确的证明。如今，中心极限定理被认为是 (非正式地) 概率论中的首席定理。

然而，正态分布真正走入人们视线的并不是由这个无聊的投硬币试验所得的二项分布的逼近，而是实实在在的工程误差分析中应用。据说 wiki 说，拉普拉斯在 误差分析试验中使用了正态分布。 勒让德于 1805 年引入 最小二乘法这一重要方法；而 高斯则宣称他早在 1794 年就使用了该方法，并通过假设误差服从正态分布给出了严格的证明。

之前我们说到高斯对测量误差研究中发现了正态分布，并且这项研究也成为了当代统计学的中重要的思想–最大似然发现的源头。下面我们来仔细看看，他是如何导出这个完美的分布的。

首先我们要解释几个概念，第一个是似然（Likelihood）。什么是似然，简单通俗的来讲就是，一系列的概率密度函数的乘积，说白了也就是还是一种特别的复合的 “概率”。比如对于正态分布，如果有独立同分布的观察值 $x_1, x_2, \ldots, x_n$，则其的似然为：

2 引言

中文 LaTeX 文档并非难题。当然这句话得站在巨人 CTeX 的肩膀上才能说，它让我们只需要一句

```
\documentclass{ctexart} % 或者 ctexrep/ctexbook
```

或者

```
\usepackage{ctex}
```

就轻松搞定中文 LaTeX 排版问题。

3 字体和选项

LaTeX 包**ctex**支持若干种字体选项，如果你是 **ctex** 老用户，请注意这里我们要求的最低版本是 2.2，你可能需要升级你的 LaTeX 包。从版本 2.0 开始，**ctex** 支持根据不同操作系统自动选择中文字体，简直是为人类进步作出了巨大贡献，我们再也不必费尽口舌向用户解释“啊，你用 Windows 啊，那么你该使用什么字体；啊，你用 Mac 啊，又该如何如何”。

下面的 YAML 元数据应该能满足多数用户的需求，主要设置两项参数：文档类为 `ctexart`（当然也可以是别的类），输出格式为 `rticles::ctex`，其默认 LaTeX 引擎为 XeLaTeX（真的，别纠结你的旧爱 PDFLaTeX 了）。

```
---
documentclass: ctexart
output: rticles::ctex
---
```

`rticles::ctex` 的参数都是普通的 `pdf_document` 参数，参见文档 **rmarkdown** 包的文档，这里就不赘述了。

Windows 和 Mac 用户应该都已经有自带的中文字体了。Linux 用户可以考虑 Fandol 字体，它号称是免费的，不过我们也没太搞清楚它的来头。如果你不想操心这些问题，我们强烈建议你卸载你当前的 LaTeX 套装（TeX Live 或 MiKTeX 或 MacTeX），换上 TinyTeX，一切将会自动化搞定。

```
devtools::install_github(c('rstudio/rmarkdown', 'yihui/tinytex'))
tinytex::install_tinytex()
```

4 R 代码段

R 代码用 R Markdown 的语法嵌入，即三个反引号开始一段代码```` ```{r} ```` 和三个反引号``` ``` ``` 结束一段代码：

```
options(digits = 4)
fit = lm(dist ~ speed, data = cars)
coef(summary(fit))
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  -17.579     6.7584  -2.601 1.232e-02
## speed          3.932     0.4155   9.464 1.490e-12
```

```
b = coef(fit)
```

上面回归方程中的斜率是 3.9324，完整的回归方程为：

$$Y = -17.5791 + 3.9324x$$

画图当然也是木有问题的啦，想画就说嘛，不说我怎么知道你想画呢？

```
par(mar = c(4, 4, .1, .1), las = 1)
plot(cars, pch = 19)
abline(fit, col = 'red')
```

请不要问我为什么图浮动到下一页去了，这么初级的 LaTeX 问题问出来信不信我扁你。

5 源代码控

这里提供的 rticles 模板可能由于种种原因不能满足客官的要求，LaTeX 用户就是这样永无止境地调格式（唉，跟 Word 用户到底有啥区别呢）。若真是需要调整，你可以复制一份默认模板去改。默认模板来自 Pandoc：https://github.com/jgm/pandoc/blob/master/data/templates/default.latex 它是一个文本文件。若熟悉 LaTeX 的话一看就明白，只不过里面有些 Pandoc 变量而已；若不熟悉 LaTeX 我们在这里说了也白说，花几天时间好好啃一啃 LaTeX 入门手册吧。

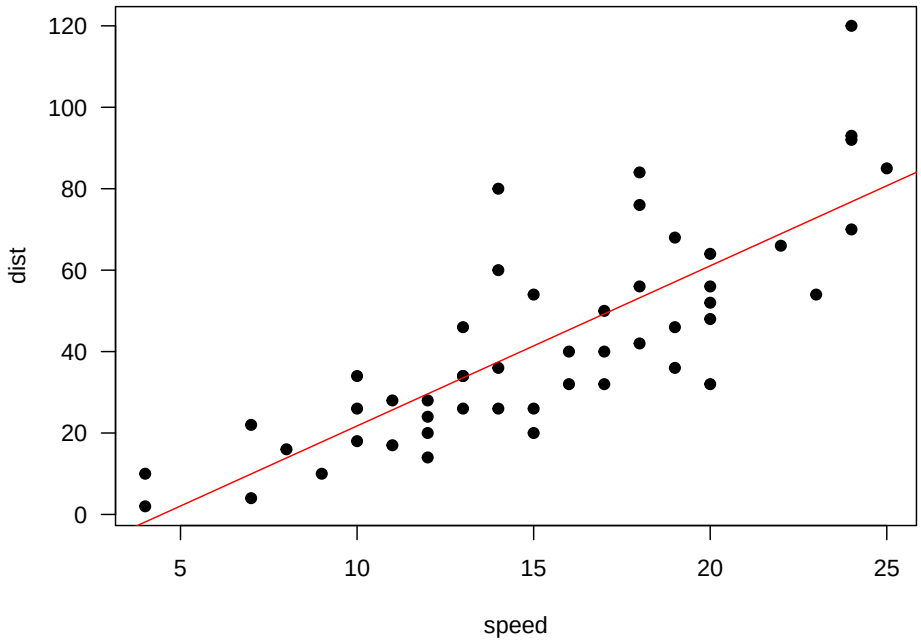

图 1: cars 数据散点图以及回归直线。

6 小结

事实证明我们可以理直气壮地通过 XeLaTeX 将中文 R Markdown 转化为 PDF 文档，麻麻再也不用担心我的论文满屏幕都是反斜杠，朕养完小白鼠之后终于不必先折腾三个小时 LaTeX 再开始写实验报告了：打开 RStudio，菜单 File > New File > R Markdown，然后从模板中选择 CTeX Documents，搞定。