统计图形和模拟视角下的模型理论解析

谢益辉*

2010年4月25日

摘要

统计模型往往来自于抽象的数学理论，但我们可以通过统计模拟和统计图形的手段去分析、解读统计模型，降低它们的抽象程度，绕过对统计使用者并不重要的数学细节，使之易于理解和学习；进一步，统计图形和统计模拟也可以作为统计建模前的启发工具和建模后的探索工具。本文的研究中心为统计模型，但研究角度并非传统的数学理论，而是分析了统计图形和统计模拟对统计模型在学习和应用上的辅助作用，并给出了大量示例。

第一节中，我们回顾了统计图形和统计模拟各自的发展和优势，并辅之以案例说明它们在建模中的不可替代的作用；第二节中，我们首先以t检验为例，用图形和模拟分析了模型方法假设条件的稳健性，其次以多元回归为例，用反例解除人们对模型的常见误解，然后以交互作用为例，提出对经典图示方法的补充，再以最小中位数平方回归为例，用图形和模拟的方法直观说明了模型的缺陷所在，最后以离群点检测为例，提出了一种模拟的思路，可作为对传统离群点检测方法的补充；第三节中，我们继续研究了图形和模拟在模型应用过程中的作用，以LOWESS方法、假设检验以及Tukey首尾计数法则为例；第四节对本文作出了小结以及未来展望。本文所有计算和作图都基于R语言。

关键词：统计模拟；统计图形；统计模型；统计理论；统计教学；R语言

Abstract

Statistical models are usually based on abstract mathematical theories, but we can analyze and interpret models from the viewpoint of statistical simulation and graphics, so that models can be less abstract and easier to

*Email：xie@yihui.name；主页：http://yihui.name。

learn, since we have ignored the non-essential mathematical details. Furthermore, we can also use statistical graphics and simulation to obtain intuition before modelling and explore models after they have been built. This paper is focused on statistical models, but the key point is not on the mathematical theories; instead, we mainly introduce the assistance of statistical graphics and simulation to learning and using statistical models, and we give several examples to illustrate our ideas.

Firstly, we give an overview to the development and advantages of statistical graphics and simulation, and argue that they cannot be replaced by modelling in some cases. Secondly, we explain the role of graphics and simulation in interpreting model theories in four aspects: to verify the robustness of model assumptions (t test under heteroscedasticiy), to explain model concepts (the meaning of conditioning and interaction in linear models), to validate model properties (least median regression), and to gain new ideas by the intuition from graphics and simulation (outlier detection in regression). Thirdly, we explore the applications of graphics and simulation in data analysis and find that they can help us know more about the relationship between variables, extract information beyond models and update the old rules-of-thumb which are no longer appropriate. Finally we conclude with a discussion on the current situation of statistical teaching and emphasize that we should make better use of the increasing computing power in statistical education and applications, besides, the powerful R language is also briefly introduced in the end.

Keywords: simulation, statistical graphics, statistical models, statistical theories, teaching, R language

目录

1 研究背景 **1**
1.1 图形和模拟的发展及优势 1
1.2 BinormCircle数据案例 4

2 阐释模型理论 **6**
2.1 检验理论假设条件 6
2.2 直观解释模型概念 11
2.3 快速验证模型性质 14
2.4 启发新型理论思路 16

3 探索模型应用 **20**
3.1 深入探索变量间关系 21
3.2 提供模型之外的信息 24
3.3 更新陈旧的经验法则 28

4 小结与展望 **30**

A MSG程序包 **33**
A.1 函数说明 34
A.2 数据说明 34

参考文献 **34**

索引 **40**

插图

1 寻找二维大数据中隐藏的特征 5
2 假设等方差和异方差对t检验结果的影响（样本量相同） . . . 8
3 假设等方差和异方差对t检验结果的影响（样本量不同） . . . 9
4 不同样本量组合下的t检验P值之差 11
5 控制变量z之后y与x的关系 13
6 连续型自变量的交互作用气泡图 15
7 LMS回归的稳健性及其缺点 17

8 用部分抽样方法诊断多个离群点 19
9 中国政府网站中的百分比数据LOWESS图 22
10 海拔高度与物种数目的LOWESS曲线 23
11 Student的睡眠增量数据：箱线图 25
12 Student的睡眠增量数据：小提琴图 27
13 两组受试者睡眠增量均值的分布 29
14 Tukey首尾计数的经验法则与常规检验的P值 31

1 研究背景

统计图形的历史源远流长，种类繁多，根据Friendly and Denis (2001)的记录，世界上最早的统计图形主要起源于地图，而史上有记载的最古老地图大约诞生于公元前6200年。众所周知，地图的作用在于提供地理位置的导航和探索。统计图形经过数千年的发展，虽然形式和工具发生了巨大的变革，但其目的始终没有改变，就是通过可视化的手段引导读者（或用户）去探索和发现信息。由于统计图形可以充分利用人的视觉系统，因此它相比起复杂的数学理论来说具有“使用简便直观、传达信息迅速”的优势。

统计模拟则通常是从计算的角度先构造一个满足数学理论假设的环境，然后按照数学理论的过程描述直接由计算得到结果。类似地，统计模拟也是一种便捷的手段，它可以用来辅助验证理论的正确性、解释理论的内在作用机理，而不需要繁琐的数学推演。

我们知道数学理论在统计学的发展中扮演了重要的角色，甚至可以说没有数学则没有统计学。历史上统计学的重大理论突破，几乎无一不是基于数学理论基础的；但从另一方面来说，在学科间合作日益加强的今天，我们却不可能要求统计学的使用者全都精通统计方法背后的数学理论，所以我们需要适当的工具来绕过数学的障碍，却又不能简单忽略数学理论的重要性。

在这样的背景下，本文提出统计图形和统计模拟这两种途径（如无特殊说明，下文的“图形”和“模拟”分别特指“统计图形”和“统计模拟”），用以分析和探索统计模型理论，并对统计建模和应用提供进一步指导。如前文所述，这两种方法都具有简便快捷的特征，因此它们尤其能为统计模型初学者构造良好的沟通媒介和探索工具。同时，统计计算和统计图形在很多情况下都紧密结合在一起，而统计模拟是统计计算的重要组成部分，所以统计模拟和统计图形的结合可作为解读统计模型的自然载体。

1.1 图形和模拟的发展及优势

学界对统计图形的研究主要限于数据的可视化：早期可追溯至历史上第一幅饼图(Playfair, 1801)，以及后来著名的“提灯女士”南丁格尔的玫瑰图(Nightingale, 1858)等；近代统计图形以Tukey (1977)的探索性数据分析为里程碑式的起点，继而诞生了大批具有数理统计意义和计算机应用的图形著作和图形种类，如我们熟知的箱线图(McGill *et al.*, 1978)，

LOWESS曲线(Cleveland, 1979)，直方图和密度曲线(Scott, 1992)，基于S语言的著作Chambers *et al.* (1983); Cleveland (1985, 1993)以及注重表达信息的著作Tufte (1992, 2001)等；现代统计图形的发展则更偏重计算机工具的开发以及高维图形和动态图形的展示，其中S语言(Becker *et al.*, 1988)为现代统计图形的发展奠定了重要的基础，随后R语言(Ihaka and Gentleman, 1996; R Development Core Team, 2009)的兴起，更是带来了数不胜数的统计图形方法，比较有代表性的如R语言的基础包**graphics**包和**grid**包(Murrell, 2005)、基于Trellis图形(Cleveland, 1993)思想的**lattice**图形(Sarkar, 2010)、基于统计图形理论著作Wilkinson (2005)的**ggplot2**图形(Wickham, 2009)、基于动态图形GGobi软件系统(Cook and Swayne, 2007)的高维数据交互图形实现**rggobi**包(Temple Lang *et al.*, 2009)、基于OpenGL的三维动态图形系统**rgl**包(Adler and Murdoch, 2010)和分类数据图示的**vcd**包(Meyer *et al.*, 2010)等，此外，还有一批新的高维图形思想被提出，如打破笛卡尔坐标系常规的平行坐标图(Inselberg, 2007)，并出现了一些R语言之外的独立交互图形软件如用于分析缺失值的MANET软件(Unwin *et al.*, 1996)和交互式图形分析软件Mondrian (Theus, 2002)等，这些动态图形和交互图形的综述可参考Symanzik (2004)。

基于数据可视化的研究提供了丰富的数据信息挖掘手段，但我们在这众多的研究中却难以找到以模型为重心的代表性成果。在这种想法的驱动下，作者于2007年开始开发R包**animation** (Xie, 2010a)，目的就在于将统计模型、方法和理论以动画的形式“可视化”，后文我们将看到一些示例；该软件包获得了2009年美国统计学会的John Chambers Award（`http://stat-computing.org/awards/jmc/winners.html`），这从一个侧面说明评委会对“可视化统计模型和理论”的高度认可以及从图形角度去探索模型的重要价值。

无论是从理论方面还是应用工具方面来说，统计图形看似都应该有广泛的使用价值，然而如谢益辉 (2008a)指出的，国内的统计分析氛围仍然是以统计模型为主，对统计图形的使用频率非常低，这也是本文写作的重要动机之一。

相比之下，统计模拟并没有如此悠远的历史，主要原因显然是因为它涉及到大量的计算，而计算机的问世已经是近代的事情了。由于本文涉及到的统计模拟主要是随机数方面的内容，所以对统计模拟的发展历程不加回顾，然而这里我们特别指出值得注意的三篇文献，来说明统计模拟

本身相比起纯理论研究的价值所在。首先是Bootstrap方法的开篇作(Efron, 1979)，它可算是统计学界最有影响力的论文之一，该文几乎从此开辟了一个全新的研究领域，究其原因大致有二：它展示了Bootstrap方法的优良理论性质，并且计算方法极为简单（传闻因为过于简单，最初被编辑拒之门外）；这说明切合实际的统计计算对于统计学的发展来说是至关重要的，若仅有理论推导而实践困难，不易被广为接受，更不必谈应用。其次是继承Bootstrap的重抽样思想的一本名为“Resampling: The New Statistics”的书(Simon, 1997)，这本书第4页给出了一个赌注为$5000的挑战，由作者Simon挑战任何一位用传统的数理统计方式教统计学的老师，该挑战内容如下：

> [A] public offer: The intellectual history of probability and statistics began with gambling games and betting. Therefore, perhaps a lighthearted but very serious offer would not seem inappropriate here: I hereby publicly offer to stake $5,000 in a contest against any teacher of conventional statistics, with the winner to be decided by whose students get the larger number of simple and complex numerical problems correct, when teaching similar groups of students for a limited number of class hours — say, six or ten. And if I should win, as I am confident that I will, I will contribute the winnings to the effort to promulgate this teaching method. [...] This offer has been in print for many years now, but no one has accepted it.

作者Simon的挑战已经公开放出来多年，但至今无人能应接；这件事情直接反映出纯粹用数学的方式学习统计学是有缺陷的，而Simon在全书的观点都倾向于使用模拟。第三篇文献(Simon *et al.*, 1976)为前一篇文献的作者Simon和其他几位作者合作，它是一篇关于统计教学的论文，其中的试验表明，无论老师的统计理论教得多么好，学生都更经常采用模拟的方式得到正确答案。

King *et al.* (2000)对统计模拟的优势作了两点总结：

1. 几乎任何一种统计理论方法能得到的结果都可以通过模拟得到，然反之则不然[1]；

[1]本文作者认为这一点优势应该限于可计算的结果，因为目前计算机还没有发展到有足够的智能可以推导数学定理的程度。

2. 统计模拟能对教学提供重要帮助；

总而言之，本文提倡的是尽量使用简单方法揭示深度规律，而不是一味偏向某一种途径。

1.2 BinormCircle数据案例

下面我们用一个简单的案例来说明图形方法对统计模型分析的辅助作用。这个例子的设计灵感来自于1986年美国统计学会的数据展（Data Exposition）[2]，他们当时所用的是一个人造数据：在大量独立的随机数中隐藏了EUREKA几个字母的轮廓坐标数据。如果只是用单纯的统计模型分析，可能无法得出任何线索，因为数据真实的信息只是在少量的数据中，它被掩盖在大量的“噪声”中，而简单的统计图形（如散点图）也无法看出数据有何不同寻常之处。此时如果在画图时使用半透明色让“噪声”的颜色透明化，真实的“信息”也就可以被清楚地观察到了。这个数据集被收录在**animation**包(Xie, 2010a)中，名为`pollen`，该数据的帮助文件中提供了详细的探索示例，此处不详述；我们也特别为pollen数据制作了一个三维探索视频，可访问网页`http://yihui.name/cn/2008/10/historical-demo-of-pollen-data/`观看。

这个例子可以看作是pollen数据的简化版：我们设计了2万个样本，其中有1万个样本点来自于两个独立的标准正态分布，另1万个样本点的坐标落在半径为0.5的圆上，最后将这2万个样本拼起来并打乱顺序。该数据收录在**MSG**包(Xie, 2010b)中，名为`BinormCircle`。虽然数据只有两个变量，但我们用普通的统计模型和数值分析几乎无法找出数据的特征，例如线性回归显示两个变量`V1`和`V2`的回归系数非常不显著：

```
> library(MSG)
> data(BinormCircle)
> head(BinormCircle)
```

```
      V1     V2
1  0.889 -1.764
2  0.072 -0.495
3  0.123 -0.180
4 -0.499  0.030
```

[2]这是每年Joint Statistical Meetings的常规竞赛：组委会事先给出一个数据，参赛者可以自由发挥，用统计图形或模型的手段将数据中的信息展示出来。

```
> par(mfrow = c(1, 2), ann = FALSE, mar = c(2, 2, 0.5,
+     0.2))
> plot(BinormCircle)
> smoothScatter(BinormCircle)
```

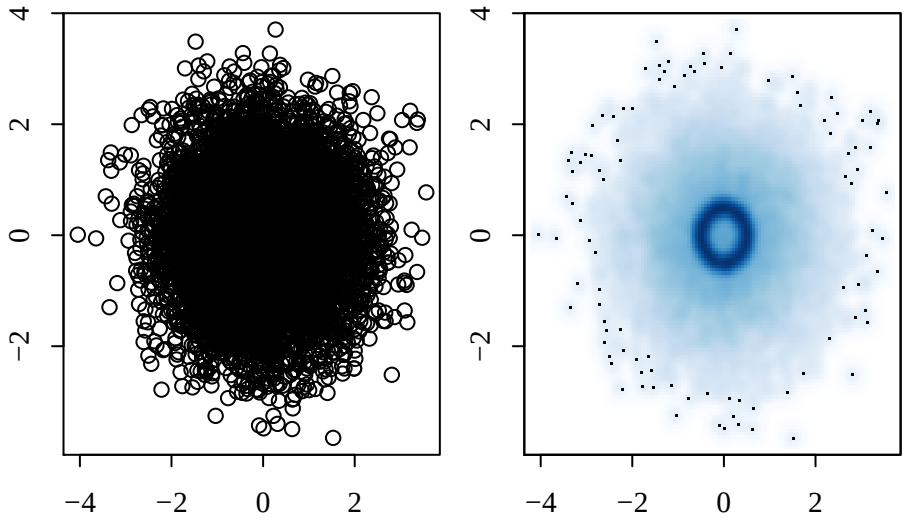

图 1: 寻找二维大数据中隐藏的特征：左图是一幅普通的散点图，图中几乎看不出数据有任何异常特征；右图使用了基于二维核密度估计的平滑散点图，颜色越深表示该处数据密度越大。

```
5  0.252  0.432
6  0.450  0.218
```

```
> coef(summary(lm(V2 ~ V1, BinormCircle)))
```

```
                Estimate  Std. Error    t value  Pr(>|t|)
(Intercept) -0.006085156 0.005258164 -1.1572776 0.2471728
V1           0.003988851 0.007015317  0.5685917 0.5696397
```

换用高阶回归的结果也类似，无论回归阶数为多少，系数均不显著，这一点从数据的构造上就可以知道（理论上两个变量的相关系数为0）。由于样本量太大，普通的散点图上点与点之间严重重叠，所以也很难看出散点图有何异常。对于这种大型数据，我们可以采取一些特殊的图形方法揭示图中的信息。图1左图是普通的散点图，右图是V1和V2的平滑散点图。平滑散点图的原理很简单，它并不直接将散点画出来，而是基于

二维核密度估计(Wand, 2009)用特定颜色深浅表示某个位置的密度值大小。右图很明显立即显示出在大量的数据点背后，还隐藏着一层关系，图中的深色圆圈揭示出有一部分数据分布在圆圈上，而且数据密度很大。除了平滑散点图之外，我们在网页`http://yihui.name/en/2008/09/to-see-a-circle-in-a-pile-of-sand/`上给出了其它五种不同的解决方案，都可以从图形的角度反映出这种规律。

虽然本例使用的是一则模拟数据，但实际应用中大数据比比皆是，所以这里提供的平滑散点图并非特定场合的特定方法，而是可以作为一种标准工具去探索数据中的聚集现象，这种“聚集现象”用统计模型和数值的方式不一定能容易发现出来。

2 阐释模型理论

统计图形和模拟可以用来辅助解释统计模型的数学理论，我们经常见到的例子就是关于中心极限定理的模拟：产生服从某些条件的随机数，查看样本均值的分布是否为正态分布。我们在**animation**包(Xie, 2010a)中已经给出了中心极限定理的模拟以及对传统演示方法缺点的说明和改进（参见*clt.ani()*函数）；由于中心极限定理的例子太普通，不足以说明统计图形和模拟的作用，本文并不采用它作演示。本节中我们首先以t检验为例，用图形和模拟分析模型方法假设条件对模型稳健性的影响，其次以多元回归为例，用一则反例解除人们对模型的常见误解，然后以交互作用为例，提出对经典图示方法的补充，再以最小中位数平方回归为例，用图形和模拟的方法快速而直观说明了模型的缺陷所在，最后以离群点检测为例，提出了一种模拟的思路，可作为对传统离群点检测方法的补充。

2.1 检验理论假设条件

在实际应用中，一些数学假设可能只是结论的充分条件，而非充要条件。某些情况下，即使违反数学假设，也不会影响结果，或者对结果的影响几乎可以忽略。这里我们用统计学中最初级的两样本t检验作为分析对象，它假设两独立样本$X_1, X_2, \ldots, X_{n_1}$和$Y_1, Y_2, \ldots, Y_{n_2}$分别来自等方差的正态

分布，均值分别为μ_1和μ_2，在零假设$H_o : \mu_1 = \mu_2$下有t统计量：

$$t = \frac{\bar{X} - \bar{Y}}{S_{XY} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

其中

$$S_{XY} = \sqrt{\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}}$$

$\bar{X}$和$\bar{Y}$分别为两个样本的样本均值，S_X^2和S_Y^2分别为两个样本的样本方差。传统统计学教科书往往会提到方差齐性的问题，方差相等的时候用上面的方法，方差不等的时候用Welch校正的方法(Welch, 1947)调整t统计量自由度，但我们可以通过模拟发现，在两组样本的样本量相近时，等方差的假设条件可能并不重要，即在真实情况为异方差时，校正（假设异方差）或不校正（假设等方差）没什么区别；而样本量相差悬殊的时候，Welch校正才会对检验结果起到较大影响。

模拟过程是：从两个正态总体中生成样本，第一个总体均值为0，标准差随机取自均匀分布U(0.5, 1)，第二个总体均值为1，标准差取自U(2, 5)，显然两个总体标准差不相等，理论上应该用Welch校正方法（以下所有计算都可用Welch校正方法得到的P值作为正确的参考标准）。在t检验时我们可以分别设定等方差和异方差的选项，看看对结果有多大影响。这个过程重复1000次，得到1000组P值，每组2个P值。

我们首先看两组样本量相等的情况（均设置为100）。用普通两样本t检验和Welch校正方法分别得到的P值如图2：左图是原始P值，可见基本在对角线上，说明两种计算方法得到的P值大致相等，右图是P值的差异，可见用Welch校正方法得到的P值普遍偏大，原因很简单，因为Welch校正的自由度小于或等于不校正的自由度，样本量相等的时候统计量的分母即标准误一样，因此统计量完全一样，自由度越小，P值越大，但实际上也并没大多少，实际应用中可忽略。

然后我们看两组样本量差别较大时的情况（第一组样本量为10，第二组样本量为100）。图3形式类似于图2，但我们很快可以发现，这种样本量设置下，两种计算方法得到的P值就大不一样了，确切地说，等方差方法得到的P值严重偏大，意即：在真实情况为异方差的情况下，若仍然使用等方差假设下的t检验，则会严重高估P值，也就是难以检测出两样本均值本来存在的差异。

```
> par(mar = c(3.5, 3.5, 1, 0.5), mfrow = c(1, 2))
> plot(pval <- t(replicate(1000, {
+     x1 = rnorm(100, mean = 0, sd = runif(1, 0.5,
+         1))
+     x2 = rnorm(100, mean = 1, sd = runif(1, 2, 5))
+     c(t.test(x1, x2, var.equal = FALSE)$p.value,
+         t.test(x1, x2, var.equal = TRUE)$p.value)
+ })), xlab = "异方差", ylab = "等方差", pch = 20,
+     asp = 1, col = rgb(0, 0, 0, 0.3))
> abline(0, 1)
> plot(pval[, 1], pval[, 2] - pval[, 1], xlab = "异方差",
+     ylab = "等方差减异方差", pch = 20, col = rgb(0,
+         0, 0, 0.3))
```

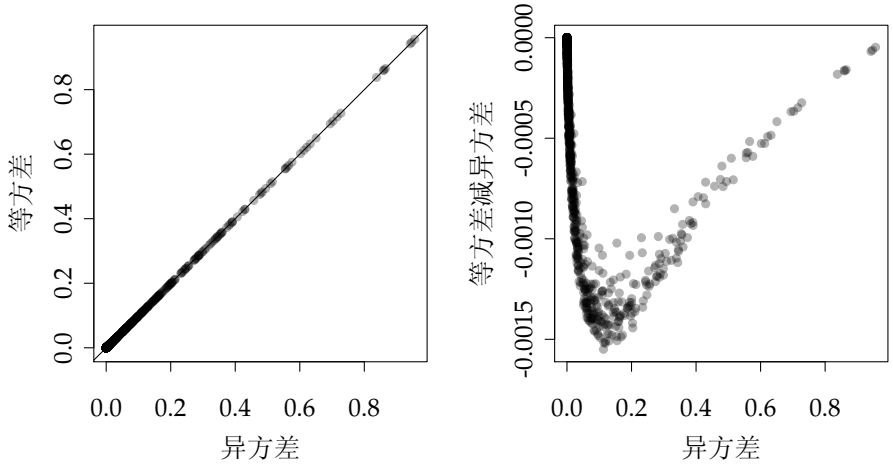

图 2: 假设等方差和异方差对*t*检验结果的影响（样本量相同）：左图为两种方法的*P*值散点图，右图是两种方法的*P*值之差与*Welch*校正后*P*值的散点图。通过设置*R*函数t.test()的*var.equal*选项，我们可以强制*R*采用普通*t*检验或*Welch*校正的方法去作*t*检验，然后分别提取*P*值并画图。

```
> par(mar = c(3.5, 3.5, 1, 0.5), mfrow = c(1, 2))
> plot(pval <- t(replicate(1000, {
+     x1 = rnorm(10, mean = 0, sd = runif(1, 0.5, 1))
+     x2 = rnorm(100, mean = 1, sd = runif(1, 2, 5))
+     c(t.test(x1, x2, var.equal = FALSE)$p.value,
+         t.test(x1, x2, var.equal = TRUE)$p.value)
+ })), xlab = "异方差", ylab = "等方差", pch = 20,
+     asp = 1, col = rgb(0, 0, 0, 0.3))
> abline(0, 1)
> abline(h = 0.05, v = 0.05, col = "gray")
> plot(pval[, 1], pval[, 2] - pval[, 1], xlab = "异方差",
+     ylab = "等方差减异方差", pch = 20, col = rgb(0,
+         0, 0, 0.3))
```

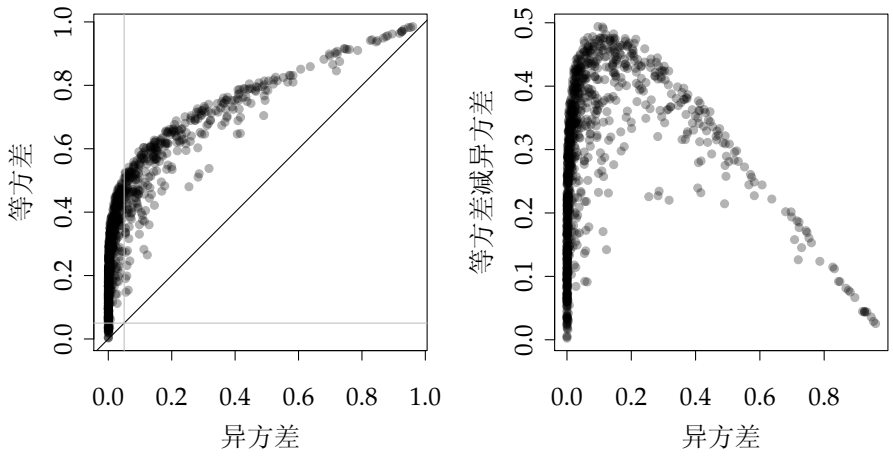

图 3: 假设等方差和异方差对t检验结果的影响（样本量不同）：两组样本量分别为10和100，方差设置与前一种情形相同，此时两种方法得到的P值有很大差异。

我们可以继续将这个模拟推广到更多样本量组合的情况，例如固定第二组样本量为30，调整第一组样本量从2变化到100（$n_1 = 2, 3, \cdots, 100$; $n_2 = 30$），看假设异方差和等方差得到的P值之差的变化。对每一种样本量组合，我们分别模拟1000次，因此得到1000个P值之差，最终我们把这个差值记录在`t.diff`对象中，这个数据对象已收录在**MSG**包(Xie, 2010b)中，完整生成过程如下：

```
> set.seed(123)
> t.diff = NULL
> for (n1 in 2:100) {
+     t.diff = rbind(t.diff, replicate(1000, {
+         x1 = rnorm(n1, mean = 0, sd = runif(1, 0.5,
+             1))
+         x2 = rnorm(30, mean = 1, sd = runif(1, 2,
+             5))
+         t.test(x1, x2, var.equal = TRUE)$p.value -
+             t.test(x1, x2, var.equal = FALSE)$p.value
+     }))
+ }
> t.diff = as.data.frame(t(t.diff))
> colnames(t.diff) = 2:100
```

图4用箱线图展示了随着n_1从2变化到60，P值之差的变化情况（受页面宽度限制，仅取了`t.diff`数据的前59列）。显然$n_1 = 30 = n_2$时，使用或不使用自由度校正方法并没有太大区别，而两组样本量相差越大，则P值相差越大；两种方法P值平均相差在0.02范围内的n_1取值范围为26–34：

```
> data(t.diff)
> names(which(abs(apply(t.diff, 2, mean)) < 0.02))
```

```
[1] "26" "27" "28" "29" "30" "31" "32" "33" "34"
```

也就是说，两组样本量相差5以下时，得到的结论大致是相同的（P值不会相差太大）。

本例中，由于模拟的两个关键步骤（生成随机数和计算等方差/异方差情况下的P值）都非常容易实现，所以它可以迅速说明Welch校正的适用场合。作者曾在Iowa State University统计系的课堂上听一位教授讲t检验中的Welch校正，他倾向于在大多数情况下不使用Welch校正，对此观点作者并不认同，因为现代计算机的发达程度已经足以忽略戈赛特（Gosset，t分布发明者）时代的计算负担。

```
> par(mar = c(3.5, 3.5, 1, 0.5))
> data(t.diff)
> boxplot(t.diff[, as.character(2:60)], ylab = "等方差减异方差",
+     xlab = "$n_1$", xaxt = "n", border = ifelse(abs(apply(t.diff,
+         2, mean)) < 0.02, "black", "green"), at = 2:60)
> axis(1)
```

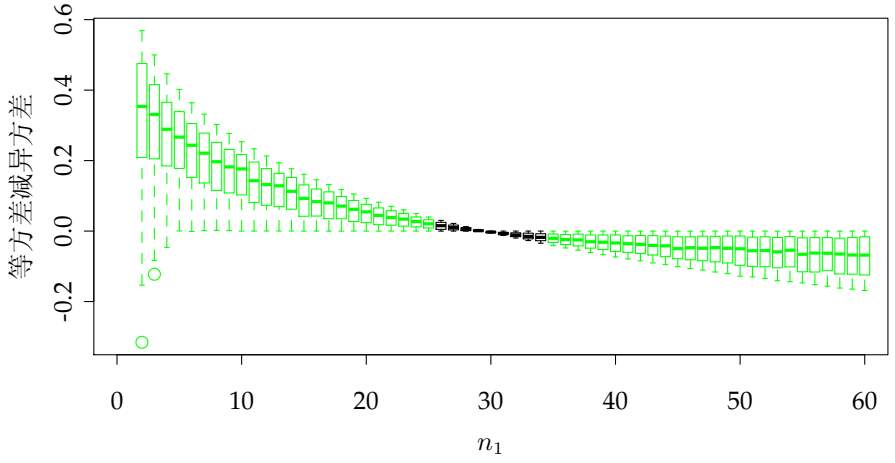

图 4: 不同样本量组合下的t检验P值之差：对于每一种样本量组合，模拟得到1000个P值差值，这些差值用箱线图表示出来，最终得到59个箱线图，反映了n_1变化时P值之差的变化。

类似地，我们还可以模拟等方差的情形，即：在真实总体方差相等的情况下，查看使用或不使用Welch校正的t检验结果。前文的代码可以很方便修改为这种情形下的模拟，限于篇幅，此处不再给出等方差情形下的结果。

2.2 直观解释模型概念

统计模型中常出现一些抽象概念，我们可以通过图形和模拟去将这些抽象的概念具体化，用事实说话，使得模型的意义直观可见，本节以回归中的一个概念为例，说明图形和模拟对模型意义的解释。

回归模型是绝大多数统计模型的基础，而一元回归又是回归的基础。一般教学中常从一元回归引入基本思想，在讲完大量的一元回归性质之后

再开始多元回归。这种顺序的优点在于它由浅入深，使初学者容易入门，但同时也会带来一些误区。多元回归与一元回归的显著不同在于，它通过控制其它自变量来检查一个自变量与因变量的关系，而这里的“控制”可能会对初学者造成理解上的困难；其次多元回归引入了“交互作用”的概念，也是一元回归中不存在的。为了使初学者走出用一元回归的视角去看待多元回归的常见误区，我们可以通过模拟和图形的方式给出两个非常直观的例子。

首先考虑“控制变量”：一元回归下我们通常用散点图观察自变量和因变量的关系，并将回归模型解释为X变化导致Y如何变化，多元回归则需要考虑其它自变量的水平，在其它自变量保持不变的条件下，看我们关心的自变量和因变量的关系。模拟的场景设计为：因变量y与自变量x在控制了第二个自变量z之后为负相关关系，但不控制z的时候为正相关关系。真实模型如下：

$$y = -x + z + \epsilon$$

其中x在$[0, 4]$区间上取值，$z = 0, 1, \cdots, 4$，$\epsilon \sim N(0, \sigma^2)$，$\sigma = 0.25$。这个模拟的关键在于让$z$的增长胜过$x$，这样看似$y$随着$x$的增大而增大，实际上控制$z$的水平之后$y$与$x$是负向关系。以下是一个示例：

```
> set.seed(123)
> x = seq(0, 4, length = 100)
> z = rep(0:4, each = 20)
> y = -x + z + rnorm(100, 0, 0.25)
> coef(summary(lm(y ~ x)))
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.3847108 0.06964946 -5.523529 2.733688e-07
x            0.2036561 0.03008323  6.769757 9.555139e-10
```

```
> coef(summary(lm(y ~ x + z)))
```

```
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.03297214 0.05475317 -0.602196 5.484488e-01
x           -0.90709758 0.09831225 -9.226699 6.272130e-15
z            0.93488438 0.08107839 11.530624 6.949985e-20
```

显然，若用y对x直接做一元回归的话，得到的回归系数是非常显著的正数，但若在回归模型中加入z变量，x的系数则变为非常显著的负数！图5用散点图进一步揭示了这个问题的本质。左图中，我们可以看到x与y是

```
> par(mar = c(3.5, 3.5, 1, 0.5), mfrow = c(1, 2))
> plot(x, y)
> abline(lm(y ~ x), col = "red")
> plot(x, y, pch = z, col = rainbow(5)[z + 1])
> for (i in z) abline(lm(y ~ x, subset = z == i), col = "darkgray")
```

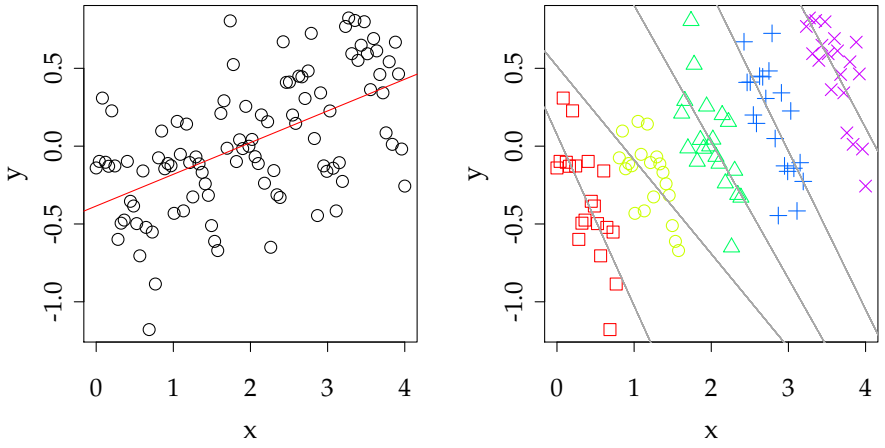

图 5: 控制变量z之后y与x的关系：左图看似x和y正向关系，而右图中控制了z取值水平之后x和y就变成了负向关系。

正向关系，而右图中我们根据z的不同取值将样本点用不同的符号和颜色标示出来，每一种符号（及颜色）代表了一种z的取值，可见每一小组数据点中，y与x都是负向关系。所谓多元回归的“控制其它变量”的意义，可以用图5清晰表达出来。本例也说明了一元回归和多元回归的本质不同，多元回归系数不能由简单的一元回归得到。

然后我们考虑“交互作用”：交互作用仅存在于模型中有多个变量时的情形，它的含义是一个自变量对因变量的影响系数受另一个自变量的取值水平影响，其基本数学形式为（以二元回归为例）：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

我们将上式稍作改写：

$$\begin{aligned} y &= (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) x_2 + \epsilon \\ &\equiv \alpha_0 + \alpha_2 x_2 + \epsilon \end{aligned}$$

若我们将x_1固定在特定水平，那么x_2的回归系数为$\alpha_2 = \beta_2 + \beta_3 x_1$，它与$x_1$有关；同理，$y$与$x_1$的关系也受$x_2$的不同水平影响。交互作用的含义在传统的统计学教科书中一般都用折线图表示，而折线图只能表示自变量为分类变量时的交互效应，对于连续自变量情况的交互作用图示，我们则很难找到任何示例。

这里我们提出用气泡图的方式来展示连续变量的交互效应。模拟场景如下：

$$y = 2 + x + 0.5z + 0.5xz + \epsilon \quad (1)$$

$$y = 2 + x + 0.5z + \epsilon \quad (2)$$

式(1)是有交互效应的回归模型，式(2)不包含交互效应。我们让x和z都从1到10取值，然后对于每一组x和z的组合，计算出y值，最终我们将x、z和y用气泡图表示出来如图6。我们无需用数值的方式去解读交互效应，只需要看图中“气泡”的大小随着x和z的取值不同如何变化即可。这样一来，交互效应的概念便一目了然。同时本例也是对交互效应的传统展示方法的一种补充。

2.3 快速验证模型性质

如第1节所说，统计模拟具有简便易行的优势，只要我们清楚数学理论假设，就可以按照假设条件设置模拟环境，以计算作为推导的一种替代。本小节以Venables and Ripley (2002)中介绍的最小中位数平方（**L**east **M**edian **S**quares，LMS）回归模型为对象，说明统计模拟在验证模型理论上的优势。

最小中位数平方回归（下文简称LMS回归）是稳健回归方法中的一种，它对离群点有良好的耐抗性，即：数据中的离群点对LMS回归系数的影响非常小。LMS回归的目标函数是残差平方的中位数，系数估计通过下式得到：

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \text{median}\left\{(y_i - \hat{y}_i)^2\right\},\ i = 1, 2, \cdots, n$$

其中$\hat{y}_i = X_i\boldsymbol{\beta}$。Venables and Ripley (2002, pp.159)简略介绍了LMS回归并提出了它的一个缺点：它对大量集中在数据中心的数据点非常敏感。这一条性质在书中并没有详细介绍，但我们可以很快用模拟的方式来验证它，而不需要真正去进行数学推导。模拟场景如下：

```
> par(mar = c(3.5, 3.5, 2, 0.2), mfrow = c(1, 2), cex.main = 1)
> sq = 1:10
> x = rep(sq, 10)
> z = rep(sq, each = 10)
> y = c(outer(sq, sq, function(x, z) 2 + x + 0.5 *
+     z + 0.5 * x * z + runif(1)))
> symbols(x, z, y, bg = rgb(0, 1, 0, 0.3), fg = "blue",
+     main = "$y = 2 + x + 0.5 z + 0.5 x z + \\epsilon$",
+     inches = 0.4)
> y = c(outer(sq, sq, function(x, z) 2 + x + 0.5 *
+     z + runif(1)))
> symbols(x, z, y, bg = rgb(0, 1, 0, 0.3), fg = "blue",
+     main = "$y = 2 + x + 0.5 z + \\epsilon$", inches = 0.2)
```

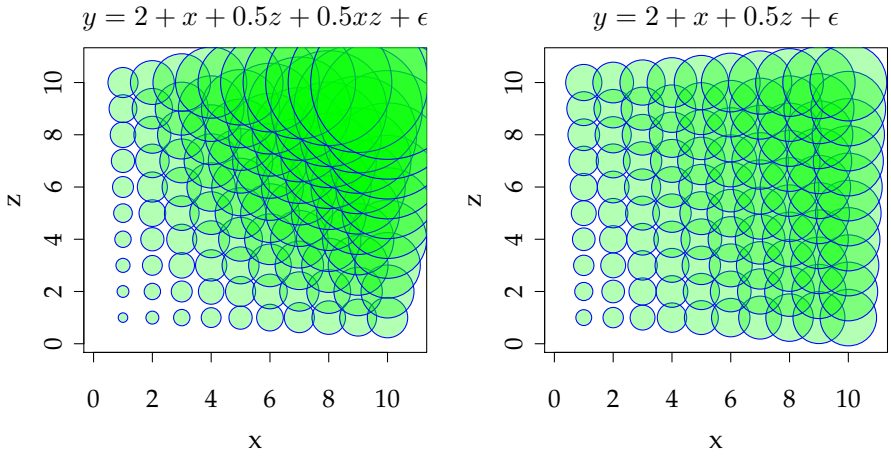

图 6: 连续型自变量的交互作用气泡图：左图中x与z有交互效应，右图无交互效应。气泡图中，气泡的大小与真实的y值大小成正比，所以如果我们要查看自变量对y的影响，只需要看自变量对气泡大小的影响即可。以左图为例：例如分别给定$x = 1$和$x = 10$，随着z的增大（从下向上看），y值在增大，但$x = 1$和$x = 10$处的增大速度明显不一样（后者快），也就是说，z对y的影响大小受x的取值水平影响。同理可以看右图，任意给定x值，y随着z的增大速度都一样，说明x与z之间没有交互效应。

首先生成具有线性关系的自变量x和因变量y，然后在各自的均值附近生成大量随机数填充进原数据，最后计算LMS回归结果，看原来的线性关系是否能被保持（理论上x与y的线性关系将受到严重影响）。为了更直观地观察计算结果，我们用散点图加回归直线的方式来表达结果。下面的R函数用来生成包含普通最小二乘（OLS）回归直线和LMS回归直线的散点图：

```
> library(MASS)
> olsLms = function(x, y, l.col = c("red", "blue"),
+     l.lty = c(1, 2), ...) {
+     plot(x, y, ...)
+     abline(lm(y ~ x), col = l.col[1], lty = l.lty[1])
+     abline(lqs(y ~ x, method = "lqs"), col = l.col[2],
+         lty = l.lty[2])
+     legend("topleft", legend = c("OLS", "LMS"), col = l.col,
+         lty = l.lty, bty = "n")
+ }
```

然后我们按照模型$y = 2+3x+\epsilon$生成两批模拟数据，第一批包含一个离群点，用以检验LMS回归相比起OLS回归的稳健性；第二批数据包含500个分布在数据中心附近的随机数，用以检验“LMS回归对中心数据敏感”的性质：

```
> set.seed(123)
> x = runif(50)
> y = 2 + 3 * x + rnorm(50)
> x1 = c(x, 2)
> y1 = c(y, 50)
> x2 = c(x, jitter(rep(mean(x), 500), 10))
> y2 = c(y, jitter(rep(mean(y), 500), 10))
```

图7中的两幅散点图及其回归直线表明了LMS回归对离群点的稳健性和对中心数据的敏感性。左图中OLS回归直线的斜率明显被右上角的离群点“拉”大，但LMS回归并没有受离群点影响，它的斜率反映了大部分数据所体现的规律；右图中OLS回归直线斜率反映了所有数据的趋势，而LMS回归的斜率则明显违背了数据的趋势。通过模拟和图形，LMS回归的优缺点一目了然。

2.4 启发新型理论思路

在某些情况下，我们也可以在统计模拟中找到解决问题的新思路，这样能避免严格的数学证明推导，更有效地利用现有的计算机资源为统计模

```
> par(mar = c(3.5, 3.5, 1, 0.2), mfrow = c(1, 2), pch = 20)
> olsLms(x1, y1)
> olsLms(x2, y2, cex = c(rep(1, 50), rep(0.1, 500)))
```

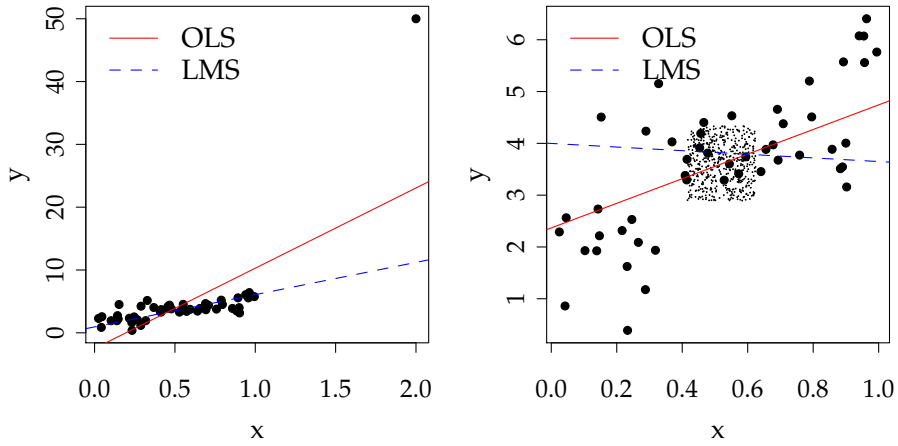

图 7: *LMS*回归的稳健性及其缺点：左图体现了*LMS*回归的稳健性，右图体现了*LMS*回归对中心数据点敏感的特征。

型理论提供发展和创新的可能性。在绝大多数情况下，统计模拟一定能得出结果（无论对错或是否有普遍意义），但数学推导则并不一定，这也是统计模拟的一大优势。下面我们基于传统的回归离群点诊断方法通过模拟和图形提出一种新的诊断方法。

我们知道传统的离群点诊断方法有一个很大的弱点，就是当数据中有多个离群点的时候，传统方法如Cook距离等测度可能会失效，因为这些方法都是基于删除一个数据点来看回归模型的变化；多个离群点有可能会在同一个“方向”上，如果只是删除其中一个，剩下的离群点仍然会影响回归模型，从而掩盖掉删除该离群点的效果。对这个问题，作者的一个直接想法是，我们可以通过重抽样或部分抽样并结合图形可以找出多个离群点。具体来说，我们可以把“删除一个数据点”的想法推广到“删除若干个数据点”，这样一来，就存在多个离群点被同时删掉的可能性了，当出现这种情况时，回归系数理论上会发生很大变化，这种变化既可以用数值指标计算出来，也可以用图形画出来。

以下是模拟场景：生成两个服从标准正态分布的独立随机变量x和y，

长度为100，理论上它们的回归系数为0，但是在样本点中加入2个距离相近的离群点，然后用Cook距离方法诊断，最后用前面的部分抽样思路诊断。以下是模拟的R代码：

```
> set.seed(123)
> x = c(rnorm(100), 20, 21)
> y = c(rnorm(100), 20, 24)
> fit = lm(y ~ x)
> fit1 = update(fit, subset = 1:60)
> betaSim = numeric(100)
> for (i in 1:100) {
+     idx = sample(c(TRUE, FALSE), length(x), replace = TRUE,
+         prob = c(0.6, 0.4))
+     betaSim[i] = coef(update(fit, subset = idx))[2]
+ }
```

图8展示了传统诊断方法与这里提出的抽样诊断方法的比较。左上图显示普通线性回归受离群点影响严重：理论上回归直线应该是水平的（斜率为0），但右上角的两个离群点将回归直线拉起；右上图画出了这个回归模型中每个样本点的Cook距离，从图中可以看到，只有最后一条数据是离群点，而事实上倒数第二条数据也是离群点，只是删除这一条数据之后回归模型不会有太大变化（受最后一条数据掩盖），所以它不能被Cook距离识别出来；左下图显示了一种抽样的可能性：我们抽取数据的前60条（图中用实心点表示），去掉数据的后42条（空心点表示），然后重新建立回归模型，并画出回归直线，此时我们可以看到，由于去掉了两个离群点，回归直线的斜率大致为0，与理论相符了；基于这种抽样的想法，我们将这个步骤重复100次，每次重新随机抽取一部分数据（可能包含离群点，也可能不包含），并重新计算回归系数，最终把100次的斜率都记录下来并画在右下图中，可以看出，这些斜率大致分为三群，这种“多群”的特征反映出原数据中有不止一个离群点（否则100个斜率只会分为两群：包含或不包含一个离群点的结果），靠近0的斜率是抽样不包含两个离群点的结果，而靠近1的斜率是包含两个离群点的结果，中间一层斜率是包含一个离群点的结果（可能是最后一条数据，也可能是倒数第二条）。这样我们就成功诊断出传统方法找不出来或者找不完全的离群点现象。这个模拟的Flash动画版本可以在网页`http://yihui.name/cn/2008/09/multiple-outliers-detection/`观看。

本例仅仅是以模拟的方法提供了一种离群点诊断新思路，沿着这种想

```
> par(mar = c(3.5, 3.5, 1, 0.5), mfrow = c(2, 2), pch = 20)
> plot(x, y, col = rgb(0, 0, 0, 0.5))
> abline(fit)
> plot(cooks.distance(fit), ylab = "Cook's distance")
> plot(x, y, col = rgb(0, 0, 0, 0.5), pch = rep(20:21,
+     c(60, 42)))
> abline(fit1)
> plot(betaSim, ylab = "$\\beta_1$")
```

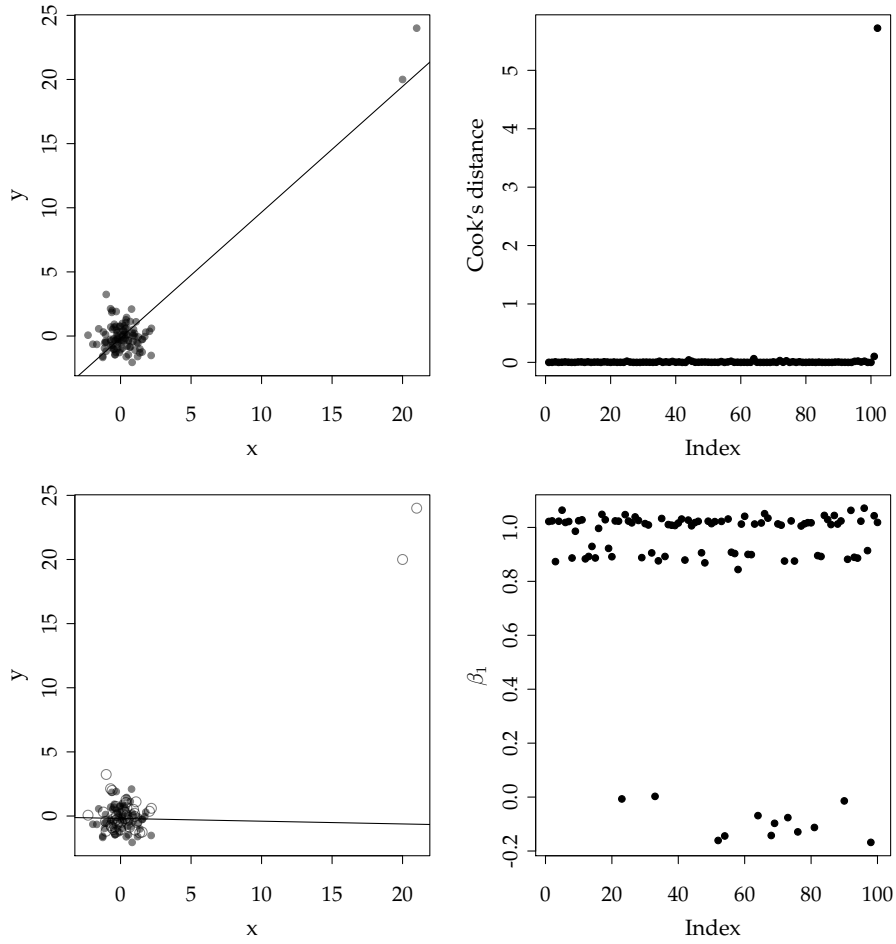

图 *8:* 用部分抽样方法诊断多个离群点：普通线性回归受离群点影响（左上），但传统离群点诊断方法如*Cook*距离并不能诊断出所有离群点（右上），如果我们对数据进行抽样（左下），则可以得到几类回归系数值（右下）。

法，我们可以继续发展新的理论，来弥补传统理论的不足。

3 探索模型应用

除了上一节介绍的解释作用之外，统计图形和模拟还可以在模型理论建立和应用过程中用来引导建模的方向，并且图形和模拟往往能提供模型理论难以揭示或者甚至不能揭示的信息。关于第二点，本文第1.2小节已经给出了一个模拟的案例作为支持，下面我们再用一个实际数据简单说明第一点。

这个案例的背景是一则名为“神奇87.53这个数字竟然走红”的新闻报导[3]，而这则新闻的导火索是“国家统计局称，在他们随机调查的100位网友中，有87.53%的网友支持封杀BTchina”，其中百分比87.53%引起了网友们的注意，进而有人继续收集了各大网站中的百分比数据，试图说明一些统计数字的荒谬。作者也对这件事情关注了一段时间，并得到了一批通过程序自动抓取的百分比数据[4]进行了一个粗略的探索。图9展示了中国政府网站（域名后缀为gov.cn的网站）中通过Google搜索得到的从0到99.99的百分比数据的搜索频数，这批数据收录在MSG包中，名为`gov.cn.pct`，以下是数据的前6行：

```
> data(gov.cn.pct)
> head(gov.cn.pct)
```

```
  percentage  count round0 round1
1       0.00 158000   TRUE   TRUE
2       0.01 171000  FALSE  FALSE
3       0.02 156000  FALSE  FALSE
4       0.03 114000  FALSE  FALSE
5       0.04 103000  FALSE  FALSE
6       0.05 201000  FALSE  FALSE
```

```
> pct.lowess = function(cond) {
+     with(gov.cn.pct, {
+         plot(count ~ percentage, pch = ifelse(cond,
+             4, 20), col = rgb(0:1, 0, 0, c(0.04,
+             0.5))[cond + 1], log = "y")
+         lines(lowess(gov.cn.pct[cond, 1:2], f = 1/3),
+             col = 2, lwd = 2)
```

[3] http://news.sina.com.cn/c/2009-12-12/034016758777s.shtml

[4] http://chemhack.com/cn/2009/12/87-53-stat/

```
+         lines(lowess(gov.cn.pct[!cond, 1:2], f = 1/3),
+             col = 1, lwd = 2)
+     })
+ }
```

图9中左上图用垂线表示了每个百分比的搜索频数大小，从中我们可以发现垂线在某些区间上显得异常得高，为了更清楚查看这些大频数的位置，我们可以把图形放大，如右上图显示了[10%, 11%]区间上的频数，这里我们可以清楚看到取整的百分比的频数明显比其它百分比的频数大，其它区间上有类似的特征（GIF动画`http://yihui.name/cn/wp-content/uploads/2009/12/percent-count.gif`展示了所有长度为1的区间上的频数，使这个特征更容易观察到）。为了进一步验证“取整”的猜测，我们可以分别将取整和不取整的百分比以不同样式的点表示出来，并且加上LOWESS曲线（见3.1小节），从图中可以看到，无论是取整到整数还是取整到1位小数，搜索频数都明显更高。注意左下图和右下图的y轴是取过对数的，因此取整和不取整的实际差异比图中看到的更大。

类似的建模前的探索性分析还可以在Cook and Swayne (2007)中找到。这种分析结果很难用数值的方式从数学模型中得到，因此在统计模型应用中，若能事先辅之以统计图形之类的探索，则可能会发现意想不到的信息。下面我们以几例数据继续讨论统计模型和模拟对于模型应用过程中的辅助作用。首先我们强调二元变量关系探索中LOWESS曲线相比起线性回归模型的重要地位，其次我们说明统计假设检验相比起统计图形和模拟的局限性，最后我们查看一则应用中的经验法则，以模拟的方式说明它的适用性。

3.1 深入探索变量间关系

我们知道线性模型只是非线性模型的特例，尤其对于二元变量，我们不应仅仅以线性模型的简便性而直接假设线性关系。局部加权回归散点平滑法（**Lo**cally **W**eighted **S**catterplot **S**moother，LOWESS）提供了一种非常方便的探索二元变量之间关系的图示方法(Cleveland, 1979)。LOWESS主要思想是取一定比例的局部数据，在这部分子集中拟合多项式回归曲线，这样我们便可以观察到数据在局部展现出来的规律和趋势；而通常的回归分析往往是根据全体数据建模，这样可以描述整体趋势，但现实生活中规律不总是（或者很少是）教科书上告诉我们的一条直线。我们将局部范

```
> par(mar = c(3.5, 3.5, 1, 0.2), mfrow = c(2, 2))
> with(gov.cn.pct, {
+     plot(percentage, count, type = "l", panel.first = grid())
+     plot(percentage, count, type = "l", xlim = c(10,
+         11), panel.first = grid())
+     pct.lowess(round0)
+     pct.lowess(round1)
+ })
```

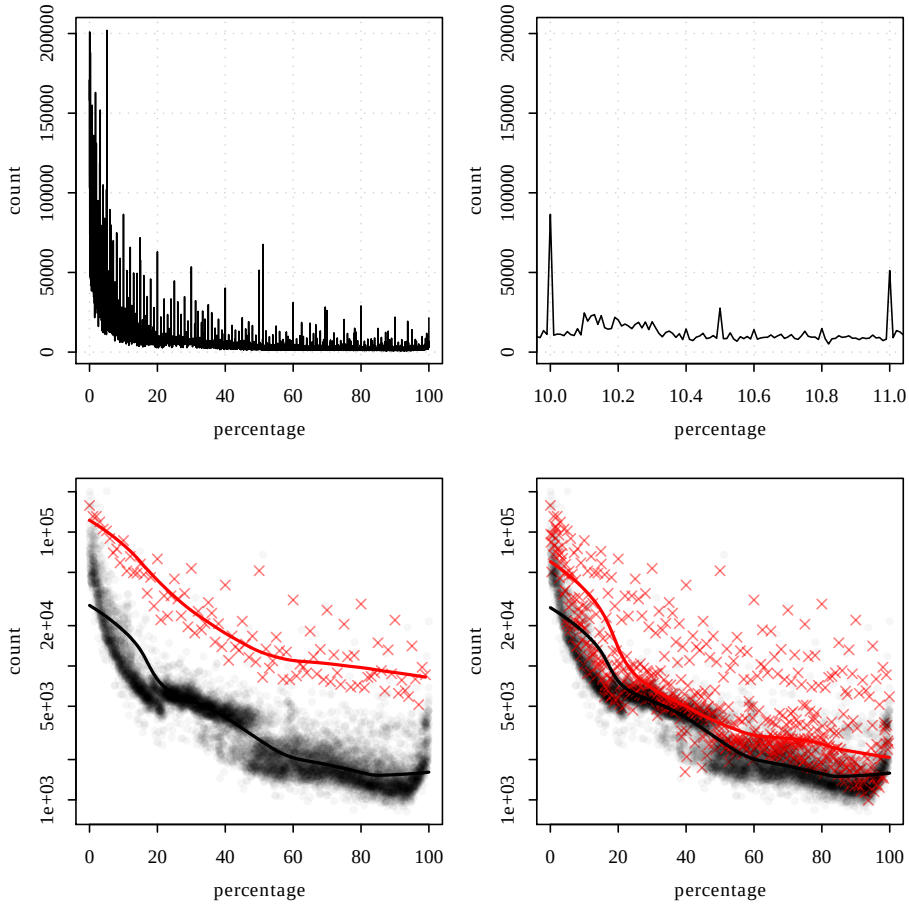

图 9: 中国政府网站中的百分比数据*LOWESS*图：首先我们画出每个百分比数据的频数（左上），然后放大$[10\%, 11\%]$区间上的频数图（右上），继而猜测数据有四舍五入的特征，所以分别对整数位和非整数位的百分比画*LOWESS*曲线（左下），最后分别对保留一位和两位小数的百分比画*LOWESS*曲线（右下）。

```
> data(PlantCounts)
> par(mar = c(3.5, 3.5, 1, 0.2), mfrow = c(1, 2), pch = 20)
> with(PlantCounts, {
+     plot(altitude, counts, col = rgb(0, 0, 0, 0.3),
+         panel.first = grid())
+     for (i in seq(0.01, 1, length = 70)) {
+         lines(lowess(altitude, counts, f = i), col = rgb(0.4,
+             i, 0.4), lwd = 1.5)
+     }
+     plot(altitude, counts, col = rgb(0, 0, 0, 0.3))
+     for (i in 1:200) {
+         idx = sample(nrow(PlantCounts), 300, TRUE)
+         lines(lowess(altitude[idx], counts[idx]),
+             col = rgb(0, 0, 0, 0.1), lwd = 1.5)
+     }
+ })
```

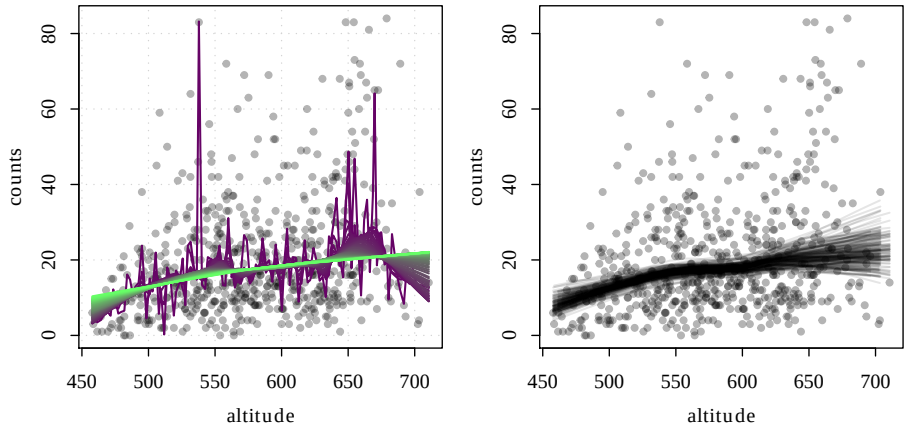

图 10: 海拔高度与物种数目的LOWESS曲线：左图为范围参数从1% 到100% 的LOWESS 曲线（深色表示范围参数小），右图为200 次Bootstrap 重抽样之后的数据分别建立的LOWESS曲线。

围从左往右依次推进，最终一条连续的曲线就被计算出来了。显然，曲线的光滑程度与我们选取数据比例有关：比例越少，拟合越不光滑（因为过于看重局部性质），反之越光滑（捕捉全局性质）。普通的线性回归可以看作LOWESS的特例：数据选取范围为全部数据，局部回归模型用一阶线性回归。

谢益辉 (2008b)提供了一个植物物种数目与海拔高度的数据，数据中记录了每个海拔高度上的某地植物物种数量。图10用LOWESS曲线对这批数据进行了初步探索。左图中，曲线颜色越浅表示所取数据比例越大。不难看出中部浅色的曲线几乎已呈直线状，而深色的线则波动较大，总体看来，图中大致有四处海拔上的物种数目偏离回归直线较严重：450米（偏低）、550米（偏高）、650米（偏高）和700米（偏低）附近。若研究者的问题是，多高海拔处的物种数最多？那么答案应该是在650米附近。如果仅仅从回归直线来看，似乎是海拔越高，则物种数目越多。但如此推断下去，必然得到荒谬的结论（地势不可能无限高）。从图中的曲线族来看，物种数目在过了650米高度之后有下降趋势，所以从这批数据来看，我们的结论将是物种数目在650米海拔处达到最大值。图10右图发挥统计计算的优势，从重抽样的角度对左图的规律作了进一步验证：我们对数据进行重抽样（在600行数据中有放回地抽取300行），并对重抽样数据画LOWESS曲线，为了得到比较稳定的规律，我们将这个过程重复200次，得到右图中的200条曲线，此处LOWESS的范围参数为默认的2/3。从Bootstrap之后的LOWESS曲线族来看，在海拔700米处的预测可能会有很大的波动，因为这一族曲线在低海拔的位置吻合较好，但在高海拔位置“分歧”比较严重，这进一步说明了我们不能简单以直线外推的方式来预测高海拔的物种数目走向。

至此我们看到了LOWESS方法的灵活性，但遗憾的是在国内大多数涉及到回归模型的图形中，我们却极少看到它的使用。本例没有任何数学推导（尽管LOWESS方法有一定的数学背景），但两个变量的所有可能关系都可以在图中的曲线中显示出来，而且LOWESS方法可以看作是一种非参数方法，不涉及到统计分布的假设，这和基于参数理论的回归也具备一定的优势。由于R语言自带*lowess()*函数可以很方便计算LOWESS曲线，所以才使得图10中的大量计算具备可能性。

3.2 提供模型之外的信息

如2.1小节所述，假设检验是各种统计学理论中最基础的理论之一，

```
> library(ggplot2)
> print(ggplot(sleep, aes(group, extra)) + stat_boxplot() +
+     coord_flip())
```

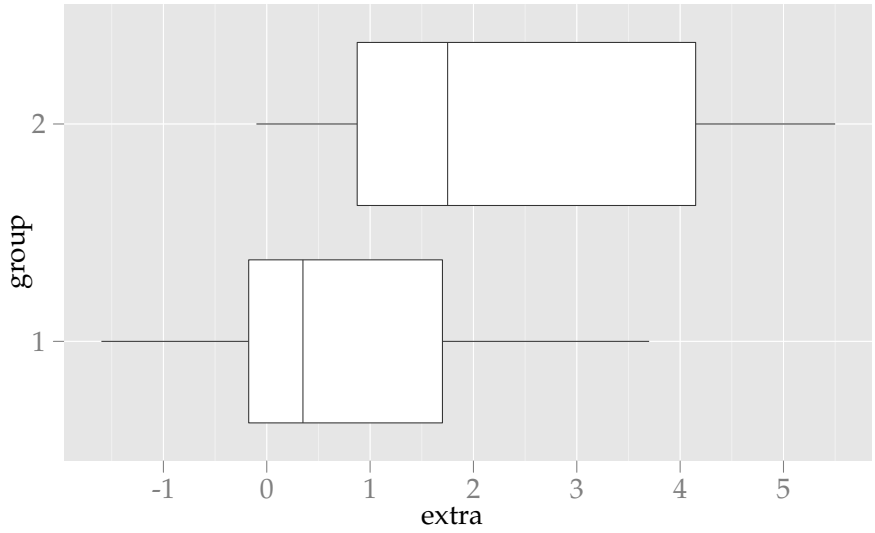

图 *11: Student*的睡眠增量数据：箱线图

但传统的假设检验给出的信息非常有限，所有的数据最终产生的只有一个P值，而且这个P值也是在诸多数学假设条件下产生的。事实上数据本身还包含着很多其它信息，我们可以用统计图形和模拟的方式去探索、表达，而不必将目光放在一个高度综合的P值上。下面我们以常见的两样本均值的检验为例，来说明图形和模拟的补充作用。

数据来自R语言自带的`sleep`数据（源于Student即戈赛特的试验），它记录了两组受试者服用不同的安眠药之后的睡眠时间增加量。数据如下：

```
> sleep
```

```
   extra group
1    0.7     1
2   -1.6     1
3   -0.2     1
4   -1.2     1
5   -0.1     1
6    3.4     1
```

```
7    3.7     1
8    0.8     1
9    0.0     1
10   2.0     1
11   1.9     2
12   0.8     2
13   1.1     2
14   0.1     2
15  -0.1     2
16   4.4     2
17   5.5     2
18   1.6     2
19   4.6     2
20   3.4     2
```

其中extra是睡眠时间增量，group是分组编号。我们感兴趣的问题是这两组受试者的睡眠时间增量的均值有无显著差异。当然我们可以用t检验或者Wilcoxon秩和检验：

```
> t.test(extra ~ group, data = sleep)

        Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.0794
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

```
> wilcox.test(extra ~ group, data = sleep)

        Wilcoxon rank sum test with continuity correction

data:  extra by group
W = 25.5, p-value = 0.06933
alternative hypothesis: true location shift is not equal to 0
```

如前文所说，假设检验提供的信息非常有限，我们除了得到一个约为0.07的P值，就几乎没有其它信息了。从统计图形的角度来说，通常两样本或多样本的比较可以用箱线图来展示如图11，箱线图除了告诉我们两组

```
> print(ggplot(sleep, aes(x = extra)) + facet_grid(group ~
+     .) + stat_density(aes(ymax = ..density.., ymin = -..density..),
+     fill = "grey50", colour = "grey50", geom = "ribbon",
+     position = "identity") + geom_point(aes(y = 0,
+     x = extra)))
```

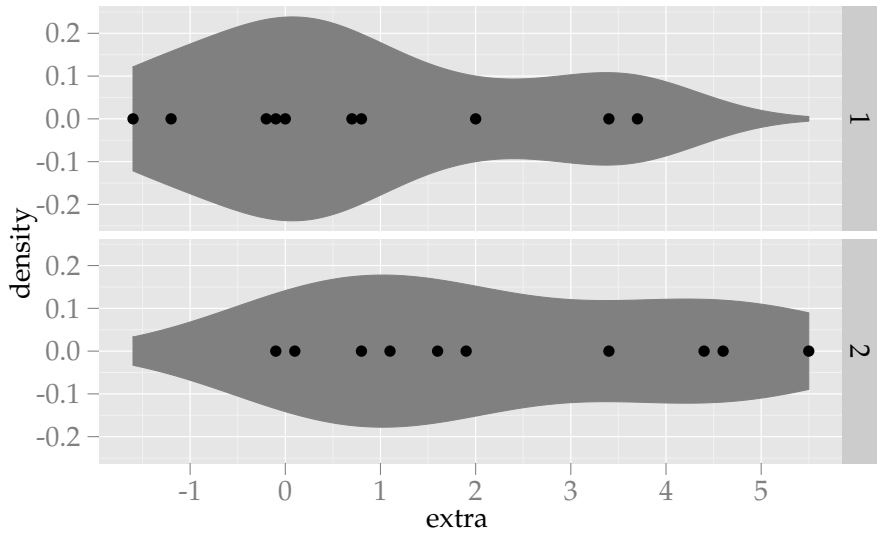

图 12: *Student*的睡眠增量数据：小提琴图。该图用密度曲线的形式展示了数据的分布，可以看出两组数据都是双峰分布，因此显然不会来自正态分布（不满足*t*检验假设）。

样本中位数有差异之外，还可以看出数据的分布是右偏的，也就是数据集中在左侧，说明更多受试者的睡眠时间增量并不大。事实上，箱线图能刻画的信息也很有限，我们可以继续用小提琴图(Hintze and Nelson, 1998)来描述两组数据的分布，如图12。小提琴图本质上就是核密度估计曲线，只不过将曲线沿x轴（或y轴）翻转，并填充曲线之间的区域，“小提琴”宽的位置表示数据的密度大，反之数据密度小。从图中我们可以看到两组数据都呈双峰分布，分布的偏度与箱线图反映出来的偏度吻合。这些图形反映出来的信息都可以促使我们进一步研究这个本来是均值比较的简单问题，例如为什么数据是右偏的？为什么两组数据都呈双峰分布？等等。

除了图形之外，我们仍然可以用重抽样的思路直接模拟两组均值的分

布。这个思路比传统的假设检验方法让人更易理解：我们要比较的是均值，那么如果把均值本身当做随机变量并且能知道均值的统计分布，也就能对均值做推断了（任何一个统计推断问题都是根据随机变量的分布作出概率推断）。那么样本均值的分布如何得到呢？对原数据进行有放回地抽样并重新计算样本均值即可—两组数据都进行这个操作，最终能得到两组数据的均值的若干“Bootstrap实现”。计算过程如下：

```
> boot.mean = tapply(sleep$extra, sleep$group, function(x) {
+     replicate(500, mean(sample(x, replace = TRUE)))
+ })
> mean(boot.mean[[2]] >= max(boot.mean[[1]]))
```

```
[1] 0.384
```

```
> mean(boot.mean[[2]] >= quantile(boot.mean[[1]], 0.95))
```

```
[1] 0.854
```

可见能够模拟两组均值各自的分布之后我们几乎能得到任意我们想知道的量，例如上面代码中的各种概率值，它们可以极大扩展单一P值所能传达的信息。两组受试者睡眠增量均值的分布如图13所示，可以看到，这两个均值的分布呈尖峰，说明均值的方差很小，加上它们的中间位置相距较远，所以可以推断两组样本均值的差异比较显著。

3.3 更新陈旧的经验法则

统计学的应用中有很多经验法则，例如用正态分布或泊松分布去近似二项分布的条件（n相对大，p不太靠近0或1），这些法则都是前人在应用过程中总结出来的，但我们须认真审视这些法则的时效性。统计学与计算密切相关，而早期的计算设备非常不发达，所以促使人们使用简化的经验去做判断，从而避免在当时看来会非常繁琐的计算。时至今日，计算对于统计学来说已经非常廉价，有些经验法则可能并没有太大意义，甚至会给出错误的指导。下面我们以Tukey快速检验的经验法则为例说明这种在早期有很大优势的经验法则可能带有很大偏误。

Tukey快速检验用于检验两样本的均值是否相同，它的做法很简单：将两组样本放在一起排序，看首尾的样本是否来自于同一组，如果是，则判断两样本均值无显著差异，否则分别从头和从尾往中间数，直到样本组别变化为止，看首尾上这样的样本的数量是多少（称为首尾计数），然后根

```
> library(ggplot2)
> boot.mean = data.frame(extra.mean = unlist(boot.mean),
+     group = gl(2, 500))
> print(ggplot(boot.mean, aes(x = extra.mean, colour = group,
+     group = group)) + geom_density(fill = NA, size = 1))
```

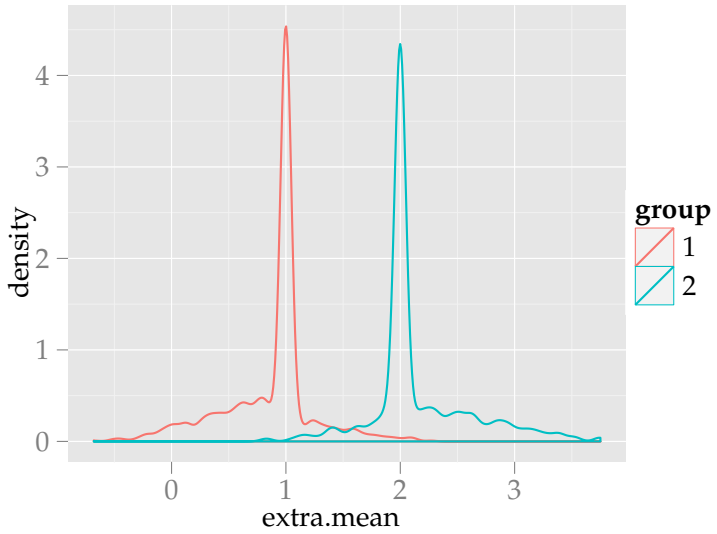

图 13: 两组受试者睡眠增量均值的分布

据Tukey的经验法则判断均值差异的显著性。首尾计数为7、10、13分别对应显著性水平5%、1%和0.1%(Basler and Smawley, 1968)。这种检验方法的操作只需要排序和数数，都可以人工完成，所以在计算机不发达的年代是一个很好的统计应用指导。然而作者通过一则模拟发现，这个经验法则可能存在很大的问题。

模拟场景如下：对于样本量分别为30的两组来自于Weibull分布（记为$W(\lambda, k)$，其中λ为尺度参数，k为形状参数）的样本，我们使用常规的均值或中位数检验计算P值，并记录Tukey首尾计数的数值，最终看首尾计数和常规统计检验的吻合程度。这两组样本分别来自$W(1, k_1)$和$W(1, k_2)$，为了使模拟进行得更充分，我们不使用固定的k_1和k_2，而是取自均匀分布：$k_1 \sim U(0.5, 4)$，$k_2 \sim U(1, 5)$。之所以选取Weibull分布，一方面是因为它常常是一些寿命数据的分布，另一方面它也是右偏的分布，我们想

查看Tukey的经验法则对非对称的数据是否适用。为了避免t检验对分布假设的依赖，我们同时也做非参数检验即Wilcoxon检验。以下是R代码（数据tukeyCount已收录进**MSG**包）：

```
> set.seed(402)
> n = 30
> tukeyCount = data.frame(t(replicate(10000, {
+     x1 = rweibull(n, runif(1, 0.5, 4))
+     x2 = rweibull(n, runif(1, 1, 5))
+     c(t.test(x1, x2)$p.value, wilcox.test(x1, x2)$p.value,
+         with(rle(rep(0:1, each = n)[order(c(x1, x2))]),
+             ifelse(head(values, 1) == tail(values,
+                 1), 0, sum(lengths[c(1, length(lengths))])))
+ })))
> colnames(tukeyCount) = c("pvalue.t", "pvalue.w",
+     "count")
```

我们可以将两种检验的P值和首尾计数的数值画在图中，看每一种计数对应的P值是否和经验法则吻合。如图14，图中大致趋势是首尾计数越大，则P值越小（两组数据均值差异越显著），但显然，这些P值与经验法则并不相符，例如首位计数为7的时候，P值大约在0.3左右，这与经验法则的0.05相差非常大。

本例模拟的动机来源于帖子http://cos.name/cn/topic/101246，作者为某工厂的6-sigma黑带，从作者的描述中我们不难看出非统计专业的专家对统计学的重视程度，但这种重视往往也容易变成迷信。如果我们从数学推导的角度去解释该作者的问题，恐怕难度会很大，但一则模拟会轻而易举给出我们的证据。本例除了说明统计模拟的教学优势之外，还表明了统计中的经验法则并不能恒久不变，在时代变迁中，我们应该适当摒弃一些旧知识，因为它们曾经的优势已经可以被更先进的技术取代。

4 小结与展望

与传统的以数学理论推导为主的统计模型解析方式不同，本文从统计图形和统计模拟的视角对统计模型理论提出了新的解析方式。这种看似不严谨的解析方式在简捷性和直观性上远胜于繁杂的数学理论，并且在统计实践操作上也可以提供建模前后的启发和指引。

```
> data(tukeyCount)
> with(tukeyCount, {
+     ucount = unique(count)
+     stripchart(pvalue.t ~ count, method = "jitter",
+         jitter = 0.2, pch = 19, cex = 0.7, vertical = TRUE,
+         at = ucount - 0.2, col = rgb(1, 0, 0, 0.2),
+         xlim = c(min(count) - 1, max(count) + 1),
+         xaxt = "n", xlab = "Tukey Count", ylab = "P-values")
+     stripchart(pvalue.w ~ count, method = "jitter",
+         jitter = 0.2, pch = 21, cex = 0.7, vertical = TRUE,
+         at = ucount + 0.2, add = TRUE, col = rgb(0,
+             0, 1, 0.2), xaxt = "n")
+     axis(1, unique(count))
+     lines(sort(ucount), tapply(pvalue.t, count, median),
+         type = "o", pch = 19, cex = 1.5, col = "red")
+     lines(sort(ucount), tapply(pvalue.w, count, median),
+         type = "o", pch = 21, cex = 1.5, col = "blue",
+         lty = 2)
+     legend("topright", c("t test", "Wilcoxon test"),
+         col = c("red", "blue"), pch = c(19, 21),
+         lty = 1:2, bty = "n")
+ })
```

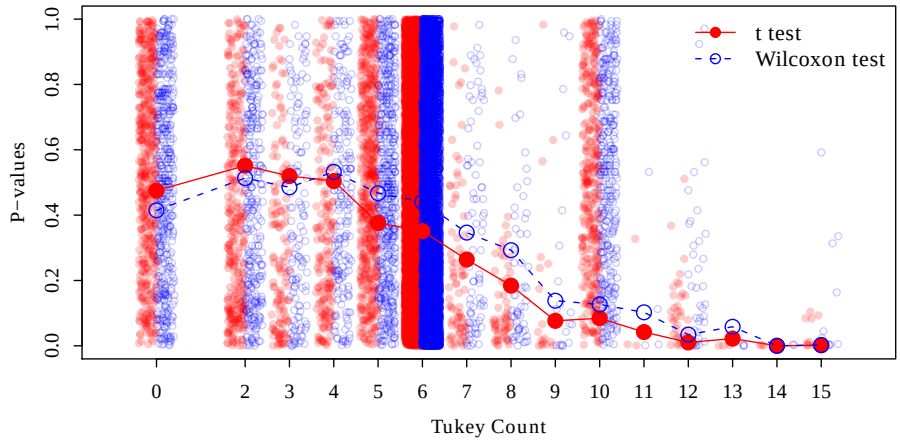

图 14: *Tukey*首尾计数的经验法则与常规检验的*P*值：对每一种首尾计数，分别画出对应的*P*值，左侧为*t*检验*P*值，右侧为*Wilcoxon*检验*P*值，每种检验方法的*P*值中位数用大点标示并连线。注意图中的点的横坐标稍作随机打乱，以免点与点过度重叠。

作者认为，现代统计教育对计算机的利用程度还远远不够，例如我们仍然在课堂上和教科书中随处可见如下内容：

- 为了计算的简便，用泊松分布近似二项分布
- 手工查统计分布表
- 通过比较临界值和统计量的值来决定是否拒绝零假设，而不直接介绍P值
- 用概率图纸的方法检验正态分布

这都是统计教学严重过时的表现。当今计算机的能力已经比皮尔逊/费歇尔时代强得太多：我们已经拥有可以计算精确分布的能力，为什么还需要使用那些早已不需要的数学近似方法？我们可以很方便画出QQ图或者用各种数值检验方法来检查数据正态性，为什么还需要在概率纸上用铅笔描点？

本文的大量实例表明，有效利用现代统计图形和统计模拟或统计计算，可以帮我们跨越很多数学障碍，让看起来并不直观的理论问题拥有直观的意义，让数学无法解决或很难解决的问题能够找到（可能并不精确的）答案。数学是统计学的理论支柱，但结合科学和艺术于一体的统计学并非仅由科学的理论构成，我们可以从多个角度解读统计学的方法、理论和模型。

由于计算机的发展和统计学理论的起源不在同一年代，使得数学理论推导的方式在统计学中拥有主要地位，然而理论推导也会受计算实现的限制，其中一个非常典型的例子就是贝叶斯统计学在近年来的复兴。贝叶斯思想几百年前就已经诞生，比频率学派早得多，它后来之所以难以取得实质性进展，原因就是遇到了计算问题（主要是数值积分和抽样），仅有理论而无法付诸实现，而现在有了高性能的计算机和算法，很多需要庞大计算量的问题可以被轻而易举解决。贝叶斯统计推断都是基于参数的后验分布，计算的方式（如MCMC）也就是从后验分布中产生大量的随机数，根据这些随机数做推断，这种方式在严谨的数学家眼中可能显得太粗略：由计算机生成的随机数也可以作为统计推断的依据？答案应该是肯定的。从某种程度上来说，与其在完美的假设下推导出精确的理论但又无法知道理论假设在实际应用中是否能被满足（典型例子如频率学派视为瑰宝的渐近理论或大样本理论：样本量多大能视为“趋于无穷”?），不如从实用角度给出近似但足够可靠的答案。

作者自2007年起开始关注统计模拟和统计图形对统计理论的解释作用，并以动画的形式实现了一部分统计方法的展示，这些工作主要收录在R附加包**animation**(Xie, 2010a)中，包括统计计算中的二分法、牛顿法、梯度下降法、蒙特卡洛积分，概率论中的高尔顿板（Quicunx）、蒲丰投针、大数定律、中心极限定理，多元统计学中的K-Means聚类，抽样调查中的简单随机抽样、分层抽样、整群抽样、系统抽样、比率估计，数据挖掘中的k-近邻方法等等，这些展示统计理论方法的动画引起了国内外一些统计学教师的关注，作者也看到它们出现在统计课堂教学上，如中科大统计与金融系的概率论和数理统计课程（`http://staff.ustc.edu.cn/~zwp/teach.htm`）等。本文将过去的一些零散的展示系统化，在一定的逻辑框架下整理了统计模拟和统计计算的优势和用途。今后作者将继续关注这个主题，使得这项工作的意义超出教学范围，能够展示更多应用价值。

最后值得一提的是，本文的所有计算与分析都是基于R语言—强大而灵活的统计分析语言，它可以说是与统计学发展结合最紧密的统计软件，几乎涵盖了所有我们能见到的统计模型方法，并且它有很好的扩展性，这使得前沿方法能以最快的速度在R中得以实现。而本文档则是基于R中的Sweave(Leisch, 2002)以及**pgfSweave**包(Sharpsteen and Bracken, 2009)生成；Sweave是R语言的一大优势，它将LaTeX文档与R计算紧密融合在一起，使得统计计算和图形能够动态嵌入文档[5]，因此文档具有可重复性（reproducibility），任何读者只要拥有R软件和LaTeX并获得本文的源文档，都可以重复生成本作者的结果（以及整篇PDF文档）。Sweave并非只是一个技术问题，它让统计报告变得可公开透明，一切统计分析过程都在文档中，因此不可能存在任何编造分析结果的可能性，这一点又是现代统计计算新技术的优势，并且我们相信它对于当前普遍的“Word+Excel/SPSS+手工复制粘贴”的统计分析报告方式会是一大重要革新。

A MSG程序包

为了配合本论文的写作，作者编写了一个R包名为**MSG**，该包目前可从作者主页下载（`http://yihui.name/cn/publication/`），在完成后会发布到CRAN供用户下载。这里简要介绍一下它包含的函数和数据。

[5] “动态”的意思是文档中包含的计算结果和图形都会根据R代码自动更新，例如更改R的输入数据之后所有图表无需手工改动，只需要重新编译文档，所有结果都会自动更改。

A.1 函数说明

暂无。

A.2 数据说明

BinormCircle 人造数据：两个独立的正态分布随机变量（10000行实现值），加上半径为0.5的圆上的点的坐标（10000行）

gov.cn.pct 中国政府网站中从0%到99.99%的每一个百分比的Google搜索频数，共10000行

PlantCounts 植物数目与海拔高度的数据，共两列，记录了某一海拔高度上植物物种数目

t.diff 异方差情况下的t检验P值差异：如果真实总体方差不同，那么使用等方差假设和异方差假设都可以计算t检验的P值，本数据记录了两组异方差而且样本量不同的数据的t检验P值差异（分别用等方差和异方差方法检验），对于每一种样本量组合（$n_1 = 2, 3, \cdots, 100$; $n_2 = 30$），重复模拟1000次，记录1000个P值差异，最终得到1000×99的矩阵

tukeyCount 对两组样本，分别用t检验和Wilcoxon检验计算P值，并记录相应的Tukey首尾计数，不断重复生成样本，得到10000行数据，前2列为P值，后1列为首尾计数

参考文献

谢益辉(2008a). “统计图形在数据分析中的应用.” In 张波(ed.), *统计学评论*. 中国财政经济出版社.

谢益辉(2008b). “用局部加权回归散点平滑法观察二维变量之间的关系.” *统计之都*. 最后访问于2010年3月21日, URL `http://cos.name/2008/11/lowess-to-explore-bivariate-correlation-by-yihui/`.

Adler D, Murdoch D (2010). *rgl: 3D visualization device system (OpenGL)*. R package version 0.90, URL `http://CRAN.R-project.org/package=rgl`.

Basler DD, Smawley RB (1968). "Tukey's Compact versus Classic Tests." *The Journal of Experimental Education*, **36**(3), 86–88.

Becker RA, Chambers JM, Wilks AR (1988). *The New S Language*. Wadsworth & Brooks/Cole.

Chambers JM, Cleveland WS, Kleiner B, Tukey PA (1983). *Graphical Methods for Data Analysis*. Wadsworth & Brooks/Cole.

Cleveland WS (1979). "Robust locally weighted regression and smoothing scatterplots." *Journal of American Statistical Association*, **74**, 829–836.

Cleveland WS (1985). *The Elements of Graphing Data*. Monterey, CA: Wadsworth.

Cleveland WS (1993). *Visualizing Data*. Hobart Press.

Cook D, Swayne DF (2007). *Interactive and Dynamic Graphics for Data Analysis With R and GGobi*. Springer. ISBN 978-0-387-71761-6.

Efron B (1979). "Bootstrap methods: another look at the jackknife." *The Annals of Statistics*, **7**(1), 1–26.

Friendly M, Denis DJ (2001). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Accessed: March 18, 2010, URL `http://www.math.yorku.ca/SCS/Gallery/milestone/`.

Hintze JL, Nelson RD (1998). "Violin plots: a box plot-density trace synergism." *The American Statistician*, **52**(2), 181–4.

Ihaka R, Gentleman R (1996). "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*, **5**(3), 299–314. ISSN 10618600.

Inselberg A (2007). *Parallel Coordinates: VISUAL Multidimensional Geometry and its Applications*. Springer.

King G, Tomz M, Wittenberg J (2000). "Making the most of statistical analyses: Improving interpretation and presentation." *American Journal of Political Science*, **44**(2), 347–361.

Leisch F (2002). "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis." In W Härdle, B Rönz (eds.), *Compstat 2002 — Proceedings in Computational Statistics*, pp. 575–580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9, URL `http://www.stat.uni-muenchen.de/~leisch/Sweave`.

McGill R, Tukey JW, Larsen WA (1978). "Variations of box plots." *The American Statistician*, **32**, 12–16.

Meyer D, Zeileis A, Hornik K (2010). *vcd: Visualizing Categorical Data*. R package version 1.2-8, URL `http://CRAN.R-project.org/package=vcd`.

Murrell P (2005). *R Graphics*. Chapman & Hall/CRC.

Nightingale F (1858). "Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army." *Technical report*.

Playfair W (1801). *The statistical breviary*. London: T. Bensley.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Sarkar D (2010). *lattice: Lattice Graphics*. R package version 0.18-3, URL `http://CRAN.R-project.org/package=lattice`.

Scott DW (1992). *Multivariate Density Estimation. Theory, Practice and Visualization*. New York: Wiley.

Sharpsteen C, Bracken C (2009). *tikzDevice: A Device for R Graphics Output in PGF/TikZ Format*. R package version 0.4.8, URL `http://CRAN.R-project.org/package=tikzDevice`.

Simon JL (1997). *Resampling: The New Statistics*. 2nd edition. Resampling Stats. URL `http://www.resample.com/content/text/index.shtml`.

Simon JL, Atkinson DT, Shevokas C (1976). "Probability and statistics: Experimental results of a radically different teaching method." *The American Mathematical Monthly*, **83**(9), 733–739.

Symanzik J (2004). *Handbook of Computational Statistics*, chapter Interactive and Dynamic Graphics, pp. 293–336. 1 edition. Springer.

Temple Lang D, Swayne D, Wickham H, Lawrence M (2009). *rggobi: Interface between R and GGobi*. R package version 2.1.14, URL `http://CRAN.R-project.org/package=rggobi`.

Theus M (2002). "Interactive Data Visualization using Mondrian." *Journal of Statistical Software*, **7**(11), 1–9. ISSN 1548-7660. URL `http://www.jstatsoft.org/v07/i11`.

Tufte ER (1992). *Envisioning Information*. Cheshire, CT, USA: Graphics Press. ISBN 0-961-39211-8.

Tufte ER (2001). *The Visual Display of Quantitative Information*. 2nd edition. Cheshire, CT, USA: Graphics Press. ISBN 0-9613921-4-2.

Tukey JW (1977). *Exploratory data analysis*. Massachusetts: Addison-Wesley.

Unwin A, Hawkins G, Hofmann H, Siegl B (1996). "Interactive Graphics for Data Sets with Missing Values: MANET." *Journal of Computational and Graphical Statistics*, **5**(2), 113–122. ISSN 10618600.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer. ISBN 0-387-95457-0.

Wand M (2009). *KernSmooth: Functions for kernel smoothing for Wand & Jones (1995)*. S original by Matt Wand. R port by Brian Ripley. R package version 2.23-3, URL `http://CRAN.R-project.org/package=KernSmooth`.

Welch BL (1947). "The generalization of Student's problem when several different population variances are involved." *Biometrika*, **34**(1/2), 28–35.

Wickham H (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN 978-0-387-98140-6. URL `http://had.co.nz/ggplot2/book`.

Wilkinson L (2005). *The Grammar of Graphics*. 2nd edition. Springer.

Xie Y (2010a). *animation: Demonstrate Animations in Statistics*. R package version 1.1-0, URL `http://animation.yihui.name`.

Xie Y (2010b). *MSG: Modern Statistical Graphics*. R package version 0.1-0, URL `http://yihui.name/cn/publication`.

索引

Bootstrap, 2, 28

Cook距离, 17

LMS回归, 14
LOWESS曲线, 21, 23, 24

OLS回归, 16

R语言, 2

Sweave, 33
S语言, 1

Tukey快速检验, 29
t检验, 6, 26, 29

Welch校正, 7
Wilcoxon检验, 29

中心极限定理, 6
二项分布, 29
交互作用, 12, 13

假设检验, 25
可重复性, 33
图形历史, 1
小提琴图, 27
平滑散点图, 6

控制变量, 12
散点图, 4, 12, 16
数学理论, 1, 30
正态分布, 29
气泡图, 15
泊松分布, 29

离群点, 16, 17
箱线图, 10
线性回归, 4, 11
经验法则, 29
统计动画, 2, 6
统计图形, 1, 5, 8, 9, 11, 13, 15, 17, 19, 23, 27, 32
统计教学, 3
统计模型, 4, 11
统计模拟, 1–3, 7, 12, 14, 27, 29, 32
统计计算, 1, 32

部分抽样, 17
重抽样, 2, 17, 27
首尾计数, 29

致谢

首先感谢我的导师赵彦云老师对我多年来在学术上的指导和生活上的关怀，自本科时起，赵老师便经常给我讲一些学术研究方向上的经验和建议，鼓励我关注国外研究动向，并提供了大量的国内外交流机会，这些经历大大拓宽了我的眼界，更为重要的是他也非常尊重我个人的意见和判断，让我有充分的学术发展自由，例如他知道我对前沿统计模型和统计图形感兴趣，便尽力为我创造便利的学习研究条件；此外，我也要感谢人大统计学院的各位老师这几年来对我的谆谆教导，特别要感谢林秋池老师对我的悉心关怀和帮助。此外，我也要感谢“统计之都”网站（http://cos.name）上各位网友的讨论（本文3.1小节源于作者发表在“统计之都”主站上的文章，3.3小节源于一篇论坛帖子）；感谢在我的个人网站上留言的各位朋友，这些讨论为本文写作提供了重要的启发和帮助（本文2.1、2.2和2.3小节源于作者发表在个人网站上的文章）。