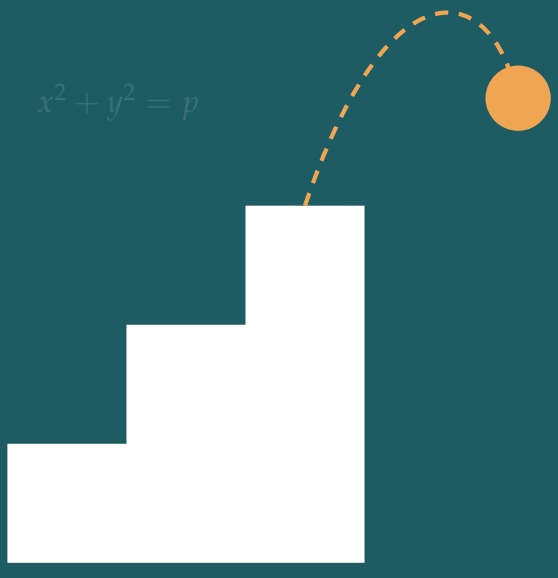
STEP 备试 · 大学数学

Beyond STEP

第 13 章

现实世界中的统计与随机性



KEITH McNULTY

Beyond STEP

以历届 STEP 试题为门
走进大学数学

第 13 章

现实世界中的统计与随机性

Keith McNulty

Beyond STEP：以历届 STEP 试题为门，走进大学数学

第 13 章：现实世界中的统计与随机性

本册是 *Beyond STEP* 全书 13 章中的第 13 章。全书、各章的单行本、每一节的可列印工作纸，以及每一道 STEP 试题的详细解答，均可在 `beyondstep.keithmcnulty.org` 免费阅读及下载。

2026 年 11 月首次于网上发表。

本 PDF 内凡属作者本人创作的内容，包括正文、解答、插图、练习及答案，均按照 Creative Commons CC0 1.0 Universal 授权条款贡献予公有领域：你可以为任何目的复制、改编及分享，毋须征求许可。详见 `creativecommons.org/publicdomain/zero/1.0`。

本书讨论的 STEP 试题，版权属 Cambridge University Press & Assessment 所有。本书并无转载这些试题：每一题均附连结，通往官方来源的题目，即 STEP Support Programme 的试题资料库或 OCR 的历届试卷。

如有错漏，欢迎到 `github.com/keithmcnulty/step_book_mcnulty` 提交 issue 指正。

章

13

统计与现实世界中的随机性

概率由模型推向数据，统计则由数据返回现实世界。

本章内容

13.1	解读新闻标题	STEP I 1992
13.2	计数田鼠	STEP II 2000
13.3	德国坦克问题	STEP III 1998
13.4	平均数还是中位数?	STEP II 2000
13.5	问尴尬的问题	STEP II 2023
13.6	累积起来的误差	STEP II 1993
13.7	制造相关性	STEP III 2010
A	练习答案	第 51 页

从数据中学习

所有模型都是错的，但有些是有用的。

— 统计学家 George Box

在概率（probability）中，我们知道骰子怎样被动了手脚，然后问会发生甚么事。在统计（statistics）中，我们看到已发生的事，然后问骰子必定是怎样被动了手脚。概率由因推果，统计则由果溯因。对总体的每一个估计、每一次民意调查、每一项临床试验、每一次科学测量，都是一个统计推断（statistical inference）：根据有限而带有杂讯的数据，对现实世界作出结论。

与数学的大部分分支相比，统计是一门年轻的学科。它的现代面貌是在二十世纪初由皮尔逊（Karl Pearson）、费雪（Ronald Fisher）、奈曼（Jerzy Neyman）等人建立的，背后的推动力往往是实际问题：酿造啤酒（t 检验就是在健力士（Guinness）啤酒厂发明的）、种植农作物、测试药物。时至今日，随着数据量爆炸式增长，统计及其近亲机器学习（machine learning）已成为数学最重要的应用之一。

统计为何重要

有关药物、公共卫生、经济政策、气候和司法的决定，全都建基于统计证据。误读这些证据会带来实实在在的后果：病人被假阳性（false positive）结果吓坏（Section 12.2），政策建基于一次偶然的波动（Section 13.1），电脑系统中的微小误差不断累积（Section 13.6）。明白数据能告诉你甚么、不能告诉你甚么，以及任何结论有多不确定，是作为一个知情公民的一部分，对科学家、医生、工程师和经济学家而言更是专业上的必需。

你在大学会遇到甚么

统计的入门课程涵盖估计（estimation，即为未知量找出好的估计值，并说明其精确程度）、假设检验（hypothesis testing，即判断数据是否与某个假设相符）、置信区间（confidence interval）、回归（regression，即把一个变量与其他变量联系起来），以及实验和调查的设计。核心概念是*抽样分布*（sampling distribution，即估计值本身是一个随机变量（random variable））、*偏差*（bias）与*方差*（variance），以及*似然*（likelihood）。其后的课程会加入贝叶斯推断（Bayesian inference）、多元分析（multivariate analysis）、时间序列（time series）和机器学习的方法，而机器学习其实就是规模极其庞大的统计。

贯穿本章的问题

三大问题

1. **是真有其事，还是只是杂讯？** 单凭偶然，我们应预期有多大的变动？一项变化要到甚么程度才值得认真看待？
2. **我们应该怎样估计？** 有了数据，一个未知量的最佳估计是甚么？我们又怎样比较不同的方法？
3. **随机性怎样进入数据？** 数据的收集、保护或处理方式，会对我们的结论造成甚么影响？

本章先从以批判的眼光阅读报章上的统计数字开始（Section 13.1），然后转向估计：用捉放法（capture–recapture）计数田鼠的数目、由序号估计总数，以及比较平均值（mean）与中位数（median）。最后是三道关于随机性怎样进入数据的题目：保护隐私的随机化调查、会累积或互相抵消的舍入误差（rounding error），以及构造具有指定相关性的变量。

本章的概念

§	课题	概念
13.1	解读统计数字	比率、自然变异、趋势
13.2	捉放法	最大似然（maximum likelihood）
13.3	序号	偏差、方差与充分性（sufficiency）
13.4	估计中心位置	效率（efficiency）与稳健性（robustness）
13.5	随机化回答（randomised response）	混合分布（mixture）与总方差（total variance）
13.6	舍入误差	系统误差（systematic error）与随机误差（random error）
13.7	相关（correlation）	楚列斯基（Cholesky）构造法

Section 13.7 会用到协方差（covariance）和相关，这两个课题已于 2019 年从 STEP 考试范围中删去；该节会解释所需的知识。

怎样阅读本章

请记住，每一个估计值都是随机变量：假如数据有所不同，估计值也会随之不同。对于本章的每一种方法，都要问一问：重复很多次之后，它的分布会是怎样的？平均而言它是否正确？它的变动又有多大？这一个问题，正是统计思维的核心。

13.1 解读新闻标题

题目

STEP I 1992，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/92-s1-q14`

引子

统计学（statistics）是从数据中学习的科学，而它的第一课就是：数字本身不会说话。“安全六倍”、“升幅惊人”、“死伤日益惨重”：新闻标题每天都把数字变成论断，而这些论断往往是错的；错不在数字本身有假，而在于比较的对象不对，或是把寻常的随机波动（random variation）误当成趋势。

这道题完全没有公式，在 STEP 中甚为罕见。它要求你像统计学家那样思考：要证明每个标题成立，我们需要知道甚么？单凭偶然，我们又应预期有多大的波动？答案要用到泊松分布（Poisson distribution）（Section 12.18），而当中的推理正是本章其余部分的基础。

动笔之前

? **想一想** 假如死亡事件以稳定的比率独立而随机地发生，每年的总数应有多大的变动？若平均每年为 1080，每年与平均值的典型偏差是多少？

独立罕见事件的计数近似服从泊松分布，而平均值为 μ 的泊松计数，其方差（variance）为 μ，标准差（standard deviation）为 $\sqrt{\mu}$。取 $\mu = 1080$，标准差约为 33：每年总数落在约 1015 至 1145 之间，都不足为奇。

逐步解题

阶段 1 “坐飞机比步行安全六倍”

想一想 每年有 1080 名行人死亡，180 人死于空难。六倍，究竟是甚么的六倍？

这个标题比较的是死亡总数，但安全与否关乎的是*风险*：每单位风险承担量（exposure）的死亡人数。几乎人人每天都步行，坐飞机的人却相对较少，而且只是偶一为之。公平的比较应采用每公里行程、每小时行程或每次行程的死亡人数，而这几种算法可以得出截然不同的答案（按每公里计，航空旅行极为安全；按每次行程计则没有那么出色，因为大多数空难死亡都发生在起飞和降落之时）。题目所给的数字根本不能“证明”两者孰更安全。此外，“在道路意外中丧生的行人”是个狭窄的类别；整体的道路死亡人数要大得多。

阶段 2 “我们的天空更安全：1990 年仅 125 人丧生”

想一想 空难死亡像独立的罕见事件吗？你预期每年总数的变动有多大？

空难死亡是成群出现的：一次坠机就可能夺去数百人的性命。因此每年的总数主要取决于坠机的次数，而这个次数很小，于是总数每年的变动极大。某一年有 125 人死亡，相对于平均的 180 人，与情况毫无改变完全吻合：只要多一次坠机，这个数字就会彻底改观。一年的数据不能显示趋势。

阶段 3 “道路死伤日益惨重：星期二 7 人丧生”

想一想 平均每天有多少名行人丧生？若每天的死亡人数服从泊松分布，一天之内出现 7 人或以上的情况有多常见？

平均每天有 $\frac{1080}{365} \approx 2.96$ 名行人死亡。若每天的计数服从这个平均值的泊松分布，则 $P(\geq 7) \approx 0.031$：纯粹出于偶然，每年就约有 11 天会有 7 名或以上的行人死亡。（而且“道路死伤”想必还包括司机和乘客，那么一天之内有 7 人死于道路意外就更不值得大惊小怪了。）单单一天的数字，对趋势毫无说明作用。这个标题从一个变动极大的数列中挑出一个极端值，是典型的谬误。

阶段 4 “升幅惊人：1990 年 1350 名行人丧生”

想一想 把 1350 与泊松标准差比较。这纯属偶然吗？若不是，还有甚么可以解释它？这又足以称为“升幅惊人”吗？

这是全题最有分量的部分。在稳定的泊松模型之下，1350 比平均值 1080 高出 270，约为 8 个标准差：这极不可能是偶然所致。所以很可能确实有些事情改变了。但这个标题仍然言过其实，原因有以下几方面：

- *被平均值掩盖的趋势。*1080 这个数字是 16 年的平均值。假如死亡人数由 1974 年的约 900 稳步上升至 1989 年的约 1260，平均值仍然是 1080，而 1990 年的 1350 只不过是

延续既有的趋势，并非突如其来的“惊人”升幅（图 13.1.1）。我们需要逐年的数字。

- 风险承担量。人口增长、步行的人增多、交通更繁忙或人口老化，都会令死亡人数上升，而道路按每次行程计却没有变得更危险。按人均计算，或按每公里步行路程计算的比率，才是恰当的量度。
- 定义。死亡个案的记录或分类方式一旦改变（例如意外发生后多久之内死亡才计算在内），就可以造成数字的跳升。
- 过度离散（*over-dispersion*）。现实中每年计数的变动比泊松模型所显示的要大，因为各种条件（天气、经济活动）都在变。对“8 个标准差”这个说法应有所保留。

升幅很可能是真实的，但它是否惊人、成因为何，则需要更多数据才能判断。

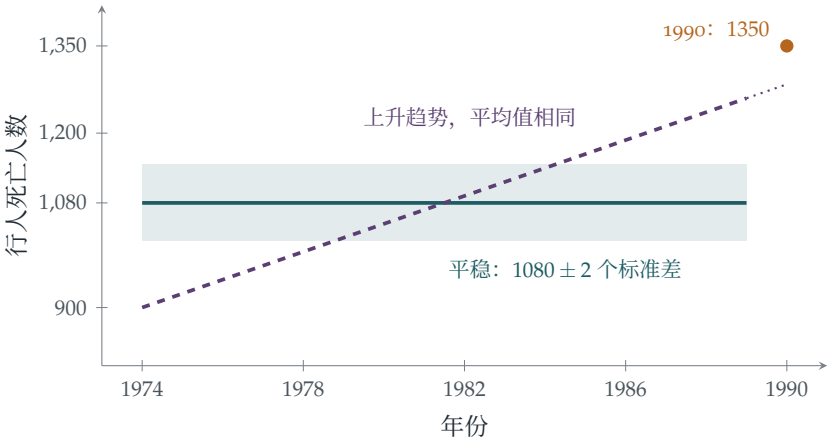


图 13.1.1 两段截然不同的历史，16 年的平均值同样是 1080。假如死亡人数一直平稳（蓝绿色，连同泊松分布典型波动的范围带），1990 年的 1350 就是急剧的跳升。假如死亡人数一直稳步上升（紫红色），1350 就只是略高于趋势。单凭平均值，无法分辨这两种情况。

回顾

- **同类才可相比。**风险需要有分母：每人、每公里或每次行程的死亡人数。
- **认识自然波动。**泊松分布的变异性，以及成群出现的事件那大得多的变异性，让我们知道甚么时候的变化才值得留意。
- **平均值会掩盖趋势。**单凭一个概括的数字，无法描述一个量如何随时间变化。

背后的数学

信号与杂讯

每一个统计问题，归根究底都是要把*信号*（signal，即真实的效应或趋势）与*杂讯*（noise，即随机波动）分开。泊松模型为独立事件计数中的杂讯提供了一把尺：标准差为 $\sqrt{\mu}$。罕见事件的*相对*波动很大：计数为 9 时，波动约为 ± 3，即其大小的三分之一；而计数为 10 000 时，波动只有约 1%。正因如此，小数目（某一天的死亡人数、某一年的坠机次数）才会屡屡被人过度解读，而小镇上的癌症“群组”最终也多半只是偶然。

均值回归

还有另一个陷阱。道路安全摄影机往往安装在刚刚经历了异常恶劣一年的地点。翌年，意外减少，功劳就归于摄影机。但异常恶劣的一年过后，本来就很可能迎来较为平常的一年：这就是*均值回归*（regression to the mean），最早由高尔顿（Francis Galton）在 1880 年代提出。要把摄影机的效果与均值回归分开，就必须与没有安装摄影机的类似地点作比较，这正是对照实验（controlled experiment）背后的理念。

南丁格尔的统计学

南丁格尔（Florence Nightingale）以克里米亚战争中的护士身分闻名于世，同时也是统计学的先驱。她在 1858 年绘制的“玫瑰图”显示，死于可预防疾病的士兵远多于死于创伤的士兵，并说服了政府改革军队的卫生状况。她坚持使用比率（每年每千名士兵的死亡人数）而非原始总数，并坚持同类才可相比，这些正是本题所考核的要点。她是首位获选为皇家统计学会（Royal Statistical Society）会士的女性。

思维技巧

问一句：“跟甚么相比？”

在接受一个统计论断之前，先问清楚：比较的对象是甚么，分母是甚么，以及单凭偶然会预期有多大的波动。

到了大学，统计学课程会把这些问题化为精确的方法：假设检验（hypothesis test）探究一项观察结果是否与偶然相符；置信区间（confidence interval）说明一个量测量得有多精确；而实验设计（design of experiments）则确保比较公平。

延伸练习

1 **泊松分布的尾部。**设 X 服从平均值为 2.96 的泊松分布，验证 $\mathrm{P}(X \geq 7) \approx 0.031$。你预期每年有多少天会有 9 名或以上的行人死亡?

2 **成群出现的死亡。**假设坠机事故按泊松过程（Poisson process）发生，比率为每年 1.2 宗，而每宗夺去 150 人的性命。求每年空难死亡人数的平均值和标准差。一年之内有 125 人或以下死亡，是否不寻常？有没有可能某一年恰好有 125 人死亡?

3 **显著的升幅?**假如 1974–1989 年的数字一直平稳地维持在 1080 ± 33，试（利用正态近似（normal approximation））求单一年份出现 1350 或以上的概率。

4 **拟一个更好的标题。**为四个标题逐一改写，各拟一个数据真正支持的标题。

延伸阅读

- **David Spiegelhalter, *The Art of Statistics***（Pelican）。讲述如何从数据中学习，并附有大量取自新闻的例子。
- **Tim Harford, *How to Make the World Add Up***（Bridge Street）。思考传媒中的数字的十条法则。
- **Darrell Huff, *How to Lie with Statistics***（Penguin）。1954 年出版的短篇经典，至今仍不过时。

13.2 计数田鼠

题目

STEP II 2000，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/00-s2-q13`

引子

一个岛上住了多少只田鼠？逐只去数是不可能的：牠们会躲藏，会走动，而且长得一模一样。不过，你可以先捉一些，做上记号，放走，然后再捉一批。假如有 200 只做了记号，而第二次捉到的 200 只当中有 11 只带有记号，那么整个种群中大约有 $\frac{11}{200}$ 带有记号，所以种群的大小约为 $200 \times \frac{200}{11} \approx 3636$。

这种*标记重捕法*（capture–recapture）用于估计鱼类、鸟类和鲸鱼的数量，也用于估计吸毒者、无家者和战争伤亡者的人数，即任何无法完全计数的总体。这道题用*最大似然*（maximum likelihood）原理为这个直观的估计打下稳固的基础。最大似然是统计估计中最重要的思想：选取使观察到的数据出现概率最大的那个未知量的值。

动笔之前

想一想 种群共有 N 只田鼠，其中 200 只带有记号。现随机抽取 200 只作为样本。恰好有 11 只带有记号的概率是多少？

从 200 只有记号的田鼠中选出 11 只，再从 $N-200$ 只没有记号的田鼠中选出 189 只，而等可能的样本共有 $\binom{N}{200}$ 个：

$$p_N = \frac{\binom{200}{11}\binom{N-200}{189}}{\binom{N}{200}}.$$

这就是*超几何分布*（hypergeometric distribution）。

逐步解题

阶段 1　似然

把含有 N 的二项式系数（binomial coefficient）展开写出，

$$\binom{N-200}{189} = \frac{(N-200)!}{189!\,(N-389)!}, \qquad \binom{N}{200} = \frac{N!}{200!\,(N-200)!},$$

所以

$$p_N = \binom{200}{11}\frac{200!}{189!}\cdot\frac{((N-200)!)^2}{N!\,(N-389)!} = k\frac{((N-200)!)^2}{N!\,(N-389)!}. \qquad \blacksquare$$

把数据固定，视之为未知量 N 的函数时，p_N 称为*似然*（likelihood）。

阶段 2　对整数求最大值

> **想一想**　N 是整数，所以用微积分并不方便。不如比较 p_{N+1} 与 p_N。两者之比何时大于 1？

大部分阶乘都会互相约去：

$$\frac{p_{N+1}}{p_N} = \frac{(N-199)^2}{(N+1)(N-388)}.$$

当 $(N-199)^2 > (N+1)(N-388)$ 时，这个比大于 1，即 $N^2 - 398N + 39601 > N^2 - 387N - 388$，亦即 $11N < 39989$，$N < 3635.4$。所以 p_N 一直递增至 $N = 3636$，其后递减：最大似然估计值为

$$\hat{N} = 3636. \qquad \blacksquare$$

这正是 $\frac{200\times 200}{11} = 3636.4$ 的整数部分，也就是那个直观的估计。

阶段 3　第三个样本

抽取第二个样本之后，其中 189 只没有记号的田鼠也被做上记号，所以现在共有 389 只田鼠带有记号。取 $N = 3636$，则有 $3636 - 389 = 3247$ 只没有记号。第三个含 200 只田鼠的样本中恰好有 j 只带有记号的概率为

$$\frac{\binom{389}{j}\binom{3247}{200-j}}{\binom{3636}{200}}.$$

这些概率（对 $j = 0, 1, \ldots, 200$）之和必定等于 1，由此即得该恒等式。 $\blacksquare$

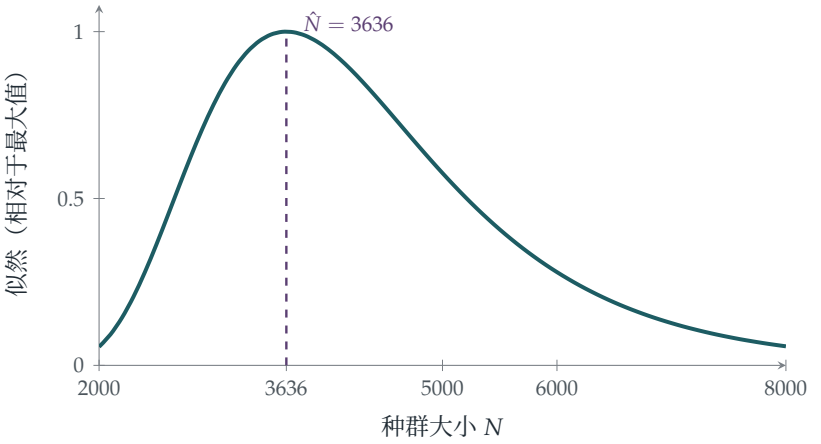


图 13.2.1 在第二个样本中发现 11 只有记号田鼠的似然 p_N，表示为种群大小 N 的函数，并按比例缩放使其最大值为 1。它在 $N = 3636$ 处达到峰值，但曲线宽阔而且偏斜：由大约 2500 至 6500 的种群大小，都与数据相当吻合。

回顾

- **把数据的概率看成未知量的函数。**似然把惯常的观点倒转过来：数据固定不变，参数（parameter）才是变量。
- **用比值求离散最大值。**比较相邻两项，便找出了一个整数变量函数的最大值所在。
- **概率之和为 1。**把一个分布对所有结果求和，便证明了一个组合恒等式（范德蒙恒等式，Vandermonde's identity）。

背后的数学

最大似然

最大似然原理由费雪（Ronald Fisher）在 1910 至 1920 年代发展出来，内容是：以使观察到的数据出现概率最大的那个值，作为未知参数的估计。它是至今应用最广的估计方法，由以曲线拟合数据，到训练机器学习中的神经网络，无一不用；神经网络的“损失函数”（loss function）通常就是负对数似然。费雪证明了，对于大样本，最大似然估计值近似服从正态分布（normal distribution），而且其精确程度不逊于任何合理的估计量（estimator），前提是模型具备某些良好性质，而这些性质对大多数日常模型都成立。不过也有重要的例外。在下一节（Section 13.3）中，未知量是可能出现的最大值，即分布的一个端点，在那里最大似然估计值的表现大不相同：它的误差并不近似服从正态分布，而且误差是以 $\frac{1}{n}$ 而非 $\frac{1}{\sqrt{n}}$ 的速度缩小。

这个估计有多精确？

图 13.2.1 显示似然曲线很宽阔：11 次重捕所提供的信息并不多。粗略计算一下：第二个样本中有记号田鼠的数目近似服从泊松分布（Poisson distribution），平均值为 $\frac{200\times 200}{N}$，所以其标准差（standard deviation）约为 $\sqrt{11}\approx 3.3$，相对误差约为 30%，而这个误差会传递到 N 的估计上。若要把 N 的估计误差控制在 10% 以内，生物学家需要大约 100 次重捕。统计学家还会修正这个简单估计的轻微偏差（bias），采用*查普曼估计量*（Chapman estimator）$\frac{(M+1)(C+1)}{R+1}-1$，其中 M 和 C 是两次捕捉的数量，R 是重捕到的数目。

拉普拉斯计数法国人口

拉普拉斯（Laplace）在 1802 年用了同样的想法，在没有人口普查的情况下估计法国的人口。他知道全国一年内登记的出生人数，并且就一批抽样的市镇而言，同时知道其人口和出生人数。他假设人口与出生人数之比在各地都相同，从而估计法国人口约为 28 百万，甚至给出了不确定程度的度量。后来这个方法在动物种群研究中被重新发现：C. G. J. Petersen 在 1890 年代于丹麦的峡湾为鲽鱼做记号，Frederick Lincoln 则在 1930 年代使用鸟环；这个方法至今仍常被称为*林肯—彼得森*（Lincoln–Petersen）估计量。

思维技巧

选取使数据最有可能出现的参数

要估计一个未知量，先把你所观察到的结果的概率写成该未知量的函数，然后求其最大值。

在大学里，最大似然是统计推断（statistical inference）的骨干，并有一套理论说明估计值有多精确（费雪信息量，Fisher information），以及如何比较互相竞争的模型（似然比检验，likelihood ratio test）。

延伸练习

1. **一般公式。**若有 M 只动物做了记号，而第二个含 C 只动物的样本中有 R 只带有记号，证明 N 的最大似然估计值是 $\frac{MC}{R}$ 的整数部分。

2. **范德蒙恒等式。**考虑从 a 名女士和 b 名男士中选出 c 人组成委员会的方法数目，以此证明 $\sum_j \binom{a}{j}\binom{b}{c-j}=\binom{a+b}{c}$。

3 **掷硬币。**一枚硬币掷出正面的概率 p 未知，掷 n 次后观察到 h 次正面。证明 p 的最大似然估计值是 $\frac{h}{n}$。

4 **没有重捕。**如果第二个样本中没有任何带记号的田鼠，最大似然估计值会变成怎样？这说明了这个方法的甚么特点？

延伸阅读

- **David Spiegelhalter, *The Art of Statistics***（Pelican），谈估计与不确定性。
- **George Seber, *The Estimation of Animal Abundance***（Blackburn）。标记重捕法的标准参考书。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury）。数理统计的入门课本，内容包括最大似然。

13.3 德国坦克问题

题目

STEP III 1998，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/98-s3-q14`

引子

第二次世界大战期间，盟军情报部门想知道德国每月生产多少辆坦克。间谍的报告显示每月约 1400 辆。统计学家则另辟蹊径：缴获的坦克在变速箱和其他零件上都刻有序号，而这些序号实际上就是从 $1, 2, \ldots, N$ 中抽出的一个随机样本（random sample）。单凭这些数字，他们估计产量约为每月 250 辆。按照一般引用的数字，战后查阅的德国纪录显示为 245 辆。统计学家的估计远比间谍准确。

本题比较两种利用序号估计 N 的方法。两者平均而言都是正确的，但其中一种运用数据的效率远高于另一种；弄清楚是哪一种，就会引出估计理论（estimation theory）的核心思想：在平均而言正确的估计量（estimator）当中，应选取波动最小的一个。

动笔之前

想一想 设 M 是 $[0, N]$ 上 n 个独立均匀分布（uniform distribution）取值中的最大者，那么 $\mathrm{P}(M \le m)$ 是多少？

$M \le m$ 当且仅当每个 $X_i \le m$，所以当 $0 \le m \le N$ 时，$\mathrm{P}(M \le m) = \left(\frac{m}{N}\right)^n$（可与 Section 12.25 比较）。求导后可知，M 的概率密度（density）为 $\frac{nm^{n-1}}{N^n}$。

逐步解题

阶段 1　两者皆无偏

对于 M：

$$\mathrm{E}(M) = \int_0^N m \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+1}N, \qquad \text{所以} \qquad \mathrm{E}(Z_1) = \frac{n+1}{n} \cdot \frac{n}{n+1}N = N.$$

对于平均值：每个 X_i 的平均值为 $\frac{N}{2}$，所以 $\mathrm{E}(Z_2) = \frac{2}{n} \cdot n \cdot \frac{N}{2} = N$。 ■

两个估计量都是无偏（unbiased）的：平均而言，它们会给出正确的答案。因子 $\frac{n+1}{n}$ 用来修正这个事实：观察到的最大序号总是比 N 略小一点。

阶段 2 两者的方差

想一想 求 $\mathrm{E}(M^2)$，从而求 $\mathrm{Var}(M)$。至于 Z_2，单一个 $[0, N]$ 上的均匀变量的方差（variance）是多少？

$\mathrm{E}(M^2) = \int_0^N m^2 \cdot \frac{n m^{n-1}}{N^n}\,\mathrm{d}m = \frac{n}{n+2}N^2$，所以

$$\mathrm{Var}(M) = \frac{n}{n+2}N^2 - \frac{n^2}{(n+1)^2}N^2 = \frac{nN^2}{(n+1)^2(n+2)},$$
$$\mathrm{Var}(Z_1) = \frac{(n+1)^2}{n^2}\mathrm{Var}(M) = \frac{N^2}{n(n+2)}.$$

$[0, N]$ 上的均匀变量的方差为 $\frac{N^2}{12}$，而各个 X_i 互相独立，所以

$$\mathrm{Var}(Z_2) = \frac{4}{n^2} \cdot n \cdot \frac{N^2}{12} = \frac{N^2}{3n}. \qquad ■$$

阶段 3 哪一个较好？

想一想 比较 $\frac{1}{n(n+2)}$ 与 $\frac{1}{3n}$。当 n 增大时，两者如何变化？

两者之比为 $\frac{\mathrm{Var}(Z_1)}{\mathrm{Var}(Z_2)} = \frac{3}{n+2}$，当 $n \geq 2$ 时小于 1（当 $n = 1$ 时则相等，此时两个估计量完全相同）。所以 Z_1 较好，而且 n 愈大就好得愈多：它的标准差（standard deviation）以 $\frac{1}{n}$ 的速度缩小，而 Z_2 的标准差只以 $\frac{1}{\sqrt{n}}$ 的速度缩小（图 13.3.1）。若有 $n = 10$ 个序号，Z_1 的标准差约为 N 的 9%，而 Z_2 的约为 18%。

选用 Z_1 还有另一个理由：Z_2 有时会得出荒谬的结果。若序号为 10、20 和 300，则 $Z_2 = \frac{2}{3} \cdot 330 = 220$，比我们实际见过的一个序号还要小。而 $Z_1 = \frac{4}{3} \cdot 300 = 400$ 永远不会小于最大的观察值。 ■

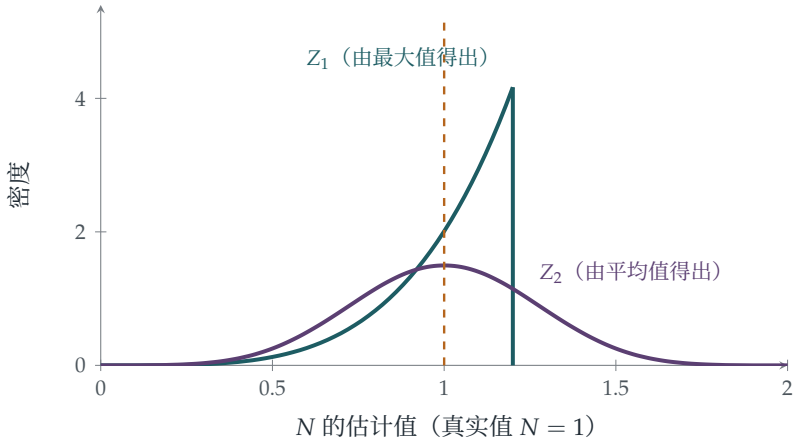


图 13.3.1 两个估计量的抽样分布（sampling distribution），其中序号数目 $n = 5$，真实值 $N = 1$。两者的平均值都恰好是 1，但 Z_1（蓝绿色）比 Z_2（紫红色）更紧密地集中在真实值附近，而且它永远不会低于观察到的最大序号。

回顾

- **无偏性。**每个估计量都经过缩放，使其期望值（expected value）恰好等于 N。
- **比较方差。**在无偏估计量当中，方差较小的一个较为可靠。
- **选用信息量最大的统计量。**最大值包含了样本中关于 N 的全部信息；平均值则浪费了其中一大部分。

背后的数学

充分性

为甚么最大值好这么多？一旦知道最大值 M，其余各值不论 N 是多少，都只是在 $[0, M]$ 上均匀分布：它们再没有带来关于 N 的任何信息。最大值是一个*充分统计量*（sufficient statistic）。C. R. Rao 与 David Blackwell 的一条定理（1945–47）指出，任何无偏估计量，只要换成它在给定充分统计量下的期望值，都可以得到改进，或至少不会变差；在这里，这个做法正好把 Z_2 变成 Z_1。*Lehmann–Scheffé* 定理进一步证明，Z_1 是方差最小的无偏估计量。还要留意，Z_1 的方差以 $\frac{1}{n^2}$ 的速度下降，比大多数估计量的 $\frac{1}{n}$ 更快：这是因为 N 是分布的端点，数据会挤向这个端点。

离散版本

真实的序号是整数 $1, 2, \ldots, N$，并且是以不放回（without replacement）方式抽样的。这时最佳的无偏估计是 $\hat{N} = m + \frac{m}{k} - 1$，其中 m 是 k 个观察到的序号中的最大者：即最大值加上各观察值之间的平均间距。盟军统计学家运用这类估计，主要根据变速箱和车轮的序号，估计德国的产量约为每月 246 辆坦克；相比之下，其后在德国纪录中找到的数字是 245 辆，而情报报告所显示的是 1400 辆。

历史上的序号

此后这个方法曾多次派上用场，例如分析员曾根据最先售出的几件产品的序号，估计某款新消费产品的生产数量。制造商汲取了教训：现在不少制造商都会把序号编码，使人无法把它们当作简单的顺序编号来解读。Richard Ruggles 与 Henry Brodie 于 1947 年在 *Journal of the American Statistical Association* 上记述了这项战时工作。

思维技巧

以抽样分布评价估计量

估计量是一个随机变量（*random variable*）。要问它平均而言是否正确（偏差），以及它的波动有多大（方差），然后选取误差较小的一个。

在大学，这会引出有效估计（efficient estimation）的理论：Cramér–Rao 下界（Cramér–Rao bound）指出无偏估计量的方差最小可以小到甚么程度，而充分性则说明如何达到这个下界。

延伸练习

1. **均方误差（mean squared error）。**有偏估计量 M 满足 $\mathrm{E}(M-N)^2 = \mathrm{Var}(M) + \big(\mathrm{E}(M) - N\big)^2$。计算此值，并与 $\mathrm{Var}(Z_1)$ 比较。求使 $\mathrm{E}(cM - N)^2$ 取最小值的常数 c。

2. **最小值也可以。**证明 $(n+1)\min\{X_i\}$ 也是 N 的无偏估计量，并求它的方差。这个估计量好用吗？

3. **离散公式。**对于序号 $1, \ldots, N$，以不放回方式抽取 k 个，其中最大者为 m。证明 $\mathrm{P}(\max \le m) = \binom{m}{k} / \binom{N}{k}$，并验证当 $k = 1$ 时 $m + \frac{m}{k} - 1$ 是无偏的。

4. **坦克。**若 5 辆缴获的坦克的序号为 19、40、42、60、98，试估计坦克的生产数量。

延伸阅读

- **Richard Ruggles & Henry Brodie, "An empirical approach to economic intelligence in World War II"**，*Journal of the American Statistical Association* 42 (1947)。最初的记述。
- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury），论及偏差、方差与充分性。
- **David Spiegelhalter, *The Art of Statistics***（Pelican）。

13.4 平均数还是中位数?

题目

STEP II 2000，第 14 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/00-s2-q14`

引子

要估计一组量度数据的中心，你可以取平均值（mean），也可以取中位数（median）。哪一个较好？答案视乎数据而定。假如量度结果偶尔会出现离谱的误差，中位数就可靠得多：一个荒谬的数值足以把平均值拉到任何地方，却几乎动摇不了中位数。至于规规矩矩的数据，通常是平均值胜出。本题要量度的，正是在数据服从均匀分布（uniform distribution）时，平均值究竟胜出多少：中位数的方差（variance）是平均值的三倍，所以要令中位数与平均值同样精确，就需要三倍数量的观测值。

这个比较引入了*效率*（efficiency）的概念，它是选择统计方法时的主要考虑之一；同时也用到 Beta 积分（beta integral），这个积分在 Section 4.4 也出现过。

动笔之前

想一想 为甚么中位数的密度（density）与 $y^n(1-y)^n$ 成正比？（要 Y 接近 y，$2n+1$ 个值之中须有多少个在其下，又有多少个大于 y？）

中位数要落在 y，就必须有一个值在 y，n 个值在其下（各自的概率为 y），另有 n 个值在其上（各自的概率为 $1-y$）。这就得出因子 $y^n(1-y)^n$；常数则是数算哪些值扮演哪个角色的选法数目。

逐步解题

阶段 1　常数

密度的积分必须等于 1：

$$k\int_0^1 y^n(1-y)^n\,\mathrm{d}y = k\cdot\frac{n!\,n!}{(2n+1)!} = 1, \qquad k = \frac{(2n+1)!}{(n!)^2}. \quad \blacksquare$$

阶段 2　平均值与方差

密度关于 $\frac{1}{2}$ 对称（把 y 换成 $1-y$ 即可看出），所以 $\mathrm{E}(Y) = \frac{1}{2}$。

> **想一想**　利用 Beta 积分计算 $\mathrm{E}(Y^2) = k\int_0^1 y^{n+2}(1-y)^n\,\mathrm{d}y$，并化简当中的阶乘。

$$\mathrm{E}(Y^2) = \frac{(2n+1)!}{(n!)^2}\cdot\frac{(n+2)!\,n!}{(2n+3)!} = \frac{(n+2)(n+1)}{(2n+3)(2n+2)} = \frac{n+2}{2(2n+3)}.$$

因此

$$\mathrm{Var}(Y) = \frac{n+2}{2(2n+3)} - \frac{1}{4} = \frac{2(n+2)-(2n+3)}{4(2n+3)} = \frac{1}{4(2n+3)}. \quad \blacksquare$$

阶段 3　与平均值比较

> **想一想**　$\bar{X}$ 是 $2n+1$ 个独立均匀变量的平均，它的平均值和方差是多少？

每个 X_i 的平均值为 $\frac{1}{2}$，方差为 $\frac{1}{12}$，所以 $\mathrm{E}(\bar{X}) = \frac{1}{2}$，而 $\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)}$。当 n 很大时，

$$\mathrm{Var}(Y) \approx \frac{1}{8n}, \qquad \mathrm{Var}(\bar{X}) \approx \frac{1}{24n}.$$

> **想一想**　假设 Y 和 $\bar{X}$ 都服从正态分布（normal distribution），平均值为 $\frac{1}{2}$，标准差（standard deviation）分别为 σ_Y 和 $\sigma_{\bar{X}}$，试用标准正态分布函数 Φ 表示这两个概率。

$\mathrm{P}\big(|Y-\frac{1}{2}| < \frac{d}{\sqrt{n}}\big) = 2\Phi\big(\frac{d}{\sqrt{n}\,\sigma_Y}\big) - 1$，而 $\sqrt{n}\,\sigma_Y \to \frac{1}{\sqrt{8}}$；$\bar{X}$ 的情况相似，其中 $\sqrt{n}\,\sigma_{\bar{X}} \to \frac{1}{\sqrt{24}}$。所以两个概率分别趋向 $2\Phi(d\sqrt{8}) - 1$ 和 $2\Phi(d\sqrt{24}) - 1$。由于 Φ 是递增函数，而且 $\sqrt{24} > \sqrt{8}$，后者较大，因此对所有足够大的 n，不等式都成立。■平均值较有可能接近真正的中心。

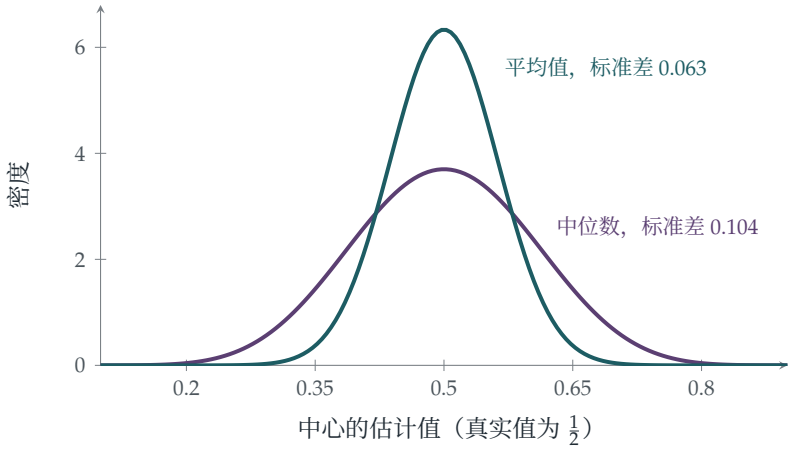


图 13.4.1 21 个独立均匀观测值的中位数（紫红色）和平均值（蓝绿色）的抽样分布（sampling distribution）。两者都以真实值 $\frac{1}{2}$ 为中心，但平均值集中得多：它的方差大约是中位数方差的三分之一。

回顾

- **Beta 积分。**$\int_0^1 y^r(1-y)^s\,\mathrm{d}y$ 这一条公式,同时给出了归一化常数(normalising constant)和各阶矩（moments)。
- **对称性给出平均值。**求 $\mathrm{E}(Y)$ 根本不用积分。
- **在合适的尺度上比较离散程度。**把标准差乘以 $\sqrt{n}$，比较的结果便与 n 无关。

背后的数学

效率

两个无偏估计量（unbiased estimator）的*相对效率*（relative efficiency）是它们方差之比；它说明其中一个需要多少倍的观测值，才能与另一个看齐。对于均匀分布的数据，中位数相对于平均值的效率是 $\frac{1}{3}$。对于正态分布的数据，这个数字是 $\frac{2}{\pi} \approx 0.64$：中位数大约需要多 57% 的观测值。但遇上厚尾（heavy-tailed）数据，形势就逆转了。对于柯西（Cauchy）分布（Section 12.22），平均值毫无用处，因为 n 个观测值的平均并不比单独一个观测值好，而中位数却表现良好。对于拉普拉斯（Laplace）分布（Section 12.24），中位数的效率是平均值的两倍。

稳健统计学

中位数是*稳健*（robust）的：即使把接近一半的数据改成任意数值，它也不会移动太远。平均值却没有这种保障：一个错误的数值就足以把它毁掉。在 1960 及 1970 年代，John Tukey、Peter Huber 等人发展出*稳健统计学*（robust statistics），即一些估计量，对干净的数据而言效率几乎与平均值一样高，却又能抵御离群值（outlier），例如*截尾平均值*（trimmed mean）（舍去最大和最小的 10%，再取余下数据的平均），跳水、体操等运动项目的评分就采用这种方法。

拉普拉斯与中位数

拉普拉斯在 1774 年证明，假如误差服从双指数（拉普拉斯）分布，则真实值的最佳估计是中位数；高斯（Gauss）在 1809 年证明，假如误差服从正态分布，则最佳估计是平均值，他并由此得出最小二乘法（method of least squares）。谁是谁非，取决于误差的表现如何；每遇上一种新的数据，统计学家至今仍会为这个问题争论一番。

思维技巧

以变异程度比较不同方法

两种方法可能平均而言都正确；要比较的是它们的答案变动有多大，以及数据增加时这种变动收缩得有多快。

在大学里，渐近效率（asymptotic efficiency），即比较大样本下的方差，是在不同统计方法之间作取舍的标准工具；与之并用的还有稳健性，以应付模型可能出错的情况。

延伸练习

1 **中位数的密度。**用数算的方法推导密度 $\frac{(2n+1)!}{(n!)^2}y^n(1-y)^n$：一个值落在 $[y, y+\delta y]$ 内，n 个在其下，n 个在其上。

2 **三个观测值。**当 $n=1$（三个均匀观测值）时，求 $\mathrm{Var}(Y)$ 和 $\mathrm{Var}(\bar{X})$ 的准确值。哪一个较小？

3 **中列数。**中心的另一个估计量是中列数（midrange）$\frac{1}{2}(\max+\min)$。证明对于均匀分布的数据，若有 $m=2n+1$ 个观测值，它的方差是 $\frac{1}{2(m+1)(m+2)}$。它与平均值相比如何？（试与 Section 13.3 比较。）

4 **一个离群值。**十一个量度结果为 9.8、9.9、10.0、10.0、10.1、10.1、10.2、10.2、10.3、10.4 和 104（打字错误）。比较平均值和中位数。

延伸阅读

- **John Rice, *Mathematical Statistics and Data Analysis***（Duxbury），论及顺序统计量（order statistics）和效率。
- **Rand Wilcox, *Fundamentals of Modern Statistical Methods***（Springer）。为非专家讲解稳健方法为何重要。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***（Harvard University Press）。从历史角度看数据的汇总、平均值和中位数。

13.5 问尴尬的问题

题目

STEP II 2023，第 11 题

继续阅读之前，请先看题目（原题为英文）。题目载于 OCR 提供下载的 2023 年 STEP 试卷（zip 文件）内的 STEP II 试卷：

`www.ocr.org.uk/Images/707102-2023-past-papers.zip`

引子

怎样才能知道有多少比例的学生曾在考试中作弊，或者有多少比例的雇员曾偷取公司财物？直接问的话，很多人都会说谎。1965 年，统计学家 Stanley Warner 想出了一个巧妙的解决办法：随机化回答（randomised response）。每个人私下掷一枚硬币或一颗骰子，由结果决定他是如实回答那条敏感问题，还是回答与之相反的问题。访问员永远不知道是哪一种情况，所以任何一个人的答案都不足以成为对他不利的证据；然而，由于随机化的概率（probability）是已知的，整体的比例仍然可以估计出来。

本题探讨的正是这背后的数学。让每个人独立地（independently）随机化（第 (ii) 部分），所得的总数性质良好，足以确定该比例。用同一次掷硬币为所有人随机化（第 (iii) 部分），所得的总数虽然平均值（mean）相同，分布（distribution）却截然不同，呈双峰形状，几乎不能告诉我们任何事。随机性放在哪里，至关重要。

动笔之前

想一想 Y 以概率 p 等于 X_1，以概率 q 等于 X_2。$\mathrm{E}(Y)$ 是甚么？$\mathrm{E}(Y^2)$ 又是甚么？

以选中哪一个变量为条件，可得 $\mathrm{E}(Y) = p\mu_1 + q\mu_2$，以及 $\mathrm{E}(Y^2) = p\,\mathrm{E}(X_1^2) + q\,\mathrm{E}(X_2^2) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2)$。

逐步解题

阶段 1　混合分布的方差

$$\mathrm{Var}(Y) = p(\sigma_1^2 + \mu_1^2) + q(\sigma_2^2 + \mu_2^2) - (p\mu_1 + q\mu_2)^2.$$

含 μ 的各项为 $p\mu_1^2 + q\mu_2^2 - p^2\mu_1^2 - 2pq\mu_1\mu_2 - q^2\mu_2^2 = pq\mu_1^2 + pq\mu_2^2 - 2pq\mu_1\mu_2$，当中用到 $p - p^2 = pq$ 和 $q - q^2 = pq$。因此

$$\mathrm{Var}(Y) = p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1 - \mu_2)^2. \qquad \blacksquare$$

混合分布（mixture）的方差（variance），等于各成分内部方差的平均，再加上各成分平均值之间的方差。

阶段 2　独立随机化

> **想一想**　$\mathrm{P}(B = 1)$ 是多少？B 的 n 个值是否独立？

$\mathrm{P}(B = 1) = \frac{1}{2}\cdot\frac{1}{6} + \frac{1}{2}\cdot\frac{5}{6} = \frac{1}{2}$。这 n 个值互相独立，每个的成功概率都是 $\frac{1}{2}$，所以 $Z_1 \sim \mathrm{B}(n, \frac{1}{2})$，其平均值为 $\frac{n}{2}$，标准差（standard deviation）为 $\frac{\sqrt{n}}{2}$。

与平均值相差不超过 10%，即是 $|Z_1 - \frac{n}{2}| < \frac{n}{20}$。以标准单位表示，这就是 $|Z| < \frac{n/20}{\sqrt{n}/2} = \frac{\sqrt{n}}{10}$，而此值趋向无穷大，所以该概率趋向 1。 $\blacksquare$

阶段 3　所有人共用一枚硬币

> **想一想**　现在只掷一次硬币，便决定全部 n 颗骰子的情况。已知掷得正面，Z_2 的分布是甚么？已知掷得反面呢？试用第 (i) 部分的结果。

已知掷得正面，$Z_2 \sim \mathrm{B}(n, \frac{1}{6})$：平均值为 $\frac{n}{6}$，方差为 $\frac{5n}{36}$。已知掷得反面，$Z_2 \sim \mathrm{B}(n, \frac{5}{6})$：平均值为 $\frac{5n}{6}$，方差为 $\frac{5n}{36}$。在第 (i) 部分取 $p = q = \frac{1}{2}$，得

$$\mathrm{E}(Z_2) = \frac{n}{2}, \qquad \mathrm{Var}(Z_2) = \frac{5n}{36} + \frac{1}{4}\left(\frac{4n}{6}\right)^2 = \frac{5n}{36} + \frac{n^2}{9}.$$

> **想一想**　大部分的概率集中在哪里？在平均值 $\frac{n}{2}$ 附近的概率多吗？

Z_2 有两个峰，分别在 $\frac{n}{6}$ 和 $\frac{5n}{6}$ 附近，每个峰的阔度约为 $\sqrt{n}$（图 13.5.1），而两峰之间几乎甚么都没有。具有相同平均值和方差的正态分布（normal distribution）则只有一个峰，中心位于 $\frac{n}{2}$，正正是 Z_2 几乎从不出现的位置。要 Z_2 与 $\frac{n}{2}$ 相差不超过 10%，一个服从 $\mathrm{B}(n, \frac{1}{6})$ 或 $\mathrm{B}(n, \frac{5}{6})$ 的变量必须落在 $[0.45n, 0.55n]$ 之内。对 $\mathrm{B}(n, \frac{1}{6})$ 而言，这比其平均值高

出至少 $0.28n$，约为 $\frac{0.28n}{0.37\sqrt{n}} \approx 0.75\sqrt{n}$ 个标准差，所以该概率趋向 0；$\mathrm{B}(n, \frac{5}{6})$ 的情况亦类似。 ■

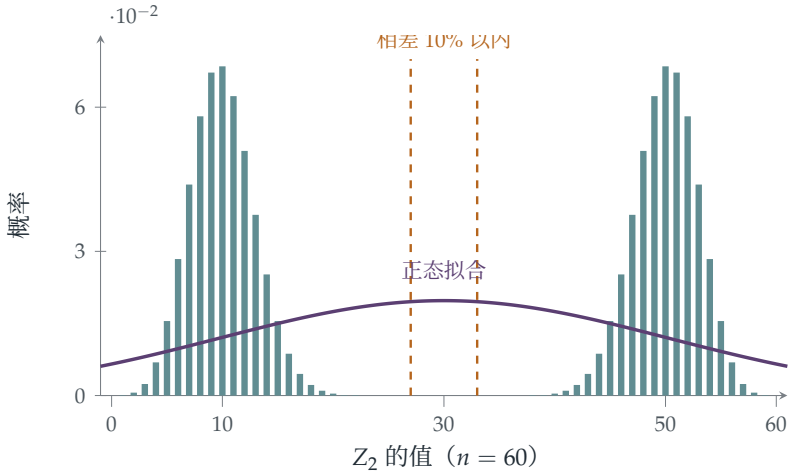


图 13.5.1 $n = 60$ 颗骰子时 Z_2 的分布（长条）：两个峰分别位于 10 和 50，来自那一次掷硬币的两种结果。具有相同平均值（30）和方差的正态分布（紫红色）拟合得很差，而且几乎没有概率落在与平均值相差 10% 以内的范围（两条琥珀色线之间）。

回顾

- **以隐藏的选择为条件。**混合分布的平均值和方差，都是以选中哪一个成分为条件而求得的。
- **内部方差加上组间方差。**额外的一项 $pq(\mu_1 - \mu_2)^2$ 量度了两个成分之间的离散程度。
- **平均值和方差并非全部。**Z_2 与 Z_1 的平均值相同，但形状完全不同。

背后的数学

随机化回答

在 Warner 的方案中，每位受访者私下作出选择，以概率 p 回答“你是否属于 A 组?”，以概率 $1 - p$ 回答“你是否不属于 A 组?”。若属于 A 的人所占比例为 π，则答“是”的概率为

$$\lambda = p\pi + (1 - p)(1 - \pi),$$

因此由观察所得答“是”的比例 $\hat{\lambda}$，我们可以估计 $\hat{\pi} = \frac{\hat{\lambda} - (1-p)}{2p-1}$（前提是 $p \neq \frac{1}{2}$）。本题第 (ii) 部分正是这种情况，由每个人自己的骰子来决定。隐私的代价是精确度：除以 $2p - 1$

会令方差增大，所以随机化调查所需的受访者比直接调查更多。不过每个人都是独立地随机化的，所以大数定律（law of large numbers）能够发挥作用。第 (iii) 部分显示了共用随机化时会出甚么问题：那一枚硬币所带来的不确定性，永远不会被平均抵销。

差分隐私

随机化回答是*差分隐私*（differential privacy）的前身。差分隐私是 Cynthia Dwork 及其同事于 2006 年提出的一个关于隐私的数学定义：如果加入或移除任何一个人的数据，只会令任何输出的概率改变至多一个很小的倍数，该程序便算是保障隐私的。科技公司用它来收集使用统计数据，美国人口普查局亦在 2020 年人口普查中采用了它，在公布的数字上加入经仔细校准的随机杂讯，令任何个人的答案都无法被还原。

全方差定律

第 (i) 部分的恒等式是*全方差定律*（law of total variance）的一个特例：$\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y \mid C)\big) + \mathrm{Var}\big(\mathrm{E}(Y \mid C)\big)$，其中 C 是隐藏的选择。它是*方差分析*（analysis of variance，ANOVA）的基础。方差分析由费雪（Ronald Fisher）于 1920 年代在 Rothamsted 为农业实验而引入，它把农作物产量的变异分拆为两部分：可由肥料或品种解释的部分，以及未能解释的部分。

思维技巧

要看整个分布，而不只是它的概括数值

平均值和方差能很好地描述单峰分布，但对混合分布却可能造成严重误导。要问一问：概率实际上落在哪里。

在大学里，混合模型（mixture model）用来描述由几个群组组成的总体，而全方差定律则是统计学的标准工具，从调查设计到阶层模型（hierarchical model）都会用到。

延伸练习

1 **双胞胎。**试做 STEP II 2005，第 12 题（`step.maths.org/questions/05-s2-q12`）。这其实是一道乔装了的随机化回答谜题：在甚么情况下，一个答案完全不能透露硬币的结果？

2 **估计 π。**在一项 $p = \frac{3}{4}$ 的 Warner 调查中，有 400 人作答，其中 160 人答“是”。试估计 A 组所占的比例 π，并求出你的估计值的近似标准误差（standard error）。

3 **全方差。**设 C 为取多个值的离散隐藏变量，证明全方差定律。

4 **共用硬币。**在第 (iii) 部分中，$Z_2 > \frac{n}{2}$ 的概率是多少？试解释为何当 n 很大时，这个概率接近 $\frac{1}{2}$，尽管 Z_2 几乎从不接近 $\frac{n}{2}$。

延伸阅读

- **Stanley Warner, “Randomized response: a survey technique for eliminating evasive answer bias”**，*Journal of the American Statistical Association* 60 (1965)。
- **Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy***（可于网上免费阅读）。现代理论；开首几章不难读懂。
- **David Spiegelhalter, *The Art of Statistics***（Pelican），谈调查及其陷阱。

13.6 累积起来的误差

题目

STEP II 1993，第 15 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/93-s2-q15`

引子

1982 年，温哥华证券交易所推出了一个新指数，起始值为 1000。每当有股价变动，指数便会重新计算，并截断（truncate）至小数点后三位。每次截断平均损失 0.0005，看来微不足道。可是每天大约要重新计算 3000 次，二十二个月后，尽管大市上升，指数却只剩下约 520。经正确重算后，指数一夜之间跃升至约 1100。

这道题目解释了个中原因。截断误差（truncation error）全部朝同一方向，因此会稳步累积，与运算次数成正比。舍入误差（rounding error）则有正有负，大部分互相抵销：它们的增长速度只相当于运算次数的平方根，这正是中心极限定理（central limit theorem）（Section 12.24）所预测的随机漫步（random walk）行为。

动笔之前

想一想 若某数的小数部分 U 在 $[0,1)$ 上均匀分布（uniform），把该数截断时误差是多少？舍入时又是多少？两者的平均值（mean）和方差（variance）各是多少？

截断是以 $\lfloor a \rfloor$ 取代 a，误差（真确值减近似值）为 U，在 $[0,1)$ 上均匀分布：平均值为 $\frac{1}{2}$，方差为 $\frac{1}{12}$。舍入所得的误差在 $(-\frac{1}{2}, \frac{1}{2}]$ 上均匀分布：平均值为 0，方差为 $\frac{1}{12}$。

逐步解题

阶段 1 总误差

加了 1500 个数之后，总误差是 1500 个独立（independent）误差之和，根据中心极限定理，它近似于正态分布（normal distribution）：

- LEP（舍入）：平均值为 0，方差为 $\frac{1500}{12} = 125$，标准差（standard deviation）为 $\sqrt{125} \approx 11.18$。
- VOZ（截断）：平均值为 $\frac{1500}{2} = 750$，标准差同样是 11.18。

阶段 2 出现大误差的概率

> **想一想** 对 LEP 而言，若 $E \sim \mathrm{N}(0, 125)$，$\mathrm{P}(|E| > 15)$ 是多少？对 VOZ 而言，误差几乎必然落在哪里？

对 LEP 而言，$\frac{15}{11.18} \approx 1.342$，所以

$$\mathrm{P}(|E| > 15) \approx 2\big(1 - \Phi(1.342)\big) \approx 2 \times 0.090 = 0.18.$$

对 VOZ 而言，误差约为 750 ± 11；其绝对值小于 15 的概率基本上是零，所以 $\mathrm{P}(|E| > 15) \approx 1$。 ■

阶段 3 可以做多少次加法？

> **想一想** 做了 n 次加法后，LEP 的误差的标准差为 $\sqrt{n/12}$。何时有 $\mathrm{P}(|E| < 10) \geq 0.9$？

我们需要 $\frac{10}{\sqrt{n/12}} \geq 1.645$（因为 $\mathrm{P}(|Z| < 1.645) = 0.9$），即 $\frac{n}{12} \leq \left(\frac{10}{1.645}\right)^2 \approx 36.95$，所以 $n \leq 443$。LEP 可以做大约 443 次加法。 ■

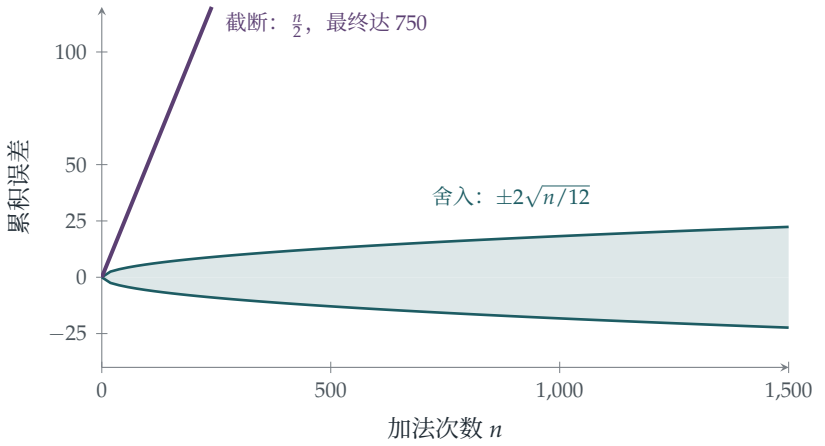


图 13.6.1 误差如何累积。截断误差（紫红色）全部同号，与 n 成正比增长：做了 1500 次加法后，总误差约为 750。舍入误差（蓝绿色带状区域，即零的两侧各两个标准差）平均而言互相抵销，增长速度只相当于 $\sqrt{n}$：做了 1500 次加法后，误差通常在约 ±22 之内。

回顾

- **为每个误差建立模型。**小数部分均匀分布，令每个误差都有简单的分布。
- **把平均值和方差相加。**平均值可以相加（得出偏差 $\frac{n}{2}$），而对于独立的误差，方差同样可以相加。
- **正态近似。**中心极限定理把 1500 个均匀误差之和化为正态分布。

背后的数学

偏差与方差

截断是*有偏*（biased）的：其误差的平均值不是零，而偏差（bias）会线性累积。舍入是*无偏*（unbiased）的，其误差的表现有如随机漫步，增长速度相当于 $\sqrt{n}$。经过一百万次运算后，两者的分别就是数量级为 500 000 的误差与数量级为 300 的误差之别。这与实验中系统误差（systematic error，例如仪器校准失准）和随机测量误差（random error）之间的区别如出一辙：把多次测量结果取平均，可以减少随机误差，对系统误差却毫无作用。

浮点数

所有电脑都只能以有限的精确度储存实数，每次算术运算都要把结果舍入。现代电脑遵从 IEEE 754 标准，把结果舍入至最接近的可表示数；若刚好位于两数正中，则取末位为

偶数的一个，目的正是令舍入平均而言无偏。即使如此，把大量数字相加时若不小心处理，仍会损失准确度；William Kahan 的*补偿求和*（compensated summation）算法（1965 年）会记录失去的低位数字，再把它们加回去。1991 年波斯湾战争期间，一个爱国者导弹连未能拦截一枚来袭导弹，原因是它的时钟以十分之一秒为单位计时，而这个数无法以二进制准确储存；连续运作 100 小时后，截断误差已增至约三分之一秒，足以令目标位置的估算偏差超过半公里。

温哥华指数

温哥华证券交易所指数于 1982 年 1 月设立，起始值为 1000.000。到 1983 年 11 月，尽管大市普遍上升，指数却已跌至约 524。顾问发现，指数每天重新计算数千次，而每次都截断（而非舍入）至小数点后三位。经过一个周末的正确重算，指数重开时约为 1098。这件事已成为数值计算课程中的经典警示。

思维技巧

分清系统误差与随机误差

正负号一致的误差会线性累积；平均值为零的独立误差则按平方根的速度累积。先认清你面对的是哪一种。

在大学，数值分析（numerical analysis）研究误差如何在计算过程中传播，统计学则区分偏差与方差；“偏差–方差权衡”（bias–variance trade-off）是现代机器学习的核心概念。

延伸练习

1. **修正 VOZ。**若 VOZ 在最后把 $\frac{n}{2}$ 加到答案上，其误差的分布是甚么？它现在是否与 LEP 一样好？

2. **舍入至小数点后一位。**若把各数舍入至小数点后一位，而被舍去的数字均匀分布，求把 10 000 个数相加时总误差的标准差。

3. **温哥华指数。**若指数每天重新计算 3000 次，每次截断平均损失 0.0005，那么在一年 250 个交易日内，指数会损失多少？

4. **精确分布。**求两个舍入误差之和的精确密度函数（density，呈三角形）。若是三个误差之和，试绘出其草图。它多快便开始貌似正态分布？

延伸阅读

- **Nicholas Higham, *Accuracy and Stability of Numerical Algorithms***（SIAM）。关于舍入误差的权威著作；第一章浅白易读。
- **David Goldberg, "What every computer scientist should know about floating-point arithmetic"**, *ACM Computing Surveys* 23 (1991)。经典之作，可在网上免费阅读。
- **Tim Harford, *How to Make the World Add Up***（Bridge Street），探讨数据中的系统误差。

13.7 制造相关性

题目

STEP III 2010，第 13 题

继续阅读之前，请先看题目（原题为英文）。题目载于 STEP Support Programme 的试题资料库：

`step.maths.org/questions/10-s3-q13`

引子

一间银行想知道，手上的股票投资组合（portfolio）有多大机会亏蚀。股价往往一起变动：一只股票下跌，其他股票也倾向跟着下跌。要模拟数以千计的可能未来，银行必须产生一批随机数，而这些随机数之间的相关程度要恰到好处。如果手上只有互相独立（independent）的随机变量（random variable），怎样才能制造出具有指定相关系数（correlation）的随机变量呢？

这道题就三个变量的情形，一步一步给出了配方。每个新变量都是先前各个独立材料的组合，再加上一份全新的独立材料，用来补足方差（variance）。这个构造其实是线性代数（linear algebra）中一个著名结果的化身，即*科列斯基分解*（Cholesky decomposition），它同时揭示了一个隐藏的限制：并非每一组相关系数都有可能出现。

关于课程范围的说明

协方差（covariance）和相关系数已于 2019 年从 STEP 的考试范围中删去。协方差是 $\mathrm{Cov}(U,V) = \mathrm{E}(UV) - \mathrm{E}(U)\mathrm{E}(V)$；独立变量的协方差为零，而且协方差对每个变元都是线性的：$\mathrm{Cov}(aU+bV,W) = a\,\mathrm{Cov}(U,W) + b\,\mathrm{Cov}(V,W)$。此外 $\mathrm{Var}(U) = \mathrm{Cov}(U,U)$。相关系数必定介乎 -1 与 1 之间。

动笔之前

想一想　Z_1 和 Z_2 互相独立。它们的协方差是多少？相关系数又是多少？

独立变量满足 $\mathrm{E}(Z_1Z_2) = \mathrm{E}(Z_1)\,\mathrm{E}(Z_2)$。所以协方差是 0，而 $\mathrm{Corr}(Z_1, Z_2) = 0$。

逐步解题

阶段 1　两个变量

$\mathrm{E}(Y_2) = 0$。由于 Z_1 和 Z_2 互相独立，方差均为 1，

$$\mathrm{Var}(Y_2) = \rho_{12}^2 + (1-\rho_{12}^2) = 1, \qquad \mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(Z_1, \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2) = \rho_{12}.$$

两个方差都是 1，所以 $\mathrm{Corr}(Y_1, Y_2) = \rho_{12}$。 ■

想一想　把 Z_1 和 Z_2 想像成互相垂直的单位向量。Y_2 在哪里？Y_1 与 Y_2 之间夹角的余弦是多少？

$Y_2 = \rho_{12}Z_1 + \sqrt{1-\rho_{12}^2}\,Z_2$ 是一个单位向量，与 $Z_1 = Y_1$ 的夹角为 θ，其中 $\cos\theta = \rho_{12}$。把变量看成向量，相关系数就是它们之间夹角的余弦（图 13.7.1）。

阶段 2　第三个变量

想一想　试以 a、b、c 写出 $\mathrm{Var}(Y_3) = 1$、$\mathrm{Cov}(Y_1, Y_3) = \rho_{13}$、$\mathrm{Cov}(Y_2, Y_3) = \rho_{23}$ 这三个条件。

利用线性和独立性：

$$a^2 + b^2 + c^2 = 1, \qquad a = \rho_{13}, \qquad \rho_{12}a + \sqrt{1-\rho_{12}^2}\,b = \rho_{23}.$$

所以

$$a = \rho_{13}, \qquad b = \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}, \qquad c = \sqrt{1 - \rho_{13}^2 - \frac{(\rho_{23} - \rho_{12}\rho_{13})^2}{1-\rho_{12}^2}}.$$ ■

（由于 $c \geq 0$，这里取非负平方根。）

阶段 3　任意的均值和方差

缩放和平移都不会改变相关系数，所以

$$X_i = \mu_i + \sigma_i Y_i \qquad (i = 1, 2, 3)$$

的均值为 μ_i，方差为 σ_i^2，相关系数为 ρ_{ij}。 ■

> **停一停，试一试** c 的公式要求平方根内的量非负。试取 $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$。哪里出了问题？三个变量不可能两两之间都有很强的负相关，为甚么这是合理的？

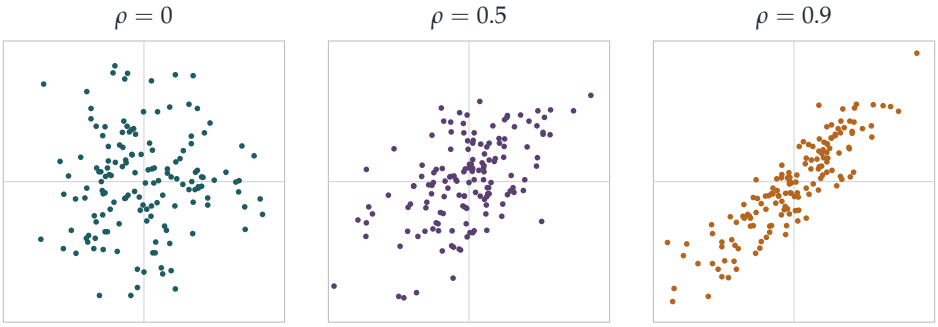


图 13.7.1 (Y_1, Y_2) 的样本，由独立的标准正态（standard normal）变量 Z_1, Z_2 按配方 $Y_1 = Z_1$, $Y_2 = \rho Z_1 + \sqrt{1-\rho^2}\, Z_2$ 制成，分别对应 $\rho = 0$、0.5 和 0.9。同样的独立材料，可以造出我们任意选定的相关系数。

回顾

- **每次只构造一个变量。**每个新变量都从先前的材料中取用适当的份量，再加入一份全新的独立材料。
- **协方差是线性的。**所有条件都化成了关于系数的简单方程。
- **相关系数是余弦。**把标准化后的变量看成单位向量，相关系数就变成了夹角。

背后的数学

科列斯基分解

用矩阵形式来写，这个构造就是 $\mathbf{Y} = L\mathbf{Z}$，其中的矩阵是下三角（lower triangular）矩阵

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & \sqrt{1-\rho_{12}^2} & 0 \\ a & b & c \end{pmatrix},$$

而 $\mathbf{Y}$ 的相关矩阵（correlation matrix）是 $R = LL^{\mathsf{T}}$。把一个对称矩阵写成 LL^{T} 的形式，就是科列斯基分解，由法国军事工程师科列斯基（André-Louis Cholesky）约于 1910 年

发现，用来求解大地测量中的方程组（他在第一次世界大战中阵亡，这项成果在他身后才发表）。只要 R 是正定（positive definite）的，这个分解就行得通，而计算过程正是我们刚才用手做的：逐行求解。在金融、工程和气候模型中，这是模拟相关变量的标准方法。

哪些相关系数是可能的？

c 的平方根内的数必须非负。展开后，条件是

$$1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} \geq 0,$$

这正是相关矩阵的行列式（determinant）非负的条件。从几何上看：三个单位向量两两之间的夹角为 θ_{ij}（其中 $\cos\theta_{ij} = \rho_{ij}$），它们必须真的能在空间中拼合起来。如果 $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$，各个夹角都约为 127°，而三个两两相隔 127° 的向量，夹角之和就要超过 360°：这是不可能的。一般而言，一个矩阵要成为相关矩阵，必须是半正定（positive semi-definite）的，即所有特征值（eigenvalue）均非负（Section 7.2）。如果各个相关系数是根据不完整的数据分别估计出来的，这个条件就可能不成立，统计学家于是要找出最接近的有效相关矩阵，这个问题由 Nicholas Higham 在 2002 年解决。

高尔顿、皮尔逊与相关系数

高尔顿（Francis Galton）在 1888 年研究父母与子女身高的关系时引入了相关的概念，而皮尔逊（Karl Pearson）在 1896 年给出了本题所用的积矩（product moment）公式。高尔顿还注意到均值回归（regression to the mean）的现象（Section 13.1）。皮尔逊于 1911 年在伦敦大学学院（University College London）创立了全世界第一个大学统计学系。“相关不等于因果”这句警告，几乎与相关系数本身一样古老。

思维技巧

逐步构造，每次加入全新的独立材料

要构造具有指定关系的对象，可以逐个加入，每一个都是先前各个对象的组合，再加上一些全新而独立的东西。

这是线性代数中的格拉姆–施密特（Gram–Schmidt）正交化过程倒过来做，也就是科列斯基分解。在大学里，它会出现在最小二乘法（least squares）、多元正态分布（multivariate normal distribution）的理论，以及随机过程的模拟之中。

延伸练习

1 **不可能的相关系数。**证明：除非 $\rho \geq -\frac{1}{2}$，否则三个随机变量不可能两两之间的相关系数都是 ρ。对于 n 个变量，相应的结论是甚么？

2 **检查行列式。**证明 $1 - \rho_{13}^2 - b^2 \geq 0$ 等价于 $\det R \geq 0$，其中 R 是 3×3 相关矩阵。

3 **投资组合的风险。**两只股票回报的标准差（standard deviation）分别为 $\sigma_1 = 0.2$ 和 $\sigma_2 = 0.3$，相关系数为 ρ。若投资组合中两只股票各占一半，求 $\rho = -1$、0 和 1 时该组合的标准差。分散投资达到了甚么效果?

4 **还原材料。**给定题目中的 Y_1、Y_2，试以 Y_1 和 Y_2 表示 Z_2。（这就是格拉姆–施密特过程中的一步。）

延伸阅读

- **Gilbert Strang,** ***Introduction to Linear Algebra***（Wellesley-Cambridge），讲述正定矩阵和科列斯基分解。
- **Paul Glasserman,** ***Monte Carlo Methods in Financial Engineering***（Springer）。讲述实际应用中如何模拟相关变量；内容较深，但开头几章不难读。
- **Theodore Porter,** ***Karl Pearson: The Scientific Life in a Statistical Age***（Princeton University Press）。

我们学到了甚么

本章开首说过，概率（probability）是由模型走向数据，而统计（statistics）则是由数据回到现实世界。我们提出了三个问题：这是真有其事，还是纯属杂讯？应该怎样估计？随机性是怎样进入数据的？以下是本章七道题目所揭示的答案。

1. 是真有其事，还是纯属杂讯？

在 Section 13.1 中，四则新闻标题把数字变成了论断。比较总数而不理会分母、单凭一年或一日的数据就看出趋势、把趋势隐藏在平均数之中，这些全都是错误；而泊松模型（Poisson model）则提供了一把尺，用来量度计数数据的自然变异。上升了 8 个标准差（standard deviation），几乎肯定是真有其事；某个星期二有七人死亡，则几乎肯定不是。再加上均值回归（regression to the mean）和公平比较的必要，整幅图画便完整了。

2. 应该怎样估计？

有三道题目发展了估计的理论。

- **最大似然（maximum likelihood）**（Section 13.2）：选取令观察所得数据出现的概率最大的未知数值。在捉放法（capture–recapture）中，它给出了自然的估计值 $\frac{200\times 200}{11} \approx 3636$，而宽阔的似然函数则显示出这个估计有多不确定。
- **偏差（bias）与方差（variance）**（Section 13.3）：潜艇数目有两个无偏估计量（unbiased estimator），一个以最大值为基础，另一个以平均值为基础。最大值是充分统计量（sufficient statistic），精确得多，其方差按 $\frac{1}{n^2}$ 而非 $\frac{1}{n}$ 的速度缩小。同样的推理，在战时估计德军坦克数量一事上，胜过了情报部门。
- **效率（efficiency）与稳健性（robustness）**（Section 13.4）：对于均匀分布的数据，平均值的效率是中位数（median）的三倍；但对于重尾（heavy-tailed）数据，则是中位数胜出，而且它不受极端误差影响，十分稳健。

3. 随机性是怎样进入数据的？

随机性可以是刻意加入的：随机化回答（randomised response）（Section 13.5）既保障每位受访者的隐私，又仍然可以对总体作出估计，条件是每个人各自独立地进行随机化；假如所有人共用同一枚硬币，便会得出一个双峰的混合分布，几乎不能告诉我们任何东

西。随机性也会在处理数据的过程中进入：四舍五入的误差会互相抵销，只按 $\sqrt{n}$ 的速度增长，而截尾误差（truncation error）则会线性地累积，温哥华证券交易所（Vancouver Stock Exchange）便亲身领教过（Section 13.6）。随机性还可以经人手设计出来：把独立的成分以 Cholesky 构造法（Cholesky construction）组合起来，便可产生具有任何一组有效相关系数（correlation）的变量（Section 13.7），但并非每一组相关系数都是有效的。

4. 统计学的工具箱

方法	在哪里用过
泊松变异	新闻标题（13.1）
似然及其最大值	田鼠（13.2）
无偏性与方差	序号（13.3）；中位数（13.4）
充分统计量	序号（13.3）
正态近似（normal approximation）	13.4、13.5、13.6
全方差定律（law of total variance）	随机化回答（13.5）
偏差与随机误差	四舍五入（13.6）
协方差（covariance）的线性性质	相关系数（13.7）

下一步往哪里去

- 贝叶斯定理（Bayes' theorem）（Section 12.2）是贝叶斯统计学的基础，而贝叶斯统计学正是本章各种方法以外最主要的另一套方法。
- 中心极限定理（central limit theorem）（Section 12.24）为本章处处用到的正态近似提供了理据；柯西分布（Cauchy distribution）（Section 12.22）则显示这些近似在甚么时候失效。
- 相关矩阵（correlation matrix）是半正定矩阵（positive semi-definite matrix），可以用 Section 7.2 的特征向量（eigenvector）来研究。

如果你想在入读大学之前再走远一点

- **David Spiegelhalter, *The Art of Statistics***：为一般读者而写的统计思维入门书之中最好的一本。
- **Tim Harford, *How to Make the World Add Up***：教你怎样解读新闻中的数字。
- **John Rice, *Mathematical Statistics and Data Analysis***：大学数理统计的第一门课程。
- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom***：统计学的重大思想及

其来源。

George Box 说过，所有模型都是错的，但有些是有用的。本章尝试说明怎样令模型变得有用：知道应该预期多大的变异，以方法在多次重复之下的表现来评价方法，并且思考数据是怎样得来的。正是这些习惯，把数字变成知识。

最后的话

全书到此结束。在这十三章里，我们以 STEP 题目为门径，走进大学一年级课程的数学：数论与分析、微积分与复数、代数与线性代数、微分方程与动力系统、组合数学、力学、概率与统计。每一道题目都是一份邀请，请你追问它背后藏着甚么，而答案几乎每一次都是一整门学科。如果你曾经手执铅笔，认真做过其中哪怕只是几节，思考页边提出的问题，并尝试每节末的习题，你就已经开始像数学家那样思考了。余下的旅程，就由你自己去走。

延伸练习答案

以下只是简短的答案和提示，并非完整解答：这里只给出结果和关键步骤，方便你核对自己的做法。每道练习都应先认真做过，然后才看答案。

13.1 解读新闻标题

1. $1 - \sum_{k=0}^{6} \mathrm{e}^{-2.96} \frac{2.96^k}{k!} \approx 0.0315$，所以一年之中大约有 11 天会出现七宗或以上的死亡个案。$\mathrm{P}(X \geq 9) \approx 0.0035$：大约一年一天。

2. 死亡人数为 $150K$，其中 K 服从平均值（mean）为 1.2 的泊松分布（Poisson distribution）：平均值为 180，方差（variance）为 $150^2 \times 1.2 = 27\,000$，标准差（standard deviation）约为 164。死亡人数为 125 或以下的年份，就是没有发生空难的年份，其概率为 $e^{-1.2} \approx 0.30$：完全是寻常事。但在这个模型中，死亡人数不可能恰好是 125，因为死亡人数只能是 150 的倍数。平均值和标准差都显示不出这一点：个案扎堆出现，令分布变得一团一团，全部概率都集中在 0、150、300 等数值上。

3. $z = \frac{1349.5-1080}{33} \approx 8.2$：概率的数量级为 10^{-16}。稳定不变的发生率无法解释 1350 这个数字；必定是有些事情真的改变了（又或者是统计死亡人数的方法改变了）。

4. 例如：(i)“单凭总数，无法判断乘飞机与步行哪样较安全”；(ii)“1990 年空难死亡人数属正常范围”；(iii)“星期二七名行人死亡：不幸的一天，但这样的日子一年约有十一次”；(iv)“1990 年行人死亡人数急升：必须追查原因”。

13.2 计数田鼠

1. $\frac{p_N}{p_{N-1}} = \frac{(N-M)(N-C)}{N(N-M-C+R)}$，而这个比值不小于 1，当且仅当 $NR \leq MC$。所以 p_N 一直递增至 $N = \frac{MC}{R}$，然后递减，而最大值就在该数的整数部分取得。

2. 从 a 名女性和 b 名男性中选出 c 人组成委员会，按其中女性的人数 j 来数：对每个 j，共有 $\binom{a}{j}\binom{b}{c-j}$ 种。

3. $\frac{\mathrm{d}}{\mathrm{d}p}\big(h \ln p + (n-h)\ln(1-p)\big) = \frac{h}{p} - \frac{n-h}{1-p} = 0$ 给出 $p = \frac{h}{n}$。

4. 当 $R = 0$ 时，p_N 永远递增，估计值是无限大。这个方法只有在第二个样本中有足够多已标记的动物时才行得通；重复捕获的数目少的话，估计值会非常不稳定。

13.3 德国坦克问题

1. 设 M 为最大值，则 $\mathrm{E}(M) = \frac{nN}{n+1}$ 及 $\mathrm{Var}(M) = \frac{nN^2}{(n+1)^2(n+2)}$，所以 $\mathrm{E}(M-N)^2 = \frac{2N^2}{(n+1)(n+2)}$。这是 $\mathrm{Var}(Z_1) = \frac{N^2}{n(n+2)}$ 的 $\frac{2n}{n+1}$ 倍：较差。最佳的倍数是 $c = \frac{\mathrm{NE}(M)}{\mathrm{E}(M^2)} = \frac{n+2}{n+1}$，其均方误差（mean squared error）为 $\frac{N^2}{(n+1)^2}$，比 Z_1 稍佳。

2. $\mathrm{E}(\min) = \frac{N}{n+1}$，而 min 的方差与 M 的相同，所以 $\mathrm{Var}\big((n+1)\min\big) = \frac{nN^2}{n+2}$。这个方差不会随 n 增大而缩小：是个差劣的估计量（estimator）。

3. 最大值不超过 m，当且仅当全部 k 个数都来自 $\{1,\dots,m\}$。当 $k=1$ 时，估计量是 $2X-1$，其平均值为 $2\cdot\frac{N+1}{2}-1 = N$。

4. $98 + \frac{98}{5} - 1 = 116.6$：大约 117 辆坦克。

13.4 平均数还是中位数?

1. 先选定哪一个值是中位数（median）（有 $2n+1$ 种选法），再选定其余各值中哪 n 个在它之下（有 $\binom{2n}{n}$ 种选法）。所求概率为 $(2n+1)\binom{2n}{n}y^n(1-y)^n\,\delta y = \frac{(2n+1)!}{(n!)^2}y^n(1-y)^n\,\delta y$。

2. $\mathrm{Var}(Y) = \frac{1}{4(2n+3)} = \frac{1}{20}$ 及 $\mathrm{Var}(\bar{X}) = \frac{1}{12(2n+1)} = \frac{1}{36}$。平均值较佳。

3. $\mathrm{Var}(\max) = \mathrm{Var}(\min) = \frac{m}{(m+1)^2(m+2)}$ 及 $\mathrm{Cov}(\max,\min) = \frac{1}{(m+1)^2(m+2)}$，由此得出 $\frac{1}{2(m+1)(m+2)}$。它缩小的速度与 $\frac{1}{m^2}$ 同阶，远快于平均值的 $\frac{1}{12m}$：对于均匀分布（uniform distribution）的数据，信息都在极端值之中，与坦克问题的情况一样。

4. 平均值约为 18.6，被一个打字错误完全破坏；中位数是 10.1，不受影响。

13.5 问尴尬的问题

1. 设 $t = ap + bq$，即答案属实的概率。回答“是”的概率为 $\frac{1}{2}t + \frac{1}{2}(1-t) = \frac{1}{2}$，所以 $\mathrm{P}(\text{正面} \mid \text{是}) = t$，而它等于 $\frac{1}{2}$，当且仅当 $2t = 1$。若由同一名孪生儿回答两次，则不论硬币掷出的是正面还是反面，先答“是”后答“否”的概率都是 $pa(1-a) + qb(1-b)$，所以后验概率（posterior probability）是 $\frac{1}{2}$。两次都答“是”的概率分别为 $pa^2 + qb^2$（正面）和 $p(1-a)^2 + q(1-b)^2$（反面），两者相等当且仅当 $2(ap+bq) = 1$。

2. $\mathrm{P}(\text{是}) = \frac{3}{4}\pi + \frac{1}{4}(1-\pi)$，所以 $\hat{\pi} = \frac{0.4-0.25}{0.5} = 0.3$。标准误（standard error）为 $\frac{\sqrt{0.4\times0.6/400}}{0.5} \approx 0.049$。

3. $\mathrm{E}(Y^2) = \sum_c \mathrm{P}(C=c)\big(\mathrm{Var}(Y \mid c) + \mathrm{E}(Y \mid c)^2\big)$；减去 $\mathrm{E}(Y)^2 = \big(\sum_c \mathrm{P}(C=c)\mathrm{E}(Y \mid c)\big)^2$ 便得 $\mathrm{E}(\mathrm{Var}(Y \mid C)) + \mathrm{Var}(\mathrm{E}(Y \mid C))$。

4. Z_2 接近 $\frac{n}{6}$（正面）或 $\frac{5n}{6}$（反面），两种情况的概率各为 $\frac{1}{2}$。所以 $\mathrm{P}(Z_2 > \frac{n}{2}) \to \frac{1}{2}$，尽管 Z_2 几乎从不接近 $\frac{n}{2}$。

13.6 累积起来的误差

1. 每个截尾误差(truncation error)都在 $(-1,0]$ 上均匀分布;各加上 $\frac{1}{2}$ 后,便在 $(-\frac{1}{2},\frac{1}{2}]$ 上均匀分布,与四舍五入(rounding)的情况完全一样。所以修正后的 VOZ 与 LEP 一样好。

2. 每个误差都在 $(-0.05,0.05)$ 上均匀分布,方差为 $\frac{0.1^2}{12}$,所以总和的标准差为 $\sqrt{10\,000\times 0.01/12}\approx 2.9$。

3. $3000\times 250\times 0.0005=375$ 点。

4. 两个误差时,是 $[-1,1]$ 上的 $1-|x|$。三个误差时:当 $|x|\le\frac{1}{2}$ 时为 $\frac{3}{4}-x^2$,当 $\frac{1}{2}\le|x|\le\frac{3}{2}$ 时为 $\frac{1}{2}\left(\frac{3}{2}-|x|\right)^2$,已经很接近正态曲线(normal curve)。

13.7 制造相关性

1. 对于标准化变量(standardised variables),$\operatorname{Var}(X_1+X_2+X_3)=3+6\rho\ge 0$,所以 $\rho\ge-\frac{1}{2}$。对于 n 个变量,$n+n(n-1)\rho\ge 0$ 给出 $\rho\ge-\frac{1}{n-1}$。

2. 取 $b=\frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{1-\rho_{12}^2}}$,则 $(1-\rho_{12}^2)(1-\rho_{13}^2-b^2)=1-\rho_{12}^2-\rho_{13}^2-\rho_{23}^2+2\rho_{12}\rho_{13}\rho_{23}=\det R$。由于 $1-\rho_{12}^2>0$,这两个条件等价。

3. $\operatorname{Var}=\frac{1}{4}(0.04+0.09+0.12\rho)$:当 $\rho=-1$、0、1 时,标准差分别为 0.05、0.18 和 0.25。除非两只股票完全相关(perfectly correlated),否则这个组合的风险低于两者风险的平均值(0.25)。

4. $Z_2=\dfrac{Y_2-\rho_{12}Y_1}{\sqrt{1-\rho_{12}^2}}$。

Beyond STEP

STEP 是大多数中学生一生中应考过最艰深的数学考试，也是最出色的考试之一：题目简短、陌生而深刻，当中最精彩的，其实是乔装了的大学数学小品。

本书精选最有趣的 STEP 试题，由 1987 年首次开考之前的样本试卷（specimen papers），到 2025 年的试卷，并以每一道题目作为一扇门。每道题目都会仔细解答，沿途提出优秀学生会问的问题，然后带你穿过这扇门：这个想法从何而来，到了大学会演变成甚么，以及接下来该读些甚么。

本书为正在准备 STEP、有天分又有志投身数学的学生而写。他们想要的不止是模拟试卷：他们想开始像数学系本科生那样思考，并一探前方有甚么正等着自己。

- 共 13 章，涵盖纯数学（pure mathematics）、力学（mechanics）、概率（probability）与统计（statistics）
- 136 道 STEP 历届试题，每题均附引导式解答
- 每道题目背后的大学数学
- 附答案的延伸练习，以及延伸阅读

全书共 13 章，本册为第 13 章。 全书各章及所有详细解答均可免费阅读、下载及分享，网址为

beyondstep.keithmcnulty.org

文字内容已贡献予公有领域（CC0 1.0）。STEP 试题 © Cambridge University Press & Assessment。

