第 3 章

统计学思维

统计学思维是一种与众不同的思维方式，它有点像是做侦探，鼓励质疑精神与另辟蹊径。

Statistical thinking is a different way of thinking that is part detective, skeptical, and involves alternate takes on a problem.

——统计学教授弗兰克·哈瑞尔（Frank Harrell）

本章的内容是统计学思维，致力于帮助读者以批判性的方式看待生活与工作中遇到的数据，为本书接下来的内容做铺垫。阅读本章之后，你也会在阅读新闻与最新的科技文章时获得一个崭新的视角——统计学的视角。

在进入正文前，先说明两件很重要的事情。

首先，本书涉及的统计学仅仅是统计学科的冰山一角，这一章的内容无法代替一整个学期的“统计学”课程（对学生读者致以歉意）；也并不能像那本经典的著作——《思考，快与慢》一样，从各个心理学方面深入分析“思维”。尽管无法面面俱到，但我们会介绍几个核心的概念，并尽可能地为统计学思维建立一个基础。

其次，接下来的几章可能会让你对数据产生疑虑，开始认为统计学不过是无稽之谈，用复杂的公式和数字遮蔽了真相；或者你会质疑读到的每篇文章，因为作者不见得有多懂统计学。

但这并不是我们的本意。我们的目标并非让你拒绝接受这些信息，而是帮助你懂得如何推敲并理解它们，知道它们的局限性，并在某些情况下欣赏它们。

3.1 学会质疑

统计学思维的核心信条之一是“学会质疑”。

很多人在日常生活中或多或少都会这样做：我们会下意识

地质疑广告商们信誓旦旦的承诺(“一个月瘦十斤!”或是“这只股票将要大涨!”)和社交媒体上耸人听闻的帖子。我们都有质疑的能力,而以旁观者的视角揭穿那些明显的骗局也富有乐趣。但当事情与人们自身息息相关时,质疑就会变得困难起来,例如每一场美国大选——仔细想想,当某一党派人士接收关于另一个党派[①]的断言或数据时,是否会下意识地感到怀疑?人们的脑海中也必然会浮现出这样的话语:“他们的信息来源是有问题的,我的信息来源更可靠。他们看到的是假的,我看到的是真的。他们只是不知道事情真相。”

这样的讨论很快就会转变成哲学问题,我们并不打算引发一场政治辩论,或是深入研究那些塑造了美国公民信念和政治信仰的因素。但这里有一个教益:当“一切”当中包含人们自己的思考和逻辑时,人们就很难质疑一切。

回到本书的话题上,思考一下我们在工作中接触到的那些与公司前途、员工表现和薪水息息相关的信息。当你看到那些分散在表格与 PPT 中的数据时,是否怀着质疑精神?从作者的经验来看,很多时候并非如此。出现在会议室中的数据往往被视为铁一般的事实,白纸黑字不容置疑。

为什么会发生这种情况?可能是因为你没有时间质疑、推敲,或收集更多的数据。很多时候,被用作展示的数据就是我们所有的数据,我们只能依据它而行动,并在发生问题时从中寻找原因。当面对这样的限制时,质疑精神就自然而然被抛在一旁了。另一个可能的原因是,即便我们知道数据中的问题,但我们的上级领导未必知道。这会造成连锁反应——每个人都认为自己的上级(甚至下级)已经板上钉钉地接纳了这些数据,最终导

① 美国是两党制国家。

致所有人都“默认”了这一数据是正确的。

数据达人则会打破这个连锁反应，这要从理解随机波动开始。

关于“统计学思维”

本书所说的“统计学思维”是一个笼统的含义，就像本章开始的引言中定义的那样。人们或许更倾向于使用其他的名称，如概率思维、统计学素养，或是数学思维。但不论选择哪个名字，它们都涉及对数据和证据的评估。

有些人或许会感到奇怪，为什么思维方式如此重要。工作和生活中，大多数情况下我们都不需要在意它，那么为什么现在要开始关注呢？为什么数据达人需要关注思维方式？

在一篇题为“数据科学：受过教育的人都该了解的事”的文章当中，哈佛大学的经济学家兼教务长阿兰·贾伯解释道：“了解数据的益处是显而易见的，在当今愈发重要。随着人们根据数据给出的预测越来越准确，数据科学的价值也与日俱增，这个领域也会吸引更多的关注。但与此同时，进步也会引发自满情绪，让我们对这一学科的缺陷视而不见。未来的工作者们不仅需要认识到数据科学将如何协助他们的工作，也需要认识到数据科学的不足……我们需要对概率思维有更深入的了解，也要具备衡量证据的能力，这会造福所有人。”

3.2 无处不在的随机波动

我们观测的数据会波动，这并不是什么石破天惊的大新闻。

股价每天都在变化；随着收集数据的公司或时间不同，民意调查的结果也会波动；油价时高时低；看医生时我们的血压会升高……即便是每天上班通勤需要的时间，如果精确到秒，也会随着交通状况、天气、是否需要接送孩子，或途中是否停下来买了一杯咖啡而有所不同。万事万物中都有随机波动，对此你有何想法？

你或许能够接受，或至少可以忍受这些日常生活中的变化，甚至也有可能享受它们（当然，股价除外）。但总体而言，我们都知道，事物在时时变化，而并不是所有时候我们都能解释清楚变化的原因。在交车辆保养费或电费时，只要是在可以理解的范围之内，我们可以接受每次的价格略有不同。但就像之前 3.1 节中解释的，当遇到与我们的职业前途或公司前景有关的数据时，我们更难批判性地看待。

一家公司的销售额每天、每周、每月和每年都在发生波动。前一天和后一天的顾客满意度有可能大相径庭。如果我们接受了随机波动的现实，就不必对每个波峰与波谷做出解释。但在商业活动当中，解释却是必需的。公司高层可能会问：某一周的销量为什么格外高？试图让我们得出答案后，又会让我们重复能够增加销量的“好”做法，减少那些“坏”做法。在涉及我们自己从事的工作时，随机波动将会带来无助感。

当涉及商业活动时，我们可能就没有自己想象的那样对随机波动习以为常了。

事实上，波动分为两种。第一种波动来自收集和测量数据

的方法，称为**测量误差**。第二种波动是过程本身带有的随机性，称为**随机波动**。乍一看，二者之间的区别无关紧要，但对于统计学思维而言，其中的差别却十分重要。当我们做决策时，是基于一个不可控的随机波动，还是一个本质上可控的潜在过程？我们都希望是后者。

简而言之，波动会带来不确定性。

接下来通过一个虚构案例和一个真实案例来说明这一点。

虚构案例：客户满意度调查[①]

假如你是一家连锁超市的管理人员，而公司总部密切关注着店铺客户满意度的数据，这些客户通过拨打小票上提供的电话号码反馈他们的意见。问卷要求顾客用数字 1～10 描述自己的满意程度，1 表示“非常不满意”，10 表示“非常满意”。虽然问卷还涉及一些其他的问题，但最终只有总体评分起作用。

除此之外，总部只希望看到 9 分和 10 分，8 分在他们看来和 0 分没有什么区别。数据以周为单位收集，最终以 PDF 文件的格式送到管理店铺的你以及总部办公室手中，文件里都是一些花花绿绿的图表，长度刚好足以传达这些信息。但这些数字将会影响你和你上司的奖金，于是每周你必须高度紧张，努力钻研这些数字，使得 9 分和 10 分的比例达到 85%。

现在我们讨论一下波动的来源——问卷调查是如何衡量结果的。用 1～10 的数字来衡量任何东西都是糟糕的做法。对一个人而言的 10 分体验（“他们不卖我想要的东西，但其中一位工作人员帮我找到了替代品！”）对于另一个人来说可能只值 5 分

① 我们之所以格外关注客户满意度，是因为它：很难精确测量；受到小部分人群的高度影响；被管理层高度看重。

(“他们不卖我想要的东西！一位工作人员不得不帮我找替代品。”)。更何况,其他可能的变量未被考虑在内,如超市里有一个粗鲁的员工,或超市里人太多,或经济下行导致每个人都心情欠佳,或其他顾客是否带着孩子购物……及无数其他的可能性。

我们并不是说问卷的形式应该被舍弃,而只是想要说明,这种收集数据的方式会带来一些常常被人忽略的误差。如果忽略这些误差,就会把预期之外的偏离归因于低质量的服务,而非评判标准本身带来的差异。于是商家就会盲目地追逐着高分(9分与10分),却不明白这些波动的真正来源是收集数据的方法。

事情可能是这样的——假设每天有50个人提供反馈,持续52周,一周就会得到350份问卷,一年就会得到18200份。

有了这样的参与度,你或许会认为这很好地反映了客户感知。于是,在每周结束时,收集结果,总部计算出9分与10分的总数,再除以350,并将结果总结成图3.1中的图表。数字在85%以上时,你会获得表彰;而数字在85%以下时,你会开始焦虑。每周一你都会收到上周的结果,并与总部通电话讨论这个数字。想象一下第5～9周的谈话该有多么紧张,只差一点点就能达标了。直到第10周,你终于突破了那条线,而这毫无疑问要归功于上司的决心。但是很快,第11周到了,你得到了一个新低。如此循环往复。

但是,图3.1中显示的是纯粹的随机现象。我们生成了18200个随机的数字,都是8、9和10,用来代表顾客反馈的打分,并把它们打乱[①]。我们从每一“周”中取出350个数字,并计算满意率。数据集中9分和10分一共占85.3%,这非常接近真

① 在我们的模拟中,得到8分的概率是15%,9分的概率是40%,10分的概率是45%。因为数据是人工生成的,所以我们知道真实的满意率,即9分和10分的概率之和,是85%。

实值 85%，并达到了总部设下的标准。但由于随机波动，每周的结果围绕着标准线上下浮动。

由于缺乏统计思维，你、你的上司，和总部的所有人都努力提升服务水准，希望提高一个随机数字，尽管你们所做的一切对这个数字完全没有影响。

我们将这种行为称作**达标幻象**，即试图提升一个缺乏统计学意义和基础的指标。

在你的工作中，是否有这种现象呢？

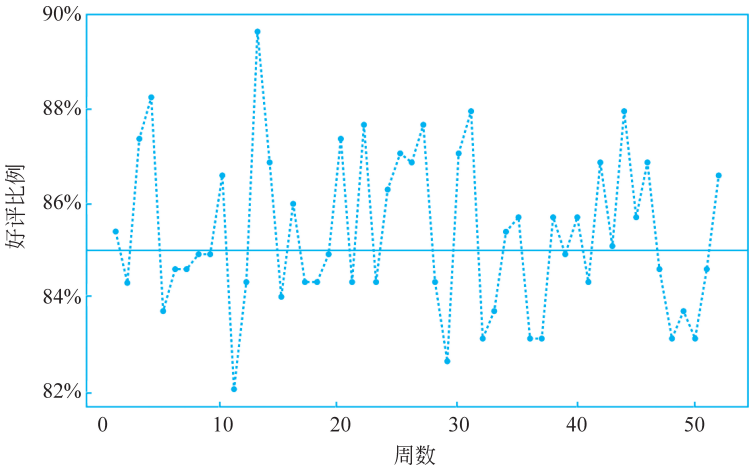

图 3.1　每周顾客反馈中好评所占的比例，水平线代表达标标准，即 85%

案例分析：肾癌发病率

按照每 10 万人中的病例数来看，美国肾癌发病率最高的地区是其中西部、南部和西部乡村地区的某些郡。

设想一下为什么会这样。

或许你会认为，在美国中部的乡村地区，医疗条件比较差。又或者，高蛋白、高盐、高脂肪的不健康饮食习惯，外加过多摄入

的啤酒和雪碧也是可能的原因。围绕事实编织叙事，这是很简单、很自然的事情。你甚至也可以想象出，研究者为了减轻肾癌发病率高的状况，已经开始制订应对措施。

但与此同时，还有另一个事实：美国肾癌发病率**最低**的地方，也同样是中西部、南部和西部乡村地区的某些郡，而且往往与那些发病率最高的地方相邻。[①]

怎么会这样呢？两个人口结构相似的地区怎么会得出大相径庭的结果？用来解释为何乡村地区具有高发病率的每个原因，在相邻的地区或多或少都能找到。所以，其中一定有一些其他的原因。

让我们从美国中西部找两个郡，设为A与B，并假设它们各有1000个居民。如果A郡没有病例，发病率就是0，很显然是最低的。而如果B郡有1个病例，它的发病率就是每10万人100个，是全美国最高。极小的人口数量导致了很大的波动，于是同时带来了最高和最低的发病率。作为**对比**，如果在拥有150万人口的纽约郡（纽约市曼哈顿区）多了一个病例，基本无法对整体结果带来任何影响。75例和76例对应的每10万人发病率分别是5和5.07。

这个现象是真实存在的，《科学美国人》杂志为此发表了一篇文章，题为“最危险的公式”。[②] 图3.2总结了美国各个郡的癌症发病率与人口的关系，那些人口最稀少的郡分布在图中左侧，它们的癌症发病率波动幅度非常大，从0到20，分别包括了全美国的最低和最高值。随着人口逐渐增多（对应着图中从左到

① 如果我们先说发病率最低的是乡村地区，你会给出怎样的理由来解释呢？试试吧，你会发现围绕数据编故事是很容易的事情。

② Wainer, H. (2007). The most dangerous equation. *American Scientist*, 95 (3): 249.

右),波动幅度逐渐减小,使得图表整体呈现三角形。右侧的人口密集地区波动很小,意味着额外的病例很难影响整体结果,最终稳定在 10 万人中约 5 例。

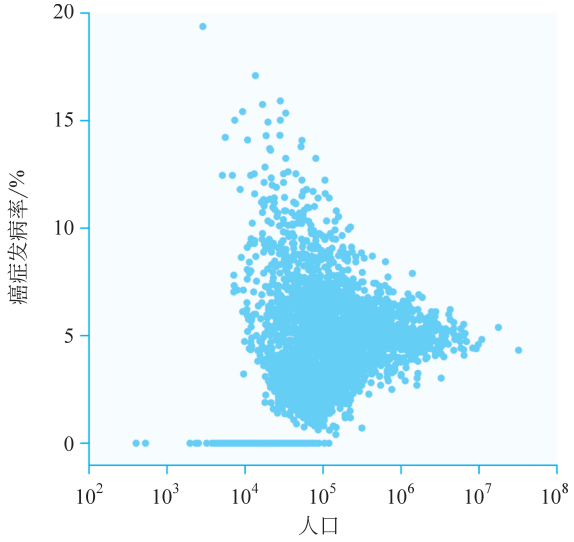

图 3.2　癌症发病率(图源:《科学美国人》杂志)

上面提到的文章中还分享了其他的例子,都与小样本带来的大波动相关。例如,在小规模的学校中,一两个不及格的学生将会使学校的整体及格率受到极大的影响。小样本往往会带来极端结果。

3.3　概率与统计

前文我们解释了随机波动,并讨论了它会为很多行业带来不确定性。而不确定性是可以被人为控制的,这就是概率与统计的用武之地。

当描述结果数字时，我们常常混用“概率论”与“统计学”这两个术语，或是把它们相提并论。但这里我们可以稍稍深入一些，理解其中的不同之处。

想象一大袋鹅卵石子，你不知道它们的颜色、形状和大小，也不知道袋子里有多少个石子。但你可以伸手从袋子里随机拿出一把。

现在我们有一大袋内容未知的鹅卵石子，外加手里的一小把，也没来得及看。你既不知道袋子里是什么状况，也不知道自己手中是什么状况。

概率论和统计学的区别之处就在这里——在概率论当中，你知道袋子中的情况，并利用这些信息去猜测手中的情况。在统计学当中，你看向自己手中的内容，并用这些信息去反推袋子中的内容。

简而言之，概率论从大及小，统计学由小见大。

现在考虑两个现实中的例子。

（1）拉斯维加斯的赌场是建立在概率论之上的。

每当你加入一个赌局，都相当于从赌场设置好的鹅卵石袋子中拿了一把。袋子里有“输”有“赢”，代表“赢”的鹅卵石数量恰好足够吸引你继续玩下去。赌场对随机波动十分了解，它们通过精心优化的输赢率使你保持兴致，借此盈利。而长期看来，赌场总会赚钱，因为他们设计了那袋鹅卵石，准确地知道那里面究竟有什么。对于赌下的每一注、桌上的每一枚筹码、每一次拉动老虎机，赌场都知道赢钱的概率是多少。如果你思考赌场中将会产生多少数据，就能够明白他们既生活在一个充满随机波动的世界，同时也能精准地把握可能发生的结果。

（2）政治民调是建立在统计学之上的。

在赌场中，鹅卵石袋子是精心设计的，人们从中反复取样。

但在美国大选中，直至选举日，当所有的鹅卵石（即选票）在被计数之前，政客们并不知道袋子中究竟有些什么[①]。在选举日之前，政客和党派都只能看到随机选取的一小部分鹅卵石（即民调结果），而且为了得到这些数据，他们需要花费大量资金。通过这个样本，他们可以推断出袋子中的模式，进而调整竞选策略。因为他们的信息并不完整，并时常伴随着偏差和错误，他们并不总能得到正确的结果。但当他们掌握了正确信息时，那往往能够成为决胜的筹码。

接下来，让我们简短地讨论一些与概率和统计密切相关的概念。

概率与直觉

我们之前提到随机波动是不可控的。但它可以被度量，而概率就是用来度量不确定性的工具。

有些情况下，概率符合我们的直觉。当我们扔一枚骰子时，我们知道得到某个结果的概率（每个结果均是 1/6）。这些与偶然性相关的小游戏背后是简单的概率论，它符合我们的直觉。事实上，正因它们看上去太简单，我们才无法察觉其背后的复杂性。一些商业广告就利用了这一点，用简单的概率迎合我们的直觉，让我们误以为自己对此有所了解。

你或许曾经见到过这样的广告词："5 名牙医中，有 4 名对 X 表示赞同。"X 可以是任何主张，比如口香糖能帮助减少蛀牙，或是小苏打能够使牙齿洁白，这无关紧要。

现在，假设你的面前有 5 位牙医。假设 80％的牙医都赞同

① 这里我们简化了问题，在美国大选中，党派都会试着影响袋子中的内容，包括鹅卵石的个数与颜色。但即便如此，他们仍然不知道袋子里的具体内容，而必须依赖取样。

X，那么这5位牙医中有4位表示赞同的概率是多少呢？[①]

100%？90%？抑或80%？

答案是41%。

直觉看来，这似乎太低了，但事实的确如此。让我们来看一看为什么。表3.1中显示了5位牙医中4位表示赞同的一个可能情形。

表3.1 牙医赞同广告内容的概率

	牙医1	牙医2	牙医3	牙医4	牙医5
是否同意	是	是	是	是	否
概率	0.8	0.8	0.8	0.8	0.2

这个组合出现的可能性$=0.8\times0.8\times0.8\times0.8\times0.2=0.08192$。

或者可以简写成：

$p=0.8^4\times0.2=0.08192$

但如表3.2所示，一共有5种组合可以达成1人反对的情形，因此我们将这个概率乘以5：$0.08192\times5=0.4096$，约等于41%。

表3.2 5名牙医中有4名赞同的可能情形

情况	牙医1	牙医2	牙医3	牙医4	牙医5
1	同意	同意	同意	同意	不同意
2	同意	同意	同意	不同意	同意
3	同意	同意	不同意	同意	同意
4	同意	不同意	同意	同意	同意
5	不同意	同意	同意	同意	同意

① 例子的来源：www.johndcook.com/blog/2008/01/25/example-of-the-law-of-small numbers。

或许 5 名牙医中有 4 名表示赞同，但这并不意味着**每 5 名**牙医中都有 4 名会认同 X。回到我们的鹅卵石比喻上，如果整袋鹅卵石中有 80％是白色，20％是黑色，当我们一把抓 5 个石子时，有可能 5 个都是白色。在另一些罕见情况下，我们会得到 5 个黑石子。这就是随机波动。

我们之所以分享这个例子，也是为了再一次强调人们往往会低估随机波动的大小，尤其是在处理小样本的时候。人们依照直觉期望看到的情况鲜少与根据概率论计算出的实际情况相符。而**低估随机波动**则会导致人们**高估小样本数据的可信度**，这被称作“小数定理”，定义是“挥之不去的信念……认为小样本能忠实地反映整体的样貌”。[①]

一位数据达人应该具有统计思维，而这意味着对直觉保持警惕，知道它有时会欺骗我们。接下来的章节将在这方面给出更多的例子。

统计发现

统计学往往被分成**描述性**统计和**推断性**统计。你可能已经对描述性统计非常熟悉，即便之前没有使用过这个术语。描述性统计指的是那些用来总结数据的数字，你会在报纸上或工作报告中看到它们——上季度平均销量、年度同比增长、失业率等。平均数、中位数、极差、方差和标准差都是描述性统计，有特定的计算公式，在统计教材中都会涉及。

描述性统计对数据进行了有意的简化，例如，为了将记录着公司产品销量的一整个表格浓缩成少数几个关键数字，用以总

① Tversky, A., Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157): 1124-1131.

结主要信息。回到那个鹅卵石的比喻上，描述性统计指的就是数一数手中的鹅卵石，并总结出结果。

尽管这一步很有用，但我们鲜少停留在这里。我们想要更进一步，研究如何通过手中的信息，对整个袋子中的情况做出有理有据的推断。这就是推断性统计，它是一个过程，“从整个世界到数据，再从数据返回整个世界”。[①] 我们将在第 7 章深入讨论这个话题。

现在先举一个例子。想象一下当你看到这样的报纸头条时会作何反应：“75%的美国人相信 UFO 存在！”接着，你发现那是在美国新墨西哥州罗斯维尔的国际 UFO 博物馆与研究中心采访了 20 名游客后得出的结果。你认为能够通过这些信息来准确地**推断**美国人中究竟有多少相信 UFO 吗？

一个数据达人会立刻引起警觉。这个 75%的统计数字并不准确，因为如下几个原因。

（1）**样本偏差**。前往国际 UFO 博物馆与研究中心的游客比普通大众更有可能相信 UFO 存在。

（2）**小样本**。我们已经看到过小样本会带来多么大的统计误差，从 20 个人的样本推断上千万人的想法是不可靠的。

（3）**隐含假设**。头条新闻提到了相信 UFO 的“美国人”，因为样本是在美国选取的。但你或许已经注意到了，该博物馆是一个国际旅游景点。你无法确定参与调查的每位游客都是美国人。

像**偏差**和**样本量**这样的概念，都是用来帮助我们衡量某个统计推断结果是否合理的工具。它们是你工具箱中的重要组

① O’Neil, C., Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. O’Reilly Media, Inc.

件。而搞清隐含假设同样重要。想要像一个数据达人一样思考，就意味着你不能按照表面含义接受结论中的那些隐含假设。

所以，当你在工作中接触数据时，不要一味接受其中的信息，也不要盲目相信自己的直觉。

掌握统计学思维和学会质疑都是数据达人应该做的。接下来的章节中，我们将会指出哪些问题有助于读者增进统计思维。

统计学思维参考资料

在本章开头，我们提到，本章所涉统计思维只是统计学中的冰山一角。幸运的是，有几本非常优秀的书籍对统计思维进行了更深入的讨论。作者最喜欢的几本是：

- 《糟糕的谎言与统计学：如何理解媒体、政客和活动家给出的数字》（*Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists*, by Joel Best）
- 《如何避免犯错：数学思维的力量》（*How Not to Be Wrong: The Power of Mathematical Thinking*, by Jordan Ellenberg）
- 《如何利用统计学撒谎》（*How to Lie with Statistics*, by Darrell Huff）
- 《赤裸裸的统计学：如何不再害怕数据》（*Naked Statistics: Stripping the Dread from the Data*, by Charles Seife）
- 《可证性：你是如何被数字欺骗的》（*Proofiness: How You're Being Fooled by the Numbers*, by Charles Wheelan）

- 《醉鬼的步伐：随机性如何掌控我们的生活》(*The Drunkard's Walk: How Randomness Rules Our Lives*, by Leonard Mlodinow)
- 《信号与噪声：预测成功与否的因素》(*The Signal and the Noises: Why So Many Predictions Fail-But Some Don't*, by Nate Silver)
- 《思考，快与慢》(*Thinking Fast and Slow*, by Daniel Kahneman)

本章小结

本章给出了统计学思维的基础内容，以此为起点，我们将展开本书的其他内容。

首先，我们提到了随机波动的重要性，并强调了我们应该了解它在测量过程中以何种方式存在。在调查客户满意度的案例中，我们看到问卷的形式往往会带来很大的误差。这种误差并不是因为服务本身有问题(虽然并不排除这种可能性)，而是因为问题本身的设计，使得本来相似的情况可能带来截然相反的结果。

我们还讨论了概率论与统计学，它们是处理随机波动的优良工具，能够显示哪些波动是可以预测的，而哪些是在长期看来无关紧要的。概率论由大及小：它建立在非常庞大的信息之上，告诉我们如果随机从中取一小部分，有可能得到怎样的结果。统计学以小见大：它通过我们手中持有的部分，告诉我们有关整体数据的某些信息。当我们想要的信息被隐藏起来时，统计学与概率论都有助于让我们对整体图景有更多的了解。

最后，我们谈到了如何利用概率论与统计学的知识来帮助我们保持正确的怀疑态度。