《简单统计学：如何轻松识破一本正经的胡说八道》，加里·史密斯著

注：读书笔记仅做记录，可读性较差。

看这本书一直出于高度谨慎的状态，因为有时候不知道前一页看到的例子，是否就会在下一页被推翻，有时候看到一个标题就知道肯定不靠谱，但是仍然想继续看下去为什么会出现这种研究观点。自选择偏差，幸存者偏差，证实性偏差，计算误差，相关性因果偏差，均值回归，偏好数据保留，平均定律等都能带来很多迎合直觉或违反直觉的问题。读这本书会有稍微的不适感，一方面是作者列举了太多一本正经胡说八道的例子，同时作者那种直言不讳的风格也非常尖锐。书名虽含有统计学，其实统计学知识并不多，如果是奔着学习统计学知识去的，不是很推荐。我生活中非严肃场合偶尔也会一些有点根据的胡说八道，不过通过作者的案例，对“一本正经”有了不一样的认识，不管证据是否可靠，但至少一本正经。

◆ 第1章 模式、模式、模式

当时“保罗”预测的时候就一直很纳闷为什么一直在选德国，只是当时没有详细考虑这个问题，原来问题出在国旗身上，如果当时对章鱼预测的国旗做个简单的相似性分析，很容易就会发现这个问题，只是大多数人更愿意相信惊艳的结果。

> 不过实验表明，章鱼能够识别明暗度，而且喜欢横向形状。德国国旗有由三块鲜艳的水平条纹组成，塞尔维亚和西班牙的国旗也是如此，

有一定道理，数据挖掘建模中80%时间都在做特征，就是为了对数据进行重塑，不过有一点需要申明，有效的挖掘模式是可以迁移的，而这种可迁移就是利用价值。

> 这些做法——选择性报告和数据搜刮——被称为数据挖掘。通过数据挖掘发现的统计显著性只能体现出研究人员的耐心。在独立检验证实或拒绝结论之前，我们无法判断某种数据挖掘马拉松到底证明了某种实用理论的有效性还是研究人员坚定的毅力。

◆ 第2章 不再神奇的超级畅销书

自选择偏差是一个无法完全拆分的变量，这种变量有一个特点，在结果出来之前，无法度量它的存在，可是当结果出来后再度量它也失去了意义。

> 当数据涉及人们的选择时（比如当人们选择上大学、结婚或者要孩子时），就会出现“自选择偏差”。在这种情况下，对于做出不同选择的人进行比较的做法是靠不住的。

这里有一个明显的问题是因果关系的界定，到底是逮捕定罪降低了投票率，还是不愿意参与投票的人的逮捕定罪率更高，而且这里只能说明观察到的两项数据变化趋势有一定的相关性，并不能直接说明它们有因果关系。这在实际中经常会被一些所谓的分析师们捉弄。

> 一份针对美国城市最边缘群体的大规模调查发现，在曾被警察拦截和盘问的群体中，投票概率降低了8%；在曾被逮捕的群体中，投票概率降低了16%；在被定罪的群体中，投票概率降低了18%；在曾经遭到拘留或监禁的群体中，投票概率降低了22%。

简单而言，这里把去过法国不止一次定为了认为法国人友好的因，实则可能只是因为这些人觉得法国人友好才去了不止一次。

> 大多数过去两年对法国进行过不止一次休闲旅行的美国人不认为法国人不友好。

幸存者偏差和自选择偏差类似，不同的是幸存者偏差不会出现嵌套循环，而且幸存者偏差似乎更容易被大家识别。

> 许多观测性研究存在幸存者偏差。例如，健康维护组织在一项调查中发现，超过90%的成员对该组织感到满意。这里存在两种幸存者偏差，它们都在推高调查的满意度：一些人由于不满意而退出了这项计划，还有一些人离开了人世。

预测一件未发生事情的出现非常难，但是给已经发生的事找理由并不难。就像在股票市场中，每天收盘后网络上一大推长篇大论的分析，似乎自己真的看透市场一样，其实都是事后诸葛亮。

> 对成功的企业、婚姻和人生进行回溯性研究的所有书籍都存在这个问题，包括成功企业、持久婚姻、活到百岁的方法/秘密/诀窍等。这类书籍存在固有的幸存者偏差。

◆ 第3章 被误传的谋杀之都

这种效应很可怕，可是设置对照组的成本又非常高（我一直对以人的未来做对照实验的方法持谨慎态度），所以实际中做很多事情并非一定要证真，只要不可证伪就有存在的必要。

> 当人们最终结束这项实验时，得到的结果令人吃惊。在接受胃冷冻治疗的患者中，34%的患者表示病情出现了好转；在接受与体温相当的液体的患者中，这个比例是38%。又是安慰剂效应！

◆ 第4章 新的经济学上帝

抛开计算机bug而言，我们认为计算机能算出正确的数和计算机算出的数是正确的有本质区别，计算机只能保证从输入到输出的正确，并不能保证全流程。而很多时候，我们在遇到错误时，直接帅锅给计算机，计算机算出来的就是这样，怎么可能错误。

> 错误的问题会导致错误的答案。如果我想知道169的平方根，实际却让计算机计算196的平方根。计算机会告诉我，答案是14。对于我所提出的问题来说，这个答案是正确的。对于我想提出的问题来说，这个答案是错误的。这类错误被称为“计算误差”，但它们实际上是人为误差。遗憾的是，一些计算误差具有极为严重的后果。

◆ 第6章 美国有多少非裔职业运动员?

不换门的话，赢得大奖的概率是1/3。换门的话，赢得大奖的概率是2/3。

> 在电视节目《一锤定音》中，你可以在三扇门之中做出选择。其中，一扇门后面是一项大奖，另外两扇门后面是山羊。在你选择一扇门以后，主持人蒙提·霍尔（Monty Hall）每次都会向你展示你没有选择的一扇门后面的山羊，并且询问你是否想要换一扇门。

的确，概率的乘积会发生很多奇妙的变化，不过说真，作者这里的用法稍微有点过激了，比如，把一个小概率事件（阳性）和大概率事件（全部调查者）放在同一个系统中对比，本身就会有争议。这里还体现了准确率，精确率和召回率的艺术，这种用数字避重就轻的方式很容易迷惑人。

> 在950个阳性检测结果中，475个结果是假阳性。高达50%的阳性员工没有使用大麻。这就是我们需要谨慎对待条件概率的原因。虽然95%的大麻使用者会检测出阳性结果，但是只有50%的阳性结果来自大麻使用者。

◆ 第7章 辛普森悖论

数据被分解，某种程度上就会为每个分类赋予一定的权重，而这个权重会直接影响到结果，总会受到数据分布的不同和异常值的影响。

> 这仍然是辛普森悖论。当数据被分解时，聚合数据中的模式遭到了逆转。

这就有意思了，到底是喝咖啡导致了疾病，还是喝咖啡导致了抽烟，然后抽烟导致了疾病，或是抽烟导致了喝咖啡，又或是存在这种疾病的人本身就更喜欢喝咖啡，当引入系统性差异是，关系就变得复杂了。很多时候相关性并不能证明因果关系。

> 1971年的一项研究发现，同没有膀胱癌的人相比，患有膀胱癌的人更愿意喝咖啡——这意味着咖啡会导致膀胱癌。不过，这里存在一个混杂因素，那就是喝咖啡的人更愿意吸烟。导致膀胱癌的到底是咖啡还是香烟呢?

◆ 第8章 状态火热的雷·阿伦

我们都希望能在一定程度上控制未发生的事，倾向于基于已知寻找可以预测未知的模式，但是有时候这可能就是一个随机事件。就像股价变化，用随机过程一模拟，似乎能发现所有的股票走势。不过换个角度，手热的时候可能就是因为自信了呢，自信了在一定程度上后面的投篮就不再是简单的随机过程，和前面投篮结果就有了直接关系。

> 类似地，如果篮球选手在5次投篮中4次命中，我们可能认为他下次投篮命中的概率是80%。如果这个选手在5次投篮中4次不中，那么他下次投篮命中的概率只有20%。根据很小的投篮样本，我们认为选手的状态由好转坏，从80%的命中率转变成了20%的命中率。我们没有意识到，即使选手每次投篮命中的可能性都是50%，他也会时而五投四中，时而五投一中。

◆ 第9章 胜者的诅咒

我们常认为供给与需求相辅相成，实际在新事物出现时，往往是需求先行，供给只是为了更好的满足需求。反观当下，很多还停留在炫技的所谓技术革命只不过是自我麻醉。

> 供给不会自动创造需求；相反，需求常常会创造供给。

回归有一个条件，影响结果的条件相对独立，这有点像形成正态分布的条件，举个小例子，有一组1到100的整数，随机从里面取两个数取平均值，当实验很多次后，会发现这个平均数趋于50。

> 异常的父母通常拥有不太异常的孩子，反之亦然；同样的道理，不管我们沿着时间前进还是后退，利润率都会出现回归。观测到的回归只能证明当利润率在能力值附近波动时，观测到的利润率差异大于实际能力值的差异。

均值回归的时间成本相对较高，五年的时间，不是所有人都能熬过的，而且五年后是这样，3年后或6年后的结果呢？均值回归的趋势并不是单调的，这似乎就能体现投机者和投资者的差异吧。

> 在被踢出道指五年半以后，西尔斯在2005年被凯马特收购。如果你在西尔斯被道指删除以后立即购买它的股票，那么到西尔斯被凯马特收购的时候，你的总回报率将达到103%。在同样的五年半时间里，取代西尔斯的家得宝下跌了22%。

◆ 第10章 如何转变运气?

世界可能是上帝随机的骰子，但是在每一个人可掌控的那一微末范围内，用上帝的视角思考往往得不偿失。还不如去拥抱现在的规律，发生了什么就勇于去接受，存在即合理也有一定的道理。

> 霉运不会提高好运的可能性，反之亦然。每一次失败不会提高成功的可能性，反之亦然。它可能仅仅是随机性的一种表现而已。

◆ 第11章 德克萨斯神枪手

细思极恐，实际中因这种对数据的过度解读产生的结论可能充斥在我们的周围，甚至是某些故意安排的伎俩，这促使我们在面对结论之前，首先思考有新数据时结论是否可被验证。

> 我们对少年棒球联合会球场附近和水塔附近的推理存在同样的问题。如果我们根据数据编造理论（少年棒球联合会球场导致癌症，水塔预防癌症），那么这些数据当然会支持这种理论！它怎么会不支持呢？我们会编造出一种与数据不符的理论吗？当然不会。根据创建理论时使用的数据来检验这种理论的做法是不公平的。我们需要新的数据。在新数据面前，这种理论也应当成立。

◆ 第13章 黑色星期一

忽略一些数据，只是为了让留下的数据更好的支持自己最初的观念，这很符合人们的观念，认知的进步基本上都是基于过去的知识来对未知进行推理，这就难免推理中会保留历史知识的惯性。就像我写下这个想法，并不是单独是因为作者的这个观点，还因为我对这个观点做出了认同或否定的态度。人的认知地图也是这样不断丰富完善的。

> 在“挑战者”号的例子中，对于重要数据的忽略是一个无心而致命的错误。在其他情形中，人们故意忽略一些数据，因为这些数据不支持他们事先形成的观念。为了相信某件事情是正确的，他们丢弃了与这种信念相冲突的数据。

◆ 第16章 彩票是一种智商税

如果只是使用过去的趋势预测未来，而不考虑实际意义，得出的结论往往是荒谬的，数据有价值，但是只有给数据赋予意义后才能对实际发展产生指导意义。

> 类似地，如果我们仅仅根据过去的趋势推测未来，而不去考虑这种趋势是否有意义，那么我们的结论可能会与众所周知的真相相去甚远。如果我们对股票价格和彩票中奖数字进行仔细检查，寻找跑赢大盘和中彩票的荒谬办法，我们几乎一定会得到更加糟糕的结果。

这句话好有意思，从后视镜中看未来，这种幽默讽刺真是让人哭笑不得。

> 你很少能够通过后视镜看到未来。

◆ 第17章 超级投资者

不完全接受也不通盘否定，通常接受的观点而言股票市场是随机的，这没有错，不过股票价格的变化是相对连续的，这就不同于一般彩票的随机，它让寻找一定的模式成为可能，只是这个模式受到的影响因素太多，完全不可控。西蒙斯的神话是有一定理论基础的，不可高估历史数据的特征，也不可低估量化投资的力量。

> 这些人的开创性工作很好地说明了计量金融分析的两个主要陷阱：天真地相信历史模式是对未来的可靠指引，并且依赖于在数学上很方便却不切实际而且非常危险的理论假设。

这个故事有点意思，金融市场不会让百元纸钞躺在马路上，这是建立在完全的经济假设上，可世界上不止有金融系统，而且在纯金融市场上，躺在马路上的也应该是真钞，因为每个人都是那位金融教授。

> 两位金融教授在人行道上看到了一张一百元的钞票。当一位教授伸手去捡钞票时，另一个人说：“别理它；如果它是真的，那么它早就被人捡走了。”金融教授喜欢说，金融市场不会让百元钞票躺在人行道上；也就是说，如果有一种轻松的赚钱方式，那么它早就被人发现了。

完 ～