

量来改变医生的确定性，或者是他认为患者得了癌症的主观概率。

主观概率和确定性程度的概念，在医学术语里似乎有好多的形式。举例来说，一个作者写道：“因为大多数恶性病变都发生在较大年龄的妇女身上，所以当一位医师面对一个年龄较大的人的时候，癌症的怀疑指数（index of suspicion）就被过高地强调了。”（Gold, 1969, p. 162）另一个作者写道，乳房射线检查可以降低进行活检检查的人数，“在很多情况下，当医生坚持认为该病属于良性的疾病，这是由乳房射线检查诊断来支持的。”（Wolfe, 1964, P. 253）第三个描述认为：“如果医生的主观印象已经给出了足够的理由怀疑是癌症，医生就会强烈地要求做组织活检，尽管乳房射线检查给出阴性的结果。”（Clark et al., 1965, p. 133）其他反映这个概念的表述包括，“信心水平”（Byrne, 1974, p. 37）、“恶性诊断的印象”（impression of malignancy）（Wolfe, 1967, p. 138）、“更为肯定的诊断”（Egan, 1972, p. 392），等等。这些陈述是不够精确的，因为很少有医生会正式地了解主观概率和决策分析的概念。但是，还是有足够的证据证明，对于医生来说，确定性程度的概念，对他们来说是很自然的，而且他们经常用它来帮助选择该采取的行动。

2. 解释乳房射线检查的准确性

现在让我们来想象一个病人，有乳房肿块，并且医生认为是良性的。假设这种可能性是 99%，你把这个句子解释为“医生认为可能（100 里有 99 个）是良性的”等。假设医生已经与很多类似的妇女接触过，她们在大部分主要的方面比如年龄、症状、家族史和查体发现等都和这个病人很类似。假设医生从他的经验中获知对这个群体的人来说，得癌症的几率只有 1%。如果缺乏其他的信息，医生会认为这种病人得癌症的主观概率只有 1%。

现在医生让这个病人去做了一个乳房射线检查，并且得到一个报告，放射科医生认为此人的病变是恶性的。这个信息是新的，并且下一步的行动，也取决于医生对于病人得癌症的可能性的一个重新的评估。如果医生求助于一些文献，他就会发现无数有帮助的陈述，比如：“乳房射线诊断的准确率大概有 90%”（Wolfe, 1966, p. 214）；“一个乳腺包块的病人的乳

房射线检查如果是癌症，通常是正确的”（Rosato, 1973, p. 491）；“乳房射线检查正确诊断乳房恶性病变的准确率为 80% ~ 85%”（Cohn, 1972, p. 98）。如果还需要更多的细节，医生可以找到很多陈述，诸如“结果显示 475 例恶性病变中的 79.2% 都被正确地诊断出来了，1105 例良性病变中的 90.4% 也被正确地诊断出来了，总的准确率达到 87%”（Snyder, 1966, p. 217）。

这样说来，通过你自己对于病人患有癌症的新的可能性的估计，你可能提高了自己的认识，认为医生的评估是有问题的。医生认为肿块可能（99%）是良性的，但是放射科医生却得出了阳性的检测报告，并且有我们刚给出的那些准确性数据。

258 贝叶斯公式可以用来估算概率。这个公式告诉我们：

$$P(\text{ca} | \text{pos}) = \frac{P(\text{pos} | \text{ca}) P(\text{ca})}{P(\text{pos} | \text{ca}) P(\text{ca}) + P(\text{pos} | \text{benign}) P(\text{benign})}$$

在这里， $P(\text{ca} | \text{pos})$ 是指患者患有癌症的概率，假设通过 X 射线获得的诊断结果是阳性（后验概率）

$P(\text{pos} | \text{ca})$ 所代表的概率，是如果病人有癌症，并且放射科医生正确地诊断出来了（正确的阳性检查，或者说是灵敏度）

$P(\text{ca})$ 是患者患有癌症的概率（先验概率）

$P(\text{benign})$ 是指这个患者是良性的病变 [$P(\text{benign}) = 1 - P(\text{ca})$] 的先验概率。

$P(\text{pos} | \text{benign})$ 所代表的概率，是病人是良性病变，但是放射科医生错误地诊断为癌症（假阳性评估）

表 18—1 总结了斯耐德的数字。方格中的每个部分，对应着确定的概率 [如 $P(\text{pos} | \text{ca}) = 0.792$]

表 18—1 乳房射线检查在检测良性和恶性病变中的准确性

X 射线的结果	恶性病变	良性病变
	癌症	非癌症
阳性	0.792	0.096
阴性	0.208	0.904

资料来源：Snyder, 1966。

用 1% 作为医生估计肿块可能是恶性的先验概率，然后加入我们后来获得的检查中的新信息，我们就可以得到：

$$P(\text{ca} | \text{pos}) = \frac{(0.792)(0.01)}{(0.792)(0.01) + (0.096)(0.99)} = 0.077$$

这样，医生就可以得出病人得病的可能性大概有 8%。

3. 错误的概率推理

不幸的是，大多数医生（根据作者的非正式的调查，大概每 100 个人中就有 95 个）都误解了关于检查和评估的准确性的陈述，他们认为 $P(\text{ca} | \text{pos})$ 有大约 75%。其他的研究者，也得到了类似的结论（Casscells et al., 1978）。当被问到这个，那些犯错误的医生通常报告说，他们认为一个 X 射线检查为阳性的人患癌症的概率 [$P(\text{ca} | \text{pos})$]，和一个癌症患者其 X 光检查为阳性的概率 [$P(\text{pos} | \text{ca})$] 是一样的。后一种概率，是在临床研究项目中检测出来的概率，并且为人所熟悉，但正是前一种概率，才是临床诊断的决定中所需要的。即使不是所有的，也是大部分的医生，似乎都混淆了这两者。

任何用来检测某一种疾病是否存在的检查，都具有两种准确性。回顾性准确性（retrospective accuracy），关注的是 $P(\text{pos} | \text{ca})$ 和 $P(\text{neg} | \text{no ca})$ （no ca 这个缩写指的是病人没有癌症。这是因为她的病是良性的，或者她根本就没有病）。这种准确性，在乳房射线检查中经常被提到，它经常在组织诊断已经做出以后，再回过头去看 X 光的诊断。让我们用预测性准确性（predictive accuracy）来描述 $P(\text{ca} | \text{pos})$ 和 $P(\text{benign} | \text{neg})$ ，这种准确性对于临床医师来说很重要，他要根据 X 光检查来预测一个尚未被确诊的病人的病情。

（1）混淆回顾性准确性和预测性准确性

一份关于乳房射线检查的医学文献评论揭示，在 X 光检测报告中，有一种强烈的将回顾性准确性和预测性准确性等同起来的倾向，即 $P(\text{ca} | \text{pos}) = P(\text{pos} | \text{ca})$ 。有很多理由让我们怀疑这种错误是存在的。首先，很多文献中的陈述的措辞，都明显地表明作者在报告他们的研究的时候将预测性准确性 $P(\text{ca} | \text{pos})$ 和回顾性准确性 $P(\text{pos} | \text{ca})$ 等同了起来。举例来说，在 1964 年的《放射学》（Radiology）杂志中有一篇文章说

道, “在 759 例关于 X 光的检验中, 正确的有 674 例, 或者说 89%” (vol. 84, p. 254)。《临床妇产科》(*Clinical Obstetrics and Gynecology*) 杂志的一位投稿者在 1966 年说道: “通过对 500 名患者进行的病理检测, 阿什发现乳房射线检查和病理发现之间的相关高达 90%” (vol. 9, p. 217), “X 光诊断和病理诊断的一致性达到 91.6%” (Egan, 1972, p. 379)。所有的陈述都表明, 如果病人得到了阳性的检测结果, 那么检测就是正确的, 并且病人有 90% 的可能性会患有癌症。这是不真实的。

其次, 一些作者的错误表现得更明显。下边的这段论述, 出现在《外科、妇科与产科》(*Surgery, Gynecology and Obstetrics*) 杂志 1972 年的一期中, 文章的标题是《乳房射线检查的恰当视角》, 它试图去纠正在一些已经存在的文献中的混淆, “大约 1/5 被确诊为乳腺癌的妇女的乳房射线检查是阴性的。如果在阴性检测结果的基础上, 我们就会延误对乳房肿块进行组织活检, 然后, 就有 1/5 的可能性延误了对恶性病变进行组织活检” (vol. 134, p. 98)。这个作者错误地认为 $P(\text{neg} | \text{ca}) = 0.2$, 也就是说 $P(\text{ca} | \text{neg}) = 0.2$ 。他的错误更加严重, 他做出结论说, “对有临床良性病变的乳房延缓进行乳房组织活检, 以期望在乳房射线检查中获得良性结果, 会耽误对我们女性乳房病症的根除”。这样的病人, 得癌症的机会, 依赖于前一种概率, 但是少于 1/100。他的分析错误的比率增加了 20 多倍。

《外科学和妇产科研究》1970 年 (vol. 131, pp. 93-98) 出版了另一个研究小组的研究成果, 他们计算了放射诊断和病理诊断之间的相关。他们的结果如下: 他们选择的病人, 都是已经有了确定的组织诊断, 然后根据 X 光诊断结果将她们分成三组——良性的、癌症和怀疑为癌症的。在 X 诊断为良性的群体中, (在我们的术语中相当于阴性的) 数据显示 84% 的人事实上是有良性的病变, 而且他们强调说 X 光诊断为癌症的人中有 87.5% 的人的组织活检证实它们确为恶性病变。这样, $P(\text{ca} | \text{pos}) = 87.5\%$, $P(\text{benign} | \text{neg}) = 84\%$ 。但是作者把预测性准确性误认为是回顾性准确性了。他们认为, “根据乳房射线检查, 对于有良性病变的人的诊断的正确率达到 84%, 对有癌症的人检查的正确率高达 87.5%。”事实上, 在这个研究里, 阳性的正确率 [$P(\text{pos} | \text{ca})$] 是 66%, 而阴性检查的正确率

$[P(\text{neg} | \text{benign})]$ 只有 54%。

在 1976 年 9 月，一封给《国家观察者》(*National Observer*) 编辑的信中，一个物理学家陈述了五个“观察和事实”(observations and facts) 来支持他们的观点：“常规乳房射线检查并不能在最大限度上代表最大多数的妇女的利益。”以下是他的观察的第一部分。

报告显示的乳房射线检查的准确性在 80% 到 90% 之间，这要依据一些因素，比如病人的年龄，她是否有纤维囊性的疾病，X 光射线检查的仪器，放射医师的经验，还有就是我们关于准确性的定义……即使我们认为准确率通常是 85% (我可以肯定并不是每一个放射专家都会在他的实践中得出这样的数字)，这就意味着有 15% 的妇女的 X 光检查以失败的解释而告终，或者更可能的是，她们的乳房射线检查不能够准确地反应她们的疾病。这就意味着，当她们确实患有癌症，当她们被告知她们的 X 光检查是正常的，她们就被给予了一种虚假的安全感。很难估计这将对这个群体造成的伤害，因为她们获得错误的信息比没有信息更加可怕。当被告知她的检查是正常的，并且一年之内不用再回来检查，这个患有乳腺癌的妇女可能会忽视掉她乳房中的肿块，而这个肿块很有可能让她马上去看医生的。

这个作者的推理中有几个错误。首先，乳房射线检查，不能用一个简单的数字来表示，假设作者认为真实的阳性和阴性比例都是 85%。

256

其次，这一概率是在用乳房射线检查对已知症状和体征进行鉴别诊断时观察出来的。这一病变通常较筛查更为晚期，而作者所指的正是筛查。更为合理的估计是，在筛查中，真阴性和真阳性比例分别是 60% 和 98%。

再次，即使用 85% 这个比例，我们发现在推理过程中还是有一些错误。考虑一下第二句话。不正确的解释有两种发生的途径：(1) 病人有癌症，但是得出了阴性的检查结果， $P(\text{ca}, \text{neg})$ ；(2) 她得到了阳性的检测结果，但是她没有癌症， $P(\text{no ca}, \text{pos})$ ^[2]。根据初级概率论，我们可以知道：

$$P(\text{ca, neg}) = P(\text{neg} | \text{ca}) P(\text{ca})$$

$P(\text{neg} | \text{ca})$ 是 $P(\text{pos} | \text{ca})$ 补数 (complement), 所以在这个例子里等于 0.15。我们不能确切地知道 $P(\text{ca})$, 但是在透视检查中我们可以比较确定地认为是少于 0.005 的, 这就是, 每 1 000 个妇女中有少于 5 个人没有症状但是却可以用射线检查出来乳腺癌。

这样:

$$P(\text{ca, neg}) \leq (0.15)(0.005) = 0.00075$$

并且,

$$\begin{aligned} P(\text{no ca, pos}) &= P(\text{pos} | \text{no ca}) P(\text{no ca}) \geq (0.15)(0.995) \\ &= 0.14925 \end{aligned}$$

错误解释的全部可能性, 是这两个数字即 $[P(\text{ca, neg}) + P(\text{no ca, pos})]$ 的和, 正如作者说的, 也就是 15%。但是这并不意味着“更可能的情况是, 她们的乳房射线检查, 不能检查出她们的疾病”。 $P(\text{ca, neg}) = 0.00075$ 比 $P(\text{no ca, pos}) = 0.14925$ 小多了。几乎是 200 倍的差距。

另一个问题, 是 85% 的准确性并不表示“有 15% 的妇女如果被告知 X 光检查结果正常就会被给予错误的安全感”。作者好像在试图估计 $P(\text{ca} | \text{neg})$ 。现在用贝叶斯的公式,

$$\begin{aligned} p(\text{neg} | \text{ca}) &= \frac{P(\text{neg} | \text{ca}) P(\text{ca})}{P(\text{neg} | \text{ca}) P(\text{ca}) + P(\text{neg} | \text{no ca}) P(\text{no ca})} \\ &= \frac{(0.15)(0.005)}{(0.15)(0.005) + (0.85)(0.995)} \\ &= 0.00089 \end{aligned}$$

也就是说, 如果 10 000 没有症状的妇女进行透视的话, 如果我们用作者对准确性的错误的估计来估计, 8 458 个妇女会被检查出来没有问题。作者认为会有 1 269 个妇女被给予了虚假的安全感。事实上, 只有 9 个。这个数字被错误地高估了 150 倍。

最后, 加上一句, “如果她们真的患有癌症”更加混淆了这句话的意思。一些短语“虚假的安全感”, “如果她们被告知她们的 X 光检查是正常的”, “如果她们已经患有癌症”, 可以象征性地被理解为 $P(\text{ca} | \text{neg, ca})$ 。这种概率是 1 而不是 0.15。

(2) $P(\text{ca})$ 的重要性

除了将两种准确性混淆起来，许多作者似乎也不明白，作为一种持续的回顾性准确性的检测，这种检查结果（即预测准确性）对于医师的意义，就在于它依赖于患者接受乳房射线检测之前患有癌症的风险。就算我们假定，在所有的研究中，真实的阳性检查和阴性检查的比例，都是恒定不变的，对于检测结果——乳房射线检测的结果是阳性（或阴性）的病人患有癌症——的合适解释，将依赖于患者被选择出来的人群中癌症的发病程度，依赖于患者患有癌症的预检查概率（pretest probability）。当人们比较临床诊断（妇女有一些乳房病变的症状）和对没有症状的妇女进行临床筛查时，这就变得非常重要。

这种重要性可以由一个例子来体现。假设一个临床医师给一个乳房查体异常的妇女做乳房射线检查。在一项研究中（Wolfe, 1964），这种妇女患有癌症的频率（frequency）大概有 8%。在这群人中做的一系列的乳房射线检查，真实的阳性结果有 92%，真实的阴性结果有 88%（Wolfe, 1964）。现在，让医师面对一个病人，他的这个病人在他的样本中是有代表性的 [即 $P(\text{ca}) = 8\%$]。假设他要求进行乳房射线检查，而且从放射科医师那里获得了阳性的诊断结果。他要求作组织活检的决定，就应该建立在这种新的可能性上，即这个患者可能有癌症。这种可能性通过计算有 40%（见表 18—2），是否一份阴性的报告就可以排除癌症，癌症的微小的可能性是不是小于 1%？这种估计的逻辑可以见表 18—2。

表 18—2 1 000 个体检有问题的妇女患有癌症和 X 光检查的结果

	患有癌症的妇女	没有癌症的妇女	总数
X 光检查			
结果为阳性的妇女	74	110	184
X 光检查			
结果为阴性的妇女	6	810	816
总数	80	920	1 000

注：一个真阳性率 0.92 [$P(\text{pos} | \text{ca})$] 表明，在 80 个患有癌症的妇女中，有 74 个人会有阳性的 X 光检查结果，6 个会有阴性的检查结果。在所有获得阳性检查结果的妇女里，74/184 有癌症，或者说 $P(\text{ca} | \text{pos}) = 74/184 = 40\%$ 。

资料来源：Wolfe 的数据（1964）。

现在，假设医师要求使用透视检查对一个没有症状并且查体阴性的妇

女进行筛查。对这种妇女，可用乳房射线检查出癌症的几率是 0.10%（例如，Shapiro et al., 1967）。为了我们这个例子的目的，让我们假设放射科医师的回顾性准确性没有改变——即在这些样本中让他的真实的阳性比例依然有 92%，并且真实的阴性（对良性病变的诊断）比例有 88%。^[3]文献提供的资料，都是关于那些有癌症的妇女或者是有良性疾病的妇女的回顾性准确性。在一个研究中，有 60% 的妇女，根本就没有病（Wolfe, 1965）。这样在这个例子里，

$$P(ca | pos) = [p(pos | ca) p(ca)] \div [p(pos | ca) p(ca) + p(pos | benign) p(benign) + p(pos | no disease) p(no disease)]$$

$P(benign)$ 、 $P(no disease)$ 和 $P(pos | no disease)$ ，在文献中所讨论的，不是很清楚。这是有益的，它引导我们怀疑在对这些问题的比较中，它们的重要性没有被充分地认识到。这个例子，我们可以用沃尔夫提供的数字（1965），并且假设 $P(no disease)$ 大约有 60%， $P(benign)$ 大概有 40%。我们要做出一个对乳房射线检查有力的假设，并且让 $P(pos | no disease)$ 等于 0%。

继续这个例子，假设放射医师认为这个没有症状的妇女的乳房射线检查的结果是阳性的。假设有一个阳性的诊断，这个病人患有癌症的可能性大概有 1/49，或者说是 2.0%（表 18—3）。在前一个例子里，有症状的妇女的 $P(ca | pos) = 40\%$ 。这样，根据被检查者的不同，一个有阳性检查结果的妇女患有癌症的可能性，会有 20 倍的差异。

表 18—3 Presence of cancer and results of X ray in
1 000 women hwo have no symptoms

	患有癌症的妇女	Women with benign lesions	没有癌症的妇女	总数
X 光检查				
结果为阳性的妇女	1	48	0	49
X 光检查				
结果为阴性的妇女	0	352	599	951
总数	1	400	599	1 000

注：A true-positive rate of 0.92 implies that the X ray will detect cancer in the one woman who has the disease. A true-negative rate of 0.88 for benign disease implies that of 400 women with beginn disease, 352 will have negative X rays, whereas in 48 the X ray will be positive. Thus, 49 women will have positive X rays, but only one has cancer, or $P(ca | pos) = 1/49 = 2\%$.

这就给医疗推理提出了一个很重要的问题——当我们试图估计一个病人的症状和体征，医师如何在对大多数的病人中运用关于一个疾病的比例的基本的信息？专家们似乎把这个问题搞糊涂了。一方面，医师们声称，259一个病情的相对普遍性，不应该影响对一个特定的病人患有此种疾病的评估。这个概念有几种相对极端的表述，比如，“这个患者就是一个病例”，“统计数字就是给死人准备的”。在讨论具体的问题的时候，这种观点通常被含蓄地表述为，“年轻一些的妇女，显然有更少数量的恶性疾病，但是它并不会对每个个体的诊断产生影响”（Wolfe, 1967, p. 138）。它也可以被清晰地表述，并且成为一个要被遵守的规定。举例来说，接下来的这句话，就出现在关于临床诊断的教科书里：“当一个病人向他的医生咨询他的未被诊断的疾病的时候，无论是医生还是病人，都不能知道这个病是不是罕见病，直到最终确诊。统计方法，只能被应用于几千人。个人或者有或者没有患上罕见的疾病；两种疾病的相对比例，跟医生最后做出诊断是完全没有关系的。”（DeGowin & DeGowin, 1969, p. 6）

另一方面，这些陈述和医生的行为并不一致，他们还是试图运用这些诊断信息，尽管不够完美。我们来看一下这些流传在医学院校里的格言：“当你听到马蹄声，想到的是马而不是斑马”，“普通的事情就会最普通地发生”，“遵守萨顿的法则：哪有钱去哪”，如此等等。医生似乎很认可关于一种疾病的前一种可能性，但是正式的概率理论的课程却没有被很好地理解。没有正规的理论，医师们倾向于在概率推理中总是犯同样的错误，这些错误已经在其他的环境中被观察到了（Kahneman & Tversky, 1973, 4; Lyon & Slovic, 1976）。

（3）含义：乳房射线检查和组织活检

这些问题有非常重要的实践意义。例如，在刚才引用的两个例子里，260两位作者就把其结论建立在了不正确的概率推理上。一个人错误地认为一个有乳腺增生的妇女，通过查体和 X 光检查都认为是良性的，仍然有 20% 的可能性会患有癌症，所以应该推荐进行组织活检。另一作者将他的根据透视做出的推荐，建立在一个粗略的错误估计的概率上，根据这个概率，妇女获得了虚假的安全感（举例来说，患有癌症但是通过乳房射线检查没

有查出来)。两个作者，通过正确的推理都可以得出同样的结论，但是他们没有。

(4) 诊断信息的价值

对于那些有症状的乳房疾病的妇女的乳房射线检查的价值，就在于它能够提供诊断信息，而这种信息会影响医师是否进行组织活检的决定。更具体地说，检查的结果，可能会改变医生对于一个患者患有癌症的可能性的估计。正如一个作者所说的：

乳房射线检查可以帮助医生区分恶性和良性的病变……一些病变，尤其是比较小的，可能会缺少让医生怀疑它值得做乳房组织活检的特征。正是这里的……乳房射线检查，可能会提供其他客观的证据。这样，在不确定的乳房病变中，乳房射线检查，就可以帮助医师决定是否要进行乳房的组织活检。(Clark & Robbins, 1965, p. 125)

对于任何有用的诊断检查而言，它必须可以提供信息来潜在地改变关于这个病人的治疗决定——要求对一些可能不进行组织活检的患者进行组织活检，并且，我们应该期望对一些已经准备进行组织活检的妇女排除这种可能性。这个概念，是在统计决策理论中被正式地发展出来的，并且在一些研究场景中被用来分析一些医疗问题（如 Lusted et al., 1977）。

许多医师承认，X光检查结果含有很多有用的信息可以用来帮助处理病人，但是准确地说，这些信息应该如何利用，并没有得到很好的说明。大多数作者给出的解释，都包含了很少的具体的方向。“乳房射线检查，不是被设计用来指导检查的步骤，而是在一些病例中可能提供一些更为精确的信息，这样，一些不必要的后遗症，因此可以被排除。”（Egan, 1972, p. 392）“乳房射线检查，对于外科医生来说，在他们诊断和治疗乳房病变的时候，是一种有价值的辅助手段。”（Lyons, 1975, p. 231）“乳房射线检查，可以用来帮助澄清被混淆了的可察觉的发现。”（Egan, 1969, p. 146）它“是一种支持和辅助的角色……”（Block & Reynolds, 1974, p. 589）这种支持的属性和程度，通常就有赖于医生的判断了。

(5) 乳房射线检查和组织活检：实践

乳房射线检查，在这些病例中的作用，似乎只是部分地被认识到了。²⁶⁷ 为了明白这一点，让我们检查一下临床调查者所预测的乳房射线检查，对于做出对有问题的乳房进行组织活检诊断的影响。前边我们引用的陈述意味着，X 射线的使用，将会帮助我们选择出那些需要组织活检的病人，另一些相同数量的陈述，则表明乳房射线检查事实上不能发挥这种功能。“任何可见的病变，都需要进行切除该组织活检的证实，无论 X 射线检查的结果是怎么样的。”（Lesnick, 1966, p. 2007）“尽管乳房射线检查具有权威性，但是它绝不是乳房组织活检的替代品。”（Egan, 1969, p. 148）“这个步骤绝不能降低组织活检的重要性。事实上，常规的乳房射线检查，将会更加确定组织活检的重要性，因为 X 射线检查关于一个恶性病变的证据，需要组织活检来进行证实……它绝不能降低组织活检的重要性……组织活检，在确定 X 射线的发现的作用上，和它在对一个物理症状进行检查时，同样都是必要的。”（Gershon-Cohen & Borden, 1964, pp. 2753, 2754）“非常明显，乳房射线检查，不是外科手术的替补品。”（McClow & Williams, 1973, p. 318）“让我们强调一下，乳房射线检查，并不是组织活检的替代品。”（McClow & Williams, 1973, p. 618）

一个关于如何使用乳房 X 射线来选择病人进行组织活检的精确的策略，出现在 1966 年的《外科档案》（*Archives of Surgery*）（vol. 93, pp. 853-856）中。对其用法说明的仔细审查，揭示出这个检查的潜力只有一半被发挥出来了。使用乳房射线检查的方案，来“决定对于一个病人的治疗和最终诊断”，涉及三种类型的病人：

A 类：“这些病人乳房中有‘肿块物’或者‘显性的病变’，这些都是需要进行手术治疗的，应该马上给病人进行组织活检。乳房射线检查，在这里严格说来是一种补充……它可能会发现那些隐性的肿瘤。”（p. 854）

B 类：“病人有一些和乳房有关的症状，但是没有散在的包块或‘显性病变’……在这种类型中，外科医生和内科医生，可以从乳房射线检查中获得最大的受益，因为形态是确定的。”在这里，乳房射线检查，会进行证实和鼓励，“如果临床的印象是良性的，它通常不能劝阻他进行组织活检的

决定”(p. 855)。

C类：这类病人没有症状和任何迹象，也没有任何临床指征 (indications) 要进行组织活检，乳房射线检查，只能增加进行组织活检的可能性。

这样，作者就勾画了一个计划，它使所选择的病人乳房射线检查的信息空洞化，而这些病人的组织活检，本来可以避免。只是加上的那一点信息，是为了暗示组织活检是必须的。而那些可能排除进行组织活检的信息，则被忽视了。

(6) 乳房射线检查和组织活检：潜能

262

为了明白概率推理中的问题如何影响到现实的医疗保健的实施，让我们来检验一下乳房射线检查在不同的诊断和选择病人进行组织活检中扮演的角色。正如前边所描述的，检查的目的，是为了改变决策者对于一个病人患有癌症的几率的主观评估。如果这种概率足够高（医生和病人一起决定的），组织活检就会被应用。我们把这种概率叫做组织活检阈限 (biopsy threshold)^[4]。现在来考虑一下检查对于两类病人治疗的重要性。

第一组包括这些病人，他们在病史和体格检查的基础上，临床医师认为他们明显有临床癌症。用弗里德曼 (Friedman) 等人的数据 (1966)，让这一组有癌症的先验概率 (频率) 为 90%。如果对这样的病人进行乳房射线检查，一个阳性的结果，会使癌症的概率 [$P(\text{ca} | \text{pos})$] 增加到 95%；而一个阴性的结果，仍然会让病人有 71% 的概率患有癌症。这种高概率，是下列这些陈述的动机：“如果临床医师的主观印象给出了怀疑癌症的足够的证据，她就会非常愿意进行组织活检。” (Clark et al., 1965, p. 133) 71% 的恶性几率，已经足够高到让任何病人要求进行组织活检了。

现在让我们来考虑第二组患者，他们有明显的包块，但并不是明显的癌症。在一项研究中，这种包块症状是癌症的概率，是 14% (Friedman et al., 1966)。在缺乏进一步信息的情况下，临床医师在这种病例中的策略，是对病变进行组织活检：“如果一个显性的肿块存在，它就应该被切除并且进行显微镜检。” (del Regato, 1970, p. 861) 把这作为方针，让我们假设病人组织活检阈限是 10%。这就是说，以医生的知识背景来看，他的病人患有癌症的概率要高于 10%，然后病人和医生就会一致要求进行组织活

检。^[5]假如用 10% 作为组织活检的阈限，我们可以知道乳房射线检查在对 1 000 个妇女的治疗中的重要性。没有这种检查，所有的病人，都应该进行组织活检，其中 860 个是徒然的。这 1 000 个有显性病变的病人们的可能的命运，在乳房射线检查被运用的时候，在图 18—1 中被表现出来。^[6]

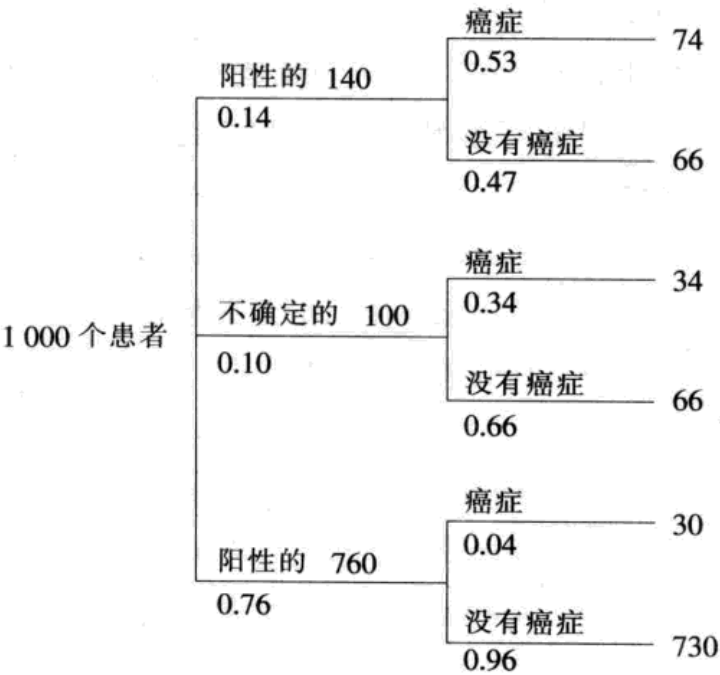


图 18—1 有显性病变的妇女患有癌症的可能性

有阳性乳房射线检查结果的病人，有 53% 的几率患有癌症，因为我们假定他们有组织活检的阈限是 10%，他们应该进行组织活检。因为那些乳房射线检查不确定的妇女有癌症的可能性是 34%，这些病人也需要进行乳房组织活检。那些有阴性的乳房射线检查的结果的病人，仍然有 4% 的可能性患有癌症，因为这个数字小于她们假定的阈限 10%，她们可能不希望进行组织活检，但是她们需要继续密切地随访。直接进行组织活检的数量，由 1 000 减少到 240。至少应该有 30 个组织活检要进行，因为在剩下的 760 个当中还有 30 个病人患有癌症。

这样，从乳房射线检查中，可以期望的收益（如将不必要的乳房组织活检的机会从 86% 减少到仅仅多于 13%），就可以和损失（如射线的损害和大约 75 美元）进行比较，并且期望的存活率，会有所下降（有 3% 的机会会使对恶性病变的诊断，可能会推迟一个月左右）。如果组织活检的阈限和

简单的概率理论在这里被运用的话，这组里很多乳房射线检查结果为阴性的病人，将会避免进行乳房的组织活检。在缺乏这类分析的情况下，“外科的共识意见，是将这组中的所有病人都进行组织活检，无论乳房射线检查的结果如何”（Friedman et al., 1966, p. 889）。

在这个例子中，乳房组织活检阈限的重要性，应该被进一步强调。如果医师和他的病人将阈限设在1%——也就是说，如果病人认为1/100患有癌症的几率就足以值得进行组织活检——那么乳房射线检查结果为阴性，就不能排除要进行组织活检的必要性（4%患有癌症的几率，将会超出这个阈限）。乳房射线检查，可能会给医师带来一些新信息，但是这些信息不一定对做出组织活检的决定有什么作用。在这种情况下，运用乳房射线检查的作用，必须要在一些其他的基础上来认识。

（7）对临床实践的反思（the practice revisited）

这种分析类型，帮助我们弄清楚乳房射线检查在对各种病变的鉴别诊断中的潜在的用途。它还帮助我们来评估以下的政策陈述：

① “乳房射线检查，为明显的乳房结节的临床诊断提供很少的信息，而根据乳房自身的特点，就可决定是否需要进行组织活检。”（摘自 *Archives of Surgery*, 1974, vol. 108, p. 589）在对那些有明显的组织包块的病人的研究中，组织活检，仅仅在临床基础上就可以被要求。而乳房射线检查的运用，将这个群体分成了几个子群体，它们有癌症的几率从53%到4%。在后一种子群体中，组织活检就可以避免，进行组织活检的数量，就会减少73%（从1 000/1 000到270/1 000）。

② “为了临床的目的，乳房射线检查，在它能提供治疗指导之前，必须提供近乎100%的准确性。”（摘自 *Archives of Surgery*, 1974, vol. 108, p. 589）在类似上面所讨论的第二组的人群中，通过乳房射线检查来选择病人进行组织活检，是非常合理的。回忆一下，在那个例子中，所运用的真阳性的比率是52%，并且一个更准确的检查，将会更有价值。

③ “乳房射线检查，不是组织活检的替代物。”（摘自 *Oncology*, 1969, vol. 23, p. 148）乳房射线检查和组织活检的目的，都是要为病人的状况提供信息。一些病人，在缺乏乳房射线检查的情况下，需要进行组织活检。

在这样的病人身上，一个阴性的乳房射线检查结果，可以排除组织活检，并且在这些例子中，乳房射线检查可以替代组织活检。

④ “在每一个进行组织活检的决定之前，都要进行乳房射线检查。”（摘自 *Oncology*, 1969, vol. 23, p. 146）考虑一下临床上明显的癌症科。所有的癌症的概率，都会高于人们的组织活检的阈限，无论乳房射线检查的结果是什么样的。在这个例子中，对这种政策最好的辩护理由，在于临床上明显是良性的机会（否则病人就必须进行乳房切除），并且这里有一个隐藏的、不明显的恶性肿瘤。这个复合事件的概率，就是这两个事件的概率的一个产物，它们是非常小的（ $1/5\ 000$ ）。

⑤ “对有临床良性病变的乳房延缓进行乳房组织活检，以期在乳房射线检查中获得良性结果，会耽误对乳腺癌的根治。”（摘自 *Surgery, Gynecology and Obstetrics*, 1972, vol. 134, p. 98）让我们认为临床检查良性的比例是 5%。在阴性的乳房检查之后，这样的病人患有癌症的可能性大约是 1%，在 100 个进行组织活检的人当中，99 个将是徒劳的。那么对组织活检的推迟，究竟是倒退还是前进呢？另一个问题是，如果采取了这些政策，所有的这些临床良性的通过临床癌症科检查，都需要进行乳房组织活检，而无论检查的结果如何。这似乎与作者的陈述相违背，“当被正确使用的时候，乳房射线检查，对于医生要检查乳腺癌来说是一个出色的助手。”（摘自 *Surgery, Gynecology and Obstetrics*, 1972, vol. 134, p. 98）

⑥ “在处理乳房的显性病变的时候，乳房射线检查绝对不可以替代组织活检，而且也不可以改变外科在检查这些疾病的时候的基本的方法，比如‘肿块就是肿块’必须进行组织活检，无论是通过切口还是穿刺。”（选自 *Archives of Surgery*, 1966, vol. 93, p. 854）那些有显性的组织病变并且其组织活检阈限高于 5% 的病人，不会同意这种说法。

⑦ “这个谬误的出现，可能在于在一些值得怀疑的例子中的乳房射线检查。在进行了乳房检查和确诊之后，很关键的就是要决定是否进行乳房组织活检，如果不能进行乳房 X 射线检查的话。如果可以进行乳房组织活检，那么就做。如果你确信没有手术指征和体格检查，那么就进行乳房射线检查。只要一个人对自己说，尤其当他对他的病人说‘我并不是很确定

这个——让我们进行 X 光检查吧，’他就下意识地依靠乳房射线检查的阴性结果了，当他只应该依靠阳性检查结果的时候。这是一个心理陷阱，我们都有可能会陷进去，而且比很多通过乳房射线检查获得的假阳性的检查结果更加严重。”（Rhoads, 1969, p. 1182）通过这种策略，是不可能避免任何一个简单的组织活检的。这是很遗憾的，因为正如前边的作者自己指出的那样，“很少有什么疾病需要进行如此多的外科检查，这些外科检查，通常是可以通过一些更好的诊断方法来避免而不是通过乳房本身”。

我们现在可以来欣赏一下下边这个故事，它出现在旧金山《年鉴》中（Kushner, 1976）。一个女记者刚刚发现她的乳房里有肿块，并且描述了她向医生进行的咨询。

“我希望你可以进行干板乳房射线检查（xeromammogram）。这是一种进行乳房射线拍片检查的新方法。”

“它可以很准确吗？”

他耸耸肩，“很有可能就和其他的片子一样吧。你知道。”他警告说，“即使结果是阴性的——也就是说肿块不是恶性的——唯一可以确定的方法，就是将肿块切除并且将它放在显微镜下检查。”

这个妇女然后和她的丈夫讨论这个问题。

“医生怎么说？”

“他希望我进行干板乳房射线检查。然后，不论结果如何都要将肿块切除。”

“既然如此，为什么还要进行干板乳房射线检查呢？”

“那是必须做的吧，我猜测。我们的医生说它的正确率可以达到 85%……所以，首先我要安排进行一个热成像。如果它是阳性的或者是阴性的，并且如果它和乳房射线检查中 X 干板图片结果是一致的，那么按照统计学来说，在诊断 95% 的情况下是可靠的。”

总结一下，我们可以合理地质问，如果乳房射线检查的目的是为了帮助医生区分良性的还是恶性的乳腺疾病，并因此使一些病人免于一些更为

广泛的并且是外伤性的程序如组织活检，那么我们就应该让这种检查发挥其功能。如果另一方面，医生总是坚持预先做出的进行组织活检的决定，而不考虑乳房射线检查的结果，那么我们就不能说乳房射线检查的目的，是为了帮助医生来区分良性的和恶性的疾病，因为毫无疑问这样的区分，可以通过组织活检来完成。最后，如果这个检查的目的，是为了在乳房的其他部位发现隐藏的和临床上没有发现的癌症（远离那些明显的需要进行组织活检的肿块），我们可以清楚地意识到这样事情的几率是非常小的，并且这个检查的效果就等于普查（筛查）。 267

我的目的，不在于讨论具体的乳房射线检查和组织活检的政策——这样做需要更多的资料和更好的对病人的价值观的评估。其实，我们还没有发展出一种比较正规的概率推理的方法来处理这类问题。临床判断可能是错误的，目前的临床政策可能是不连贯的或者是不正确的。

三、讨论

这里所呈现的例证，是为了说明医疗决策的复杂性，并且展示了临床医师是如何来应对这些问题的一个方面——对概率的操纵。我们研究的例子，是一个相对简单的例子，它用简单的检查诊断来将病变分成两种类型，良性的或者是恶性的。有关这个问题的数据库，相对来讲，是比较完备的。多年来，检查的准确性及其诊断的价值，在很多机构中被研究和分析过。正如一个研究者所说的，“我认为没有任何医疗程序像乳房射线检查一样被如此多地检验再检验”（Egan, 1971, p. 1555）。

这篇文章中所讨论的概率工具，已经有几个世纪的历史了。在过去的 20 年里，它们被更多地应用于医学问题中（比如，Lusted, 1968），并且用系统的方法来处理不确定性，在医学院的课程中，在医疗杂志中和研究生教育的项目中，也不断增长。然而目前，这些技术的应用，还是非常零散的，并且还没有能够影响大多数从业者的思维。正如这个案例所证明的，医疗问题是非常复杂的，并且形式的概率推理的威力，提供了极大地提高医疗保健的质量和效率的可能性。

注释

[1] 在这里和接下来的引用中，都会有斜体字。

[2] $P(A, B)$ 是指事件 A 和 B 同时发生的联合概率。

[3] 这并不是一个很好的假设，因为准确性随着被检查者的改变而改变。举例来说，当我们以没有症状的妇女作为样本的时候，真实阳性的比例就低得多，因为癌症似乎小得多而且不容易发现。这个假设只是用来说明 $P(ca)$ 的重要性。

[4] 任何人，如果想要确认组织活检阈限的存在，就可以按照下边这样来推理。如果患有癌症的可能性是 30 万亿分之一，我们是否会同意没有人愿意进行组织活检？如果患有癌症的几率是 98/100，我们是否会同意事实上每个人都希望确诊并且进行治疗？如果是这样，在 30 万亿分之一和 98/100 之间的某个地方，每个人都有组织活检阈限。当然，如果一个妇女在肯定患有癌症的时候，拒绝组织活检和治疗，那么她就没有组织活检阈限。

[5] 组织活检阈限，是一个非常奇妙也非常重要的数字。转变它，将会对如下因素产生极大的杠杆作用：如，妇女进行组织活检的数量、无效的组织活检的频率、治疗需要的花费，还有病人的预后。因为风险的转移和为他人进行决策的事实，医生们通常将阈限设置得非常低。这个陈述“如果有一线希望这个病变是恶性的，就要进行组织活检”是非常典型的。“如果医师并不完全肯定病变是良性的，那么马上就要进行组织活检”（Allen, 1965, p. 640）。有证据表明，妇女的阈限，通常比医生所设定的要高——虽然通常这里有很大的变化。

比如，我们可以检查一个非常大型的临床试验中的数据，在此，乳房射线检查和乳房的体格检查，是用来透视没有症状的妇女是否患有乳腺癌（Shapiro et al., 1971）。根据乳房病变是如何被检查出来的（比如，通过何种检查或者各种检查的组合），妇女所患的乳腺疾病是乳腺癌的概率，在 15%~54% 之间浮动。在阳性检查的结果基础上，医生建议 545 个阴性检查结果的妇女进行组织活检。尽管这个群体中患有乳腺癌的可能性，只有 15%，有 31% 的妇女拒绝了进行组织活检的建议。妇女乳房物理检查为阴性但是乳房射线检查结果为阳性的几率是 20%，但是在这个群体中有 29% 的妇女拒绝了进行乳房组织活检。在两项检查都是阳性的妇女那里，患有癌症的可能性有 54%，但仅有 5% 的妇女，在被建议的情况下，还是拒绝进行乳房组织活检。这样，根据这些粗略的信息，似乎有 31% 的妇女的乳房组织活检阈值超过了 15%，29% 的妇女的阈值高于 20%，有 5% 的妇女的阈值高于 54%。

[6] 为了大致了解乳房射线检查在这些病人中的重要性（和有其他的表现和症状的病人），我们需要更多的信息，但是这些信息在文献中是找不到的。幸运的是，在一个研究中（Friedman et al., 1966），患有癌症的几率和乳房射线检查的回顾性准确性，被分别对三个群体进行陈述——那些有明显的癌症状的、有显性的组织包块的，和其他有乳房疾病的症状和体征的。所发表的数据是不完全的，并且关于在恶性和良性的病变中不确定的 X 光诊断的数据没有包括进来。我们使用了弗里德曼的研究中的可用的数据，并且在这个例子里，有下边的一些假设：（1）那些没有进行组织活检的病变，事实上是良性的；（2）没有进行组织活检的病变，被认为是阴性的；（3）一半没有被认为是阴性的良性病变，被认为是阳性的（另一半被认为是不能确诊的）；（4）一半没有被确诊为阳性的恶性的病变，被认为是阴性的。前两个假设，是对于乳房射线检查的准确性的最为乐观的解释。第三个和第四个假设非常重要，并且因为假阳性（或假阴性）比例趋近于 0，阳性或阴性的 X 光的检验报告，在确定（或排除）癌症的时候，其作用就被提高了。同样的，因为假阳性和假阴性的比例提高了，检验就失去了预测的效果了。通过乳房本身的症状比通过病人的解说来解释弗里德曼的数据更难。但是，这个报告中有很多信息，并且把它们用到这个例子里，因为读者明白这些是一个例证，并不是形式的分析。对这些问题的形式分析需要更多的数据。在对第二组病人的估计中，所用数据的准确性计算如下： $P(\text{pos} | \text{ca}) = 0.52$, $P(\text{uncertain} | \text{ca}) = 0.24$, $P(\text{neg} | \text{ca}) = 0.24$, $P(\text{pos} | \text{benign}) = 0.75$, $P(\text{uncertain} | \text{benign}) = 0.75$, $P(\text{neg} | \text{benign}) = 0.85$ 。

本论文的写作，受凯瑟家庭基金会（The Henry J. Kaiser Family Foundation）的项目资助。



第 19 章

从经验中学习与决策中的次优规则

爱恩荷恩 (Hillel J. Einhorn)

268

目前关于决策的研究，显然已经从使用规范模型（及其修正）描述选择的过程转向了从认知心理学和信息加工理论的一般框架中发展出来的启发式过程的强调（Payne, 1980; Russo, 1977; Simon, 1978; Slovic et al., 1977; Tversky & Kahneman 1974, 1, 1980）。从强调人们表现得如何到强调人们如何表现，问题的转移无疑很重要（例如，Hogarth, 1975）。但是，对于同时研究这两个问题的益处，没有比在启发式规则和策略的研究中更加明显的了。之所以如此，是因为启发式规则与规范规则之间的比较，使人们可以检验实际行为和理想行为之间的差异，而这又提出了一个问题，就是为什么这种差异会存在。在本章中，我们关注的是人们如何从经验中学习这两种规则。对于从经验中学习的关注，有许多问题，这些问题还没有得到充分的解决，例如，启发式是在什么条件下习得的？在经验中它们怎样被检验和保存？在什

么条件下, 我们不能了解到偏差以及偏差的使用所导致的错误?

通过考虑以下几点, 我们可以了解学习对于理解启发式和选择行为的重要性:

(1) 人们预测一种特定的规则什么时候被运用的能力, 通常不够完备 (Wallsten, 1980)。但是, 关注一种规则怎样以及在什么条件下被习得, 将会提高人们预测这种规则什么时候可能被使用的能力。例如, 如果一种规则是在几乎没有时间做出选择的情境中被习得的, 那么通过了解任务涉及的时间压力, 就可以提高对这种规则的使用的预测。 269

(2) 伴随 (1) 的一种情况是, 通过设计一些任务包含或者模拟初始学习条件的情境, 我们将可以影响人们的判断和决策。这一点对于帮助以及操纵人们都有很大意义 (Fischhoff et al., 1978; 1980)。

(3) 关于学习的考虑, 注意力主要集中在环境变量和任务结构上。因此, 为了理解判断和决策行为, 我们应该考虑诸如强化的数量和计划、试验的次数 (等于经验的量) 等变量 (比较 Estes, 1976)。尽管任务对于理解行为的重要性已经被不断地强调 (Brunswik, 1943; Castellan, 1977; Cronbach, 1975; Dewes, 1975b; W. Edwards, 1971; Einhorn & Hogarth, 1978; Simon & Newell, 1971), 但是, 心理学家似乎和其他人一样, 容易犯罗斯 (1977) 所说的基本归因错误 (在归因中低估环境因素的作用)。

(4) 启发式理解中的一个主要变量是结果反馈 (outcome feedback)。因为结果反馈是评估我们的决策和判断规则的质量的主要信息来源, 因此关于任务变量如何作用于结果以及如何影响结果在记忆中的编码和储存方式, 对于解释启发式如何被习得和使用变得非常关键。

(5) 对于权衡关于选择行为的心理学解释和经济学解释的相对优点而言, 学习领域是个关键。一些经济学家认为, 尽管人们不会在所有的时候都“理性地”行动, 但是人们将会通过与环境的互动学习最优规则。他们提出了一堆含混不清的关于平等、效率和进化概念的断言支持这种观点。因此, 研究人们如何从经验中学习, 对于澄清心理学和经济学的选择理论的相对优点, 非常重要。

一、从经验中学习：如何学习？

显然，决策是行为导向的。为了满足基本的需求和欲望，人们必须选择采取哪一种行为。因此，对于任何有机体，了解哪一种程度的行为将会导致想要的或者不想要的结果，都是重要的。这意味着，从经验中学习，大部分必须包括行为—结果链的学习。而且，因为行为和结果在时间上是连续的，所以人们倾向于认为它们之间的联系，代表了原因和结果之间的关系（Michotte, 1963）。因此，要发现因果关系这种强烈倾向，就可以看作是采取行动满足基本需要这种必要性的结果。此外，正如卡尼曼和特沃斯基（1979b）指出的，学习因果关系，以及按照因果“图式”组织事件，使得人们能够给予他们的经验前后一致的解释。最后，行为—结果链的学习，对于理解人们如何了解他们的口味和效用非常重要。例如，假设一个孩子选择吃某一种蔬菜，尝到了一种令人不愉快的味道，因此学会将一种负效用与这种食物联系起来。注意，人们通常是根据体验的结果和习得的效用来进行选择。因此，行为—结果链的学习与效用学习是紧密联系在一起的。

尽管我们是通过采取行动从经验中学习的，但是最初人们是如何学习应该采取哪一种行为的呢？毫无疑问，大部分最初的习得，是通过试错法获得的。也就是说，人们随机地选择，然后观察结果（对比 Campbell, 1960）。试错法让位于发展策略或者规则的过程，还不是很清楚（对比 Siegler, 1979）。但是，人们可以推测，来自强化试错法的学习和归纳（刺激和反应）扮演了重要角色（Staddon & Simmelhag, 1971）。在任何情况下，我们发展的规则，似乎都与学习某种行为导致何种结果直接相关。正如上文所描述的，从经验中学习，在本质上是归纳的，也就是说，人们经历一些特殊情况或者个案，然后发展出启发式，以提供某种一般的解决方法。就启发式而言，从经验中学习的归纳本质，有几点意义：

1. 规则的特殊性（specificity of rules）

如果学习是经由归纳特殊个案发生的，那么启发式规则就应该非常依赖于情境。现在许多证据都表明事实确实如此（Grether & Plott, 1979; Lichtenstein & Slovic, 1971; Simon & Hayes, 1976; Tversky & Kahneman,

1980)。问题被表达或者展示的方式,或者对于所有人某个回答被提问的方式,似乎都对信息的处理和回答产生的方式产生了重要影响。西蒙等人 (Simon & Hayes, 1976) 关于“问题同构体”(problem isomorphs) 的研究,给出了一个这种特殊性的生动的例子。他们指出,结构上相同的问题(例如,可以用同样原则解决的问题)使用不同的表面措辞,结果完全改变了人们在记忆中再现这个问题的方式,因此也改变了解决问题的方法。这个结果的一个重要意义在于,为了使启发式模型更具有预测性,人们必须全力关注对任务的描述,而不是实验者所理解的问题结构。关于这一现象在行为预测中的重要性,下面的观察提供了一个非常合适的例子,通过观察人们把减税看作是一种收益还是较小的损失来预测人们的行为 (Kahneman & Tversky, 1979b)。

2. 规则的普遍性

如果启发式是通过归纳习得的规则,那么我们就必须根据相似性分类任务,否则就会有多少种情境,就有多少种规则。既然后一种情况不能接受,因此对于任务而言,启发式必须具有一些普遍性。但是,这一结论,违背了我们上文所说的规则的情境依赖性和特殊性。如果我们考虑一下规则可能适用的任务范围,这个悖论就可以解决了。例如,考虑一下这个规则,“不要在肉食店里叫鱼。”虽然这样一种规则对于某种类型的餐馆普遍适用,但是与这个规则“通过它们的相似程度判断事件 B 产生于程序 A 的概率”相比,毫无疑问,它是更加特殊的 (Tversky & Kahneman, 1974, 1)。显然,后一种启发式,具有更高程度的普遍性。事实上,很有可能,诸如代表性、便利性、初始化和调节这样的启发式,都是“元启发式”,也就是说,它们是一些关于如何产生规则的规则。因此,当遇到以前没有遇到过的问题(比如判断事件的概率),或者一些其特殊性使得它们看上去比较新奇的问题的时候,元启发式指导人们形成具体的规则解决问题。元启发式的观点,使得人们可以保留任何规则都必然包含的普遍性,它还同时考虑到情境、措辞、回答模式等的重要作用。为了详述这一点,可参考斯洛维奇等人 (1976; 也参见本书第 33 章) 的研究。在研究中,他们要求人们判断因为意外原因死亡的相对概率。例如,哪一种可能性更高:被闪电

271

击死还是死于肺气肿？当面对这种问题时，有很多方法可以得到答案。可能被使用的一种规则是：“想一想所有我认识的死于这两种原因的人，然后找出哪一种原因导致了更多的死亡。”就我自己而言，我会选择肺气肿（它确实有更高的可能性，尽管大多数人选择了死于闪电）。但是，我本来可以同样容易地形成另一种规则，而它会导致相反的答案；例如，“想一想所有我听说的被闪电击死和死于肺气肿的例子（从报纸、电视等）。”如果这是我的规则，那么我会选择死于闪电的可能性更高。注意，在这两个例子里，我都使用了便利性启发式。很明显，一个问题被表述的方式，可能导致特定的规则，而它又导致了不同的结果。但是这些特殊的规则，可以被归类到一个更为普遍的策略，或者是元启发式之下（也参见 Einhorn et al., 1979）。

3. 启发式的强度

272 如果启发式是通过归纳习得的，那么学习的发生，就要经过许多次的试验和强化。正如将要讨论的，由于反馈发生的方式以及我们经由经验检验规则所使用的方法，因此即使是对于错误的规则，也可能出现积极的强化（Wason, 1960）。此外，除了我们所经历的强化的数量很大以外，强化的范围和强度可能也很大。例如，遵循某种选择股票的规则获得了一笔数量可观的收入，这应该会产生非常大的强化效应。因此，与关于人类学习的实验室研究不同，在那里，伦理的因素限制了大量的积极和消极强化，而我们自己的经验则没有这些限制。

二、从经验中学习：学得如何？

我们从经验中学习得如何的问题，把注意力集中在了比较启发式规则和最优规则上。因此，必须要问的是，最优规则是如何被习得的，在我们自己的经验中应用它们会有什么意义？最优规则，是通过演绎的方法习得的，例如贝叶斯定理、最优化等。事实上，许多可以被称作是正规的学习都具有演绎的性质，也就是说，我们学习的是科学规律、逻辑原理、数学和统计学规则等。这些规则，从本质上来说是抽象的，并且独立于情境。而且，当情境可能影响规则的形式时，人们经常被告知的是这个规则有效，

“其他事情相同”。当然，在我们的经验中，其他事情几乎没有相同过，这就使得经由归纳学习最优选择非常困难（最优规则的最初发现者或者发明者，克服了这些困难；但是，这些困难将他们和我们区分开来）。

就人们在特定的情境中应用最优方法面临的困难而言，演绎规则的抽象性质，具有重要的意义。这种困难，主要集中在要发现隐藏在繁多的细节中的任务结构。因此，当人们面对一个特定问题的时候，并且这个问题细节烦琐，而且许多细节可能是无关紧要或者千篇一律的，那么，对于细节的关注，就很可能把注意力从问题的一般结构上转移开来。事实上，正是通过演绎习得的最优规则的抽象性质，阻碍了人们从记忆中检索最优规则（比较 Nisbett 等人，1976；并参考本书第 7 章）。因此，抽象规则，可能并不是很容易获得的。但是，这是在回避问题，因为重要的是了解为什么不能获得这些规则。

考虑一下行为—结果组合在记忆中可能的组织和存储方式，尤其要考虑一下这种信息更有可能根据内容还是任务结构被组织和存储。看起来，似乎根据主体而非结构组织行为—结果组合更加容易，也更加“自然”；例如，在学校的经历、与父母和异性的经验等，而不是贝叶斯式问题，选择情境、最优化问题等。人们很难注意到这个事实，即尽管结构相同，但是问题的内容可能不同（Einhorn et al., 1979; Kahneman & Tversky, 1979b; Simon & Hayes, 1976）。因此，我认为，很可能大多数人不是根据任务的结构组织他们的经验的。但是，这并不是说人们不可能被训练做到这一点。事实上，大部分专业训练就是这样的；例如，训练人们学会识别问题，把它们归类到某一类问题中去，这类问题具有特殊的结构和（有些时候）已知的解决方法。因而，通过充分的训练人们“可以使用”最优规则。当然，这种太容易获得的规则也有危害。也就是说，因为某种结构的问题存在解决的方法，就把一些问题强行归类到这种结构中去，而这些结构对于这些问题可能是不合适的。当面对一个问题的时候，专业人士会根据他们经由训练发现的结构看待这个问题，这是不言而喻的。因此，尽管专业训练并不包含某种结构，但是这种训练通常会被限制在一个比较狭小的内容领域内。

278

对于组织和教授关于世界的公共知识的方法的思考，提供了进一步的证据，证据表明，需要根据内容而不是结构分类问题。例如，分门别类的教育、专业化的训练、图书馆和百科全书的信息分类等，都是根据内容而非结构组织信息的。尽管以这种方式组织信息有很大的优越性，但是同样也要付出代价。把在一种内容领域中发展出来的最优规则应用到结构相似的其他内容领域的问题上所遇到的困难，就是一种代价。但是，在个体学习者这个层次上，我们将要考虑的其他困难可能代价更为高昂。

尽管任务的结构很难识别，但是结果却是显而易见的；它们更明显，容易获得，并且通常很明确。对于理解启发式在经验中如何被维持而言，经由结果反馈考虑强化非常重要。而且，如果结果在很大程度上是任务结构的函数，并且决策者缺乏关于这种结构的知识，那么通过积极的反馈，那些不相关的甚至是糟糕的规则仍然有可能被强化（例如，动物学习中的“迷信行为”；见 Staddon & Simmelhag, 1971）。

现在我们给出两个例子，在这两个例子中，不合乎标准的启发式可能导致良好的结果，并且人们可能没有意识到这种规则的低质量。考虑一个在超级市场购物的情境，在一堆果汁的前边，有如下的价格和整体质量标签（根据 Tversky, 1969 的资料改编）：

品牌	价格	质量
X	60 美分	高
Y	55 美分	中等
Z	50 美分	低

274 假设我使用下述规则在这三个牌子中进行选择：如果价格的差异在 5 美分之内，那么就选择质量好的那个牌子；如果价格差异大于 5 美分，那么就根据价格选择。这样一种简单的规则（它属于词典编纂学上的半序）导致： $X > Y$ ， $Y > Z$ ，但是 $Z > X$ 。因此，这个规则导致了选择的不可传递，显然这不符合逻辑。但是注意，当我选择了 X 放弃 Y 之后，我可能从余下的集合中排除 Y，然后比较 X 和 Z。因此，最后我选择了 Z，品尝之后我发现它十分令人满意。于是我就会庆幸自己是一个多么优秀的购买者——我节省了钱，并且买到不错的产品。在这里要注意的重要一点就是，我没有

比较 Y 和 Z，因此我仍然没有意识到我的规则会导致选择的不可传递。我现在意识到的全部就是我做了一个明智的选择，麻烦既少，又不费力，而且结果令人满意。因而，积极的结果反馈，强化了一个不符合标准的规则，并且没有意识到有可能出错。

第二个例子是或然性的（比较 Schum, 1980）。设想你是一名陆军上将，在一个政治环境比较紧张的地方，你担心你的敌人可能会入侵你的国家。此外，根据以往的经验可知，一旦敌人的部队在边境集结，入侵的概率是 0.75。但是，你不能直接得到关于敌军的情报，你必须依靠你的情报部门提供关于敌军行动的情报。结果证明，每一次你的情报部门报告说敌军正在集结，敌军就确实在那里。假设现在你从你的情报部门那里收到一份报告说敌军正在边境。那么入侵的概率是多少？更正式的表达是，让

H = 被入侵的假设

D = 敌军在边境集结

D^* = 关于敌军在边境集结的报告

问题中给出了 $p(H|D) = 0.75$, $p(D|D^*) = 1.0$ ，让你回答 $p(H|D^*)$ 是多少。如果你和大多数人一样，那么你可能会回答 0.75。但是，所提供的信息，对于以一种通常正确的方式回答这个问题来说，是不充分的。事实上，在上述问题中，很可能 $p(H|D^*) = 0$ ！因为大多数人认为这令人难以相信，考虑图 19—1，它使用维恩图（Venn diagram）阐明了这个问题。

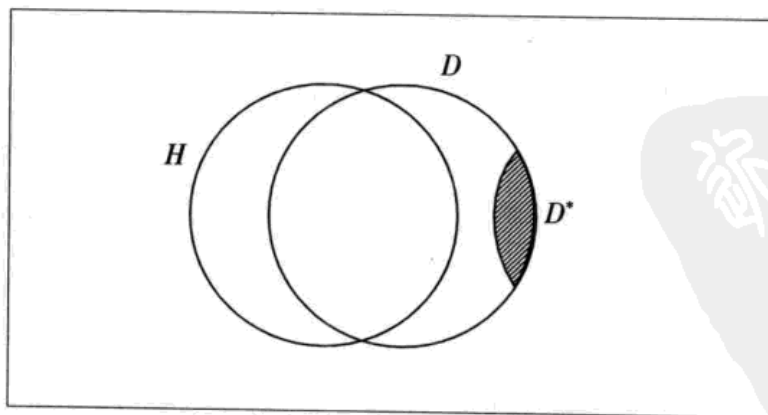


图 19—1 表示假设 (H)、数据 (D) 和报告数据 (D^*) 之间关系的维恩图

注意， H 和 D^* 的交集为空，因此条件概率 $p(H|D^*)$ 等于零。之所以人们认为这个结果非常令人吃惊，是因为他们犯了一个这种形式的逻辑错误：如果 $D^* \rightarrow D$ ，那么 $D \rightarrow D^*$ 。尽管只要 D^* 存在 D 就会发生，但是反过来却不一定成立。事实上，看待这个问题的一种直觉方法，就是认为敌人特别狡猾，结果使得只有当他们不准备入侵时，你的情报部门才能发现他们的部队。但是，当他们计划入侵并且在边境集结的时候，他们隐藏起来，结果你的情报部门没有报告发现他们。

275 这个例子说明了在具体任务中应用最优规则（在这个例子中是形式逻辑的规则）的困难。虽然当这种逻辑错误以一种可识别的方式呈现出来时，几乎没有人会犯错误。但是，这个例子的重要性就在于指出问题的具体细节如何掩盖了它的真实结构，以致轻易地就违背了最优规则（对比 Tversky & Kahneman, 1980）。关于这个例子，我们还可以得出第二个结论。假设这位将军犯了这种逻辑错误，估计战争的可能性为 0.75。然后他把他的部队派到边境上去，由此导致了敌军的入侵。因此，这位将军的错误推理被结果反馈强化了：“毕竟，”他可能会说，“那些畜生确实是要入侵我们，这和我想的一样。”

这两个例子阐明了本章的基本观点：如果没有关于任务结构的知识，对于糟糕的启发式的修正而言，结果反馈可能是无关紧要的，甚至是有害的。并且，在缺乏对任务的了解的情况下，积极的结果反馈，可能会使我们意识不到我们的规则的拙劣，因为几乎没有什么动机鼓励我们质疑我们是怎样获得成功的。这种结果反馈对于启发式和策略不起修正作用的情况，被称作结果无关的学习结构（outcome-irrelevant learning structure, OILS）。这种结构可能比我们想得还要更加常见。在具体检验一个这种结构之前，考虑一下 OILS 框架内的概率判断，因为大多数关于启发式的研究，都直接关注这种类型的判断。假设你认为某个事件的概率是 0.70。让我们假设这个事件并没有发生。那么，关于你使用的产生这个判断的规则的质量，这个结果告诉你什么？人们可能会认为任何单个的结果对于估计概率判断的“优度”（goodness）（即校准的程度）都是无关紧要的。因此，在一种重要意义上，直接的结果信息对于糟糕的启发式的修正，是无关紧要的。

只有当人们判断指定概率的事件时，记录下了相对结果频次的“总成绩”，人们才能从结果中获得有用的反馈。但是，就做出良好校准的判断而言，这很可能是一个必要条件而非充分条件。首先，在决定判断是否被校准之前，人们必须记录下多长时间段的总成绩？此外，为了说明判断是精确的（在被良好校准的意义上），多近才是“足够接近”？注意，这整个的结果评估模式就包括了强化，它因为时间过长而被耽搁了。因而，我们还不清楚这种反馈是否会有很大的自我修正作用。其次，为了了解概率估计规则的优度，我们的总成绩必须不仅包括人们的估计以及产生的结果，还必须包括人们推导这些估计的规则。例如，如果 100 次估计我给出的都是 0.70，并且记录下结果，那么 100 次中有 53 次事件发生了，关于我使用的规则的质量，这一信息能够告诉我些什么呢？因为很可能在 100 种不同的情境中，使用了许多不同的规则估计概率，因此结果信息并不是很重要，而结果反馈也不是非常有用，除非人们知道它使用的规则，并且记录下了规则的使用情况（比较 Nisbett & Wilson, 1977，关于我们是否意识到了我们自己的认知过程）。276

我的意思并不是说，人们不可能学会做出良好校准的概率判断。如果人们在同一种情境下做出许多概率判断，如天气预报员和赛马优胜预测者所做的，并且可以立即得到结果反馈，那么这种情况下的结果，就不会是无关紧要的了，而反馈也可能是自我修正的。但是，对于我们中的大多数人来说，这种情况看来只是个例外，而非常态。

尽管概率判断通常出现的情况是结果无关的学习结构（OILS），但是非概率判断的情况又是怎样的呢？可以肯定，如果人们预测某件事情，那么他可以通过检查，确定这个预测是正确还是错误。因此，对于提供自我修正的反馈而言，结果似乎应该是重要的。本章余下的部分，就在一种普通的流行的任务结构情境下讨论这个问题，尽管这种任务的具体内容可能相当不同。

三、选择任务^[1]

既然对于提供自我修正的反馈来说，结果信息似乎是重要的，并且是

可以使用的，那么现在我们就考察一种非常普通的涉及非判断概率的任务。将要考虑的任务，是做出判断从而在给出的行为选项中进行选择。例如，假设一种情境有两种可能行为 A 和 B 。用 x 代表总体评价性判断，它本身可能是各种类型和量的信息的函数。此外，令 x_c 为临界值，使得

$$\begin{aligned} &\text{如果 } x \geq x_c, \text{ 采取行为 } A; \\ &\text{如果 } x < x_c, \text{ 采取行为 } B; \end{aligned} \tag{1}$$

277 尽管非常简单，方程 (1) 适用于许多判断（决策）情境，例如：介绍工作、推销商品、入学选拔、贷款和信用的发放、安排补习计划、社会计划的筛选、期刊文章的验收、授予奖励等。在这些情况中，因为被选择的行为不可能面向所有人，因此对于“应得的”等级的判断，通常决定了采取哪一种行为。

为了比较判断和标准，假定存在一种标准作为评估判断的精确性的基础，用 y 表示。虽然发现以及发展适当的标准面临巨大的实践困难，但是这里关注的焦点是理论上的：一种标准的概念对于这种分析是必须的。为了与判断的构造保持一致，进一步假定标准有一个临界值 (y_c) 使得 $y \geq y_c$ 和 $y < y_c$ 作为评估判断的结果的基础。因此，就对判断的了解而言，结果在记忆中的再现通常是采取类别的形式，即成功和失败（比较 Estes, 1976）。

278 注意这种任务的结构是判断（预测）导致差别行为，然后结果用作反馈决定预测的精确性，这一点非常重要。考虑 y 对于 x 的回归以及 x_c 和 y_c 的交叉产生的 4 个象限，我们可以看出这种结构的形式，如图 19—2 所示。用正向和负向命中表示正确的预测，用错误肯定 ($y < y_c \mid x \geq x_c$) 和错误否定 ($y \geq y_c \mid x < x_c$) 表示两种错误类型。为了估计 x 和 y 之间的关系（即 x 和 y 之间的相关关系 ρ_{xy} ），我们必须得到每一个判断—结果组合的资料。首先假设随着时间（即按时间顺序），这种资料变得可以得到，然后考虑在这种情况下获得 x 和 y 之间的关系要涉及的实验证据。关于根据 2×2 表格的信息判断 x 和 y 之间列联关系的能力的研究（Jenkins & Ward, 1965; Smedslund, 1963; 1966; Ward & Jenkins, 1965）表明，人们通常是根据正向命中的（使用图 19—2 的术语）频次来判断相关关系的强度，而忽略其

他 3 个单元的信息。这些结果非常重要，因为它们指出，甚至是在所有这些相关的结果信息都可以得到的情况下，人们也不会使用它们。这意味着在那些具有结果相关的学习结构的实验室研究中，人们把它们转化成了与结果无关的学习结构。怎样才能解释这种现象呢？

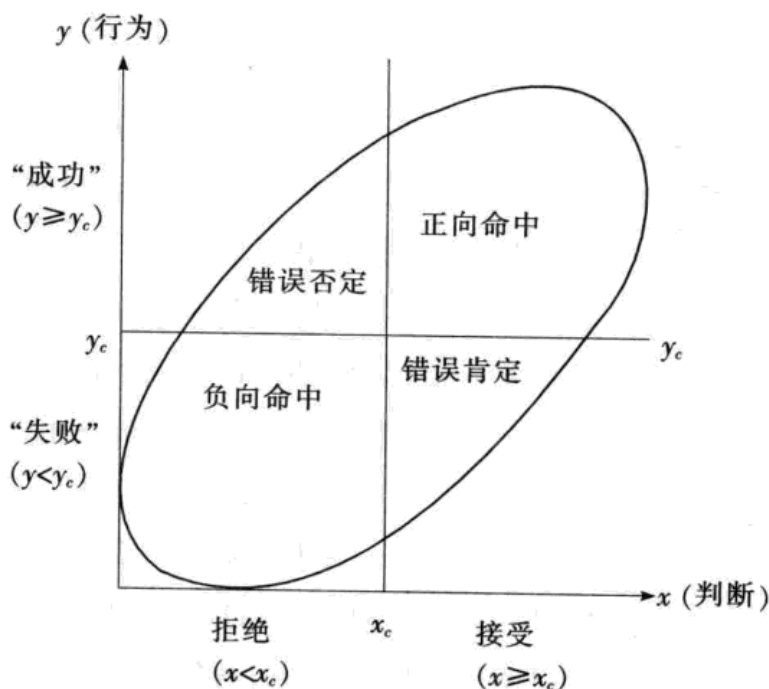


图 19—2 根据判断决定接受或者拒绝时的行为—结果组合

这里给出的解释是，因为我们在现实世界任务中的经验表明，我们发展出的规则和方法似乎“作用”得相当好。但是，这些规则可能非常糟糕，并且我们深刻认识到这些规则的不能胜任。由于不管我们的预测能力如何，都可能出现积极的结果反馈，而不是因为我们的预测能力，因此我们经常会意识不到。为了证明这一点，可参考瓦森（Wason, 1980）的研究。在这个研究中，他向被试出示一个 3 个数字组成的数列，例如：2, 4, 6。他要求被试找出这 3 个数字符合的规则（例如 3 个数字是递增的）。为了发现这个规则，允许被试给出一些组的 3 个数字，然后研究人员根据是否符合这个规则分类这些数字。当被试认为他们已经发现了这个规则的时候，他们可以随时停止。完成这个任务的正确方法应该是寻找否定性的证据，而不是累积更多的肯定性证据。例如，如果人们认为这个规则与偶数

有关，那么他只有试验一个包含奇数的数列才能检验这个规则（也就是说，累积大量符合偶数数列的例子并不能推导出这个规则）。事实是 29 名被试中只有 6 人认为他们第一次就发现了正确的规则，这说明了简单枚举归纳法的危害。正如瓦森（1980）指出的，完成这个任务的方法，必须包括“一种试图证伪假设的自我意愿，*检验那些通常可以确信的直觉观念*”（p. 139，斜体是作者添加的）。

279 瓦森的实验没有涉及行为，因此寻找否定性的证据是可能的，强调这一点很重要。但是，当行为是基于判断的时候，根据否定性的证据学习就变得更为困难了。考虑一下人们可能怎样错误地学习了一种错误的判断规则，然后集中关注这样一种假设情况，一位经理了解他预测求职者的“潜力”的能力。这里关键的因素，是行为（例如，录用/不录用）需要根据判断来决定。在接下来的日子里，这位经理只能考察被录用的那些求职者的表现来判断有多少是“成功的”。如果有很多是成功的，这很有可能，那么这些例子就证明了规则。实际上，这里重要的问题，是我们很难去证伪这个规则，即使这个规则可能是错误的。检验这个规则的一种可能方法，是录用一小组他判断为潜力低的求职者，然后观察他们的成功率。如果他们的成功率与那些他判断为潜力高的组一样，那么这个规则就被证明是错误的。但是，系统地寻找否定性证据的情况很少见，而且可能会遭到实用和伦理方面的反对。也就是说，人们不得不放弃选择一些他们认为最值得考虑的行为，选择一些他们认为最不值得的行为。因此，实用的和/或者伦理的因素，可能阻止人们甚至是去考虑收集可能的否定性资料。注意，不使用否定性证据检验假设这种倾向，是基于判断采取行动这种任务结构的一个直接后果。瓦森（1980）指出，“在现实生活中，没有什么权威宣布基于推理的判断：推理只能根据证据被检验”（p. 139）。结果，大量的积极反馈，可能导致一种无效的规则被强化。

尽管人们可能不会寻找未被采取的行为导致的结果，但是人们仍然有可能通过检查正向命中和错误肯定的数目来衡量预测的精确性。因此，虽然这种信息对于精确地估计预测与结果之间的关系而言是不完整的，但是这种信息是大多数人可以得到的。因此考虑影响这些变量的因素就很重

要了。

四、影响正向命中和错误肯定的因素

再次考虑图 19—2，注意影响正向命中率和错误肯定比率的 3 个因素： x_c 和 y_c 的位置和椭圆的“倾斜”（它是 x 和 y 的相关关系）。例如，如果 x_c 向右移，同时保持 y_c 和 ρ_{xy} 不变，那么就有一个点不再属于错误肯定。当然，错误否定方面也会有相应地增加。但是，如果人们没有关于这些情况的资料（大多数情况都是如此），那么人们的成功经验，就可能使他们十分确信他们的判断质量很高。因此，当采取所选择的行为的标准被提高时（增加 x_c ），概率 $p(x \geq x_c)$ （也称作选择比率， Φ ）就被降低了，这导致了高的正向命中率和低的错误肯定比率。第二个因素 y_c 显然影响结果，因为 y_c 的水平定义了成功和失败。注意，当 y_c 被降低时，概率 $p(y \geq y_c)$ （也称作基率， br ）被提高，人们的成功经验增多，无论判断能力如何；也就是说，如果一个人随机地指派给人们各种行为，他就会得到一个等于 $p(y \geq y_c)$ 的成功率。因此，为了判断一个人的预测能力，我们应该比较正向命中率和 $p(y \geq y_c)$ ，并且根据成功数目的边际增长估计判断能力。第三个因素 ρ_{xy} 以一种非常直接简单的方式影响结果；也就是说， ρ_{xy} 越大，正向命中率就越大。 280

这 3 个因素对于正向命中率的作用已经广为所知。例如，泰勒等人（Taylor & Russell, 1939）已经指出，对于指定的 ρ_{xy} 和基率，人们可以通过减低选择比率（ Φ ）增加正向命中率，也就是说，给予所选择的行为更小的百分比（假设 $\rho_{xy} \geq 0$ ）。因而，即使 ρ_{xy} 很低，也有可能得到一个高的正向命中率，依赖于 Φ 和 br 的值。泰勒等人（1939）提供了一些表格，给出 ρ_{xy} 、 Φ 和 br 的大部分值对应的正向命中率。对这些表格的检查表明，判断与标准之间的低相关关系，并不与高的正向命中率不相容。

除了已经提到的这 3 个因素，还必须考虑第 4 个因素。设想下述的实验，我们可以说明这个因素。假设针对一些人做出一系列判断。对于那些被判断为大于 x_c 的人，随机指派给其中的一半人行为 A，另一半人行为 B。类似地，对于被判断为小于 x_c 的人进行同样的处理。在稍晚一些时候，测

量表现，然后计算每个单元中 $y \geq y_c$ 的人的比例（每个人被指派数字 0 或者 1，表明他或者她是低于还是高于 y 的临界值——高于 y_c 的比例就作为该单元的平均值）。这是一个 2×2 因子设计，一个因子是“判断”，另一个因子是“行为的类型”。注意，因为在决策之前不能直接观测到标准（实际上，如果可能的话，就没有必要判断了），所以接受行为 A 和 B 的人，也应该接受不同的实验处理。如果这个实验成功，那么人们就可以检测判断的主效应（它测量了判断的精确性）；行为的主效应，也就是说是否接受 A 或者 B 本身导致了表现上的不同；以及判断与行为之间的相互作用。注意，这个实验的优点，在于它使得人们能够分清判断的精确性和行为的处理效应。但是，这样一个实验几乎没有成功过，即使是在概念上，尤其是对于实验设计中没有接受充分训练的人。因此，判断精确性，几乎总是与应该归于行为的可能的处理效应混淆在一起。此外，就前面的讨论而言，这个实验允许人们考察否定性的证据。因此，与大多数现实的判断任务相反，这个实验使得人们能够证伪关于判断精确性的假设，以及估计应该归于行为的任何处理效应。

图 19—3 给出了处理效应的一个例子。图中由点组成的椭圆，就是图 19—2 中的椭圆，它代表了判断与结果之间的“真实”关系。阴影部分代表可以观测到的结果，因此只给出了 $x \geq x_c$ 的值。处理效应的出现，是因为所有那些被指派了行为 A 的结果（即表现）被增加了一个常量，结果正向命中的数目，大于不存在处理效应的情况下应该观测到的数目。从一种心理学的角度来看，图 19—3 的关键在于判断者的反馈的性质被污染了，正向命中的数目被扩大，而错误肯定的数目被减少。

为了量化上述讨论的 4 种因素对于正向命中率的作用，爱恩荷恩等人 (Einhorn & Hogarth, 1978) 进行了一个模拟实验，实验使用一种因子设计来改变处理效应、选择比率、基率和预测能力的水平。因变量是正向命中率。模拟实验的结果可以总结如下：(a) 一般来说，正向命中率大于 0.50。当处理效应存在时，正向命中率可能很高，即使是在 $\rho_{xy} = 0$ 时。(b) 当 $\Phi < br$ 时，正向命中率特别高。此外，当 ρ_{xy} 的值低的时候，正向命中率对于处理效应很敏感。这意味着在高选择性情境中，糟糕的预测能

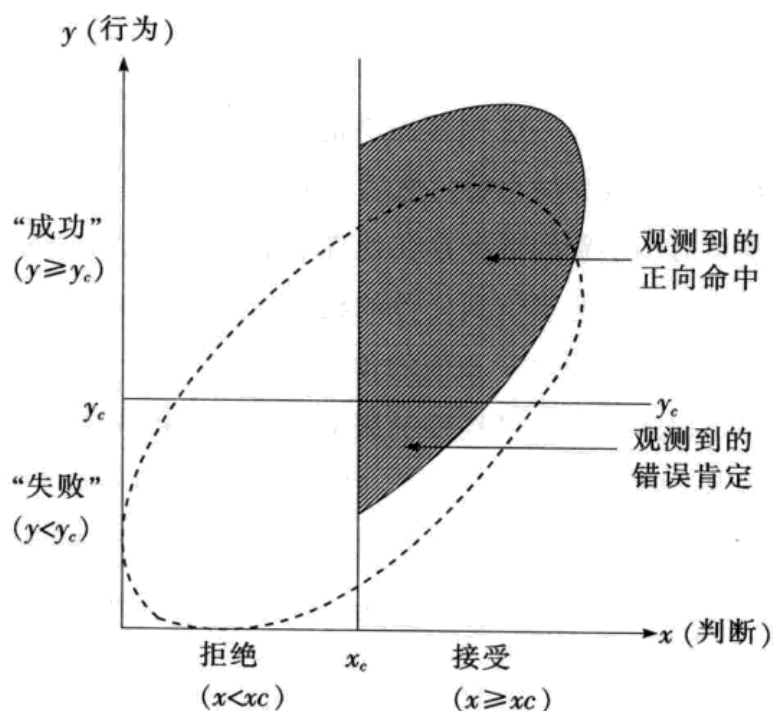


图 19—3 实验处理对观测到的正向命中率的作用效果

力最有可能被积极的结果反馈强化。(c) 当 $\Phi > br$ 时, 正向命中率最低。但是, 在这些情境中, 弱的处理效应, 对于正向命中率的提高有实质性的作用。

模拟实验的结果证明, 当预测能力很差时, 可能存在结果反馈, 并且由于不能充分地理解任务结构, 人们通常很少意识到这一点。因此, 尽管人们可能认为非概率判断是在一种结果相关的学习结构中做出的, 但是如果判断的做出是为了决定行为, 那么, 结果信息对于提供自我修正的反馈可能就是无关紧要的了。

五、结论^[2]

本章的基本主题一直是, 如果没有关于任务结构的知识, 那么对于提供关于糟糕的启发式的自我修正的反馈而言, 结果信息可能是无关紧要的。由于我们采取归纳的方法从经验中学习, 因此我们很难获得关于任务结构的知识 (比较 Hammond, 1978, 关于伽利略和亚里士多德的思维模式的比较)。这些结论提出了两个问题, 我们需要简单地讨论一下。

可能会出现这种情况，即使拥有关于任务结构的知识，但是人们所选择的行为方式排除了学习的可能。例如，考虑一个繁忙的餐厅中的一位侍者。由于他没有时间向每一位顾客提供优质的服务，因此他会预测哪一位顾客可能给较多或者较少的小费。然后根据这个预测提供不同质量的服务。如果服务的质量对于小费的数量存在处理效应，那么结果就会“证明”最初的预测。注意，如果他意识到任务的结构，那么这位侍者可以进行一个实验区分服务质量的处理效应和他的预测。也就是说，他可以向那些他判断为会给较多小费的顾客提供糟糕的服务，向那些他判断为会给很少小费的顾客提供优质的服务。但是注意，这位侍者必须自愿冒可能损失收入（如果他的判断正确会得到的收入）的风险了解他的判断的质量。后一种资料可能有长期收益，因为它可能激励人们尝试做出更好的预测，或者如果这不可能的话，鼓励人们使用随机提供优质或者糟糕服务的策略，因而减少更多的精神损失。对于组织性的决策，了解他们的预测的精确性可能获得的长期收益是巨大的。例如，如果选择面试并不能预测表现（独立于处理效应），那么为什么还要在它们上面花费金钱和时间呢？因此，我们需要更加全面地研究短期行为策略和长期学习策略的成本和收益。

289 第二个问题可以通过下面的问题提出：如果人们学习并且继续使用糟糕的规则，那么这不违背适者生存的进化观念吗？我是这样理解这个问题的，它的意思是说，与那些使用更好规则（它们更加适合）的人相比，那些使用糟糕的规则的人应该更难存活下来。但是，使用更好的规则，距离使用最优的规则仍然相差很远。最“适合”的概念，涉及的是一种相对次序，而最优则意味着某种绝对的等级。因此，次优规则在经验中被保存下来这一事实，并不违背达尔文的理论。或许最简明的表述方式是引用伊拉斯谟（Erasmus）的话：“在瞎子的世界中，独眼龙就是国王。”^[3]

本章是收录在 T. S. Wallsten 主编的《选择与决策行为中的认知过程》（*Cognitive Processes in Choice and Decision Behavior*. Hillsdale, N. J. : Lawrence Erlbaum Assoc. , Inc. , 1980）一书中的一篇论文的节选。允许重印。

本研究得到了伊利诺伊大学精神健康与发展性障碍学院的资助。研发

编号 No. 740-02。我要感谢 Robin Hogarth 对本章初稿的评论意见。

注释

[1] 这一节的大部分选自 Einhorn & Hogarth (1978)。

[2] 我要感谢 J. E. R. Staddon 提出的在本节讨论的观点。

[3] 引言的内容是为了指出相对于人们面对的环境，相对优势更加重要。绝对没有任何诋毁盲人的意图或者想法。汤姆·沃斯顿 (Tom Wallsten) 曾经给出下述评论：“在瞎子的世界里，独眼龙只有闭上他的眼睛才能活下来，因为这个世界的环境会被安排成依赖其他感官。”尽管这个评论非常有趣，但是我不同意，因为除了眼睛的优势以外，独眼龙仍然保留了他的其他感官。



11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

第 六 部分
PART SIX

过分的信心



第 20 章

在个案研究的判断中的过分信心

奥斯坎普 (Stuart Oskamp)

287 随着有关另一个人的个案材料的积累，心理学家逐渐感到对此人已非常了解，这在临床诊断中是非常普遍的现象。因此，在信息的收集过程中，心理学家迟早会带着足够的信心进行临床诊断，描述被诊断者的主要动力机制，甚至可能会冒险去预测他未来的行为。尽管心理学家的结论可能依然是试探性的，但从第一次接触个案至写出报告的期间内，他的信心的增长，往往是十分明显的。

本研究欲考察信心的增长是否因为结论的准确度的相应增长而被合理化。尽管心理学家对结论的信心，通常被称为科学探求的重要主题 (Meehl, 1957)，但科学家对其进行的集中研究，还是不充分的。另外，在对它的研究中，常常会出现非常令人惊讶的发现。比如，戈登伯格 (Goldberg, 1959) 和奥斯坎普 (Oskamp, 1962) 的研究就表明，富有经验的心理学家的诊断信心，要低于欠缺经验的心理学家的

信心。同样的研究以及其他的许多研究也表明,职业心理学家相比于未受过训练的人,并不会是更好的人际问题的判断者,有时甚至会表现更差 (Taft, 1955)。

另外一个极少被研究的要素,是判断者的信心水平和准确度水平之间的关系,它可以作为判断者的专业能力的一个很好的指标。比如,这项衡量指标可以表明判断者在做出他的决定中是过分自信还是不够自信。在这一可被称为信心合理度 (appropriateness of confidence) 的衡量指标上,研究者发现,富有经验的判断者要远远优于欠缺经验的判断者 (Oskamp, 1962)。

大量的研究 (Hamlin, 1954; Hathaway, 1956; Kostlan, 1954; Soskin, 1954; Winch & More, 1956) 都考察了不同数量的刺激信息对临床判断的影响。本实验对该因素进行了考察,我们给每位判断者四组数量逐渐增加的信息作为他们判断的基础,由此模拟心理学家分析一个典型个案时信息的逐渐积累的过程。

本研究的假设如下:

(1) 在信息收集过程的一个较早的时点之后,预测准确度,就达到了它的一个顶值 (ceiling)。

(2) 不过,随着更多信息的获得,对自己的判断的信心继续稳定地上升。

(3) 因此,在信息收集过程结束时,大多数判断者对他们的判断都是过分自信的。

一、程序

由于要尽可能地模拟通常的临床情境,我们挑选了一个真实的个案作为呈现给判断者的信息。最终被选中的个案之所以入选,是因为它信息比较详尽,描述了许多日常相关事件,而且它牵涉到的是一位相对正常的人 (一位适应失调的青年的个案,他没有接受过精神治疗)。它是有关克德 (Joseph Kidd) 的个案,被记载在怀特 (White, 1952) 的《成长中的生命》 (*Lives in Progress*) 一书中。^[1]

该个案的历史背景资料，被汇总和组织为按年代排序的几组信息，它们在四个连续的阶段中被呈现给判断者。为检验预测准确度的“心理学几率”（psychological chance）水平，阶段1仅包括了下面一些简要的人口统计信息（Patterson, 1955）：

克德（化名），今年29岁。他是白人，未婚，是二战中的老兵。他大学毕业，现在在一家花卉装饰工作室做业务助理。

阶段2加入了1.5页的不空行的打印资料，是关于克德直到12岁期间的童年生活的。阶段3（2页）覆盖了他的高中和大学生活，阶段4（1.3页）覆盖了他的服役生涯和随后直到29岁期间的活动。

1. 个案研究测试

为拥有一个决定判断者准确度的基础，我们使用了与苏斯肯（Soskin, 1954）类似的方法，即设计一份个案研究的多项选择的测试。问题涉及克德的日常行为模式、态度、兴趣和对真实生活事件的典型反应。表20—1给出了一部分这些问题的例子。

表 20—1 个案研究测试中的问题样本

-
5. 在大学期间，当克德处于一个熟悉和友好的社交场合中，他通常：
- a. 力图引导群体，并将他的意愿施加其上。
 - b. 独坐一隅，脱离群体。
 - c. 丝毫不介意他人如何对他做出反应。
 - d. 在群体中扮演积极的角色，但是是以安静和谦谨的方式。
 - e. 搞噱头，哗众取宠。◎
10. 在他服役后期，作为一名军官和分遣队指挥，克德对施加惩罚的态度是：
- a. 他深受其困扰，因为他更愿意和其他人站在同一位置上，而不是凌驾其上。◎
 - b. 他对此很讨厌，因为他对该做什么无法做出决定。
 - c. 他尽可能地避免惩罚，因为他认为，无论人们干了什么，去惩罚人们都是错误的。
 - d. 他很高兴，因为这给他一个机会去控制局面，并受人尊敬。
 - e. 他在教训他人当中获得一种虐待狂似的愉悦，这可以弥补他曾经受到的惩罚。
15. 克德现在对他母亲的态度是一种：
- a. 对她的观念的尊重和爱。
 - b. 对她的缺点出于挚爱的宽容。
-

续前表

- c. 兼有尊重和愤恨。◎
- d. 反感她和她的所有观念。
- e. 孝敬，但是情感上一般。

20. 在与他人的谈话中，克德：

- a. 更愿意让他们谈他们的工作或者经历。◎
- b. 喜欢在大多数时间内谈论他所熟悉的主题。
- c. 更愿意与他们讨论宗教或者他们的生命哲学。
- d. 喜欢吹嘘他的军队生涯或者大学时期的风采。
- e. 将他的谈话主要局限在体育、性和粗俗的笑话上。

25. 克德对他作为业务助理的生活的态度，可以由他最近_____的决定表现出来。

- a. 至少在最近几年内待在现有的职位上
- b. 通过在附近的一个镇上增开一家商店来扩大业务
- c. 离开该职务，开一家自己的花店
- d. 在与他目前所从事的相似领域中，向数家大型公司提交应聘资料
- e. 追求自我，寻找另一种工作◎

注：◎ 为正确答案。

只有在该个案中存在非常客观的标准信息，即客观的数据或者可被充分证明的结论的时候，我们才会设计出问题。每个问题中的其余四项错误的替代答案，是由心理学的研究生通过有关项目的句子——完成测验而得到的。这些问题的设计，以这样的方式进行，即基于公布的个案材料，它们显然是错误的，但它们又是令人信服和具有“诱惑性”的替代答案。没有哪个问题的答案，包含在汇总的个案材料中；相反，我们预期被试会遵循临床诊断中的一般程序（McArthur, 1954），从得到的材料中形成克德的人格图（personality picture），然后根据人格图来预测他的态度和典型行为。

2. 判断者

判断者是从三组有不同水平的心理学经验的小组中抽取的：（1）由加州州立医院聘请的 8 位临床心理学家，他们都有数年的临床经验，其中 5 位有博士学位^[2]；（2）18 位心理学的研究生^[3]；以及（3）6 位人格课程上的本科高年级学生。没有一位判断者以任何方式了解本研究的假设。

判断者以小组的形式参加实验，它们的规模由 4 到 9 不等，但每位判

断者都以自己的节奏处理他手头的材料。在阅读完每一阶段的个案材料后，判断者完成个案研究测试中的 25 道问题，然后阅读下一阶段的材料。除了回答问题之外，判断者还要写明他对每一问题回答正确的信心。

3. 信心判断

我们使用亚当斯 (Adams, 1957) 设计的量表进行信心判断，它用预期的正确判断的比例来定义信心。因为每道测试问题有 5 个备选答案，量表起始于 20% (代表了信心完全是几率水平上的)，直至 100% (代表的是对回答正确的绝对肯定)。除了提供一个可被清晰理解的信心的主观含义之外，该量表的最大优点在于，它可以让研究者在准确度水平和信心水平之间进行直接比较。因此，比如说，如果一位判断者有 28% 的问题回答正确，而平均的信心水平为 43%，则他可以被明白无误地称为过分自信。

二、结果

291 这项判断任务，被证明是非常困难的，至少就提供的个案材料的数量来说是这样的。没有一位判断者达到 50% 的准确度，最终的平均准确度低于 28%，而几率为 20% (一个非显著的差异)。不过，这一低水平的准确度，为我们研究的假设提供了更为鲜明的检验。

尽管我们没有形成对三个组的相对表现的假设，我们仍进行了一项初步的分析，比较三组判断者的分数。这些结果明显表明，三组判断者在准确度、信心或者改变答案的总数量上没有显著的差异。阶段 4 的信心分数与先前的研究是一致的 (Goldberg, 1959; Oskamp, 1962)，它们都表明，更富经验的判断者，比欠缺经验的判断者，更缺乏信心；但在本研究中，这些结果是不显著的。

我们在表 20—2 中给出了本研究的主要结果，其中连续的竖列是判断者连续收到更多数量的信息时的平均分数。由于已经有先前的统计检验，所有 32 位被试的结果都汇总在这个表中。

表 20—2 的第一行表明，在该个案的四个阶段中的准确度的波动是十分显著的。不过，邓肯多重域检验 (Duncan multiple-range test) (Edwards, 1960, p. 136) 说明，这一显著性主要是由于第二阶段中的准确度的下降。

比较阶段 1 的准确度和阶段 4 的准确度,说明不存在显著的变化 ($t = 1.13$, $df = 31$)。因此,有关准确度的顶值的第一个假设,不仅得到了支持,而且在实验中,随着信息的增加,根本就没有出现准确度的显著增加。

表 20—2 32 位判断者在个案研究测试的 25 道问题中的表现

衡量指标	平均分数				F	p
	阶段 1	阶段 2	阶段 3	阶段 4		
准确度 (%)	26.0	23.0	28.4	27.8	5.02	0.01
信心 (%)	33.2	39.2	46.0	52.8	36.06	0.001
改变答案的数量	—	13.2	11.4	8.1	21.56	0.001

假设 2 在表 20—2 的第二行中得到检验。我们在那里发现,与预期的一样,信心出现了惊人的十分显著的增长,它从阶段 1 的 33% 增长至阶段 4 的 53%。

最后,假设 3 的结果通过比较表中的第一行和第二行得到说明。在阶段 1,过分信心的平均数量是 7 点;在阶段 4 是 25 点,这一差异在远远超过 0.001 的显著性水平上是显著的 ($t = 5.14$, $df = 31$)。

有时候,小组的均值是显著的,但却是误导性的,这是因为它们掩盖了那些违背预期的单个被试的判断。本研究不存在这个状况,这可以由下面的单个判断者的数据清楚地说明。在 32 位判断者中,有 14 位的准确度从阶段 1 到阶段 4 有增加,有 6 位是保持一样的,还有 12 位有所下降——这完全是一个随机的结果。与之相反,除两位判断者之外,所有其他的判断者的信心都有增加,而且大多数是显著的增加。^[4]在阶段 1,差不多半数的判断者(32 位中的 13 位)不是过分自信的;到了阶段 4,仅仅只有两位仍然是不够自信的——一个高度显著的变化 ($\chi^2 = 9.1$, $p < 0.01$)。

本研究的另外一个有趣的结果,是在表 20—2 中的最后一行,它是判断者在个案的每一阶段改变答案的问题的平均数量。这一衡量表明,随着更多信息的获得,改变答案的数量显著下降。这一发现说明,从最初的零碎的信息起,判断者就常常非常牢固地形成了刻板印象式的结论,而在随后不断获取新的信息时也不愿改变他们的结论。不管怎样,最后阶段的信息看来主要是确认了先前的印象,而不是使他们修补克德的整个人格图。

三、讨论

我们必须要避免对这些发现进行随意的概括化。在本研究中，有三个主要的因素，很可能会制约这些结果的普遍性。(a) 该个案可能与大多数心理学家在工作中所习惯的个案，是不相似的。(b) 测试的问题所体现的行为的类别，可能不是心理学家所习惯于预测的。(c) 判断者可能不是心理判断的决定者的好的代表。为应答这些可能的异议，我们必须指出，选择该个案、测试问题和临床诊断者，都是本着尽可能接近实际心理学实践的意图。

即便这些可能的异议是真实存在的，我们仍然可以获得一些明确的结论。无论这个问题是否看起来奇怪，或者说个案材料是非典型的，判断者的信心分值表明，他们对自己逐渐增长的对个案的理解变得自信。当他们获得更多的信息时，他们的信心就会飙升。另外，他们对自己判断的确定性，与判断的实际正确率完全不成比例。

因此，尽管这一结果并不是对所有心理学家以及所有判断的类型都成立，²⁹⁹ 但我们可以清楚地得出结论，心理学家在他个案分析过程中不断增长的信心，并不是确切地标志了他的结论的准确度在不断增加。所谓的临床确认，它基于临床医生对信心的个人感受，并不是诊断或者预期人类行为的临床判断的有效性的充分证据。

本章原载《咨询心理学杂志》 (*The Journal of Consulting Psychology*, 1965, 29, 261-265) 允许重印。

注释

[1] 使用该个案的缺点在于，少数判断者记得曾经在他们培训过程的某一时候读过这份材料，但除一位判断者外，其他人都报告说早期的接触对本研究中的判断没有任何帮助。因为他们的准确度的分数证实了这一印象，所以他们的结果仍然保留在数据分析中。

[2] 还有一位临床心理学家参加了测试，但由于他未能理解和遵循说明，其结果不得不被舍弃。没有任何学生发生这一问题。

[3] 约有半数的研究生有一定的临床或者咨询经验，还有一两位很可能有等

同于临床心理学家的心理学经验水平。

[4] 两位信心下降的判断者中的一位，是位本科生，他随后说，他的信心一般来说应当是会增加的，但因为他刚刚参加了一项计算机的研究项目，其间计算机反复给出了错误的结果，以至于他甚至完全丧失了对电脑的信心。



第 21 章

关于概率评估训练的进展报告

阿尔波特等 (Marc Alpert and Howard Raiffa)

294

关于不确定状况下的决策 (decision under uncertainty) 的规范分析, 通常要求决策者和他们的专家顾问估计一些变量的判断概率分布 (judgmental probability distribution), 而这些变量的值是他们所不知道的。本章所讨论的一些经验研究发现涉及下列问题: 未经训练的个体完成这种估计任务的水平如何? 他们表现出某种经常性的偏差了吗? 怎样校准评估者? 怎样把他们训练成为更优秀的评估者?

本章仅讨论关于不确定变量的估计, 可以认为这种变量是在一个可能值的连续统上取值。因此我们只研究一元密度函数和它们的累积分布函数。可以使用几种不同的程序估计连续一元随机变量的概率分布, 但是我们将只考虑一种特殊程序, 它是我们以及我们的同事在日常实践中所经常使用的, 被称作直接分位数估计法 (direct fractile assessment)。

一、直接分位数估计的程序

假设 x^* 是某个变量的客观真实值, 假定评估者不知道 x^* 。评估者的第 k 个 x^* 判断分位数是 x_k (k 的取值区间从 0 到 1), 因此他赋予事件 $\{x^* \leq x_k\}$ 的判断概率是 k ; 用符号表示是 $P\{x^* \leq x_k\} = k$ 。数值 $x_{0.50}$ 、 $x_{0.25}$ 、 $x_{0.75}$ 分别被称作判断为中位数、下分位数和上分位数。为了确定 $x_{0.50}$, 被试必须给出一个值, 使得他认为 x^* 落在 $x_{0.50}$ 之下和 $x_{0.50}$ 之上的可能性正好相同。因而 $x_{0.50}$ 把连续统分为两个等判断概率区间。下分位数 $x_{0.25}$ 把区间 $(-\infty, x_{0.50})$ 分为两个等判断概率区间, 上分位数 $x_{0.75}$ 把区间 $(x_{0.50}, +\infty)$ 分为两个等概率判断区间。对于每一个不确定变量, 我们的实验要求被试估计它们的中位数和四分位数。连贯性 (consistency) [或者是一些作者更喜欢说的“一致性” (coherency)] 要求被试相信 (a)

$$(-\infty, x_{0.25}), (x_{0.25}, x_{0.50}), (x_{0.50}, x_{0.75}), (x_{0.75}, +\infty)$$

四个区间中的每一个区间包含 x^* 的真实值的可能性正好相同, 和 (b) 与区间 $(x_{0.25}, x_{0.75})$ 不包含 x^* 的真实值的可能性相同。因此, 我们应把区间 $(x_{0.25}, x_{0.75})$ 称为四分位差 (interquartile range)。参加我们训练练习的被试, 都被教会核查这种连贯性要求, 并且在不符合连贯性的情况下, 被指导重新评估他们的分位数估计以符合连贯性。

除了判断中位数和两个四分位数之外, 还要求被试估计他们的分布的低端和高端的各种分位数。更多内容见下文。一旦评估者确定了他的累积左偏判断概率分布, 那么他就能够运用直观方法, 不借助仪器“修整”他的曲线的其余部分。但是在本章中我们将只关注直接估计的 (x_k, k) 点, 而不是整条曲线。

二、外部效度的概率

对于每一位指定的被试, 如果我们所知道的只有他对一个不确定变量的概率分布, 那么说他的分布是“错误的”对于我们毫无意义。我们可能希望被试对于该变量能够知道得更多, 但是他的概率分布正好表明了他知道什么——或者不知道什么。例如, 我们不能说他的分布“太高耸”、“太低平”或者“太偏右”。但是, 与这种情况相反, 假设我们的被试给出了

1 000个不同的不确定变量的1 000种分布。如果每一个实际真实值或者落在他相应的0.01分位数之下,或者落在他的0.99分位数之上,那么我们就可以说他没有被外部校准,他的分布倾向于太高耸。或者,与此相反,如果碰巧他的每一个四分位差估计区间都包含真实值,那么这表明他的分布倾向于太低平。即使这些都是极端情况,但是它们证明,讨论一组概率分布的外在效度是可能的,并且至关重要。

三、练习的目标和被试组的构成

296 让我们简单介绍一下我们所做的事情。通过使用一系列问卷,我们要求大量被试(大约1 000名)在估计时记录他们关于几个未知变量的判断分位数。然后我们比较他们的估计值和真实值,确定了一些持续存在的偏差,研究了修正他们的判断输入数据的形式程序,告知每位被试小组的回答以及他自己的回答的质量。最后,我们提供一些方法使得每位被试能够注意到他自己过去的特有反应,因而使他能够修正他的下一组估计。

在1968至1969学年期间,我们进行了4次独立但是相关的练习。为了区分它们,我们现在列举如下:

第1组:选修不确定性下的计划模型课程的哈佛大学MBA项目二年级学生,共139人。

第2组:包括哈佛大学MBA项目一年级的全体学生,共800人。

第3组:来自哈佛大学商学院高级管理项目的“志愿者”,共67人。

第4组:选修哈佛大学文理学院开设的决策分析课程的学生,共60人。这些学生来自于哈佛大学和麻省理工学院的各个院系:经济学、统计学、工程学、数学、法学、政府管理、设计专业、社会关系、操作研究。

在参加这个训练练习之前,这四个组的所有被试都曾学习过基础的决策分析:决策树的构造和分析,先验和后验概率分布,效用(或者偏好)

理论和信息价值。

第 1 组作为先导研究 (pilot study)。要求每个学生独立评估 20 个不确定变量的分布。这一组被试, 我们调整了关于尾部概率的指示; 后文会讨论这一点。所有 20 个估计都在一个部分内完成, 中间没有任何反馈。

对第 2、3 和 4 组的处理完全相同。在第一部分, 每位被试估计给出的 10 个不确定变量。然后被试被简单地告知小组的总体表现以及他们自己的表现。然后进行第二回合的 10 个不同的不确定变量。

第 1 组与第 2、3、4 组在他们第一回合中的表现惊人地相似。第 2、3、4 组在他们的第二回合, 即简报与反馈之后部分中的表现, 再次惊人地相似; 但是他们在第一与第二回合之间的明显进步值得注意。由于所有组在行为上的相似性, 因此我们的讨论将把注意力集中在最大一组的具体表现上。

四、练习与反馈前的结果

在此, 我们复制了分发给 MBA 一年级班 (第 2 组) 半数学生的指导语。这一半学生得到的表我们称之为表 B, 另一半学生得到表 A, 表 A 与表 B 内容相同, 但是使用不同的不确定变量。

1. 指导语: 表 B

这个练习的目的, 是为了了解你作为一个个体和班级作为一个整体评估不确定变量 (uq) 的概率分布的水平。下面我们将给出 10 个不确定变量, 然后要求你估计每个不确定变量的中位数、0.25 分位数、0.75 分位数和极值 (为了方便你, 下面给出了这些术语的简单定义)。由于使用的不确定变量的类型, 你有机会比较你的估计和真实值, 因而你能够发现在某些类型的问题上, 你的分布是否倾向于“太高耸”、“太低平”、或者偏高、或者偏低。之后将使用其他不确定变量要求你重复这个练习。你必须独立回答这个练习中的问题, 不得参考任何来源的资料, 尽管对于一些问题一定的简单计算是允许的。我们对你的知识非常感兴趣, 无论它多么不明确或者你感到多么不舒服。

分位数的定义…… (省略)

下面是 10 个变量,但愿它们对你来说是不确定变量。你要在提供的每张答题纸上填写你关于它们的估计。根据你现有的知识尽可能地回答每个问题,但是不要查阅其他资料或者与其他人讨论,即使在交上答案以后。我们希望你在明天上课之前交一份答题纸的复印件,保留另一份,这样你就拥有一份你的答案记录(学生使用他们保留的复印件为自己计分。后文会描述)。

一些变量的量值,需要根据一年级学生对下列问题的回答确定:

2. 问题

- A. 波旁威士忌和苏格兰威士忌,你更喜欢哪一种?
- B. 你支持在校的所有研究生无论专业都应实行兵役延期这一草案吗?
- C. 如果有一次胜负机会各半的赌博,你可能会输掉 50 美元或者赢得 100 美元,你会接受吗?

3. 不确定量值列表

(1) 除了那些不喝酒的学生之外,回答喜欢波旁威士忌的一年级学生的百分比。

298

(2) 回答支持在校所有研究生无论专业都应实行兵役延期的一年级学生的百分比。

(3) 回答愿意接受上述问题 C 中的赌博的一年级学生的百分比。

(4) 1968 年 6 月,盖洛普民意测验调查了一个有代表性的成年美国人样本,他们认为如果在中东地区爆发全面战争,美国应该出兵帮助以色列。对此发表意见的被调查对象的百分比。

(5) 1968 年 3 月,盖洛普民意测验调查了一个有代表性的成年美国人样本,他们认为应该允许公立学校教师加入工会。对此发表意见的被调查对象的百分比。

(6) 1968 年波士顿及周边地区的黄页电话号码簿中列出的“内科

医生和外科医生”的数目。

(7) 哈佛商学院目前选修博士课程的学生总数。

(8) 1965 年美国生产的鸡蛋的总数，以百万计。

(9) 1967 年美国进口的外国汽车的数目，以千计。

(10) 1967 财政年度巴拿马运河征收的通行税，以百万美元计。

[新的一页]

表 B (本页需要保留)

部分_____

学号_____

问题 A、B 和 C，每一题请选择一个答案：

A. 饮料 波旁威士忌_____ (1) 苏格兰威士忌_____ (2)

从不饮酒_____ (3)

B. 延期兵役 支持_____ (1) 反对_____ (2)

C. 胜负各半的赌博 接受_____ (1) 拒绝_____ (2)

请估计下列 10 个变量中每个变量的所有 5 种值。注意最高的数值在每一行的右边 (答案中可以使用小数。但是在需要使用百分比时，答案例如 97.2% 必须是 97.2 的形式，而不是 0.972)。

不确定变量 分位数

1. 波旁威士忌 (%) (0.01) (0.25) (0.50) (0.75) (0.99)

4. 反馈备忘录

大约在学生完成表 B 一个星期之后，发给学生下面的备忘录。这个备忘录也在发给学生要完成的表 A 之前在班上讨论。

这个备忘录的目的是：

(1) 描述完成表 B 的学生的总体表现，

(2) 指出回答中存在的系统偏差，

(3) 使你在再次使用表 A 重复这个练习之前能够校准你自己。

229

当你回忆每个问题时，你估计了 5 个分位数：0.01、0.25、0.50、0.75、0.99。这些分位数把整个区间分成 6 个类别：

类别 1：0.01 分位数之下的所有数值

类别 2：0.01 分位数到 0.25 分位数之间的所有数值

类别 3：0.25 分位数到 0.50 分位数之间的所有数值

类别 4：0.50 分位数到 0.75 分位数之间的所有数值

类别 5：0.75 分位数到 0.99 分位数之间的所有数值

类别 6：0.99 分位数之上的所有数值

对于 10 个问题中的任何一个问题，一旦你指定了你的 5 个分位数（或者说，一旦你指定了 6 个类别），就可以指出实际的真实值落在哪一个类别。例如，对于波旁—苏格兰威士忌问题，学生约翰·杜给出下列的分位数：

估计：(0.01) /5 (0.25) /15 (0.50) /30 (0.75) /40 (0.99) /45

喜欢波旁威士忌的实际百分比结果是 42.5；因此在这个问题上，杜先生的答案落在类别 5 上。

在我们进一步讨论班级在这个练习中的表现如何优秀（或者拙劣）之前，现在请检查你的答题纸，对于每个问题，标出你的 6 个类别中哪一个类别包含了真实值。

表 21—1 中我们列出了真实答案。

表 21—1 表 B 中问题的答案

题号	标题	真实答案
1	波旁威士忌	42.5
2	兵役延期	65.5
3	胜负机会各半的赌博	55.2
4	以色列	10.4
5	教师工会	63.5
6	列出的医生	2 600
7	博士课程学生	235
8	生产的鸡蛋（百万）	64 588
9	进口的汽车（千）	697
10	运河税收（百万）	82.3

5. 四分位差区间的分析

让我们首先考虑表 21—2 的第 3 栏和第 4 栏。对于任何一个问题，你应该选择你的 0.25、0.50、0.75 分位数，这样就你而言，真实答案落在 0.25 到 0.75（也就是类别 3 和 4）这个范围之内和落在这个范围之外的可能性正好相同。那么我们会（数学）期望在总数 1 000 个（ 100×10 ）问题中有 500 个回答会落在类别 3 和 4 中。我们只得到了 334 个这样的回答，这还不是太糟糕。这种不一致很可能是一种统计偏差，但是我们表示怀疑。例如，几个星期以前，让一个二年级 MBA 的大班填答了类似的问卷——他们回答 20 个而不是 10 个问题——结果他们的回答有 33% 落在类别 3 和 4 中。无论如何，我们并不是想暗示 33% 是一个“普遍的半常量”（universal semi-constant），但是它是一个相当令人吃惊的巧合，需要加以密切注意 [那些首先完成表 A 的学生相应的百分比是 33%；AMP 组（第 3 组）得到的数值也是 33%，但是艺术与科学研究院的学生（第 4 组）记录的回答是 36%]。

表 21—2 被试对表 B 的回答按类别的分布

题号	标题	低于 0.01	0.01 到 0.25	0.25 到 0.50	0.50 到 0.75	0.75 到 0.99	高于 0.99	总计
1	波旁威士忌	3	16	20	40	11	10	100
2	兵役延期	15	12	35	19	10	9	100
3	胜负机会各半的赌博	11	8	28	29	13	11	100
4	以色列	51	41	6	1	1	0	100
5	教师工会	1	1	13	28	29	28	100
6	列出的医生	24	14	12	13	10	27	100
7	博士课程学生	1	3	11	9	15	61	100
8	生产的鸡蛋	9	2	13	10	8	58	100
9	进口的汽车	25	15	18	9	7	26	100
10	运河税收	18	8	8	12	16	38	100
	总计	158	120	164	170	120	268	1 000
	期望频次	10	240	250	250	240	10	1 000

作为总体，0.25 到 0.75 分位数区间过于高耸。多数真实值，落在四分位差区间（0.25 到 0.75 范围）之内和之外的可能性本应相同，但事实上，落在四分位差区间之外的是落在之内的两倍。但是，我们还不能提议说，“延长你的四分位差区间”，因为从一个问题到另一个问题，从一个个体到另一个个体存在许多变动因素。

让我们比较一下问题 1 和问题 10。对于问题 1，你可能觉得你对于你的同学的饮酒偏好相当了解——至少与关于运河税收你所知道的相比非常多。因此，你可能觉得与真实的运河税收相比，你的四分位差区间包含真实的波旁威士忌比例的（事先的）可能性更大。但是这是你应该寻找的方法吗？如果你事先找到了这种方法，那么你本应该延长你的运河税收四分位差区间。当你做表 A 的时候，一定要保证，无论你的资料状况如何，你相信：

（1）对于任何一个问题，真实值落在你的四分位差区间之外和之内的可能性正好相同；并且

（2）对于一个你了解很多的问题和一个你了解得很少的问题，真实值落在你的四分位差区间之内的可能性正好相同。

（在他们完成第一回合的估计之前而不是之后，文理学院的学生无聊地讨论了关于问题 1 和问题 10 的最后一点。这可能解释了 33% 与 36% 之间的不一致。当我们考虑尾部概率时，我们会发现更加戏剧性的变化。）

你们中的一些人的分布确实太高耸了。例如，100 个被调查对象中只有 13 个人，他们的 10 个四分位差区间包含了 0 或者 1 个真实值。每个个体，假设他的四分位差分数（interquartile score）是真实值落在他的四分位差区间之内的次数。分数接近于 0 意味着他的分布“太高耸”；分数接近于 10 意味着他的分布“太低平”；分数靠近 5 意味着他的分布“恰好合适”。表 21—3 给出了分数的分布。假设任何一个四分位差区间都有 0.33 的概率包含真实值，并且各个问题之间这种二分回答是独立的。在表格的第 3 栏，我们给出了这种简化假设情况下，会给出任何指定的四分位差分数的被调查对象的期望人数。确保你得到了你的四分位差分数。

表 21—3

表 B 四分位差分数的分布

四分位差分数	实际人数	期望人数 $p = 0.33$
0	3	1.9
1	10	9.0
2	22	19.9
3	20	26.1
4	23	22.5
5	11	13.3
6	9	5.5
7	2	1.5
8	0	0.3
9	0	0.0
10	0	0.0
总计	100	100.0

6. 极值分析

你可能指定了你的 0.01 和 0.99 分位数，因此你会“惊奇地”发现一个非常低的答案（类别 1 的回答）和一个非常高的答案（类别 6 的回答）。我们本来“期望”1 000 个问题中一共会有 20 次惊奇，10 次在左边，10 次在右边（见表 21—2）。结果一共有 426 次惊奇！这不应该是使你感到惊奇，而是让你震惊！然而，你可以感到一些安慰，因为你还没有像我们前面提到过的二年级班（第 1 组）那么糟糕。就和你一样，大约 35 名学生被要求列出关于 20 个问题的 0.01 和 0.99 分位数。他们记录下的是 46% 的惊奇——比你们的 42% 稍微糟糕一些。另外 35 名学生被要求记录 0.001 和 0.999 分位数，他们记录下的不是“期望的”0.2% 的惊奇，而是 40%。还有 35 名学生被指示根据相当含糊的术语列出“最小值”和“最大值”——我们确实不知道这意味着什么——结果他们记录下的是 47% 的惊奇。最后 35 名学生被指示给出“惊人的低”和“惊人的高”的数值，他们记录下的是 38% 的惊奇。显然，“惊人的低”低于“最小的”。

看在上帝的分上，延长那些端点分位数吧！对你自己诚实一点吧！承认你不知道的东西！

让我们看看问题 6。波士顿及周边地区黄页电话号码簿列出 2 600 名医生。你们中的一半人会对这个结果感到惊奇。这一半人中，一半人吃惊于

结果的低，一半人吃惊于结果的高。你们中有 58% 的人吃惊地发现生产的鸡蛋的数目大得超出想象——但是你们中仍然有 9% 的人吃惊地发现生产的鸡蛋的数目小得超乎想象。

让我们为每个人定义一个惊奇指数 (surprise index), 表示他记录的类别 1 或者类别 6 回答的次数 (每 10 次中)。平均而言, 对于这许多不同的、独立的问题, 每 50 次你应该会有一次惊奇。惊奇指数 0 或者 1 是我们愿意看到的, 如果惊奇指数达到 3 或者更大, 就值得注意了。表 21—4 给出了惊奇指数的分布。假设每个问题有 0.43 的概率得到令人惊奇的结果, 问题与问题之间独立, 在表格的第 3 栏, 我们给出了这种简化假设情况下, 会记录下指定的惊奇指数的被调查对象的期望人数。10 个问题中至少有一半的问题, 45 个人得到惊奇的结果。

表 21—4

表 B 的惊奇指数分布

惊奇指数	实际人数	期望人数 $p = 0.43$
0	1	0.4
1	9	2.7
2	15	9.3
3	13	18.7
4	17	24.6
5	15	22.3
6	16	14.0
7	4	6.0
8	9	1.7
9	1	0.3
10	0	0.0
总计	100	100.0

至于典型的不确定变量，我们设想你们中的大多数人愿意评估一条单峰的密度曲线（或者等于一个S形的左偏累积分布），这条密度曲线的最高点（或者S形累积曲线最陡峭的部分）出现在你的0.25到0.75分位数区间中的某一处。我们设想，由于疏忽，你们中的许多人记录下的是双峰曲线。如果你的类别2区间^[1]的长度小于类别3区间，并且你的类别5区间的长度小于类别4区间，就会是这种情况。下列的一组分位数估计就是一个适例。

在上述例子中，0.98 概率区间（从 0.01 到 0.99 分位数）与 0.50 概率区间（从 0.25 到 0.75 分位数）的长度的比率是 $(45 - 5) / (40 - 15)$ 或者 1.6。对于单峰曲线来说，这是一个非常小的数值。例如，正态钟形曲线的这个比率大约是 3.5。比率 2.25 到 4.5 可以作为一个粗略的标准——但是标准应该不时改变。记住：延长那些分布！

五、反馈后的结果

在班上讨论反馈备忘录之后，我们改换表格重复这个练习。在分析了回答之后，我们分发下面的备忘录：

分位数：	0.01	0.25	0.50	0.75	0.99
估计：	5	15	30	40	45
区间长度：		10	15	10	5

备忘录 2

现在，你们每个人按照下述的顺序参加：

a. （第 1 回合）：你被给予两种表格（A 或者 B）中的一种，然后你给出 10 个不确定变量的分位数估计。

b. （反馈）：接着给出你这 10 个四分位数的真实值，然后要求你根据两种因素计算你的分数：（1）你的答案有多少落在你的四分位差区间之内，（2）你的答案有多少落在你的 0.01 和 0.99 范围之外。你也被告知班级的总体表现：落在四分位差区间之外的答案大约是落在之内的两倍，并且惊奇结果的数目非常令人震惊——大约 40% 的答案落在 0.01 和 0.99 范围之外。

c. （第 2 回合）：接下来你被给予其他的表格，然后要求你估计分位数。

现在，我们已经计算了你（c）部分的答案，准备评论你的总体表现。但是首先我们给出了不确定变量的真实值，使得你可以校准你自己（备忘录中出示的一个图表给出这些资料）。

班级作为一个整体表现如何呢？反馈部分有帮助吗？它起了一点作用，但是没有我们期望的那么多。你们的分布大多数太高耸。表 21—5 概括了

这些结果。

表 21—5 第 1 回合与第 2 回合结果的比较

	表 A		表 B	
	第 1 回合	第 2 回合	第 1 回合	第 2 回合
落在四分位差区间内的值 (%)	32.9	40.3	33.4	46.4
落在 0.01 至 0.99 区间外的值 (%)	38.8	24.9	42.6	22.2

如果我们总计表 A 和表 B，那么真实值落在四分位差区间之内的次数的百分比从 33% 跃升到 43%。变化的方向很好，但是我们仍然差一些才能达到我们本来想要的 50%。真实值落在极端值（也就是 0.01 和 0.99 范围）之外的次数的百分比从令人震惊的 41% 下降到令人郁闷的 23%。这距离我们的目标 2% 还有遥远的距离。还必须考虑那些尾部值。我们需要另外一个反馈部分和另外一个回合的练习；但是很不幸，这在操作上不可行。我们建议以其他组开始试验，以得到更加现实的尾部概率估计。

[为了比较，表 21—6 给出了 AMP（第 3 组）和文理学院的学生（第 4 组）的前一后表现。]

表 21—6 第 1 回合与第 2 回合结果的比较

	第 3 组		第 4 组	
	第 1 回合	第 2 回合	第 1 回合	第 2 回合
落在四分位差区间内的值 (%)	33.4	44.8	35.6	42.5
落在 0.01 至 0.99 区间外的值 (%)	35.6	22.8	20.8	8.7

六、关于极端尾值的更多讨论

在先导研究（第 1 组）中，为了了解它对于回答的统计分布的作用，我们调整了关于极端尾值的指示。但是要求我们的所有被试给出他们的判断中位数和两个四分位数。表 21—7 和表 21—8 总结了我们的结果，我们将依次解释。

表 21—7 对“较低”和“较高”的不同解释的惊奇比例

对“较低”的解释	对“较高”的解释	样本规模	惊奇比例
0.01 分位数	0.99 分位数	44	0.46
0.001 分位数	0.999 分位值	25	0.40
最小值	最大值	35	0.47
惊人的低	惊人的高	35	0.38

表 21—8 问题 10 的“较高”评估的分布

获胜队伍得分	对“较高”的解释			
	0.99	0.999	最大值	惊人的高
<29	8	1	6	2
29~35	15	6	7	6
36~42	10	5	7	11
43~49	2	6	5	4
50~56	7	3	8	7
57~63	2	0	2	4
64~70	0	4	0	1
总计	44	25	35	35

139 名被试中 44 人被要求给出 0.01 和 0.99 分位数，这一小组记录得到 46%（不是 2%）的惊奇；25 人被要求给出 0.001 和 0.999 分位数，这一小组记录得到 40%（不是 0.2%）。35 名被试被要求给出一个“最小值”和一个“最大值”，并且当一些被试要求说明清楚这些术语意味着什么时，我们用一些含糊的意见搪塞他们，例如，“它们确实意义不明确，但是无论如何试着回答问题”。剩下的 35 名被试，我们给予同样模棱两可的建议，要求他们给出“惊人的低”和“惊人的高”的数值。这最后两组分别记录得到 47% 和 38%。不同的指示有一些作用，但是如同你所看到的，作用不是非常大。 305

另一种检验这四种极值指示的作用的方法是仔细考查表 21—8，它给出了对下述变量的“较高”值的回答的分布：下一届哈佛—达特茅斯橄榄球赛获胜队伍的得分（在比赛前一个星期回答）。例如，被要求给出一个 0.999 分位数的 25 个人中，有 1 个人记录的分数低于 29，6 个人记录的分数在 29 到 35 之间，5 个人在 36 到 42 之间，等等。看起来，似乎意义不明

确的术语“最大值”被解释为一个 0.99 分位数，而术语“惊人的高”被解释为一个 0.999 分位数。

至于第 2、3 和 4 组，我们没有重复这些关于极端值的指示；相反，我们始终要求这些被试给出 0.01 和 0.99 分位数。

本章最初作为一篇未发表的报告于 1969 年分发。在收录入本书之前特别进行修改。

注释

[1] 假设 l_i 代表类别 1 的长度。如果 l_2 和 l_5 每一个都小于 $(l_3 + l_4) / 2$ ，那么我们就得到了双峰分布。



第 22 章

概率的校准：从当前的 状况（1977）到 1980

利希坦斯坦等 (Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D. Phillips)

根据主观主义的观点 (de Finetti, 1937/1974), 306
概率是对某个命题的信任程度。它表达了一种纯粹的
内部状态, 不存在什么“现实中的”“正确”、“错
误”或者“客观的”概率, 能够与人们的信任程度比
较。但是在很多情况下, 我们可以赋予命题一个概
率, 证明命题的真或假。今天, 我们估计命题“明天
会下雨”的概率。明天观察雨量计, 我们便可以确定
是否降过雨。只要可能, 这种证明就可以用来判断概
率估计的恰当性。

温克勒和默菲 (1968b) 已经确定了两种类型的
概率估计“优度”: 规范优度 (normative goodness)
和实质优度 (substantive goodness)。前者反映了估计
中所包含的评估者的真实信念以及与概率理论的公理
的符合程度; 而后者反映了估计中所包含的该领域的
知识量。本章所评论的文献, 涉及的是优度的另一个
方面, 称作校准 (calibration)。

307 如果一个人估计某个命题为真的概率是 0.7, 但是后来发现这个命题为假, 那么就其本身而言, 我们并不能否定该估计。但是, 如果对 10 000 个命题的估计均为 0.7, 结果发现只有 25 个命题为真, 那么这些估计就是有问题了。它们所缺乏的性质就是校准。校准也被称作现实性 (Brown & Shuford, 1973)、外部效度 (Brown & Shuford, 1973)、信心的现实性 (Adams & Adams, 1961)、信心的合适度 (Oskamp, 1962)、次级效度 (Murphy & Winkler, 1971) 和信度 (Murphy, 1973)。形式上, 如果赋予所有命题一个指定的概率, 而从长期来看, 命题为真的比例等于所赋予的概率, 那么就可以说判断者被校准了。通过观察他们的概率估计, 证明相关的命题, 然后观察每个反应类别的正确比例, 我们就可以经验地评价判断者的校准。

本章第一部分回顾的实验文献是关于评估者的校准, 他们针对离散命题给出概率判断。第二部分则考虑所估计的非确定数值型变量的概率密度函数的校准。虽然校准本质上是一种个人属性, 但是为了确保稳定的校准估计所需要的大量数据, 这里回顾的大多数研究所描述的, 都是成组的评估者数据。

一、离散命题 (Discrete propositions)

根据所提供的选项的数目, 离散命题可以分为:

无选项型: “什么是苦艾酒?” 评估者给出一个答案, 然后给出他认为这个答案正确的概率。概率回答的范围在 0 到 1 之间, 这是合适的。

单选项型: “苦艾酒是一种珍贵的石头。这种说法为真的概率是多少?” 同样地, 相关的概率尺度的范围是 0 到 1。

二选项型: “苦艾酒是 (a) 一种珍贵的石头; (b) 一种液体。” 当使用半距方法 (half-range method) 时, 评估者首先选择一个更有可能的选项, 然后给出该选项为真的概率 (≥ 0.5)。当使用全距方法 (full-range method) 时, 被试给出他认为事先指定的选项为真的概率

(0 到 1)。

三或多选项型：“苦艾酒是 (a) 一种珍贵的石头；(b) 一种液体；(c) 加勒比海上的一个小岛；(d) ……” 可以使用这个任务的两种变种：(1) 评估者选择一个最有可能的选项，给出该选项正确的概率，对于 k 个选项，要求回答大于等于 $1/k$ ，或者 (2) 评估者给所有选项分配概率，概率在 0 到 1 之间。

就所有这些变种而言，可以使用校准曲线来描述校准。这种曲线是这样得到的：

(1) 收集大量的概率估计，对实验者而言，正确答案已知或者在短期内可知。

(2) 通常在一定范围内（例如，所有在 0.60 和 0.69 之间的估计被分在同一个类别下）分组相似的估计。

(3) 计算每个类别内的正确比例（例如，命题为真或者选项正确的题目的比例）。 308

(4) 对于每个类别，以平均回答为横轴、正确比例为纵轴作图。

如果所有点都落在标识线上，就表明是完美校准。

对于半距任务，不良校准的估计有两种，或者是过度自信的 (overconfident)，即正确比例小于估计概率，因此校准曲线落在标识线下方，或者是信心不足的 (underconfident)，即正确比例大于估计概率，进而校准曲线位于标识线上方。

对于无选项或者单选项的全距任务，过度自信可能有两种含义。评估者对于答案的真实性过度自信；这种过度自信表现为校准曲线总是落在标识线之下。要么就是评估者对于他们辨别命题真假的能力过度自信。这种过度自信表现为校准曲线总是处于标识线上下各 0.5 的区域之内。

现在已经提出了几种数值型的校准测量方法。默菲 (Murphy, 1973) 探讨了有 k 个选项题目的一般情况，以布利尔分数 (Brier Score, 1950) 为

起点,它是一种测量总体优度或者概率估计的一般方法,其数值越小越好。对于 N 道题目,布利尔分数 (Brier Score) 为:

$$B = \frac{1}{N} \sum_{i=1}^N (r_i - c_i)(r_i - c_i)'$$

其中, r_i 是关于题目 i 的 k 个选项的估计概率的向量, $r_i = (r_{1i}, \dots, r_{ki})$, c_i 是相关的输出向量, $c_i = (c_{1i}, \dots, c_{ji}, \dots, c_{ki})$, 当选项为真时, $c_{ji} = 1$, 否则 $c_{ji} = 0$ 。上标符号 (') 表示其前括号中的向量被转置为列向量。默菲指出, Brier 分数能够被拆分为三个可以相加的部分。因此把 N 个回答向量排序为 T 个子集, 使得子集 t 中的所有回答向量 r_t 都相等。令 n_t 为子集 t 中回答的数目, 令 $\bar{c}_t$ 为子集 t 的正确比例向量:

$$\bar{c}_t = (\bar{c}_{1t}, \dots, \bar{c}_{jt}, \dots, \bar{c}_{kt}), \text{ 其中 } \bar{c}_{jt} = \sum_{i=1}^{n_t} c_{ji} / n_t$$

令 $\bar{c}$ 为所有回答的正确比例向量,

$$\bar{c} = (\bar{c}_1, \dots, \bar{c}_j, \dots, \bar{c}_k), \text{ 其中 } \bar{c}_j = \frac{1}{N} \sum_{i=1}^N c_{ji}$$

309 最后, 令 u 为单位向量, 一个行向量, 它的 k 个元素都等于 1。

这样, 默菲把 Brier 分数分解为:

$$B = \bar{c}(u - \bar{c})' + \frac{1}{N} \sum_{t=1}^T n_t(r_t - \bar{c}_t)(r_t - \bar{c}_t)' - \frac{1}{N} \sum_{t=1}^T n_t(\bar{c}_t - \bar{c})(\bar{c}_t - \bar{c})'$$

第一项并非概率估计的函数; 相反, 它反映了 k 个选项中真实事件的相对频次。例如, 假设所有待估的题目都只有两个相同的选项(下雨, 不下雨)。那么公式的第一项就是 N 个题目(或者天)中下雨的基率函数。如果总是(或者从不)下雨, 这一项就等于零。它的最大值 $(k-1)/k$ 表明下雨这一事件出现的最大不确定性。第二项是校准的一个测量尺度, 它是某个类别中的回答数与该类别的正确比例之差的平方的加权平均数。第三项称作解答项, 它反映了评估者把一系列事件排序到子类之中的能力, 子类的正确比例与总体正确比例不同。

默菲的分解设计是为了对相同的事件集(例如, 下雨和不下雨)进行重复预测。当不同题目的选项没有共同意义时(例如, 在多项选择题测试中), 第一项所指的就是正确答案出现的频率, 它出现在第一、第二等选项中的频率相等。

如果每个题目只有一个回答被计分，那么默菲的公式 (Murphy, 1972) 化简为：

$$B' = \bar{c}(1 - \bar{c}) + \frac{1}{N} \sum_{i=1}^T n_i (r_i - \bar{c}_i)^2 - \frac{1}{N} \sum_{i=1}^T n_i (\bar{c}_i - \bar{c})^2$$

其中 $\bar{c}$ 是总体正确比例， $\bar{c}_i$ 是子类 i 的正确比例。当被计分的回答是大于或者等于 0.5 的回答时（如同二选项半距任务），第一项反映了被试选择正确选项的能力，因而可以称作知识。和前面一样，第二项测量校准，第三项是解答项。

亚当斯等人 (Adams & Adams, 1961) 以及奥斯坎普 (Oskamp, 1962) 提出了类似的校准测量方法。不过没有一种校准测量方法区分出过度自信与信心不足。这些测量方法的抽样性质至今不明。

1. 气象学研究

1906 年，西澳大利亚的官方天文学家库克提议，在每个天气预报后面附上一个数字，这个数字“近似地表达了预报员自己赋予这些预报的概率或者权数” (Cooke, 1906b, p. 274)。他报告了 (Cooke, 1906a, 1906b) 根据 1 951 个预报得出的结果。那些他分配了最高程度的概率的预报（“几乎可以确定会被证明”）中，98.5% 是正确的。至于中等程度的概率（“常规可能性”），93.8% 是正确的，而最低程度的概率（“怀疑的”），78.7% 是正确的。

1951 年，威廉姆斯 (Williams) 要求 8 名盐湖城气象局的专业预报员使用 0、0.2、0.4、0.6、0.8 和 1.0 这 6 个数字中的一个估计 1 095 个 12 小时天气预报的降水概率。对于可能性区间的大部分，降水天数的比例低于分配的概率。这可能反映了公开声明中一种相当自然的回避责任的形式。人们更可能去批评一则使他们在下雨天没有带伞的天气预报，而不是使他们在晴空万里的天气携伞外出的天气预报。

默菲等人 (1974) 的研究再现了类似的结果。他们的预报员在得到电脑天气预测系统 (PEATMOS) 的输出结果之前和之后两次估计第二天降水概率。7 188 个估计 (PEATMOS 之前和之后) 与威廉姆斯观察到的一样，表现出对下雨概率的过度估计。

桑德斯 (Sanders, 1958) 收集了 12 635 个预测，他使用 0、0.1、……、

0.9、1.0 这 11 种回答估计一系列二分事件：风向、风速、狂风、温度、云量、最大飞行高度、能见度、降水事件、降水类型和雷暴。这些数据显示，天气预报员的概率估计倾向于超过实际出现的天气事件的比例。^[1] 罗特 (Root, 1962) 报告了一种对称模式的校准，是关于 4 138 个降水预报的。相对于观测频次，当降水的可能性本来就低时，预报员给出的估计概率过低；当降水的可能性本来就高时，预报员给出的估计概率过高。

温克勒 (Winkler) 和默菲记录了美国康涅狄格州首府哈特福德地区全年降水预报的校准曲线。每个预报或者是 6 小时或者是 12 小时的天气预报，并且都是提前 5 到 44 个小时预报。不幸的是，我们不清楚预报员的预测是否包括“微量降雨”（少于 0.01 英寸）这种情况。这些数据被分析了两次。一次我们假设“降水”包括微量降雨，一次假设没有包括。是否包括微量降雨和预报时段一样，对校准有实质性的作用。包括微量降雨的 6 小时预报和不包括微量降雨的 12 小时预报，表现出完美的校准。包括微量降雨的 12 小时预报的校准曲线，位于标识线上方；不包括微量降雨的 6 小时预报的校准曲线，完全位于标识线下方。预报提前期的不同对校准没有影响。

311

自 1965 年以来，国家气象服务局的预报员就一直使用概率术语发布他们关于降水事件的预报。已经发表了对这个庞大数据库的一些部分进行的校准研究 (Murphy & Winkler, 1977a; 美国气象局, 1969)。这些年校准不断改善。图 22—1 给出了截止到 1976 年 6 月的四年中，芝加哥地区 24 859 个降水预报的校准结果。图 22—1 表现出非常优秀的校准；默菲 (1980) 认为近几年的数据甚至更好！他把这种优越表现归功于预报员这些年所掌握的概率估计经验以及这些数据收集自真实的现场工作表现。

2. 早期实验室研究

1957 年，亚当斯描述了使用 11 - 分 (11 - point) 自信量表的被试的校准：被试“被指导在特定的信心水平上，就某件事的预期正确性，以回答百分比的形式表达他的信心水平……在信心水平 p 上给出的回答中，大约 $p\%$ 应该是正确的”。(pp. 432-433)

在亚当斯的任务中，使用视觉记忆测试镜向 10 名被试依次出示 40 个

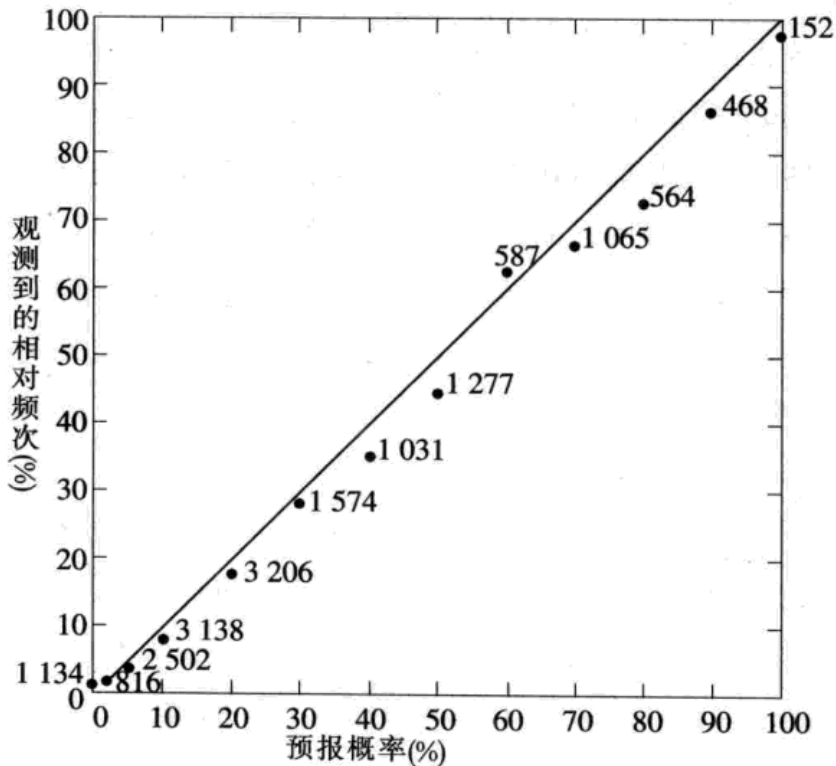


图 22—1 降水预报的校准数据，每个点给出了预报数值

资料来源：Murphy & Winkler, 1977a。

单词，每个单词出示 10 次，每一次出示的亮度逐渐增加。每一次出示之后，被试写下他们自认为所见到的内容，给出一个置信判断。结果得到的校准曲线表明，正确的比例远远超过基于全部回答量表（除了为 100 的回答）的信心等级。在解释这些数据时，必须注意：因为每个单词被出示了 10 次，所以这些回答之间相互依赖很严重。这种相互依赖对校准会产生什么影响，仍不清楚。当他们知晓同一个单词会重复出现几次时，被试可能会选择“收回”先前给出的表述，不愿意给出一个高的回答。

随后的几年中，亚当斯等人（1958）报告了一个训练实验，使用同样的回答量表，但是使用全新的三选项单回答任务：每一节 156 对单词，要求被试回答每一对单词是同义词、反义词，还是无关。一共 5 节，每一节之后被试会被给出校准记录和校准曲线，14 名实验被试的平均校准分数（基于绝对差值 $|r_i - \bar{c}_i|$ ）从第一节到最后一节下降了 48%。6 名控制组被试只有关于他们未被评分的回答记录的反馈，因此他们的差值

分数平均上升了 36%。

亚当斯等人 (1961) 讨论了校准的很多方面 (使用“信心现实性”这个术语), 预先完成了其他人在近年内所做的这些工作, 提供了很多数据, 包括那些自认为是耶稣基督的精神分裂症患者的过度自信校准曲线。在无意义音节学习任务中, 他们发现, 初试存在很严重的自负, 在 16 次试验之后有所改善。他们也简单描述了训练实验的转移现象: 第一天, 被试判断蓝点和红点组成的矩阵中蓝点所占的比例, 共 108 次。第二天和第四天, 被试判断 250 个常识性陈述的真伪。第三天, 他们在被蒙上眼睛之后提起重物。第五天他们判断成对单词 (同义词、反义词, 或者无关词), 共 256 次。8 名实验组被试, 在前 4 天的每一天之后被给出校准反馈, 第五天他们的平均绝对差值分数显著地低于 8 名控制组被试 (没有反馈), 这证明了训练的转移。最后, 亚当斯等人发现, 对于 56 名参加基础心理学课程期末考试 (多项选择题型) 的被试, 不良校准与对失败的惧怕正相关 ($r=0.36$)。而知识和过度自信, 都与对失败的惧怕无关。

奥斯坎普 (1962) 向被试出示了作为刺激物的 200 份明尼苏达多项人格测验 (MMPI) 的人格剖面图。^[2] 其中一半是来自因精神原因而进入老兵管理 (VA) 医院的病人; 另一半来自因纯粹的药物原因进入该院的病人。对于每一份人格剖面图, 被试的任务是判断该病人是属于精神疾病还是药物所致的疾病, 并给出他们认为决策正确的概率。根据一种精确推导的分类系统, 每一份人格剖面图被独立地归到以下三类: 严重 (61 份)、中等程度 (88 份) 和轻微 (51 份), 该分类系统对严重、中等程度和轻微的正确识别率分别为 57%、69% 和 92%。

所有 200 份人格剖面图由 3 组被试加以判断: 28 名心理学专业的本科生、23 名在医院工作的临床心理学实习医生和 21 位经验丰富的临床心理学家。这 28 名经验不足的被试, 后来被分为两个匹配组, 再给他们同样的 200 份人格剖面图。在第二回合期间, 一半人被训练提高精确性, 其余的被训练改善校准。

奥斯坎普使用三个指标测量被试的表现: 准确性 (正确百分比)、自

信水平（平均概率回答）和信心合适度（一个校准分）：

$$\frac{1}{N} \sum_i n_i |r_i - \bar{c}_i|$$

所有 3 个组都倾向于过度自信，尤其是第一节中本科生的表现（准确率 70%，自信水平 0.78）。但是，对于轻微程度的人格剖面图，所有 3 个组都显得信心不足（准确率 87%，自信水平 0.83）。

参加提高精确性训练的被试，他们的精确性从 67% 提高到 73%，信心水平达到 0.78，作为训练的结果，信心水平没有变化。^[3] 参加改善校准水平的被试，他们的信心水平由 0.78 降到 0.74，精确性接近 68%，保持不变。和根据这些变化得出的预期一样，两个组的校准分数都有所提高。

信号检测研究

在早期的信号检测研究中，研究人员观察使用信心评定等级而非是一否反应的概率，以求减少判定稳定的接收器操作特性曲线（ROC）所需要的数据量。在 1 200 次试验中，有学者（Swets et al., 1961）要求 4 名监听人员给出他们接收到信号加噪音而不是只有噪音时的信心水平。尽管 4 个人中有 3 个人被精确地校准，但是 4 条校准曲线却大相径庭。一名被试表现出指派非常小的概率的严重倾向（例如，被试使用回答类别“0.05~0.19”描述一个出现时间超过 70% 的信号）。

克拉克（Clarke, 1960）使用耳机向被试出示 5 个不同单词中的一个，混有噪音。被试选择他们认为自己所听到的单词，然后通过指示 5 个类别中的一个评估他们的信心水平，这 5 个类别根据一个划分为 5 个部分的概率量表定义。75 个单词，12 次练习测试，在每次测试之后，被试为他们自己的结果打分，并记录下每一个评级类别中正确识别的百分比，因而使得被试在下一次测试中能够改变他们的策略。克拉克发现，当数据在 5 个刺激单词中平均以后，所有 5 个被试似乎都被良好校准，但是对每个单词的分析表明，对于简单的单词，被试倾向于过度自信，而对于相对困难的单词则信心不足。

泊拉克等人（Pollack and Decker, 1958）使用文字定义的 6 分等级自信量表，取值从“我听到的信息绝对正确”到“我听到的信息绝对错

误”。使用这个等级量表，虽然不能确定个体是不是被良好校准，但是却可以让我们发现校准是如何随着外部条件变化的。简单单词的校准曲线，一般位于困难单词的上方，无论信号—噪音比（signal-to-noise）如何，并且无论单词难易，高信号—噪音比的校准曲线，位于低信号—噪音比的上方。

在大多数的这些研究中，校准都是次要的兴趣所在；重要的问题是，信心等级是否可以得到与是一非程序一样的 ROC 曲线。直到 1966 年，格林等人（Green & Swets）才得出结论，一般来说，等级量表与是一非程序得到几乎相同的 ROC 曲线。从那以后，对校准的研究，就在信号检测方面的文献中消失了。

3. 近期实验室研究

(1) 过度自信

在近期研究中，最具普遍性的研究发现就是，对于中等程度或者极难的常识性题目，人们显得过于自信。图 22—2 给出了一些典型结果，他们表现出过度的自信。哈扎德等人（Hazard & Peterson, 1973）要求 40 名在国防情报学校学习的军事人员使用概率或者几率回答 50 道二选项常识性题目（例如，“1970 年，哪个杂志的发行量更大，《花花公子》还是《时代周刊》？”）。利希坦斯坦（Lichtenstein, 未发表）发现了类似结果，她的研究使用相同的题目，但是只使用概率回答，她的被试是 19 名俄勒冈研究院的雇员。而另外两位学者（Phillips & Wright, 1977），使用不同题目对英国本科学生进行的研究，也得到类似结果。

其他众多使用常识性题目的研究，也都发现了同样的自负现象（Fischhoff et al., 1977; Koria et al., 1980; Lichtenstein & Fischhoff, 1977, 1980a, 1980b; Nickerson & McGoldrick, 1965）。堪布理奇等人（Cambridge & Shreckengost, 1978）对中央情报局分析人员的研究，也发现了自负现象。费希荷夫等人（1980）使用一系列不可能或者几乎不可能达成的任务 [例如，预测 6 弗隆（6-furlong）赛马的获胜者，诊断溃疡的恶化程度] 进行研究，发现了严重的自负现象。而皮茨（Pitz, 1974）使用全距方法也发现了自负现象。

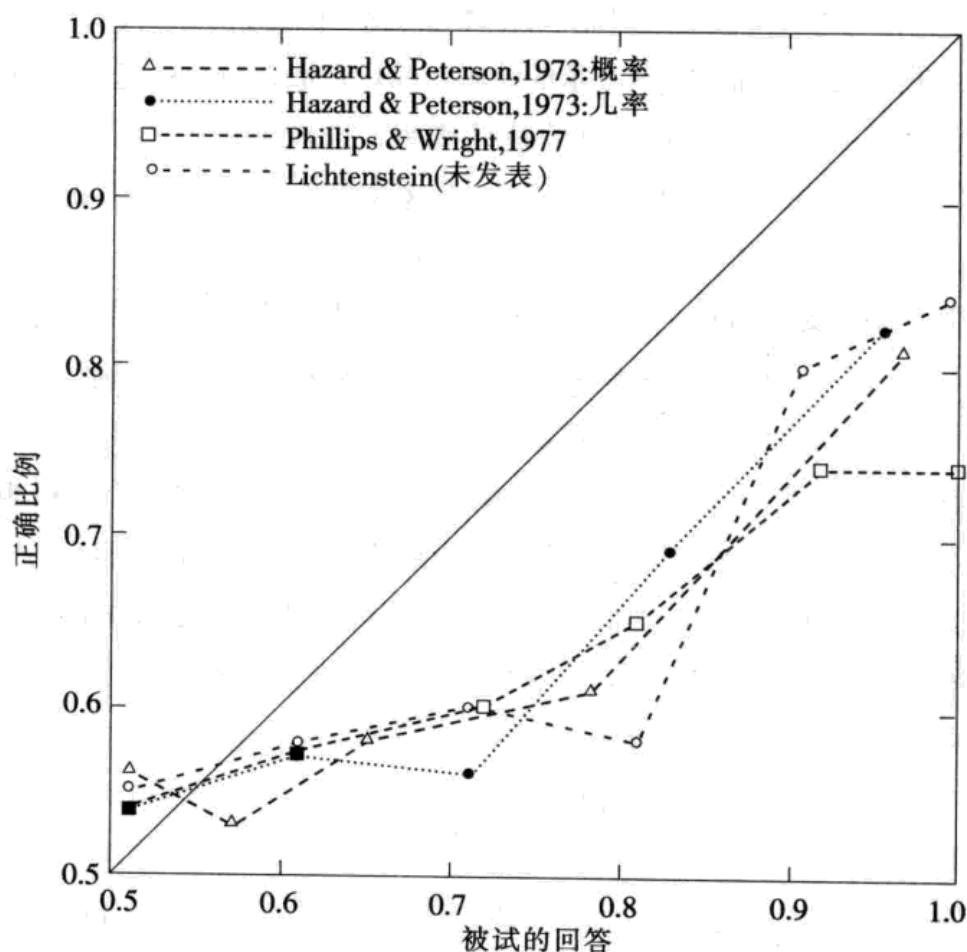


图 22—2 半距常识性题目的校准

费希荷夫等人 (1977) 把注意力集中在确定性表述的恰当性上。他们使用了多种方法 (无选项、单选项, 以及使用半距和全距法的二选项问题) 发现, 回答为 1.0 的题目中, 只有 72% 到 83% 是正确的。在全距任务中, 被指派了另一个极端 (即为 0 的) 回答的题目, 有 20% 到 30% 的时候正确。使用几率回答并不能校正自负现象。被指派几率 1 000 : 1 的回答, 只有 81% 到 88% 是正确的; 而对于几率 1 000 000 : 1 的, 只有 90% 到 96% 的机会选择到正确选项。被试并不抗拒使用极端的几率; 在一个实验中, 几乎有四分之一的回答大于或者等于 1 000 : 1。进一步的分析表明, 极端自负并不局限于少数被试或者少数题目。

(2) 难度效应

难度大的任务, 自负现象最极端 (Clarke, 1960; Nickerson &

316 McGoldrick, 1965; Pitz, 1974)。对于本质上不可能的任务（区分欧洲人和美国人的书法、亚洲和欧洲儿童的绘画作品以及股票价格的起落），校准曲线根本就不会上升；对于所有的估计概率，选择到正确选项的比例接近 0.5（Lichtenstein & Fischhoff, 1977）。在这些任务中，被试并不抗拒使用高的概率，所有回答的 70% 到 80% 都大于 0.5。

随着任务变得简单，自负也下降。在书法识别中，利希坦斯坦等人（1977）允许一组被试在做出概率估计之前，先去学习一组正确标识的刺激样本。这一经历使得任务变得更简单（有 71% 正确，而没有参加学习的小组只有 51%），并且学习组只有轻微程度的自负现象。利希坦斯坦他们（1977）对根据常识性二选项半距任务得到的两大组数据进行事后分析，研究难度对于校准的影响。他们区分了简单问题（那些大多数被试选择了正确选项的题目）和困难问题，有学识的（那些选择了大多数正确选项的被试）和学识不佳的被试。他们发现，当正确率上升时，自负相应地减低。实际上，最有学识的被试，在回答最简单的问题时表现出信心不足（例如，当回答概率为 0.80 时，正确率为 90%）。根据前面被试的表现，选择困难或者容易的题目分配给新的两组被试，在他们身上重复得到这一发现。图 22—3 给出了作为结果的校准曲线和事后分析得到的相应校准曲线。

在刚刚引用的研究中，难度是根据被试的表现定义的（Clarke, 1960; Lichtenstein & Fischhoff, 1977）。最近，受奥斯坎普（1962）的启发，利希坦斯坦等人（1980a）开发出一个由 500 道二选项常识性题目构成的问题集，这些题目的难度是独立界定的。题目分为三种类型：两个城市、州或者国家（例如，洛杉矶和迈阿密），哪一个人口更多？两个城市到第三个城市的距离，哪一个更远（例如，“墨尔本到罗马的距离比到东京更远吗？”）以及哪个历史事件最先发生（例如，大宪章的制定和穆罕默德的诞生）。因而每个题目把难度与两个数字（人口、距离或者距今的时间）联系起来。两个数字中较大的和较小的比率被作为难度的尺度：250 道有最大比率的题目被指定为简单的，其他的被指定为困难的。这种事先分类相当成功，35 名被试简单题目的正确率是 81%，困难题目是 58%。这些结果也表明，被试在困难题目上自负，在简单题目上信心不足。

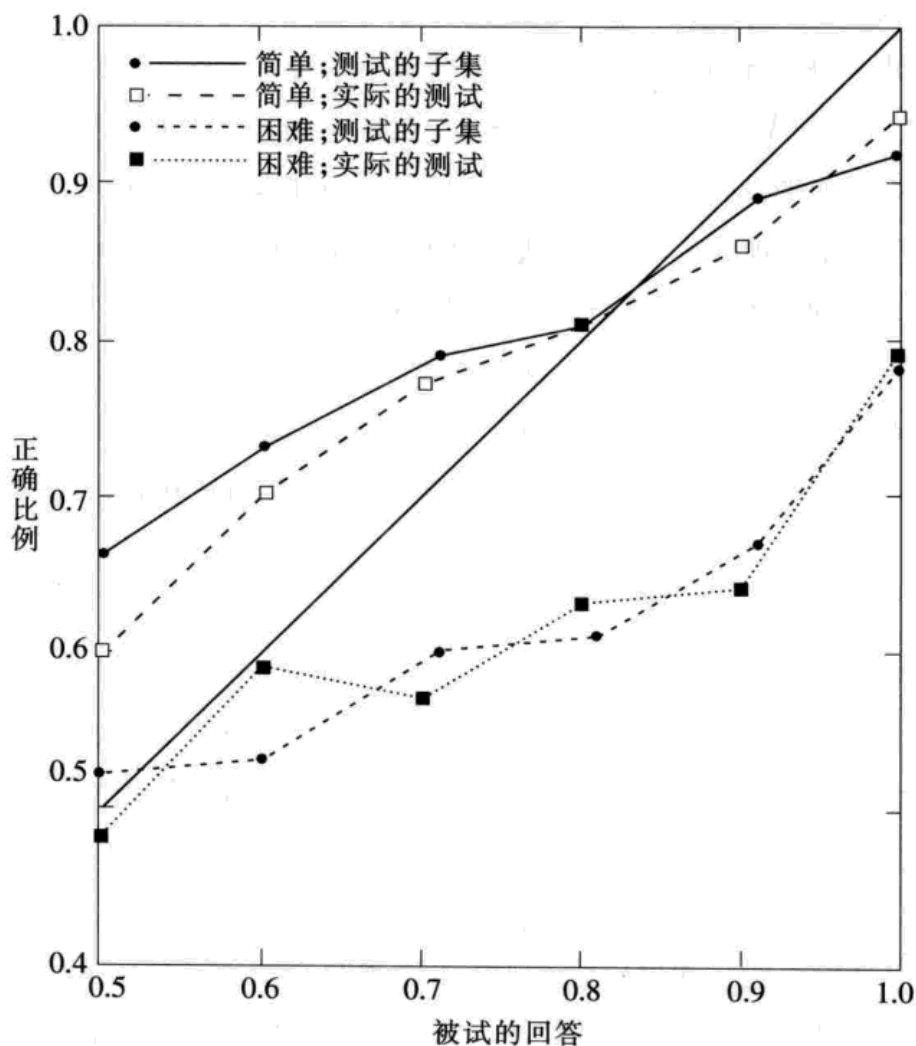


图 22—3 困难、简单测试的校准与一个困难、简单测试子集的校准

这种难—易效应，似乎源自于评估者难以正确估计任务到底有多难或者多简单。菲利普斯等人（未发表）发现，对于所有被试，正确率和使用 11 - 分等级量表给出的刚刚完成的题目的评级之间不存在相关。但是对于不同难度的任务，被试确实给出不同的回答分布；利希坦斯坦等人（1977）发现，在 16 组不同的数据中，正确率和平均回答之间的相关关系为 0.91。但是回答分布之间的差别比预期的要小：对于这 16 组数据，正确比例的变化范围是 0.43 到 0.92，而平均回答的变化范围只是 0.65 到 0.86。

法瑞尔等人（Ferrell & McGoey, 1980）最近开发出一个研究难—易效

应的离散概率估计校准模型。模型基于信号检测理论，它假设评估者把他们的主观不确定性转化为一个决策变量 X ，使用分界值（cutoff values） $\{x_i\}$ 可以拆分 X 。不论何时 X 位于 x_{i-1} 和 x_i 之间，评估者给出一个概率 r_i 。他们假设，如果没有关于校准表现的反馈，当任务难度变化时，评估者不会改变分界组的值 $\{x_i\}$ 。这一假设使得我们预期，对于困难题目，人们会自负，而简单题目人们会信心不足。假设当难度变化时，分界值保持不变，把利希坦斯坦等人的大部分数据应用这一模型，结果表明我们很好地拟合了校准曲线和回答分布。因而，难一易效应，可以看作是人们没有能力改变从对确定性的主观感受到概率回答的转化过程中所涉及的分界值。

(3) 基率效应

单选项判断（真一假）任务，可以根据问题集中真实陈述的比例加以分类。对于一组特定的题目，为了被良好校准，必须把这种基率信息考虑在内。法瑞尔等人（Ferrell & McGoey, 1980）的信号检测模型假设校准会受到两个因素的独立影响：（a）真实陈述的比例，（b）被试辨别真假陈述的能力。假设分界值保持不变，模型预测与可辨别性的变化（同时保持真实陈述的比例不变）相反，真实陈述的比例的变化（同时保持可辨别性不变）会对校准产生完全不同的作用。法瑞尔等人给出数据支持他们的模型。3 门工程学课程的学生估计他们的测验回答会被评分老师判定为正确的概率。事后分析，把被试分为 4 组（高或者低正确回答百分比，以及高或者低的可辨别性），显示出模型所预测的差别。费希荷夫等人收集的未发表数据，也支持这一模型，如图 22—4 所示。4 组被试回答 25 道常识性题目（例如，“埃涅阿斯纪是荷马写的”），这些题目的正确陈述的比例不同：0.08、0.20、0.50 和 0.71。4 个组给出了截然不同的校准曲线，与法瑞尔等人预测的基率变化而可辨别性不变这种情况下得到的曲线形状，大致一致。

(4) 个体差异

基于两个原因，我们很难得出绝对的结论说一个人比另一个人被更好地校准。首先，为了得到稳定的校准测量方法，至少需要几百个回答记录。

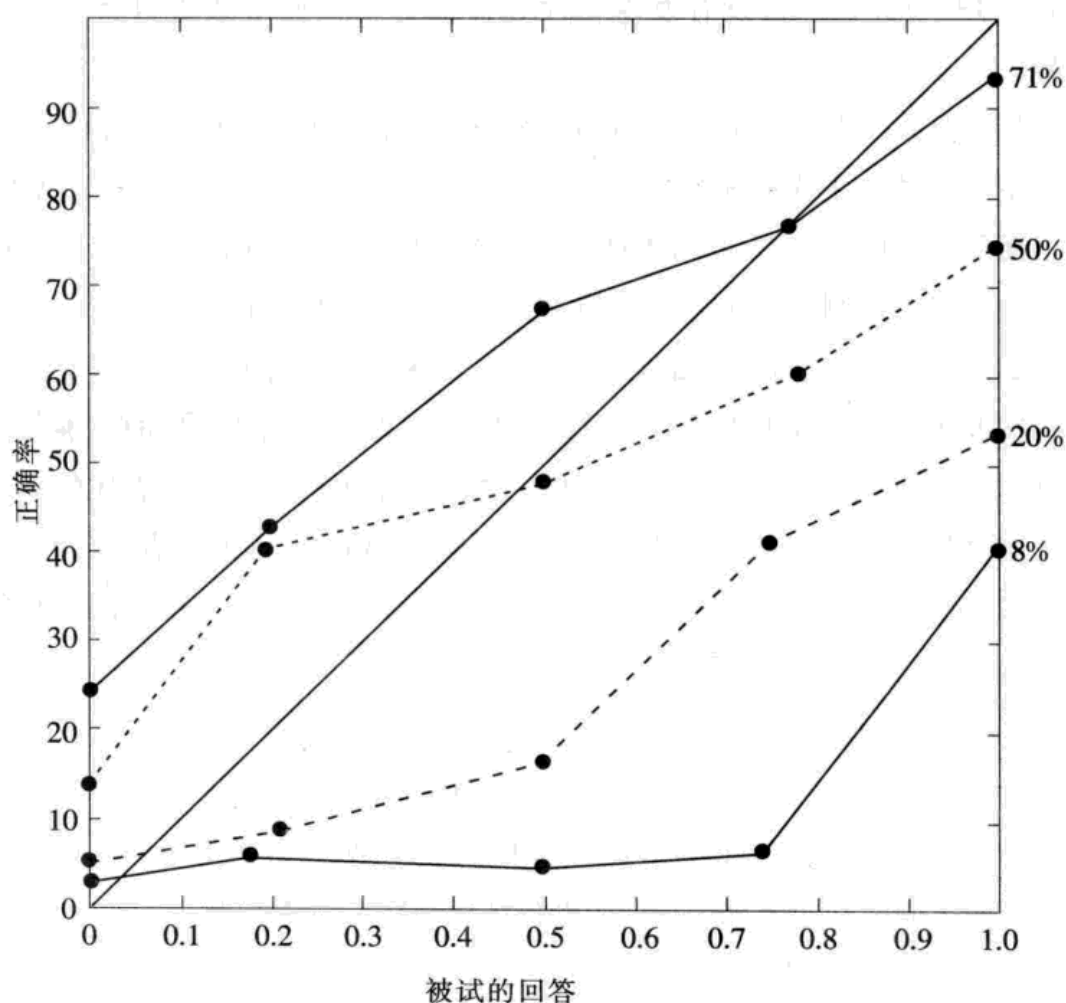


图 22—4 真实陈述的比例的变化对校准的作用

资料来源: Fischhoff & Lichtenstein 未发表。

其次, 校准似乎十分依赖于任务本身, 尤其是任务的难度。实际上, 利希坦斯坦等人 (1980a) 已经证明, 很可能对于每一个人存在一种“理想”测试 (即存在一个测试, 它的难度使得被试既不会过度自信, 也不会信心不足, 因而根据这个测试被试可以被最佳校准), 但是“理想”测试的难度因人而异。因此, 即使一个人比另一个人更好地被校准, 但是对于困难一点或者容易一点的题目, 相反的情况也可能出现。

当难度被控制以后, 不同被试组之间的比较, 通常还是会表现出一些差异。心理学专业的研究生理应比普通被试 (回答校报上刊登的一则广告的本科学士) 更有智慧, 但结果在校准方面没有差异 (Lichtenstein & Fis-

chhoff, 1977)。我们也没有发现男性和女性之间在校准或者自负上存在差异。

怀特等人 (Wright & Phillips, 1976) 研究了几种人格测量 (独裁主义、保守主义、教条主义和不能忍受含糊性)、不确定性的文字表述 (例如, 以书面形式简单回答 45 道问题时所使用的某些单词——如不可能——的次数) 和几种校准测量方法之间的相关关系。他们发现, 在 6 种人格量表和 7 种校准测量方法之间存在的唯一的相关关系, 是独裁主义量表情况下得到的适度相关 (0.41 和 0.34)。对确定回答 (即回答为 1.0) 的校准与对不确定回答的校准 (即回答小于 1.0) 之间不存在相关。文字表述的不确定性测量与任何数值型校准测量方法之间不存在相关。因此作者得出结论, 就人格测量而言, 对于个体差异, 概率思维方式既不是一个单独的影响因素, 也不存在强相关关系。

怀特等 (1978) 还研究了校准中的跨文化差异。图 22—2 给出了他们所选的英国样本的校准。他们的其他样本来自香港、印度尼西亚和马来西亚的学生。亚洲被试组给出的校准曲线, 从本质上看是平坦的。作者猜测, 宿命论倾向的亚洲哲学也许能解释这种差异。

(5) 校正的努力

通过使用明显打击被试自信心的指导语, 费希荷夫等人 (1980) 希望能够在区分亚洲、欧洲儿童绘画作品的任务中消除自负:

所有绘画作品都来自于 Rhoda Kellogg 博士的儿童艺术作品选, 他是不同国家、文化的儿童在绘画方面非常相似这一理论的主要支持者……记住, 做出这种类型区分几乎是不可能的。尽力而为。但是如果在极端情况下, 对于所有这些作品的来源你完全没有把握, 那么请不要犹豫, 每一个都回答 0.5 好了。(p. 792)

这些种指导语使得平均回答下降了大约 0.05, 但是仍然没有消除实质性的自负。

动机的增强会改善校准吗? 有学者 (Sieber, 1974) 针对一组与课程相

关的四选项问题，比较了两组学生的校准。其中一组被告知，这个测试将作为他们的期中考试。另一组则被告知这个测试不是期中考试，但是会用作期中考试前的训练。两组在选择的正确选项的数目上没有差别，但是设想中动机理应更强的一组，即他们的表现将决定他们的成绩的一组，校准更为糟糕（更强的自负）。

通过提供他们的校准的反馈来训练评估者，表现出混合的结果。正如前面所提到的，亚当斯等人（1958）发现，在经过 5 节训练之后，校准有了适度的改善，并且在其后的研究中，得到一些关于训练的一般性结论。有学者（Choo, 1976）使用 75 道二选项常识性问题，训练只有一节，结果发现校准几乎没有改善，也没有任何一般性结论。

利希坦斯坦等人（1980b）训练两组被试，训练是 2 节或者 10 节，由 200 道二选项常识性问题组成，在每一节之后，给出充分的、个性化的校准反馈。他们发现，校准有相当可观的改善，所有的改善都出现在第一节和第二节之间。具有不同难度水平、情境和回答模式（四选项而非二选项）的任务，得到温和的一般性结论。但是，分位数估计任务（fractile assessment task）（后面部分会介绍），或者欧洲人和美洲人书法作品的区分任务，却没有任何改善。

另一种改善校准的方法是重构任务，它以一种方式打击被试的自负。在一项研究（Koriat et al., 1980）中，被试先以通常方式回答 30 道二选项常识性问题，然后再做 10 道附加题。对于这 10 道题中的每一题，他们首先要写出支持或者反驳 2 个可能选项中任一选项的所有理由，然后和通常一样做出选择，并给出概率估计。这个过程显著地改善了他们的校准。另一项研究帮助我们确定这一技术中起作用的因素。在按通常方式回答了最初给出的 30 道题目之后，再给被试 30 道题目。对于每道附加题，他们首先选出自己偏好的选项，然后写出（a）一个支持所选答案的原因，（b）一个反驳所选答案的原因，或者（c）一个支持和一个反驳的原因。然后他们估计所选答案正确的概率。只有被要求写出反驳原因的一组，校准有显著改善。这一结果和根据第一项研究的数据所做的相关分析得出的结论证明，对于自负，一种部分有效的校正方法，就是找出为什么人们会做出错

误选择的原因。

(6) 专家

选修大学某门课程的学生，至少在与其课程相关的问题方面，可以暂时看作是专家。有人 (Sieber, 1974) 报告了参加期中模考的学生 (即该组学生被告知他们所参加的测试并非期中考试) 的优秀校准。他们的回答为 1.0 的有超过 98% 正确，而回答为 0 的只有 0.5% 正确。皮茨 (1974) 要求他的学生预测他们这门课的成绩；这些学生也被良好地校准。

对于难度相同但是不在其专业领域之内的问题，这些学生还会被良好地校准吗？利希坦斯坦等人 (1977) 要求心理学专业的研究生回答 50 道二选项常识性问题以及 50 道和心理学相关的题目 [例如，“Ishihara 测验是 (a) 一种能力测验，(b) 一种社会焦虑测验”]。这两小组题目难度相同，并且两个任务的校准也相似。

另外两个学者 (Christensen-Szalanski & Bushyhead, 1981) 报告了 9 位医生估计 1 531 名因咳嗽而就医的病人患肺炎的可能性。他们的校准非常差，校准曲线上升得如此缓慢，以至于对于最高程度自信 (接近 0.88) 而言，实际患有肺炎的病人的比例也低于 0.20。关于头骨骨折和肺炎的诊断研究 (Lusted, 1977)，以及关于头骨骨折的诊断研究 (Thornbury, 1979)，报告了类似的结果。对医生的田野研究的结果，与天气预报员的降水预测表现出的完美校准完全不同。我们推测有几种因素可能有利于天气预报员。首先，他们已经有多年使用概率语言进行预报的经历。其次，他们的任务是重复性的；需要回答的问题 (天会下雨吗) 总是相同的。与此相反，一个临床医生总是要不停地考虑各种各样的可能性 (是头骨骨折吗？他是链球菌感染吗？他需要进一步的入院治疗吗？)。最后，天气预报员的结果反馈，是良好定义的，并且可以很快得到。而这对于医生来说并不总是如此，病人没有回来复诊，或者另请高明，或者是难以确诊。

那些能够建立赛马胜率的估算方法，或者干脆去赌一赌运气的人，也可以看作是专家。对于同注分彩赌博法 (the pari-mutual or totalizator meth-

od) (或者赛马赌金计算法), 最终的几率决定于下在每匹马身上的赌注, 这使得一种类型组校准曲线的计算成为可能。这些曲线 (Fabricand, 1965; Hoerl & Fallin, 1974) 表现出优秀的校准, 以及人们在风险大的赌注上压更多的赌金这一轻微的倾向。但是, 这些数据仅仅在推理上与概率估计相关。另一学者杜威 (Dowie, 1976) 报告的校准结果更为重要, 他研究了英国体育报刊上每天更新的价格预测。这些预测由一个人以几率的形式给出, 预测指定的一场比赛中所有马匹获胜的几率; 在所研究的这一年中, 大约有 8 个人进行这种预测。对 29 307 匹赛马的预测的校准表明: 对于大于 0.4 的概率, 有轻微的信心不足, 而对于小于 0.4 的概率 (占总体数据的 98%), 得到完美的校准。

校准研究的迅速发展, 导致一种新型专家的成长: 校准专家, 他们对于人们在估计概率时常犯的各种错误了如指掌。利希坦斯坦等人比较了 8 名校准专家、12 名素朴被试和 15 名受过良好校准训练的人。结果是, 正规的校准专家, 不仅克服了素朴被试所易犯的典型的过度自信, 而且有明显的过度补偿, 这是他们信心不足的结果。与其他两组相比, 专家对于题目的难度要稍微敏感一些。

(7) 未来事件

怀特等人 (1979) 推测, 对未来事件的校准, 可能与常识性问题的校准有所不同。若真如此, 这就会限制我们根据常识性问题研究的结论推断预测未来事件。不幸的是, 他们的常识性问题的难度要比未来事件更难, 这可能解释了后者的完美校准。

费希荷夫等人 (1975) 要求 150 名以色列学生估计 15 个那时—未来 (then - future) 事件的概率, 其中包括尼克松总统备受关注的中苏之行的可能结果 (例如, “尼克松总统将与毛泽东至少会晤一次”)。结果得到的校准曲线, 与标识线相当接近。但是, 费希荷夫和利希坦斯坦 (未发表) 最近发现, 和具有可比性难度的常识性问题一样, 未来事件的校准表现出严重的自负。菲利普斯等人 (未发表) 得到 3 组问题的校准曲线: 常识性、未来事件和过去事件 (例如, “过去 30 天内坠毁的一架喷气式飞机, 有超过 100 人遇难”)。所有 3 条曲线都表现出过度自信。未来事件和过去事件

的校准相同，都稍稍好于常识性题目的校准。3组问题的难度水平不能解释这种现象。

杜威 (Dowie) 及其同事现在正在英国开放大学收集几百名选修风险方面课程的学生的校准数据，他们使用与课程相关的问题、常识问题和未来事件问题。学生们被给予有关校准概念的一般性介绍，还会在实验中得到关于他们的表现和校准的反馈。初步结果 (Dowie, 1980) 表明，被试有中等程度的自负。常识问题的校准最好，而与课程相关的问题的校准最差，不过这种差异的显著性和原因都还有待研究。

二、连续命题：不确定变量

1. 分位数法

关于不确定连续变量的值的不确定性 (例如，喜欢苏格兰威士忌胜于波旁威士忌的学生的比例是多少？从英国到澳大利亚的最短距离是多少？)，可以用包含该变量的所有可能取值的概率密度函数表示。但是，通常不会要求评估者给出完整的函数。相反，最常使用的是引出程序 (elicitation procedure)，它是分位数法的一个变种。在这种方法中，评估者给出一些不确定变量的值，这些值与几个事先指定的分布的分位数相对应。例如，对于中位数或者 0.50 分位数，评估者给出一个值，使得真实值落在这个值之上或者之下的可能性相等；0.01 分位数的值，使得真实值有 1/100 的概率小于这个值。通常要估计 3 或 5 个分位数，包括中位数。在这种方法的一个变种三分位数法 (tertile method) 中，评估者给出两个值 (0.33 和 0.67 分位数)，使得整个概率区间被分成相等的三个概率部分。

324

通常报告两种校准测量方法。四分位差指数 (the interquartile index)，是指真实值落在四分位差区间之内的题目所占的百分比 (即在 0.25 到 0.75 分位数之间)。从长期来看，被完美校准的个体的四分位差指数为 50。惊奇指数 (the surprise index)，是指落在两端的估计分位数之外的真实值的百分比。当两端的估计分位数为 0.01 和 0.99 时，完美校准的个体的惊奇指数为 2。较大的惊奇指数，表明评估者的置信区间太狭窄，以至于不能容纳足够的真实值，因而表现为自负 (或者过度准确) (Pitz, 1974)。信心

不足，表现为四分位差指数大于 50，以及较低的惊奇指数；不过目前还没有文献报告过这样的数据。

阿尔泼特等人 (Alpert & Raiffa, 1961, 21) 的论文推动了关于概率密度函数的校准的研究。他们的被试是熟悉决策分析的哈佛商学院的学生。第一组，所有被试估计 5 个分位数，其中 3 个是 0.25、0.50 和 0.75。但是，4 个小组的两端的分位数各不相同：0.01 和 0.99 (A 组)；0.001 和 0.999 (B 组)；“最小可能值”和“最大可能值” (C 组)；“惊人的低”和“惊人的高” (D 组)。表 22—1 给出了这 4 个小组的四分位差指数和惊奇指数。由于为数众多的惊奇值，阿尔泼特等人又使用了 3 个附加组 (2、3 和 4 组)，这 3 个组在估计了 10 个不确定变量之后，得到了以充分的结果报告和解释形式给出的反馈，和一个结束语“延长你两端的分位数！”然后，被试回答 10 个新的不确定变量。表 22—1 给出了反馈前和反馈后的结果。被试虽有改善，但是仍然表现出相当可观自负。

表 22—1 连续变量校准情况总结：真实值落在四分位差区间内和端点分位数之外的比例

	N	观测到的四分 位差指数 ^a	意外指数	
			观测值	理想值
<i>Alpert & Raiffa (1969)</i>				
组 1—A (0.01、0.99)	800		46	2
组 1—B (0.001、0.999)	500		40	0.2
组 1—C (“最小值”及“最大值”)	700		47	?
组 1—D (“惊人的高/低”)	700		38	?
第 2、3、4 组训练前	2 270	34	34	2
训练后	2 270	44	19	2
<i>Hession & McCarthy (1974)</i>	2 035	25	47	2
<i>Selvidge (1975)</i>				
5 个分位数	400	56	10	2
7 个分位数 (包括 0.1、0.9)	520	50	7	2
<i>Mokowitz & Bullers (1978)</i>				
比例				
3 个分位数	120	—	27	2
5 个分位数	145	32	42	2
道-琼斯				

325

不确定状况下的判断

续前表

	N	观测到的四分 位差指数 ^a	意外指数	
			观测值	理想值
3 个分位数	210	—	38	2
5 个分位数	210	20	64	2
<i>Pickhardt & Wallace (1974)</i>				
组 1				
第 1 回合	?	39	32	2
第 5 回合	?	49	20	2
组 2				
第 1 回合	?	30	46	2
第 6 回合	?	45	24	2
<i>T. A. Brown (1973)</i>	414	29	42	2
<i>Lichtenstein & Fischhoff (1980b)</i>				
测试前	924	32	41	2
测试后	924	37	40	2
<i>Seaver, von Winterfeldt, & Edwards (1978)</i>				
分位数	160	42	34	2
几率—分位数	160	53	24	2
概率	180	57	5	2
几率	180	47	5	2
对数几率	140	31	20	2
<i>Schaefer & Borcharding (1973)</i>				
第 1 天, 分位数	396	23	39	2
第 4 天, 分位数	396	38	12	2
第 1 天, 假设样本	396	16	50	2
第 4 天, 假设样本	396	48	6	2
<i>Larson & Reenan (1979)</i>				
“合理地确定”	450	—	42	?
<i>Pratt (1975)</i>				
“惊人的高/低”	175	37	5	?

续前表

	N	观测到的四分 位差指数 ^a	意外指数	
			观测值	理想值
<i>Murphy & Winkler (1974)</i>				
端点是 0.125 & 0.875	132	45	27	25
<i>Murphy & Winkler (1977b)</i>				
端点是 0.125 & 0.875	432	54	21	25
<i>Stael von Holstein (1971a)</i>	1 269	27	30	2

注：N = 被评估的分布的总数。

a 除了 Brown (1973) 以外，所有实验中，落在四分位差区间之内的事件的理想比例是 50%，Brown 得出了 0.30 和 0.70 分位数。因此理想比例是 40%。

另外的研究 (Hession & McCarthy, 1974) 所收集的数据，可以和阿尔泼特等人的第一个实验比较，他们使用 55 个不确定变量，36 名研究生作为被试。他们的指导语鼓励被试确保 0.25 和 0.75 分位数之间的区域确实包含一半的概率。“后来与单个被试的讨论使我们明白：在大多数情况下，这种一致性核查导致重新调节 (readjustment)，减小最初估计的四分位差区间” (p. 7) ——因而，事情变得更糟了！这种指导性的强调 (阿尔泼特没有使用)，或许可以解释为什么他们 (Hession & McCarthy) 的被试被如此糟糕地校准，如表 22—1 所示。

他们 (Hession & McCarthy) 也对其被试进行了许多个体差异测量：权威主义、教条主义、僵化性 (rigidity)、皮提格的范畴外延量表 (Pettigrew's Category-width scale) 和智力。被试的测验分数，与他们的四分位差指数、惊奇指数之间的相关相当低，尽管权威主义量表与四分位差指数的相关为 -0.31，与惊奇指数的相关为 +0.47 ($N = 28$)。这与怀特等人 (Wright & Phillips, 1976) 的发现一致，权威主义量表与校准之间存在适度的相关。

塞尔威奇 (Selvidge, 1975) 扩展了阿尔泼特等人的工作，首先要求被试回答 4 个有关他们自己的问题 (例如，“你是更喜欢苏格兰威士忌而不是波旁威士忌吗?”)，他们的回答决定了这些组内生成比例 (group-generated proportions) 的真实答案 (例如，回答喜欢苏格兰威士忌胜于波旁威士忌的被试的比例是多少?)。一组给出 5 个分位数，0.01、0.25、0.5、0.75 和

0.99。另一组给出这5个分位数，但是加上0.1和0.9两个分位数。如表22—1所示，7—分位数组表现得稍微好一些。而5—分位数组的结果，与阿尔波特等人的结果不像看上去的那么不同。阿尔波特等人的实验中有3个不确定变量，是与塞尔威奇的问题类似的组内生成比例。对于这3道题目，阿尔波特等人发现57%在四分位差区间内，以及20%的惊奇指数。最后，对于其中的一道题目，5—分位数组一半的被试，被要求首先给出0.25、0.5和0.75，然后再给出0.01和0.99；而另一半被试，被要求首先估计两端的分位数值。塞尔威奇发现前一种顺序的惊奇（8%），少于后一种顺序（16%）。

又有两个学者（Moskowitz & Bullers, 1978）也使用了组内生成比例，但是他们比塞尔威奇发现更多的惊奇。其中一个组给出塞尔威奇使用的相同的5个分位数（顺序为0.5、0.25、0.75、0.01和0.99）。另一组只要求3个估计（分布的众数值，0.01和0.99分位数）。在进行估计之前，3—分位数组接受了一次关于一些典型的参照事件的讲座和讨论（例如，“考虑一种彩票，有100人参加，你赢得奖金的可能性是1/100”），这种设计是为了使被试更好地了解概率为0.01的意义。如表22—1所示，3—分位数组比5—分位数组相比，惊奇更少。在另一个使用同样的2种方法的实验中，这两个学者（Moskowitz & Bullers）要求44名贸易专业的本科生估计1977年、1974年、1965年、1960年和1950年道-琼斯工业指数的平均值。每名被试分别在参加3人一组的讨论之前和之后给出估计。由于没有发现任何源于讨论的系统性差异，这些数据被包含在表22—1之中。同样地，3—分位数组（他们接受了什么是概率为0.01的讲座）与5—分位数组相比，惊奇更少。而5—分位数组的表现是最糟糕的。

皮克哈德等人（Pickhard & Wallace, 1974）重复了阿尔波特等人的实验，并加入一些变动因素，结果发现，几个组在反馈前，惊奇指数为38%到48%，反馈之后也不小于30%。两个变动因素，即是否使用课程成绩学分作为良好校准的奖励和是否使用评分规则反馈，它们对惊奇的数目没有影响。皮克哈德等人还研究了扩展训练的效应：分别为18和30人的两组被试（不确定变量的数值没有报告）回答5节和6节题目，每一节之后给

出校准反馈。如表 22—1 所示，发现了一定的改善。

最后，皮克哈德等人 (1974) 研究了在被称为 PROSIM 的生产模拟游戏情境下，知识的增加对校准的作用。在一个模拟的 17 “天” 生产计划期间，32 名研究生被试每人做出 51 个估计。每个估计关于一个从现在开始的 1、2 或者 3 “天” 后将发生的事件。估计时间越接近事件时间，被试对该事件了解得越多。随着这种信息的增加，自负下降：提前 3 天，惊奇为 32%；提前 2 天，惊奇为 24%；提前 1 天，惊奇为 7%。在模拟的 17 “天” 中，没有观察到任何改善。

布朗 (Brown, 1973) 要求 31 名被试估计 14 个不确定变量的 7 个分位数 (0.01、0.10、0.30、0.50、0.70、0.90 和 0.99)。因为每个问题都附有充分的历史数据 (例如，对于 “1970 年 12 月的消费价格指数是多少” 的问题，被试被给出 1962 年 3 月和 1970 年 6 月每个季度的消费价格指数)，所以结果令人特别郁闷，结果如表 22—1 所示。对于其中的 11 个问题，如果被试以历史最小值作为 0.01 分位数，历史最大值作为 0.99 分位数，那么他们应该根本不会感到惊奇。其他 3 个问题，也表现出严格的增 228
长或者衰退的发展趋势，并且真实值与任何一个对历史趋势的简单近似都非常接近。被试一定是严重依赖他们自己的有误的知识给出分布，他们给出的分布如此紧凑，结果导致 42% 的惊奇。

利希坦斯坦 (1980b) 使用 12 名被试，77 个不确定变量，在被试接受关于二选项离散问题校准的充分训练之前和之后，得到 5 个分位数 (0.01、0.25、0.5、0.75 和 0.99)。如表 22—1 所示，被试对不确定变量的校准，没有显著性的改善。

2. 其他方法

西福等 (Seaver et al., 1978) 研究了 5 种不同的回答或反应模式对校准的作用。2 个组使用分位数法，或者是 5 个分位数 (0.01、0.25、0.5、0.75 和 0.99)，或者是与这些分位数相当的几率 (1:99, 1:3, 1:1, 3:1 和 99:1)。其他 3 个组使用概率、几率或者基于对数几率量表的几率回答单选题问题，每个问题指定了不确定变量的一个特定值 (例如，“1973 年，加拿大人口超过 2 500 万的概率是多少?”)。每个不确定变量被给出 5 个这

样的指定值, 研究人员根据这些回答估计四分位差指数和惊奇指数。对于每一种方法, 7 到 9 个学生回答 20 个不确定变量。如表 22—1 所示, 与那些使用分位数法的组相比, 给出概率回答和几率回答的组的惊奇指数明显要好。这种情况的发生, 是否应归因于实验人员选择的数值所传递的信息, 仍不明确。对数几率回答模式的表现也不好。

夏福等人 (Schaefer & Borcharding, 1973) 要求 22 名学生估计 4 节中每一节的 18 个组内发生比。每名被试使用两种估计技术: (a) 分位数法 (0.01、0.125、0.25、0.5、0.75、0.875、0.99), (b) 假设样本法。在后一种方法中, 被试给出假设样本的样本容量 n 和其中成功的样本数 r , 该假设样本最准确地反映了评估者对不确定变量的了解 (例如, 我确定真实值的比例与我在观测到 n 个样本中 r 个成功情况下得到的一样)。较大的 n 值, 表明对于比例的真实值更有把握。比例 r/n 反映了概率密度函数的平均值。被试在这种方法上遇到很大的困难, 尽管指导语中包括了这种方法所隐含的 β 分布的例子。每一节之后, 被试得到充分的反馈, 反馈的重点是他们自己以及所在组的校准。表 22—1 给出了第一节和最后一节的结果。两种方法都有所改善。结果表明, 与分位数法相比, 假设样本法一开始表现较差 (50% 的惊奇, 并且只有 16% 落在四分位差区间之内), 但是结束时表现较好 (6% 的惊奇, 并且 48% 落在四分位差区间之内)。

巴克莱 (Barclay & Peterson, 1973) 比较了三分位数法 (即分位数为 0.33 和 0.67) 和“分点法”, 在“分点法”中, 被试被要求给出不确定变量的众数值, 然后再给出两个值, 一个大于、一个小于众数, 每一个出现的可能性是众数值的一半 (也就是说, 概率密度函数值等于众数值一半的点)。他们使用 10 个年鉴方面的问题作为不确定变量, 70 名国防情报大学的学生作被试内设计中的被试, 发现对于三分位数法, 29% (而非 33%) 的真实答案落在中间区域内。对于分点法, 只有 39% 落在两个半概率点之间, 而对于大多数分布, 接近 75% 的密度落在这两个点之间。

皮茨 (Pitz, 1974) 报告了使用三分位数法得到的几种结果。他发现, 对于估计 23 个国家的人口的 19 名被试, 只有 16% 的真实值落在分布的中间 $1/3$ 区域内。在另一个实验中, 他根据问题的深度和他假设的被试应该

有的知识量改变题目。对于国家的人口问题（知识量低），23% 的真实值，落在中间 1/3 区域之内；对于著名建筑的高度（知识量中等），27% 落在这一范围之内；对于知名人物的年龄（知识量高），47% 落在这一范围之内，最后一种情况高于预期的 33%。在另一项研究中，他要求 6 名被试估计三分位数，并且在几天之后要求他们根据自己的三分位数值选择赌注。他发现，被试严重倾向于选择包含中间区域的赌注，这与他们过于紧凑的区间本应引导他们选择的目标恰好相反。

拉尔森等人 (Larson & Reenom, 1979) 的被试，首先给出其对真实答案的最佳猜测（即众数），然后再给出两个值，使得他们能够“合理地确定”这两个数定义的区间包含正确答案。42% 的真实值落在这个区域之外。请注意，这一惊奇指数与阿尔泼特等人那些被给予文字表述“最小/最大”（47%）和“惊人的高/惊人的低”（38%）的被试的惊奇指数接近。

3. 有专家参与的真实任务

普拉特 (Pratt, 1975) 要求 1 位专家预测一年多内 175 部电影和当地影院上映的 2 部影片的上座率。专家估计了中位数、四分位数、“惊人的高”和“惊人的低”4 项指标值。如表 22—1 所示，四分位差区间倾向于太小。即使在实验过程中，专家得到了结果反馈，但是也只有开始几天里，有证据表明校准随着时间有所改善。

有 3 个实验使用天气预报员作为被试。在 2 个实验中，默菲等人 (Murphy & Winkler, 1974, 1976b) 要求天气预报员给出明天的最高气温的 5 个分位数 (0.125、0.25、0.5、0.75、0.875)。如表 22—1 所示，结果表现出完美的校准。与阿尔泼特等人的被试的惊奇指数大多数落在分布两端的 2% 范围内相比，这些被试落在分布两端的 25% 范围内的惊奇更少。默菲等人发现，2 个实验中使用分位数法的 5 个被试，比之其他 4 个使用固定宽度法的被试，更好地校准。对于固定宽度法，预报员首先估计温度的中位数（即他们认为有 0.5 的概率气温会超过这个值）。然后他们给出在气温会落在中位数 $\pm 5^\circ\text{F}$ 和 $\pm 9^\circ\text{F}$ 范围内的概率。这些预报员是自负的；与落在这个区间之内的气温相关的概率相比，倾向于过高。分位数法相对于固定宽度法的这种优越性，与西福等人的研究发现相反，他们发现定值法更

有优势，这也许是因为默菲等人所使用的固定区间（5F 和 $\pm 9F$ ）不能提供信息。

荷斯坦（Staël Von Holstein, 1971a）使用 3 种定值任务：（a）明天和后天平均气温（把整个回答区间分成 8 类），（b）从今天起的 4 至 5 天的平均气温（分为 8 类），（c）今后 5 天下雨天数之和（分为 4 类）。根据每一组回答（相加和为 1 的 4 个或 8 个概率），他估计隐含的累积密度函数。然后，他综合了由 28 名被试给出的 1 269 个函数。根据文章中给出群累积密度函数（group cumulative density function），我们可以估计出惊奇指数和四分位差指数（见表 22—1）。与其他天气预报员相反，这些被试被相当糟糕地校准，也许因为他们对于这种任务还不熟悉。

4. 小结：不确定变量的校准

根据使用分位数法估计不确定变量的研究，有压倒性的证据证明，人们的概率分布倾向于过于紧凑。对于两端的分位数的估计尤其容易产生偏差。训练使得校准有所改善。专家的表现则时好（Murphy & Winkler, 1974, 1977b）时坏（Pratt, 1975; Staël Von Holstein, 1971a）。有一些证据表明，对于连续命题，难度与校准相关。皮茨（1974）、拉尔森等人（1979）发现了这个效应；而皮克哈德等人（1974）发现在他们的模拟游戏中，提前 1 天比提前 3 天更有利于减少惊奇，这一发现很重要。几项研究（例如，Barclay & Peterson, 1973; Murphy & Winkler, 1974）发现，估计分布的散布程度，与估计的中位数和真实答案之差的绝对值之间，存在相关，这表明被试对于他们的所知和未知有着部分的敏感。这一研究发现，与离散命题情况下正确率和平均回答之间的相关关系相似。

三、讨论

1. 为什么要被良好校准？

为什么概率评估者应该关心是否被良好地校准？有学者（Von Winterfeldt & Edwards, 1973）指出，对于现实生活中绝大多数带有连续性决策选项（例如，投资 X 美元）的决策问题，相当大的决策失误，只会使预期收益发生相对很小的变化。但是，有几种考虑反驳这种可靠的观点。首先，

在二选项决策情境中，奖惩函数在关键部分可能会非常陡峭。假设你的医生必须判断你患有疾病 A ，因此应该接受疗法 A 的概率，和你患有疾病 B 应该接受疗法 B 的概率。假设效用情况是这样的，如果你患有疾病 A 的概率大于等于 0.4，那么疗法 A 更好；反之疗法 B 更好。如果医生估计你患有疾病 A 的概率 $p(A) = 0.45$ ，但是校准很差，使得适宜的概率是 0.25，那么医生就会选用疗法 A 而非 B ，因此你就会损失很大一部分效用。有人 (Fryback, 1974) 给出了正好是这种类型的现实生活中的效用函数。

此外，当奖惩之物非常大，误差也非常大的时候，或者当这种误差被复合的时候，预期损失就会更加严重。例如，在反应堆安全研究中（美国核能管理委员会，1975），“在分析的每个层次上，假设失败率数据的对数正态分布（the log-normal distribution）是根据 0.05 和 0.95 分位数范围定义的。”（Weatherwax, 1975, p. 31）这里所回顾的研究指出，基于 0.05 和 0.95 这两个分位数估计建立起来的分布可能有严重偏差。如果在一个分析的几个层次上都做出这种估计，由于每个估计分布都过于狭窄，那么误差就不会相互抵消，而是被复合。而且由于核电厂失败的成本巨大，这种误差所产生的预期损失也就是天文数字了。

如果说良好的校准很重要，那么怎样才能得到呢？考克斯 (Cox, 1958) 建议，通过使用一个模型拟合一组答案已知的问题的估计，我们可以从外部重新校准人们的估计。从那以后，模型就被用来校正或者调节评估者给出的回答。外部重新校准面临重大的技术性困难。当引出建模需要的估计时，我们一定要注意切不可给予评估者比他们通常所接受到的更多的反馈，以免在测量时他们改变自身的校准。正如萨威奇 (Savage, 1971) 所指出的，“借助经验你可能发现，你的专家可能在某些方面乐观或者悲观，因此缓和了他的判断。但是，他应该这样猜测你吗？你和他可能同在一部通向毁灭的扶梯上。” (p. 796) 此外，由于研究已经指出，所观测到的错误校准的类型依赖于任务的难度，我们也不得不相信，我们的未来事件会符合重新校准所使用的事件的难度。

关于外部重新校准，理论方面的反对意见可能比实践方面更加严厉。一般来说，重新校准程序产生的数字，不会遵从概率理论的公理 [例如，

与互斥完备事件相关的数字之和并不总是等于 1，而且对于独立事件， $P(A) \cdot P(B) = P(A, B)$ 也不总是为真]；因此，这些新的数字不能被称为概率。

一种更为有效的方法，就是训练评估者使得他们被良好地校准。在什么条件下我们可以认为评估者能够达到这个目标呢？当外在的或者隐含的奖励都不足以激发评估者在他们的估计中诚实表现的时候，我们就不应该期望他们会被良好地校准。作为一个极端的例子，一个评估者被威胁，如果任何一件他估计为概率小于 0.25 的事件出现，他就会被斩首，那么对于概率为 0.20 的估计，他就完全有理由认为他不会被良好地校准。尽管这个例子看起来很荒谬，但是更为微妙的压力，诸如“避免被人看成是傻瓜”或者“讨好老板”，也可能为不良校准提供很强的动机。对痴心妄想或者拒绝接受这两种情况的任何奖励，也都能够使估计产生偏差。

在每次估计之后给出结果反馈，是训练成功的最有效的条件。戴伟德 (Dawid, in press) 已经指出，在这种条件下，诚实的、坚定的主观主义评估者会期望被良好地校准，无论要被估计的题目之间是否相互依赖。与此相反，有人 (Kadane, 1980) 则指出，在没有一次次的实验反馈的情况下，只要并且只有所有要被估计的题目之间相互独立，诚实的、坚定的主观主义者才会期望被良好地校准。这一理论严格限制了一些情境，在这些情境中，我们可以合理地期望评估者通过学习会被良好地校准。即使训练能够在只使用评估者认为是相互独立的题目的前提下进行，我们仍然有理由怀疑现实任务的独立性，而评估者会把他们的训练应用到这些任务上。重要的未来事件之间可能相互依赖，这或者是因为它们会受到共同因素的影响，或者是因为评估者利用共同的知识积累评估所有事件。在这种条件下，我们不会希望或者期望被良好地校准。

人们的偏差作为任务难度的函数变化，这种可能性为缺少即时反馈情况下的校准制造了进一步的障碍。未来任务的难度可能是无法预期的，这也使得训练变得没有效果。

2. 作为认知心理学的校准

校准方面的实验也可以被用来研究人们如何思考。即使每个研究的即时

的实践意义仍有局限，但是它们仍然可以帮助我们深入了解人们是如何发展并表达他们对于确定性和不确定性的感受。但是，这里所评论的大部分文献的一个吸引人的方面是，“笨拙的经验主义” (dust-bowl empiricism)。心理学的理论很缺乏，不论是作为研究的启动因素，还是对结果的解释。

并非所有的作者都避免理论上的讨论。斯洛维奇 (Slovic, 1977a) 以及特沃斯基和卡尼曼 (1974, 1) 认为：由于信息处理能力有限，人们总是采用简单化规则或者判断启发式 (heuristics)。尽管一般来说这很有用，但是这些判断启发式也可能导致严重的系统误差。例如，当估计不确定变量时，人们倾向于给出过度紧凑的分布，这可能反映了一种被称作“初始化和调节”的启发式策略。当被问及一个不确定变量时，人们很自然地首先想到一个点估计，例如中位数。然后这个值作为一个初始点。为了给出 0.25 或者 0.75 分位数点，被试在初始点附近浮动。但是这种初始点具有如此强的支配力，以至于此后的调节很难起到作用；因此，分位数点总是离得非常近，导致自负。

皮茨 (1974) 也接受人类信息处理能力和工作记忆能力均受到限制的这种观点。他建议人们序列性地解决一个复杂的问题，一次只完成一个部分。为了减少认知约束，人们会忽略他们前面部分的任务的解决过程中的不确定性，以便减轻后面部分的工作中计算的复杂性。这可能导致过于紧凑的分布和自负。皮茨还提出一条估计不确定性的方法，就是了解他们有多少种不同的方法能够获得答案，也就是说，他们能够建构多少种不同的解答序列。如果可以找到很多，人们就会认清自己面临的不确定性；如果只找到很少一些，他们就做不到这一点。建构不同的可选框架的知识基础越丰富，自负倾向就越弱。

菲利普斯等人 (1977) 提出一种 3 - 阶段序列模型。他们的模型区分了善于运用概率思想考虑不确定性的人和那些以黑—白对立模式做出回答的人。他们还研究了文化和个体方面的差异 (Wright & Phillips, 1976; Wright et al., 1978)，取得了部分成功，他们试图识别出这种类型信息的处理过程中存在的独特认识类型。

而科利亚特等 (Koriat et al., 1980) 也采取了一种信息处理的方法。他

RA 们讨论了概率估计的3个阶段。第一阶段从记忆中寻找相关证据。下一阶段估计这些证据从而得到一种确定感或者怀疑。最后，人们把这种关于确定性的感觉转化为一个数字。科利亚特等人使用的操作设计是为了改变前两个阶段，通过迫使人们寻找和留意相反的证据，因此减轻了他们的自负。

另一方面，法瑞尔等人（1980）的模型纯粹处理第三阶段，把关于确定性的感觉转化为数值型回答。通过假定不使用反馈，人们就不能够改变他们的转化策略，或者是因为题目的难度，或者是因为事件变化的基率，模型给出了较强的预测，这些预测得到了校准数据的支持。

概率估计的结构和过程理论，现在正在开始；我们希望这种理论的进一步发展，会有助于把这一相当专业化的领域整合到更为广阔的认知心理学研究中去。

本章原载 H. Jungermann 和 G. deZeeuw 所编的《人类事务中的决策制定和变迁》（*Decision Making and Change in Human Affairs*. Dordrecht-Holland: D. Reidel, 1977），论文有修改。允许重印。

注释

[1] Raiffa (1969) 的研究，使我们注意参考 Cooke (1960)、Williams (1951) 和 Sanders (1981) 的研究。

[2] MMPI (明尼苏达多项人格测验) 是一种在精神病诊断中广泛使用的人格测验。测验得到的13项分数制成的一幅曲线图就是一个人格剖面图。

[3] MMPI 爱好者可能注意到，这种最小训练的使用，使得本科生表现出与最优秀的专家或者与最优秀的精确预测系统一样高的准确性。



第 23 章

致那些注定要研究过去的人：事后 诸葛亮的启发式和偏差

费希荷夫 (Baruch Fischhoff)

本森 (Benson, 1972) 确认了 4 种研究过去的理由：为了消遣；为了创造群体（或者国家）认同；为了展现人类可能性的极限；为了发展关于我们这个世界的系统的知识，这些知识将可能最终提高我们预测和控制的能力。在一个有意识的层面上，至少我们这些行为科学家将自身限定在最后一个动机上。在这一追求的过程中，我们进行个案研究、项目评估和文献评论。我们甚至进行实验以创造人为的历史，然后对其进行事后的剖析。

在我们回顾过去的过程中，似乎产生了三个基本的问题：(a) 是否存在这样一些可资利用的模式，它们将使我们在将来变得更为明智？(b) 是否有这样的愚蠢的实例，我们可借其确认并避免错误？(c) 如果我们不研究过去，我们是否会真地重蹈覆辙而受人谴责？也就是说，通过回顾过去，我们真地可以学到什么东西吗？

无论我们提出的是什么问题，人们通常假定，过去乐于展现它所拥有的答案。在后知之明或事后诸葛亮（hindsight）和前瞻之见中，后者更显得是一种棘手的视角。人们可以解释和理解任何旧事逸闻，只要他付诸合理的努力。然而，预测，众所周知地要远为复杂。本章通过具体分析一些典型的探求过去的尝试来考察这一推定。也许最一般的结论是，当我们试图获取过去的秘密时，我们应该对过去更为敬重。过去非常轻易地给人以消遣、提升和扩充，但只有用心良苦地奉承，它才会给人启迪。

336 一、寻找智慧

尽管过去从不在细节上重复自己，但它通常被认为包含着重复的成分。人们做出同样的决定，面临同样的挑战，遭受同样的不幸，它们如此之多，以至行为科学家相信他们可从中发现反复出现的模式。这种信念，促使心理测量学家去研究第一流的临床医生的诊断秘密，去研究寻找反常行为之间的相关关系的临床医生，猎求价格上涨前兆的经纪人和寻思爆发革命条件的独裁者。他们的研究，往往有着与多元回归或多元相关类似的逻辑。一系列相关的个案被收集，每位成员都在多种维度上被予以特征化。最后对得到的矩阵进行周密彻底的研究，意图获取重要的相关关系来帮助我们预测未来……

正式模型

比如说，《每日赛马通讯》（*Daily Racing Form*），给最早的预测赛马胜负的人提供大约 100 页的在一场比赛中的每匹赛马的信息。具有数据处理才能的人，可能会借助计算机的记忆，存入一定数量的通信的内容，试图得到一个在不同维度上的分数的加权总值作为预测速度的公式。比如：

$$\bar{Y} = b_1x_1 + b_2x_2 + b_3x_3 \quad (1)$$

这里 $\bar{Y}$ 是我们对马的速度的最优估计， x_1 是在以前的比赛中的胜率， x_2 是赛手的胜率， x_3 是马在本场比赛中所要承受的重量。假定使用的是标准分^[1]，权重（ b_i ）表示的是不同要素的重要性。如果 $b_1 = 2b_2$ ，则既定的马的胜率的变化对速度预测的影响，是相同的赛手胜率变化的影响的两倍，这是因为过去的表现，已证明速度对 x_1 的敏感程度是 x_2 的两倍。

听起来简单，但前面有一千个陷阱。其中一个，就是预测量 x_i 之间是相关的这一情况，它在获胜的马能够吸引获胜的赛手（或与之相反）时，可能会（事实上确实会）发生。在这种多重共线性（multicollinearity）的情况下，每一变量对过去的表现，都有一定的独立解释能力，它们两者又具有一定的共同解释能力。当确定权重时，这一共同的解释能力，必须在两者之间进行分割。一般而言，这一分割，将使研究者无法对 b_i 的精确性做出明确的说明。因此，回归方程，无法被视为可以体现不同要素的重要性的赛马理论。

一个更为适度的理论目标，基于要素在多大程度上增加了我们对 y 的理解，来简单地确定哪个要素是重要的，哪个要素是不重要的。这里的逻辑，是逐步回归（stepwise regression）；只要额外的变量，能够某种程度地增加方程的总体预测（或解释）能力，它们就会被加入方程中。但即便是这样的最低要求的策略，也会遭遇多重共线性的纠缠。如果某一要素的诸多方面（比如饲养的不同方面）都被考虑，则它们的共同的解释能力，会被分割为非常小的部分，以至没有一个方面会成为“重大”的变量。

当然，只要公式足够帮助他们一定程度上战胜随机概率，上面这些微妙的差别，对预测马赛胜负的人就相对没有重要性。但是，我们这些科学家，希望从我们的技术中获得智慧的同时也获得效力。放弃对权重的解释，对我们来说是非常困难的。回归程序不仅仅表达了，而且以一种机械和重复的形式产生了理解（或者至少是结果）。这样，尽管它们存在局限，但科学家仍然对它们孜孜以求是不足为奇的。在对它们的追求中，其中资料保存最为完备的一种，是在对临床判断的研究中。临床判断的做出，可能是一位放射学家要将溃疡的 X 光分为“良性”和“恶性”的，或者是一位人事官员要从一组候选者中选出最佳人选，或者是急救中心的咨询员判断威胁要自杀的来电者中谁是比较严重的。在上面这些例子中，诊断涉及的，是在一组线索或者特征的基础上做出判断。在这些例子中，当判断是重复的，而且所有个案都可以用相同的线索予以描述刻画的时候，我们可能以统计的方式对判断者的决策进行模型建构（建模）。人们可以收集一组专家已经做出总结判断的个案（如良性的、严重的），然后得到回归方

程，如方程（1），它的权重，将表现出判断者对每一要素所赋予的重要性。

20 多年的这种政策—主导的研究（policy-capturing studies），始终产生出这样一组令人不安的结论：（a）使用线索的加权总和的简单线性模型，在预测判断者的判断上是非常成功的，虽然说（b）判断者声称他们使用的是远远要复杂的策略（L. R. Goldberg, 1968b, 1970; Slovic & Lichtenstein, 1971）。一种被普遍提及的复杂判断形式，被称为完型判断（configural judgment），其中，一种线索的诊断含义，取决于其他线索的含义（比如，“这种声音的音调，让我觉得‘不是自杀型的’，除非说电话是在早晨的前几个小时打进来的”）。

研究者在后续的研究中，找到了测量的和报告判断程序有冲突的两个理由，它们每一个，都对“抓住”过去判断中的智慧的回归模型，有负面的含义。其中之一，已被人们日益认识到，即这类公式要求一个人所要储存的庞杂的信息，超过任何一个人的计算能力，除非他是白痴天才（idiot savant）。一位试图采用复杂策略的判断者，显然不能保持很好的一致性。事实上，当存在大量的线索或者在线索和被预测的变量之间存在特殊的关系时，要学习和使用即便是非完型的、加权总和的判断程序，也是很困难的（Slovic, 1974）。

第二个认识，源于对临床判断的研究，即简单的线性模型，是非常强大的预测手段（见本书第 28 章）。一个能够体现出人们做出判断时会注意什么变量的实质理论，可以完全满足想很好地预测他们行为的人的需要。如果有些迹象（sign）鼓励一种判断或诊断，而另一些迹象阻碍它，则只要简单地计算一下鼓励和阻碍的迹象的数目，就可以对个体的行为提供非常好的估计。然而，结果却是比那些华而不实的模型更为适度的理论（Fischhoff et al., in press）。因此，虽然过去似乎就放在那儿等待着我们去理解，但标准的统计程序，并不总能告诉我们所想要知道的。如果使用不细心，它们可能会误导我们，让我们比起初更不明智。我们总是被诱惑着完全去信奉那些高度复杂的理论，却没有意识到，它们的能力是源于非常简单的隐含概念，而不是它们抓住了什么过去的精髓。

二、寻找蠢事

1. 关注失败

在历史事件中寻找智慧，需要一种信念——相信存在这样的反复出现的模式，它们等待我们去发现。而在历史人物的行为中寻找智慧的人，需要某种不同的信念——相信我们的先辈知道一些我们不知道的东西。第一个信念扎根于哲学，它有别于那些将历史学视为社会科学而不是对独一无二的事件进行的特例研究（ideographic study）。第二个信念是基于宽容和谦卑。它使那些希望站在先人的肩膀上而看得更远的人，有别于那些满足于眼下视野的人。诸如“那些不研究过去的人是会被诅咒而重复它的人”的格言，意味着有关我们先辈智慧的信念，相对较少。

积极寻找蠢事（folly），当然不是没有可取之处。不仅仅是那些没有做对事情的人，需要经常做出解释，而且此类解释，对从他们的经历中进行学习，也是至关重要的。通过分析事情为什么会出错，我们希望在将来把它们办得更好。去寻求需要进行解释的不幸，是丝毫不难的。人们的眼球、记者和历史学家，都被混乱和无序吸引着。没有发生事故的驾驶或者是没有战争、压制或者地震的领土，对他们来说，都太平淡了。

尽管它有合理的目的，但关注失败，很可能会因为创造出厄运泛滥的扭曲的世界而误导我们。事件所觉知的可能性，部分取决于它们能被想象或者被回忆的容易程度（Tversky & Kahneman, 1973, 11）。因此，喋喋不休地探讨失败，将过分地强化他们所感知的失败在过去（可能还有将来）的发生频率。

它也很可能促成对我们先辈的功业进行有失均衡的评估。我们每个人身上的喜好搜集和传播丑闻的部分，总是被欺诈社会福利的故事或者对某一特定的环境危害的“过度管制”（例如，就业人员安全与健康管理委员会对工作场所的马桶座圈设置的声名狼藉的标准）吸引着。然而，我们却倾向于忘记，任何错误但凡不是十恶不赦的决策体系，都会产生这两类错误。在每一个囤积非法所得的欺诈者的身后，都伴随着一个、数个或者是一部分的被欺骗者，他们的权利，被同一不完备的体系所拒绝。事实上，这两种错误的比率，是以某种非直觉的方式联系在一起的，它取决于判断的准

确度和可供使用的总资源，也就是说，合格的贫困人口的比例或者需要处理的危害的数量（Einhorn, 1978）。在我们气势汹汹地批判福利系统容忍欺诈者之前，我们必须考虑，如果给定职责的错误和遗漏的错误的比率，是否并没有太少的此类骇人听闻的故事。

总而言之，当我们孤立地去考察基于比例而得到的决策和判断时，我们很有可能会被误导。

2. 问题是什么？

我们还有这样一些情境，当较小情境下的错误放置在较大的情境中考虑时，人们将要换种目光对其评价。比如说，我们被告知，一旦存在不一致的证据，科学理论就应该接受灭亡的命运，而让别的科学理论从身上碾过去。因此，对那些不顾被“证伪”而依然坚持其理论的科学家的愚蠢行径，我们总是急于去谴责。然而，库恩论证道，这些局部的蠢事与寻求科学知识的更为全局的智慧是一致的。事实上，其他人（如 Feyerabend, 1975; Lakatos, 1970）已经高度评价了学科的自由放任在增进理解中的作用，并对严格恪守某一受人青睐的研究方法来获取智慧的可能性表示质疑。他们认为，顽固地拒绝去看相反的证据或者是摒弃显然被驳斥了的理论，对科学的发展通常都是必要的。

340 在平托车（Pinto）这个案例中，向福特汽车公司索取 1.25 万亿美元
的金额，使得公司在燃料水槽的设计上节省几美元的决定，显得像件蠢事。但是，根据纯粹经济的计算，一笔有保证的节约——比如说是 1 000 万辆平托车的每一辆都节约 15 美元——这就使得即使面临几宗大的诉讼风险，也是更为合理的赌博。根据诉讼书（appeal），这宗广为人知的案子判决的金额下降到 600 万美元，因此按照严格的经济计算，福特公司是赚了钱的，尽管情况是越来越糟。福特公司的错误在于，考虑了一个较大的情境（它可以节约费用的车辆的数目），而没有考虑另外的情境（这一决定的非经济的后果）。福特公司似乎没有考虑到它对福特作为有安全意识的汽车制造者的形象的负面公众影响，以及对使用了的平托车的价格的影响（尽管这一价格是由 Pinto 的所有者而不是生产者承担的）。

如果说责备（reprobation）是这一游戏的名字，则错误还是错误。但

是，如果人们想要从他人的经历中学习东西，判断出他们试图要解决的问题是很重要的。通过仔细的考察，许多明显的错误，被证明是对错误问题的巧妙解决方式。比如说，如果要对福特进行批判的话，福特必须对其战术上的聪明和战略上的愚蠢（或者是将组织的利益置于社会公益之上）感到负罪。

这一区分不仅仅对评估过去是重要的，而且对了解在将来应该采取什么样的正确措施也是重要的。通常来说，战术上的错误，相比策略上的误解，要更容易纠正。一旦我们合理地勾勒出一种情形，那里就会有一本“书”，它记录了在试误过程中累积起来的惯常性智慧，或至少是以最佳方式包含了我们掌握信息的公式（Hexter, 1971）。比如说，棒球总教练可能会知道，在一场实力相当的比赛里，为了无人出局而第一垒有跑垒员的局面时，让击球员击出牺牲打，对比赛很有帮助；或者是根据所需的统计学计算出如何“跟着比率跑”。但是，如果真正要解决的问题是要维持士气（跑垒员在盗垒中有机会领先）或者是有利于球市的（球迷想看一些更丰富多彩的东西），这些指南是没有帮助的，或者是误导性的。对国际关系中的突然袭击的研究表明，被袭击的国家往往根据他们自己的一本“书”做了很好的准备，但他们搞错了竞技场（Ben Zvi, 1976; Lanir, 1978）。也就是说，他们阅读的是一本错误的书；他们阅读得越好，他们面临灭顶之灾就越快。

战略问题引起困难的一个原因是，人们必须对它予以分析性的“彻底思考”，而且是在没有累积经验的帮助之下。第二个制约是，在决策群体内或者说决策圈内，往往广泛共享着一些误解。一个人只有完成了“书”中的“问题集”后，人们才会在决策问题上向他咨询。为那些意图避免意外的组织反复提供的经验有：（a）设立几个独立的分析部门，这样可以对同一个问题提供多元的和独立的视角，或者是（b）指定一位成员，扮演“恶魔的倡导者”的角色，他提出一些不受人欢迎的观点（Janis, 1972）。在实际生活中，因为共享的误解将使一个组的观点与另一组的观点类似，这只会造成冗余而不会带来多元视角，这样的话，第一种策略就会失败（Chan, 1979）。第二种策略也可能失败，这是因为“倡导者”或者是迫于

341 群体的压力，或者是把他们不受欢迎的观点看得太认真而被排斥出去——即使这些“极端”的观点，并没有严重挑战群体的先入之见。

没能对战术决策和战略决策进行区分，也会导致本不该有的智慧的幻觉。银行和保险公司通常被认为在决策过程中极端理性和老练。但仔细察看就会发现，这一声誉，主要来自高度重复的战术决策，在这些决策中，它们是不太可能会失误的。家庭信贷和人寿保险政策的制定，是建立在对大量试误经验中得到的统计表格进行的保守表述基础上的。那些闯入更具投机性的决策领地中——它们需要分析性的战略决策——的机构表明，它们并不比我们当中的其他人更聪明。在20世纪60年代对房地产（real estate）信托的不明智投资中，商业银行损失了大笔钱。在20世纪70年代中，它们所有决策中的微不足道的一部分，使美国经济与第三世界的一些只有半偿还能力的国家的未来连在了一起，它向这些国家提供了大笔贷款（超过600亿美元）（尽管这一联系有长期的人道价值，但这并不是银行必然要解决的问题）。保险公司对伤病保险经济中的变化做出的缓慢和错误的反应，以及处理许多非常规的风险中的近乎危险的、非分析性的做法，让我们这些人在和那些洋洋自得的机构进行比较时，不会感到太愚蠢。

3. 后知之明：往回考虑？

如果我们知道发生了什么，或者一个人以前一直要解决的问题是什么，我们就处于一个可以挖掘我们的后知之明来解释和评价他（她）的行为的位置。然而，通过仔细考察，知道事情的来龙去脉的优势，可能被吹捧过高了（Fischhoff, 1975）。在后知之明中，人们始终夸大了在前瞻之见中本该预料到的东西。他们不仅将已经发生的视为是不可避免的，他们也认为在这些事件发生之前，就已经显得“相对不可避免”了。人们相信，其他人本来应该比现实当中对事件有更好的预测。他们甚至记错了他们自己的预测，因而在后知之明中夸大了他们在前瞻之见中就知晓的东西（Fischhoff & Beyth, 1975）。

正如一位历史学家（Georges Florovsky, 1969）所描述的：

决定论的倾向，在回顾方法本身中就有所暗含。在回顾中，我们

似乎察觉出一种事件的逻辑，这一逻辑按照一种被声称具有内在必然性的可识别的模式，以规则的或者是线性的方式将这些事件一一展开。因此，我们得到这样一个印象，即它确实不可能以其他方式发生。（p. 369）

对这种将报告了的结果视为本来就是相对不可避免的倾向，一个合适的名字就是蠕动决定论（creeping determinism）；而与之相对的则是哲学决定论，其有意识的信念是，存在的就是合理的。

一个推定的倾向，就是遥控（telescope）历史进程的速率，夸大那些“不可避免”的变化累积的速度（Fischer, 1970）。比如说，人们也许能够指出大庄园主消亡的历史时刻，但却没有意识到，他们是在 250 年中慢慢消失的。另一个倾向，是在回忆中将人们想得要更像现在的他们自己，而不是他们本来的面目（Yarrow et al., 1970）。第三个倾向，可以在一位历史学家（Barraclough, 1972）对法西斯主义的意识形态根基的历史编纂法的批判中看到。从第三帝国往回看，人们可以将它的根源追溯到许多作者的著作中，而他们的著作是不可能谋划过法西斯主义的。第四个是揣测一个历史情境中的参与者会非常清楚地意识到它的历史重要性（《珍贵的日记，百年大战从今日始》，Fischer, 1970）。第五个倾向，是判别性实验（critical experiment）的神话，它毫不含糊地解决了两种理论之间的矛盾，或者确立起其中一种理论的有效性。事实上，“判别性实验被视为关键，只是数十年后的事情。理论并不会放弃，因为一定的异常和偏差总是被允许的。的确，打败由天才的和富有想象力的科学家支撑的研究计划是非常困难的。”（Lakatos, 1970, pp. 157-158）

在短期内，没能忽略自己对结果有了解这一点，将带来实质性的利益。让自己相信，或者让别人相信，我们知晓事情的一切发展——这是只有在对结果有了解的情况下才能达到的，也就是说，我们拥有如后知之明一般的前瞻之见，这是多么令人愉悦的事情啊。然而，在长期内，未察觉的蠕动决定论，将严重损害我们判断过去及从中学习的能力。

试想这样一些决策者，他们被事情的突发性转变弄了个措手不及，于

是试图通过重新构造出他们在事件发生前的心智状态，来查明他们在哪里出错了。如果在回顾中，事件明显很可能发生，他们也别无他法，只能责备自己当初没有采取行动，而这种行动是他们的知识本来应该让他们去执行的。人们会说，他们给事件本身业已造成的伤害又添加了一笔屈辱。当一位带着后知之明的旁观者给出事后的忠告时，他们的不幸，就成为自身能力不济和愚蠢的代名词，甚至更糟。

在信息不完全和不能确定的情况下，偶尔的意外和相应的失败是不可避免的。去谴责在一个错误的体系中犯了错误的决策者，而矢口否认体系的错误并投身于改造这个体系的行动当中，这既是不公平的，也是自拆墙脚的举动。按照另一位历史学家（Wohlstetter, 1962）的说法，美国在珍珠港袭击中吸取的教训应该是，我们必须“接受不确定性的事实，并学会在其中生存。因为没有什么魔力能够提供确定性，我们的计划必须不借助它也能实施”（p. 401）。

SAS 当我们试图理解过去的事件时，我们是在间接地检测那些我们用来表达和预期周遭世界的假设或者规则。如果我们在后知之明中系统地低估了过去曾经和正在给我们带来的意外的话，我们是将这些假设置于非常弱的假设之下，也因而不会去改变它们。因此，恰恰是那些让我们觉得我们对过去已非常理解的有关结果的知识，阻碍着我们从过去中汲取营养。

避免这种偏差，需要对它的产生所涉及的心理过程有所了解。当我们获得有关结果的知识时，我们似乎很快将其纳入我们对事件已经有了的了解中，从而弄清它的意义。进行这一重新表述后，已被报告的结果，多少就显得是被重新表述的情形中不可避免的发展结果。这样，“弄清”我们被告知的过去的“含义”是如此自然，以致我们可能没有意识到结果的知识已经对我们产生了影响。即使我们知道它对我们产生了影响，我们也不知道这个影响具体是什么。在我们重新构建我们前瞻之见时的心智状态时，我们依然固着在后知之明的视角里，而把报告的结果看得非常可能发生。

因此，仅仅告诫人们后知之明的偏差的危害，是没有多大用处的（Fischhoff, 1977b）。一个更为有效的办法，是迫使他对报告的结果的不可避免性提出质疑，也就是说，要让他相信，结果有可能以别的方式出现。

对你用来解释结果不可避免的理由的有效性提出质疑，这可能是个很好的起点（Koriat et al., 1980; Slovic & Fischhoff, 1977）。由于即使这一非常规的步骤也不是完全充分的，所以我们需要进一步地从过去事件的原初形态中寻觅到它周遭中的不确定性。是否有信息的手抄本，在 12 月 7 日上午 7 点前到达珍珠港的指挥官手里？在你处理沃尔瑟姆工业这一股票前，是否有一本笔记本记载着你考虑了的股票？有没有日记记下了张伯伦在 1939 年对希特勒的印象？一个有趣的特例就是，在弗里曼（Douglas Freeman）撰写的有关李（Robert E. Lee）的权威性传记中，当他在写任何一个既定时期时，他都坚决不去了解随后发生的事情（Commager, 1965）。虽然这是令人钦佩的，但这一策略，需要一些与人们对那些在 Appomattox 投降的人的普遍知识相关的天真的假设。

三、寻找一切

1. 为什么寻找？

对过去的研究秉承着这样的信念，即如果我们在其中寻找，我们就可以发现一些可被表述的模式。大量研究表明，这一信念是根深蒂固的。人们似乎具有非凡的在哪怕是随机产生的数据中寻找一些秩序或含义的能力。最为熟悉的一个例子就是赌徒谬误。我们总有这种感觉，当抛掷一枚均质硬币时，如果连续 4 次都是“头像”，则接下来的应该是“背面”（Lindman & Edwards, 1961）。因此，在我们的脑海里，即使是随机的过程，也会受制于有序的内部特征。卡尼曼和特沃斯基（1972b, 3）的研究已表明，6 个二元事件存在 32 种可能的序列，但只有 1 个真正看起来像是“随机”的。

尽管赌徒谬论通常是在引人入胜但却是琐碎的例子的语境中被引用，但在更为严肃地对历史事件的解释的尝试中也能发现。譬如，在非常巧妙地说明最高法院的空缺多多少少是以一种随机（按照泊松过程）的方式——在任何给定年内，至少有一个空缺的概率是 0.39——出现后，莫里森（Morrison, 1977）声称：

从他在白宫中的第五个年头后不久，罗斯福总统就于1937年2月宣布了扩充最高法院的计划。同时，1937年也是他对最高法院首次做出任命的一年。他在1937年有这样一个机会是丝毫不奇怪的，因为他连续五年不任命一位或是多位法官的概率仅仅是0.08，或者说是1/12的机会。换句话说，当罗斯福决定通过增加额外的席位来改变最高法院时，在当年他能够以传统的方式任命一位或多位法官的几率已经是11比1了。(pp. 143-144)

然而，如果空缺确实是随机出现的话，则这一推导就是错误的。它假设制造空缺的概率过程是和抛掷硬币一样，是有记忆和公正意识的，它好像知道，进入罗斯福当政的第五个年头，它“欠了”FDR^①一个空缺席位。然而，在1937年1月1日，过去的四年已成为历史，而该年至少有一个空缺的概率还是0.39 (Fischhoff, 1978)。

费乐 (Feller, 1968) 提供了下面一则逸闻，它的赌注更高：闪电战中的伦敦人，花了大量心血来破解德国人的轰炸模式，并发展出非常完备的关于德国人的轰炸目标（以及什么时候轰炸）的理论。然而，当伦敦被分割为相邻的地理区域小块时，每一区域的轰炸次数的频率分布，基本上是泊松分布的一个完美近似。凯茨 (Kates, 1962) 也提到，自然灾害构成了另外一类重大事件，在这些事件中，（被威胁的）普通老百姓看到了秩序，而专家看到的则是随机性。

固守这样的信念的一个秘密在于，我们没有占有完全充分的记录来迫使我们自己直面不规则性。历史学家承认，缺少的证据有利于他们的解释，他们有诸如此类的评论：“维多利亚时代的历史将不再会被人写。我们对它了解得太多了。众所周知，无知是历史学家的第一需要——无知以至高的艺术亦即无法企及的不动声色的完美方式，进行简化和疏清，拣选和删除。” (Strachey, 1918, 序言)

即使当记录是可被利用而且不可绕开时，我们也似乎有着对任何我们

① 罗斯福的英文缩写。——译者注

见到的东西进行解释或者是提供因果阐释的非凡能力。当事件是由具有直觉特征的概率过程产生的时候，随机变异，对我们来说甚至不是一个可能的假设。比如说，因为表现糟糕而受到惩罚的运动员倾向于在下一次表现时有所提高，这一事实正符合我们的奖励和惩罚的天真理论。这一唾手可得的解释，蒙蔽了我们的双眼，让我们没有看到，提高是由于向那些运动员的平均表现回归造成的这一可能性（Furby, 1973; Kahneman & Tversky, 1973, 4）。

法玛（Fama, 1965）强有力地论证说，对股票市场波动的最好理解是，它反映了一个随机游走（random walk）的过程。然而，随机游走相比于与之在形式上相关联的二元过程，有着更加非直觉的特征（Carlsson, 1972）。因此，我们会发现，市场分析员对任何一个价格的变化都能给出解释，无论是有目的还是没目的的。有些解释，如图 23—1 给出的，是不一致的^[2]；另外一些解释，似乎否认了任何随机组成部分的可能性，比如说，那个最后的可以胡说八道的要素——“技术性调整”。

我们给出的解释，其解释力的徒有其名，在与回归分析的类比中就可以表现出来。给定一组事件和足够大或者说足够丰富的可能的解释因素，你可以得到你想要的任何严格度（degree of tightness）的“马后炮”解释。按照回归的说法，通过扩大自变量的数目，我们总能找到与自变量有任何你想要的相关关系的这样一组预测指标（predictor）。你为过度拟合（overfitting）付出的代价，当然就是解释力的降低，你得到的规则，应用到新的个案样本上时将会失效。方法论中针对过度拟合而给出的不断的和有力的告诫表明，在相关关系上的矫枉过正，是非常难以对付的偏差，即便对受过广泛的专业训练的人士也是如此（参见 Fischhoff & Slovic, 1980）。

一个过度拟合的理论，就犹如一套被如此精确地裁剪的衣服，它只在某一个人以某一种特定的姿势站立时才合身，它不适合于其他任何人，甚至不适合于将来的这个人，乃至不适合于现在这个人，只要他的姿态发生一定的改变（比如说，他呼出了一口气，然后露出了大肚子）。一位历史学家在解释布尔什维克为什么会取得胜利时，构建了一个可以对现有的所有线索给出说明的滴水不漏的理论，但如果苏联公布出一些被封存的档案，

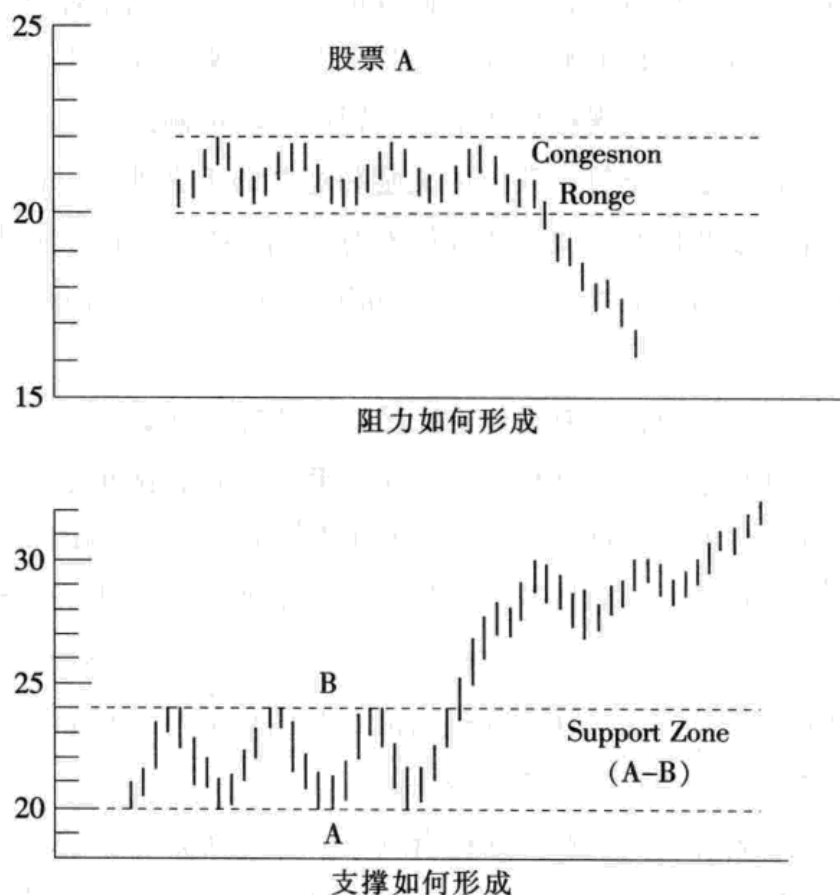


图 23—1 两例用于识别股市价格的变动前兆的线索：阻力的形成与支撑的形成。但是，有人或许会争论说，在它们各自的结尾处的巨幅变动之前，这两种模式基本是一样的，如此看来，这种波动模式既不能作出预测，也无法解释这些数据中的任何东西。

而且这些档案说明孟什维克相比于人们过去的想象是更为危险的敌对力量时，这位历史学家就处于一个非常沮丧的境地。投资分析员为过度拟合付出的代价是，在长时期内，他的预测能力并不会好于市场中的普通人 (Dreman, 1979) ——尽管愤世嫉俗的人会说，他们实际上就是靠不断地产生希望（还有佣金）来吃饭的。^[3]

过度拟合的产生是由于人们对偶尔的波动加以利用。如果计量是充分完备的，两个在某一变量上有差别的事件，也基本上会在任何其他变量上也有所差别。因此，人们可以在这两个变量之间计算出非零（在这个情况中将是完美的）的相关系数，从而获得一个“有趣的”实质理论。由我们

给出的可能的解释概念而构成的 n 维空间中，在其中进行任意的 m 次观察，这一过程与二维事件的操作过程是一样的。

在这些例子中，数据是固定的和不可否认的，而可能的解释组成的集合却相对是无限的；一个人不断寻找，直到找到一个恰好中意的解释。另外一种对偶然性加以利用的普遍方式是，解释的集合是固定的（通常集中在一个候选者上），然后对数据进行筛选，直到支持的证据找到为止。虽然这种比较明显的程序形式是广为人知的，但其他的一些形式更为微妙甚至有些隐晦。比如说，你做一个实验，但没有得到预想的结果。你在思考过程中，注意到在实验程序的一个部分中，你没有消除操作变量的影响。你纠正过来了；但是还是没有结果，只是又发现一个问题。最后，你（或者你的被试）把它纠正过来，然后得到了预想的效应。现在，对第 n 次的样本（它的结果是显著的）或者是所有的样本进行统计检验是合理的吗？如果你第一次就正确地操作了实验，这些问题可能就不会产生。或者是这样一种情况，作为一名毒物学者，你“确信”接触化学物 X 对人的健康是有害的，于是你对加工含该物质的作物的工人和不加工该作物的工人患膀胱癌的情况进行对比，但没有结果。于是你又尝试了肠癌、肺气肿和晕眩等，最后终于发现在皮肤癌上存在显著的差异。这一差异是有意义的吗？当然，检验这些解释或者理论是通过在新的样本上进行复制实验。但不幸的是，这一步骤很少被采用，而且由于技术和伦理的因素经常是不可能的（Turkey, 1977）。

甚至在理论固定、数据也固定的情况下，相关的困扰也会产生。有学者（Diaconis, 1978）发现，在超感知觉（ESP）的结果中，对惊奇的分量进行估量，是很难的事情，即便是在适度控制的情境中获得的少有的个案中，因为被寻找的事件的定义始终在变化。“神学士成功的一个关键要素在于，他事先不给出明确的将被视为惊奇的结果。与某一些结果相符合的几率要显著地小于任何事先明确指出的特定的一种结果的几率。”（p. 132）^[4]

有学者（Tufte & Sun, 1975）发现，是否存在领导者选区（bellwether precincts）取决于定义该事件所允许的创造性和灵活性（是什么政府？在什么选举中？怎样好才算好？始终被遗漏的选区包括进去了吗？）。它们通

常被认为是存在的，这是因为我们有一种不同寻常的能力，可以在哪怕是彻底的噪声（noise）中发现信号（signal）。

2. 我们寻找够了吗？

考虑到我们几乎确信能够在研究过去当中发现一些可被表述的东西，我们接下来的问题就是，“我们理解它了吗？”我们前面描述的后知之后的研究表明，我们不仅很快地去发现秩序，而且倾向于认为，我们始终都以某种方式了解它，或者说，如果我们被及时地询问，我们能够预测结果。事实上，轻易地低估我们被告知的东西中的信息量，使下面这个问题有点令人惊讶，即我们是否向过去或者其他渠道询问许多问题。这一倾向因为下面的一些倾向而愈趋严重：（a）没有意识到我们了解的或者是被告知的有多少，这让我们不知道在寻找奇怪的答案时该问什么问题（Fischhoff et al., 1977, 1978）；（b）要从哪怕是少量的不可靠的数据中获取影响深远的结论（Kahneman & Tversky, 1973, 4; Tversky & Kahneman, 1971, 2）。

浅尝辄止的品性，也受到将历史写成好故事的规范的怂恿，它要求对所有相关的细节都予以清楚地解释，将事件在结局之前把围绕它的所有不确定性，以及作者感到的任何混乱，都统统埋掉（Gallie, 1964; Nowell-Smith, 1970）。这些策略其中的一个秘密，如坦尼（Tawney, 1961）所揭露的那样：“历史学家为某一存在着的秩序披上不可避免的面纱，其做法是把已经取得胜利的力量放置在显著的位置上，而把耗尽了的力量置于后台。”^[5]（p. 177）

尽管在直觉上这是非常吸引人的目标，但构建连贯的叙事，将把读者置于一些有趣的偏差悬崖边。一个完成了的叙事，是由一系列的某种各自独立的连接组成的，其中每一个都是被很好地构筑了的。叙事的真实性取决于这些连接的真实性和真实性。一般而言，连接越多，每个连接中的细节就越多，故事作为一个整体的真实可能性就越小。然而，斯洛维奇等人（Slovic et al., 1976）发现，对事件的描述添加细节，显著地强化了主题的一致性，从而会增加该事件被感知的发生可能性。而巴-希勒尔（Bar-Hillel, 1973）则发现，人们一贯地夸大了一系列事件组合的发生可能性。比如说，她的

被试要在两种情形中做出选择，一种情形是 7 个独立的事件且每个事件的发生概率均为 0.90，如果该事件发生，被试将获得奖品；另一情形是一枚均质硬币抛掷出“头像”时，被试将获得同样的奖品，她的被试普遍选择前者。组合事件的概率小于 0.50，而单一事件的概率是 0.50。换句话说，不确定性似乎是以一种过慢的速度积累着。

如果序列包括的是一个少量的弱的或者是不太可能的连接时，那将会发生什么呢？最弱的连接的概率，为整个叙事的概率设定了一个上限。然而，连贯性的判断，则是具补偿性的，它让强连接的连贯性“平滑掉”弱连接的不连贯性。这一效应被律师加以利用，他们在总结陈词的开始一段时间内，就把弱连接掩盖下去，而以一阵令人信服、难以辩驳的论证结束。

有人 (Coles, 1973) 提供了一个令人异常兴奋的例子，其中，故事的整体连贯性，掩盖了他的连接的不可能性：这就是弗洛伊德在心理历史上最为认真的一次尝试，是他写的达·芬奇的传记。在几年的时间内，弗洛伊德都在寻找理解达·芬奇的秘密，后者的童年和青年时期几乎是无人知晓的。后来，他发现达·芬奇的一个参照，它是关于达·芬奇在摇篮时一只秃鹫碰到他的嘴唇这一反复出现的回忆。弗洛伊德注意到在埃及象形文字中“秃鹫”和“母亲”相等及附属的一些证据，然后他开始发展出一套令人叹为观止的连贯的对达·芬奇的分析。然而，当编辑在整理弗洛伊德的著作的定稿时，他发现弗洛伊德使用的达·芬奇文选的德文翻译本（原文是意大利文）有一个错误，是风筝而不是秃鹫碰到了他的嘴唇。尽管弗洛伊德的分析的关键被破坏了，但这位编辑认为留下的大厦足够牢固而可以独自挺立。正如有学者 (Hexter, 1971) 观察到的，“书写好的历史是比较罕见的，这部分是由于书写坏的历史是非常简单的。” (p. 59)

四、结论

除了知道理解过去比我们通常认为的要更为艰深这一事实之外，在对过去的研究中，我们还能学到什么呢？

1. 厚今说 (presentism)

不可避免地，我们都是我们现在的个人视角的俘虏。我们知道生活在

过去的人们所不知道的东西。我们使用的分析类别（如封建主义、百年大战）只有在回顾中才有意义（E. A. R. Brown, 1974）。当我们对过去进行解释时，我们会去证明自己的观点，而过去从来就不是充分明晰，以至可以避免我们自己的意识形态视角的入侵（Degler, 1976）。历史学家确实是“在每一代的死者身上把玩着新的把戏”（Becker, 1935）。

我们还没有已被验证有效的针对厚今说的解药。在对过去进行事后揣测中如何避免后知之明的偏差的讨论中，我们可以得到一些局部的药方。其他的一些药方，几乎在任何致力于培养历史学家的文本中都可以找到。最为普遍的几点可能是下面这些：（a）尽可能地了解我们自己和现在；“最了解他自己处境的历史学家，也是最能超越这一情境的历史学家”（Benedetto Croce, 引自 Carr, 1961, p. 44）；（b）对我们的先辈尽量要仁慈；“历史学家不是法官，更不是一位判决死刑的法官”（Knowles, 引自 Marwick, 1970, p. 101）。

2. 方法决定论 (methodism)

除了我们自己的时代所设下的不可逃离的牢笼外，我们也自愿地戴上眼罩，严格遵从着单一的科学方法，从而进一步束缚了我们的视角。即便这种科学方法的使用是非常明智的，但也没有任何一种方法足够解决我们关于过去的诸多问题。每一种方法都告诉了我们一些东西，但也在某种程度上误导了我们。当我们不知道如何得到一个问题的正确答案时，另一种认识论就是亟须的：尽可能使用广泛的技术或者视角，每一种视角都能让我们避免某类错误。这意味着一种学科间的合作和互相尊重，它不同于大多数混合两种方法的尝试中所碰到的情形。搭配或者搭配不当，如心理历史学，往往是那些对他们所选定的领地中的陷阱不敏感的倡导者所尝试的（Fischhoff, in press-b）。有学者（Hexter, 1971）将进行此类冒险的历史学家描述为“自命不凡地跳上即将沉没的船只上的老鼠”。

3. 学习

回到本森（1972）所说的，如果我们想让过去服务于未来，我们就不能孤立地看待它。我们用来解释过去的规则，必须也是我们用来预测未来的规则。在积累我们的经验的时候，我们必须仔细留意所有与我们的假设

相关的检验。这种措施，第一是要收集所有将被系统地予以统计分析的记录。第二是要铭记在我们自己的决定之前的深思熟虑，要意识到现在很快就会变成过去，而保存良好的记录，将是后知之明的偏差的最好解药。第三，做出的预测必须是可以被评估的。从三里岛核事故中，我们吸取的一个令人不安的教训是，我们尚未完全清楚，这一表面上可被诊断的事件，让我们知道了多少核反应堆安全的研究——它试图评估核能的风险——的有效性 [美国核监察委员会 (U. S. Nuclear Regulatory Commission), 1975]。第四，我们对自己感觉到的信心的合理性要有更好的认识，因为对现有知识的信心控制着我们对新信息和表述的寻求 (Fischhoff et al., 1977)。因此，我们要规划我们的生活来促进学习。

4. 含糊性

最后，我们提出的许多问题可能是没有答案的。有些问题本身，就是被错误地构造出的。其他的问题，用现有的或者可能的工具是无法解答的。虽然我们都想尽可能地了解“职业人士是怎么处理的”，但现实地说，没有一种统计方法可以以理想的准确度对专家的判断方式进行模型化。我们的理论，通常是“如此的复杂，以致没有哪个单一的数量工作 (quantitative work) 可以检验它的有效性” (O’Leary et al., 1974, p. 228)。当我们想在一个变量上对小组进行比较，而它们在其他变量上也各不一样时，我们没有在逻辑上合理的程序让它们在那个讨厌的变量上相等 (Meehl, 1970)。当我们对一组固定的数据尝试了诸多的可能的解释后，我们没有牢靠的方法去了解我们用了多少自由度，我们在多大程度上利用了偶然性 (Campbell, 1975)。当我们使用多种方法时，它们所产生的知识从来都无法很好地融合在一起。最后，我们也许要采纳 Trevelyan 的哲学观点：“对历史的数次不完美的阅读，总胜过于从不阅读。” (引自 Marwick, 1970, p. 57)

本章是《致那些为避免重蹈覆辙而研究过去的人：对历史性判断的反思》一文的改编，原文出自 R. A. Shweder 和 D. W. Fiske 主编的《行为科学方法论的新方向：行为研究中的错误判断》 (New Directions for Methodology

of Behavioral Science: Fallible Judgment in Behavioral Research. San Francisco: Jossey-Bass, 1980)。允许重印。

注释

[1] 为使某一特定的变量的分数标准化，我们可以从每一分数中减去所有分数的均值，然后除以标准差。得到的结果是一组均值为0、标准差为1的分数。

[2] 我最感兴趣的一个反差是，当市场在一个利好经济消息之后上涨时，它被说成是对该消息做出的反应；如果市场下跌了，又可以说好消息的影响已大打折扣来对此做出解释。

[3] 一位朋友曾经在当地的一位经纪人那里参加了一门如何阅读图表的课程。每一节都会讲授10~12个新的线索。当课程结束时，讲完5节和57个线索，讲授者还远远没有传授完他的东西。

[4] Diaconis 继续说：“使分析更为复杂的是，通常同时进行着数个这样的定义不完备的实验，它们之间相互作用。年轻的表演者首先让他的观众变得兴奋起来。他频繁的完全错误的猜测带来的只是同情，而不是怀疑；而且对大多数观看者来说，他们似乎只会去确证神学士具有不同寻常的能力的真实性。”

[5] 这种策略，主观地强化了等级秩序的力量和稳定性，而降低了那些改变秩序的外观上的能力（apparent capacity），这既会影响精神也会影响心智（Markovic, 1970）。



第七部分
PART SEVEN

多阶段评估



第 24 章

序列选择中复合概率的估计

科恩等 (John Cohen, E. I. Chesnick, and D. Haran)

经常发生这样的情境，其中包括几个或多或少相互独立的阶段，而一个成功的结果 (outcome)，依赖于在每一个阶段都能够独立地做出一个正确的选择。一个例子是在旅途的各个阶段选择交通方式。类似的困境，出现在职业、管理、政治和军事生活中，也通常出现在通信网络中。几次选择的时间顺序，并不总是一个关键的因素。并且，由于从一个阶段到另一个阶段，正确选择的概率可能会发生变化，那么情境作为一个整体，可能具有一种随机性质 (a stochastic character)。

我们将描述一个实验，它模拟了这种情境的一般类型，显然，对于一个成功的结果，它需要对概率作乘法运算。关于在一个 $m \times n$ 的队列中，或者一个划分为同心区域的显示屏上定位的偏好研究指出，在主观上，队列或者同心区域中的单元 (cells) 与目标所在的位置并不是等概率的 (Cohen et al., 1969)。似

乎, 被试并不是随机地猜测未知的目标位置, 相反, 他们寻找策略的特点, 是“推测”实验者藏匿目标的地方, 或者在他们期望目标会出现的地方寻找目标。其他的实验 (Cohen & Hansel, 1958) 暗示, 当面对涉及复合概率 (compound probability) 的情境时, 许多人倾向于把不同阶段的概率相加, 而不是相乘。

相应地, 我们可以预测, 对于一个 $m \times n$ 的队列, 其中 m 行的每一行中有一个未知的目标, 那么, 被试会夸大他自己正确猜测出全部目标的概率, 如果按照目标定位的复合概率 n^{-m} 来判定。我们的实验设计, 就是为了检验这一预测, 并且阐释这种情境将产生的现象。

实验的装置是一块板子, 上面设置了 m 行, 每一行有 n 个小盒子, m 和 n 各自取值 2、3、4、5 和 8。所有的小盒子都是空的, 除了每一行中有一个小盒子盛有一张票之外。因此, 独立阶段的数目是从 2 到 8, 相应地, 在每一阶段正确猜测出的概率是从 2^{-1} 到 2^{-3} 。

被试是 50 名 14 岁~15 岁的中学男生, 他们的智力可以假定为至少与成年人的平均水平相当。告知被试, 为了获得奖品, 他必须猜出每一行中票的正确位置, 每一行只允许猜测一次。他的任务, 等同于从一组彩票中抽一次而中奖的机会。

一共有 14 种抽奖方式。其中 10 种抽奖方式中, 每一种包括 100 张彩票 (tickets), 根据特定的抽奖方式中含有的获奖彩票的数目, 抽中一张获奖彩票的概率分别为 0.01、0.1、0.2、0.3、……、0.9。剩下的 4 种抽奖方式分别包括 500、1 000、5 000 和 10 000 张彩票, 抽中一张获奖彩票的概率分别为 0.002、0.001、0.000 2 和 0.000 1。

行数和小盒子数的取值范围容许 25 种不同的情境。利用随机化的拉丁方阵, 50 名被试每人被随机地指派其中的 5 种情境, 这样, 10 名被试可以判断全部的 $m \times n$ 种情境。

定位目标实际所需要的信息量是 $m \log_2 n$ 比特 (bits), 其中 m 是行的数目, 行, 我们现在称之为“阶段”; 而 n 是小盒子的数目, 我们称之为“选项” (alternatives)。对于被试选择的抽奖方式, 我们把抽中一张获奖彩票的概率, 看作是对被试获奖的心理学概率 (psychological probability) Ψ

的一个间接估计。

表 24—1 给出的结果表明, 如复合概率 p 所判定的, 所有的 Ψ 值都被高估了, 因而证实了我们的预测。表 24—2 中 Ψ 对 p 的比率指出了高估的量值 (magnitude of overestimate), 它表明, 一种忽略阶段的数目而重视每一阶段的选项数目的明显倾向, 干扰了实事求是的估计。方差分析显示, 与剩余方差相比, 阶段间和选项间的方差, 都很显著 ($p < 0.01$)。

表 24—1 基于抽奖方式选择的心理概率 (Ψ)

每一阶段中 选项的数目	队列中阶段的数目					M
	2	3	4	5	8	
2	0.45	0.51	0.36	0.38	0.37	0.41
3	0.41	0.21	0.21	0.21	0.17	0.24
4	0.31	0.23	0.21	0.12	0.11	0.19
5	0.30	0.22	0.17	0.18	0.08	0.19
8	0.07	0.16	0.11	0.13	0.05	0.11
M	0.31	0.27	0.21	0.20	0.16	

注: 表中数字代表 10 次观察的平均值。

表 24—2 心理概率与复合概率的比率 Ψ/p

每一阶段选 项的数目	队列中阶段的数目				
	2	3	4	5	8
2	1.8	4.0	5.7	12.2	95.0
3	3.8	5.9	17.8	53.2	10^3
4	4.9	14.8	54.0	118.0	7×10^3
5	7.5	27.8	106.0	563.0	3×10^4
8	4.3	82.4	465.0	4×10^3	9×10^5

如果我们以相对高估的值的对数 $\log_{10} \Psi/p$ 为纵轴, 以定位目标所需要的信息量 $m \log_2 n$ (以“比特”计) 为横轴作图, 如图 24—1, 那么这种差分效应 (differential effect) 就会变得明显。关系是线性的, 由下面的方程给出:

$$\log_{10} \Psi/p = 0.26 m \log_2 n - 0.31 \quad (1)$$

方程可以改写为:

$$\Psi/p = e^{(0.86 m \log_2 n - 0.72)} \quad (2)$$

根据方程我们得到, 如果我们使阶段的数目为常数, 那么 Ψ/p 与每一

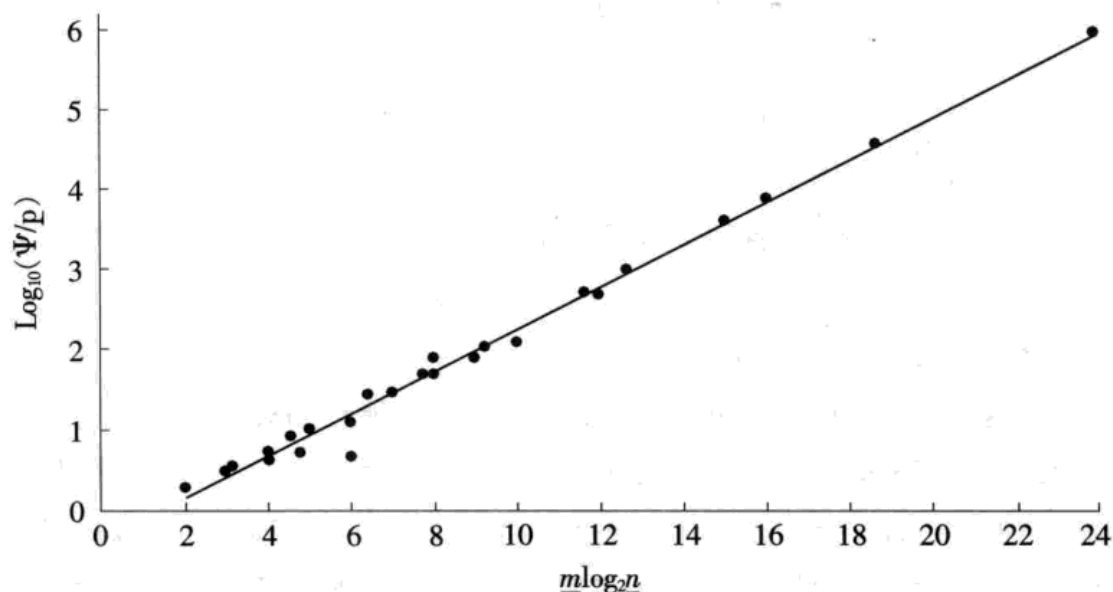


图 24—1 纵轴为相对高估值的对数 $[\log_{10} (\Psi/p)]$ ，横轴为信息的数目 $(m \log_2 n)$ 。 p = 复合概率； m = 阶段数； n = 选项数。

个阶段选项的数目的乘方成正比，而如果我们使选项的数目为常数，那么 Ψ/p 随阶段的数目成指数变化。

给定 m 和 n ，我们可以预测 Ψ 的均值，因为

$$p = n^{-m} \quad (3)$$

把 m 个阶段中各个阶段正确猜测出的概率相乘就得到了复合概率 p ，每一阶段正确猜测出的概率都等于 $1/n$ 。

因此，从 (2)

$$\Psi = e^{-0.72} n^{-0.14m} \quad (4)$$

($e^{-0.72} = 0.49$)

因此，总的来说，对 Ψ/p 的相对高估，是由于对倍乘系数 (multiplicative factor) m ——阶段的数目——的主观弱化 (subjective attenuation)。在我们的实验条件下，这种弱化是阶段数目的 $6/7$ 。这使得我们推断，复合概率中的倍乘元素，绝不是“原始的”或者直觉的，这可以帮助解释人们在统计学的学习中所遇到的特殊困难。

我们的解释，得到了类似实验的支持，这些实验使用的被试年龄有的 9 岁多，有的 10 岁多。实验中，除了对 Ψ 的相对估计严重偏高之外，没有

发现任何伴随 m 和 n 的值变化的趋势，这也证明倍乘元素不是原始的。

我们使用的方法涉及对 Ψ 的间接估计。通过要求被试在不同类型的队列中间进行选择，能够得到更加直接的估计。但是，那时，选择的效用可能成为一个重要的因素。

这个实验阐明了在各种多阶段选择情境中的一种明显倾向，决策者错误地判断了他自己成功的可能性，因此采取了一种不合适的策略，对此他以后会感到后悔。

在这方面，值得注意的历史事实是，尽管可能性的观念引起了古希腊最敏锐的思想家的极大兴趣，尤其是在斯多噶学派的哲学中，但他们从来没有掌握过组合分析的方法 (combinatorial analysis)，这种方法的发展必须要等到 16 世纪。很明显，亚里士多德自己对于可能性的概念只有很少的兴趣。无论他以及其他可能了解多少主体的直觉，长期确立的思考习惯还是淹没了它。

这个实验所揭示的相对高估复合概率的现象，可能是决策与选择领域内一个相当普遍的现象。如果是这样，那么值得特别指明。我们建议把它命名为“心理概率的惯性效应” (inertial Ψ effect)。

本章发表于《自然》(Nature, 1971, 32, 414-416)。允许重印。

第 25 章

人类信息处理中的保守主义

爱德华兹 (Ward Edwards)

大量研究表明,对于不准确的信息 (fallible information) 359,人类是一个保守的处理者。这些实验比较了人类行为和贝叶斯定理的结果。贝叶斯定理是关于如何根据新信息修正意见 (也就是说,概率) 的形式上最优的规则 (the formally optimal rule)。结果表明,意见的变化非常有规律,通常与贝叶斯定理计算出的数值成比例。但是这并不够。一种方便的、基本接近于资料的估计指出,为了使被试改变意见,一般需要 2 至 5 个观察资料来完成只需一个观察资料的工作。许多实验致力于解释这种现象。它们表明,保守性现象的一个主要原因——可能这个原因就是主要的——在于人类整合资料的能力不足 (misaggregation of the data)。也就是说,人们准确地感知 (perceive) 每一条资料,充分意识到 (aware of) 每一条资料的判断意义,但是,在修正他们的意见时,人们却不能把每一条资料的判断意义和其他资料的判断意义有效地结

合起来……

概率量化了不确定性。根据和我们一样的贝叶斯学者 (Bayesians) 的观点, 概率仅仅是 0 至 1 之间的一个数字, 它代表关于某个陈述, 一个某种程度上理想化的人 (a somewhat idealized person) 相信其为真的程度。之所以说某种程度上是理想化的, 原因在于, 他关于两个互斥事件的概率的和, 必须等于他关于其中任一事件发生的概率。概率的这种叠加性质 (additivity property), 对于逻辑推论的要求如此严格, 以至于现实中几乎没有人能够完全遵守它的所有推论。由于这种概率主要是描述持有意见的人, 而非意见所指向的事件, 所以被称为个人概率 (personal probability)。

360 贝叶斯定理, 是这种叠加性质的一个一般推论, 为所有的概率学者认可, 不会引起争议, 无论是贝叶斯学者还是其他概率学者。它的一种文字表达方式如下。如果 $P(H_A | D)$ 是资料 D 被观察到之后假设 A 的后验概率, $P(H_A)$ 是资料 D 被观察到之前假设 A 的先验概率, 如果假设 A 为真, 资料 D 被观察到的概率是 $P(D | H_A)$, $P(D)$ 是资料 D 的非条件概率, 那么

$$P(H_A | D) = \frac{P(D | H_A) P(H_A)}{P(D)} \quad (1)$$

$P(D)$ 最好被看作是一个标准化的常量 (Normalizing constant), 用来使所有假设的后验概率相加为 1, 这些互斥的假设穷尽了一切可能性。如果必须计算, 可以依据下面的公式:

$$P(D) = \sum_i P(D | H_i) P(H_i)$$

但是更经常的是消去而不是计算 $P(D)$ 。一种简便的方法是把贝叶斯定理转换成几率可能性比率的形式 (odds-likelihood ratio form)。考虑另一个假设 H_B , 它与假设 A 互斥, 根据使你改变你关于假设 A 的意见的相同资料, 修正你关于假设 B 的意见。贝叶斯定理表示为:

$$P(H_B | D) = \frac{P(D | H_B) P(H_B)}{P(D)} \quad (2)$$

现在, 用方程 2 去除方程 1; 结果是:

$$\frac{P(H_A | D)}{P(H_B | D)} = \frac{P(D | H_A)}{P(D | H_B)} \cdot \frac{P(H_A)}{P(H_B)}$$

$$\text{或者 } \Omega_1 = L \cdot \Omega_0 \quad (3)$$

其中, Ω_1 表示支持假设 A 而非假设 B 的后验几率, Ω_0 是先验几率, 而 L 是一个为统计学者所熟悉的量——可能性比率。方程 3 和方程 1 一样, 是贝叶斯定理的一种合适的表达方式, 并且方程 3 经常是更有效的, 尤其是对于包含两个假设的实验。贝叶斯统计学者认为, 贝叶斯定理, 是关于如何根据证据修正意见的形式上最优的规则。而根据证据修正意见, 恰恰是统计推论的一个组成部分, 因此, 统计推论, 就应该以贝叶斯定理为基础构造。由此, 许多方面与经典的统计学实践发生分歧。就实验心理学家而言, 一个基本的阐述可以参见爱德华兹等 (1963)。但是, 我们不是统计学家, 今天, 无论如何, 我们中没有谁愿意去戴统计学家这顶笨蛋高帽。相反, 作为心理学家, 我们感兴趣的是比较贝叶斯定理所确定的理想行为和事实上的人类行为。

367

为了使了解下面的内容, 让我们以你为被试做一个实验。有一个书包, 里面有 1 000 张扑克牌。我以两个这样的书包开始, 一个书包有 700 张红色和 300 张蓝色扑克牌, 另一个书包有 300 张红色和 700 张蓝色扑克牌。我抛掷一枚均匀的硬币来决定使用哪一个书包。因此, 如果你的意见和我一样, 那么这时, 你认为是红色牌居多的书包的概率是 0.5。现在, 你有放回地随机抽样。12 次抽样, 你得到 8 张红色牌和 4 张蓝色牌。现在, 根据你所知道的一切, 这个书包是红色牌居多的书包的概率是多少? 很明显, 概率大于 0.5。但是, 请不要继续阅读下去, 直到你给出你的估计值。

如果你是一个典型被试 (typical subject), 那么, 你的估计值会落在 0.7 至 0.8 的范围内——尽管在前面的段落中, 我们反复陈述人类是保守的信息处理者, 这可能会影响你的回答, 使答案偏高。然而, 如果我们进行适当的计算, 答案将是 0.97。事实上, 如果不是先前接触了保守性现象这一研究发现, 那么人们几乎不可能给出如此之高的估计, 即使他相当熟悉贝叶斯定理。

大约在 1960 年, 海斯 (William L. Hays)、一位名叫菲利普斯 (Lawrence D. Phillips) 的研究生和我, 对发现人类行为和贝叶斯定理确定的行为之间的不一致很感兴趣。我们并没有想到前面的段落所描述的简单例子;

相反，我们确信，为了得到非贝叶斯似的行为（non-Bayesian behavior），我们需要使用一种相当复杂的情境。因此，我们使用了一种假想的计算机化雷达系统。有 12 种可能的观察，4 种可能的假设，因此被试必须理解和使用一个包括 48 种不同的 $P(D|H)$ 值的展示（display）。被试在两种条件下工作。在一种条件下，被试看到一个单个的刺激，出现在雷达显示屏某个部分的一个光点；然后根据这个资料，调节 4 个控制杆直到他认为的后验概率估计值水平，如此被试修正自己关于 4 个假设的先验概率，然后再把控制杆重置为零状态，等待下一个刺激出现。第二个刺激由原来的光点和一个新的光点组成，被试调节他的控制杆来报告两个光点的累积影响。如此，直至累积到 15 个光点。在第二种条件下，刺激被打乱，因此实际上，每一个新的刺激，被试都要重新开始。令实验者吃惊的是，实验证实了贝叶斯定理的预测：实验条件的差异，并不会对行为产生影响。此外，还有一种实验条件，在这种条件下，每一个新的光点都是单独展示的。但是，从一个刺激转到下一个刺激时，允许被试保留他们的估计，而不是在每一次估计之后使控制杆重置为零。再一次，实验条件的变化对行为影响很小。

362 菲利普斯—海斯—爱德华兹（Phillips-Hays-Edwards）实验，可以确定的发现有三个。首先，被试极度保守；其次，面对第一个光点时，被试最不保守，光点越多，被试变得越保守；最后，在概率估计的值不受限制时，被试的概率估计之和通常大于 1，并且当被试逐步通过按规则顺序出现的相继刺激时，总和也随之增加。显然，被试发现，确定一个刺激支持哪一种假设，因此增加该假设的概率，要比决定为了增加受支持的假设的概率，应该削减那些假设的概率更加容易。

我们拖延了很久才发表最初这个关于保守性现象的实验。尽管到 1962 年就已经完成了资料，然而，菲利普斯—海斯—爱德华兹的这篇论文，直到 1966 年才出版（Phillips et al., 1966）。

研究所发现的保守性现象的强度和一致性使我们震惊。看来，试验更加简单的任务，是合适的。因此，菲利普斯和我进行了一次正式研究前的预先测试或前测，性质上类似于前面你所接受的书包和扑克牌的例子，尽管信心不足。让我们吃惊的是，实验非常成功。目前，大多数的比较人类

行为和贝叶斯定理的研究,都能够追溯到这个前测以及随后的实验。

如果书包中红色牌的比例是 p , 那么, 在 n 次抽样中, 得到特定顺序的 r 张红色牌和 $(n-r)$ 张蓝色牌的概率是 $p^r (1-p)^{n-r}$ 。因此, 在一个典型的书包和扑克牌实验中, 如果假设 H_A 代表红色牌的比例为 p_A , 假设 H_B 代表蓝色牌的比例为 p_B , 那么可能性比率就是

$$L = \frac{p_A^r (1-p_A)^{n-r}}{p_B^r (1-p_B)^{n-r}} \quad (4)$$

注意: 虽然方程 4 的推导, 考虑了抽样过程中红色和蓝色牌的实际顺序, 然而, 对于任何顺序的 r 张红色和 $(n-r)$ 张蓝色牌, 我们同样能够推导出方程 4。二项系数 (the binominal coefficient) 代表在 n 次抽样中得到 r 张红色牌的不同方法的数量, 它在分子和分母中都出现了, 因此可以从可能性比率中约去。方程 4 是贝叶斯统计学可能性原理的一个例证 (见 Edwards et al., 1963), 实际上, 它表明, 一个贝叶斯统计学者仅仅需要考虑他实际所做的观察的概率, 而不用考虑其他他可能做但没有做的观察的概率。这一原理对贝叶斯定理的所有统计的和非统计的应用具有决定性的影响; 它是贝叶斯思想中最重要的技术工具。

在 $p_A = 1 - p_B$ (对称二项式情况) 这种特殊情况中, 可能性比率简化为

$$L = \left(\frac{p_A}{1-p_A} \right)^{2r-n} \quad (5)$$

注意: $2r-n = r - (n-r)$ 是抽样中红色牌和蓝色牌数量的差数; 在这种对称情况中, 仅仅是差异而非观察的总数, 才与推理有关。统计学传统上用成功数减去失败数或者 $s-f$ 表示这个差数; $s-f$ 是书包和扑克牌实验中的常规 (usual) 自变量。为了理解常规因变量的逻辑根据, 我们把方程 5 代入到方程 3 中, 然后取对数, 整理各项, 结果是:

$$\log L = (2r-n) \log \frac{p_A}{1-p_A} = \log \Omega_1 - \log \Omega_0$$

用后验几率的对数减去先验几率的对数, 能够推断出对数可能性比率。如果被试完全按照贝叶斯定理思维思考, 那么对数可能性比率应该与自变量 $s-f$ 成比例。因此, 以被试的推断对数可能性比例为纵轴, $s-f$ 为横轴作图是合适的, 被试的推断对数可能性比例可以根据他的后验几率 (反过

来, 如果被试估计的是概率, 可以根据后验概率计算出后验几率) 计算得到。

在密歇根实验室进行的大多数书包和扑克牌实验使用的, 是一种由 48 个编号的位置组成的显示屏, 每一个位置包括一个按钮、一束红光、一束绿光。当某个位置的按钮被按下时, 一种光束出现并持续。被试被告知这等同于从书包中有放回地抽出一张相应颜色的扑克牌。被试也被告知, 控制光束的程序是按照书包抽样的程序设计的。实际上, 大多数实验都相当仔细地设计光束的控制程序, 因此, 用光束的显示顺序代表书包实验是合适的, 特别是在每个实验中, 对于使用的 p_A 值和样本规模 n 的所有取值, 规模为 n 的样本经常支持错误假设。

菲利普斯和我 (1966) 使用 20 张扑克牌的序列研究 p_A 的作用效果, p_A 的值取 0.55、0.7 和 0.85。被试把 100 张白色木制圆盘分配到两个木盆中, 以此来估计后验概率。图 25—1 给出了实验的典型结果, p_A 值为 0.7 与各种先验概率的情况。图 25—1 所阐明的 3 个研究发现, 所有被试都存在。首先, 推断的对数可能性比率与 $s-f$ 大致成比例; 其次, 先验概率被适当地使用, 也就是说, 连接各个数据点的最佳拟合线通过原点; 最后, 被试是保守的, 因为最佳拟合线比代表贝叶斯式最优行为的直线, 更加平缓。推断的对数可能性比例与 $s-f$ (或者, 等同于贝叶斯对数可能性比率) 之间的近似线性这一发现暗示, 还存在另外一个因变量: 通过被试的估计值点的最佳拟合线的斜率与贝叶斯线的斜率的比率。皮特森等 (Peterson et al., 1965), 把这一比率命名为精确性比率; 他们还发现, 精确性比率大致上等于 $s-f$ 。

364 图 25—2 给出了菲利普斯—爱德华兹资料三种 p_A 值下的精确性比率。资料的判断价值最低时, 与贝叶斯定理相比, 被试更为极端。(戴尔得到同样的发现, 见 W. Edwards, 1965)。但是, 当资料的判断价值达到适当的高度时, 被试变得保守, 精确性比率恰好等于 $s-f$ 。注意当判断价值增加时, 保守性也增强了。这是此类实验的一个共同发现; 任何增加单个观察 (一张牌或者几张牌) 的判断价值的程序, 也同时增强了保守性 (见 Peterson et al., 1965)。

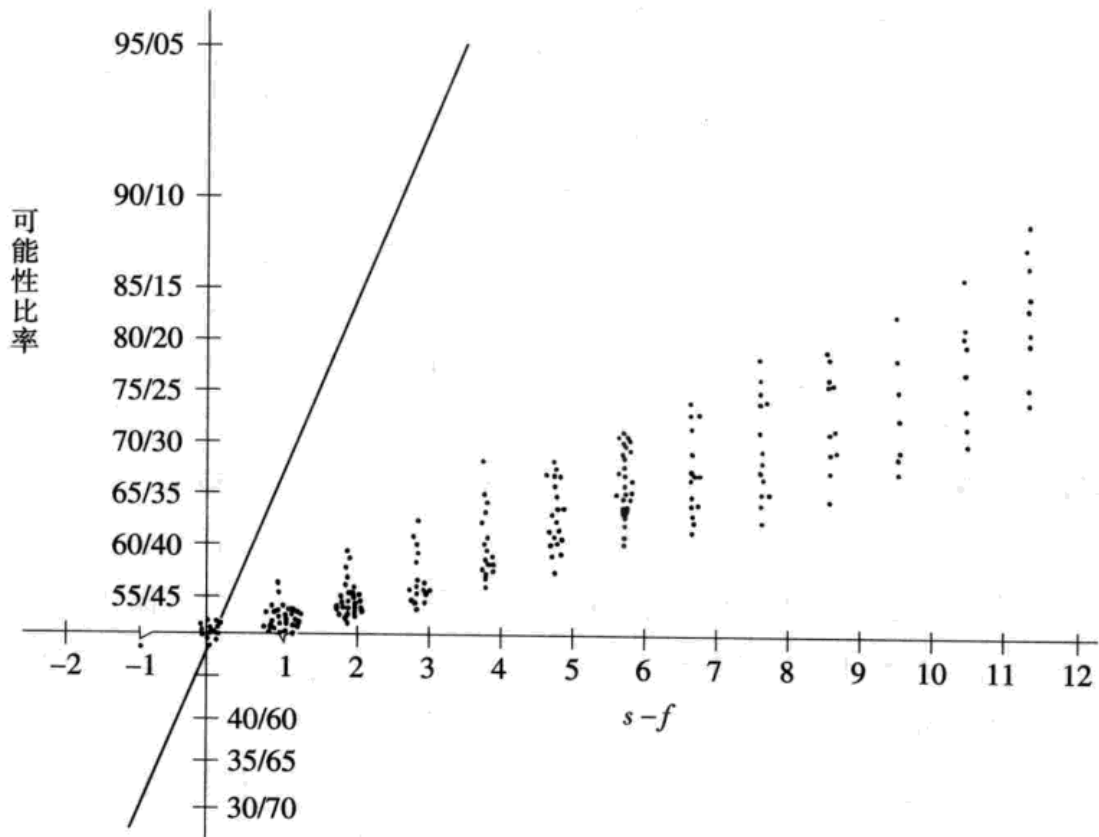


图 25—1 $p_A = 0.7$ 时，一位被试的估计表示为推断的对数可能性比率与样本中成功数和失败数之差的函数。

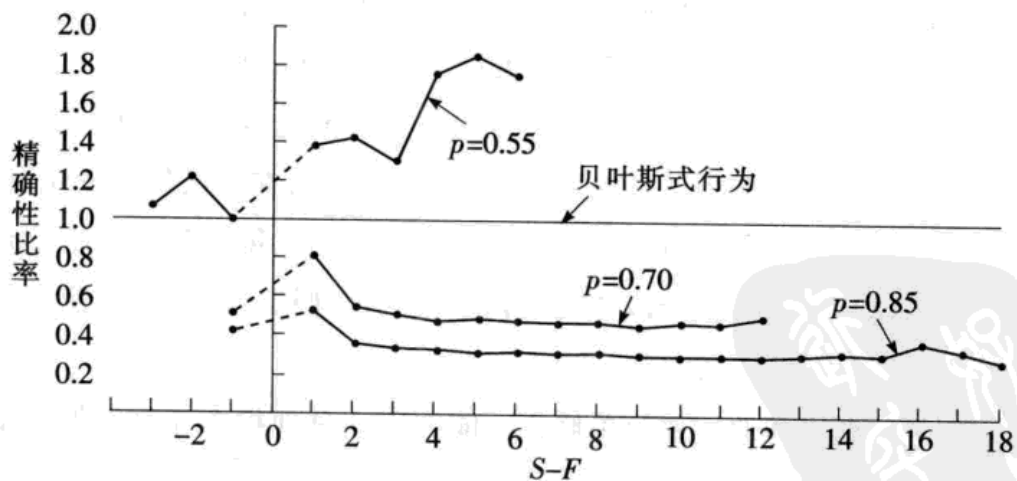


图 25—2 三种 p_A 取值时各种样本构成的精确性比率

得到这些结果之后，菲利普斯和我推测保守性现象的一个原因，可能是被试知道概率的数值大小有范围，他们注意到证据的数量可能持续上升，

365

因此压低他们的估计。如果是这样，那么一个显然的解决方法，就是使用一种没有范围限定的反应模式，如几率。因此，我们进行了一个有 4 个实验组的研究。和以前一样，控制组的被试把 100 张圆盘分配到两个木盆中，以此估计概率值。口头几率组仅仅给出几率的口头估计；由于我们总是把几率定为等于或者大于 1 的数值，因此给出几率的口头陈述时，总是同时说明该几率支持哪一种假设。使用对数尺度衡量几率组的被试，沿着一个几率标度盘移动指针，从而给出他们的估计值。几率标度包括 4 种对数循环，因此能够估计从 1:1 到 10 000:1 的所有几率值。第四个组也使用一种对数尺度装置衡量几率，但是标度盘上的刻度对应的数字是概率，而不是几率（因此是 0.5 而不是 1:1，是 0.67 而不是 2:1，是 0.80 而不是 4:1，等等）。我们称之为对数几率尺度的概率组。研究发现所有组都很保守。概率组最保守，对数几率尺度的概率组次之，两个几率组大致相当，其中对数尺度的几率组较为保守一些。

这一研究发现，仅仅强调了在贝叶斯研究过程中变得越来越清楚的一个事实。关于不确定性，概率是一种相当拙劣的度量标准，除了在必须重新分配或者其他必须直接使用叠加性质的情况下。几率或者对数几率是更好的度量标准。对于没有参加过实验的对象，几率最具有直觉性质，能够非常容易地与简单操作（如赌注的选择）联系起来。博彩行业以几率而不是概率为基础来组织其所有的陈述和展示（statements and displays）这一事实，既是对几率巨大的直觉性价值的认可，也可能是其原因。在不确定性的所有常用的度量标准中，对数几率的独特性质，就在于这种度量体系中，证据是累计性的。如果使用对数几率度量意见，那么每一条证据所导致的意见变化的总量，就独立于该意见的初始值。这种精美的性质，使得对数几率特别适用于贝叶斯式实验。

运用贝叶斯定理的一种简单修正，能够更充分地拟合菲利普斯—爱德华兹的资料：

$$\Omega_1 = L^c \Omega_0$$

366

常数 c 是精确性比率，它是使用贝叶斯定理处理可能性比率之前，每一个可能性比率自乘的指数。不幸的是，它依赖于一些重要的自变量，包括资

料的判断性和反应的度量标准。尽管如此，这样简单的一个描述性模型能够拟合地如此之好，这一事实我们还是必须根据关于保守性的某种理论来解释。

……概率信息处理系统，或者 *PIP*，……是一种关于如何设计人一机系统的理念，该系统必须处理信息以推论事情所处的状态。这种信息必须得到处理的情况的一些例子包括医疗诊断、军事指挥（其中，一个指挥官可能需要断定他是否正被敌人进攻，如果是，敌人的计划是什么）和商业管理（例如，企业家要决定是否制造一种新产品）。*PIP* 的理念过于复杂，以至于不能在此详细地解释。关于它的一个最新说明可以参见爱德华兹等（1965）或者爱德华兹（1966）。它的核心观点认为处理判断性信息的任务可以分解为两类子任务。判断每一条资料对每条假设或者每对假设的判断性影响，构成第一类子任务。对于成为许多现实的判断情境的特征的口头的、定性类型的资料和假设而言，这似乎是人们必须完成的任务，越熟练越好。但是，第二类子任务是要累积所有资料与假设之间这种独立的判断性影响，从而说明根据目前所有可资利用的资料，所有这些假设怎么成立。如果是用 $P(D|H)$ 值或者可能性比率的形式来判断每一条资料的判断性影响，那么使用贝叶斯定理，这种累积任务很容易机械化操作。[在大多数情境中，尽管不是全部，出于形式的原因，显然判断可能性比率比判断 $P(D|H)$ 值更可取。]

大约有 15 个合作者，他们和我很想弄清楚 *PIP* 是否起作用。因此我们精心设计了一个虚拟的 1975 年的世界。在这个世界中，我们列出 6 条被试将要考虑的假设，指定 3 个资料来源（弹道导弹预警系统、侦察卫星系统、情报系统）提供与这些假设有关的资料，并且设计了 4 种信息处理系统来处理这些资料。这四种系统分别命名为 *PIP*、*POP*、*PEP* 和 *PUP*。在 *PIP* 中，每一条资料，被试需要估计 5 种可能性比率。6 条假设中的一条是“和平将继续获胜”，其他的 5 条假设是关于各种可能的战争的假设；每一条战争假设与和平假设配对，这 5 对假设指定了将要估计的 5 种可能性比率。其他 3 种信息处理系统与 *PIP* 共同之处在于被试估计后验几率，或者概率，或者类似的后验量度；这样，在 *PIP* 中，计算机使用贝叶斯定理累

367 积资料，而在所有其他 3 种系统中，被试不得不依靠自己来累积资料。为了帮助他们完成任务，在 *POP*、*PEP* 和 *PUP* 中，当被试考虑第 $(n+1)$ 条资料时，他们可以使用估计完第 n 条资料后他们得到的估计值，这样他们只需要修正那些受第 $(n+1)$ 条资料影响的估计。

一共有 18 个场景，每一个场景有 60 条资料。所有的资料都是以短文的形式给出，除了来自弹道导弹预警系统的资料。34 名被试接受全面的训练，熟悉这个世界的特征，假设，3 个资料来源和将要操作的每一种信息处理系统。

显然，*PIP* 最佳，*POP* 其次，因此我只给出它们两种之间的比较（*PUP* 再次，*PEP* 最差，关于这种信息处理是如何进行的，*PEP* 是目前我们能够达到的极限）。图 25—3 给出的是 *PIP* 和 *POP* 中，每个场景的第 60 条信息给出之后，支持每一种战争假设与和平假设比较的最终几率。关于这个图，需要注意的最重要两点是：从定性的方面来看，两个组非常一致（它们之间的相关系数是 0.895），但是从定量的方面来看，它们不一致。*PIP* 对资料比 *POP* 更加敏感；同一个场景，*PIP* 非常肯定和平或者某种战争状况，而 *POP* 却更不确定。换句话说，*PIP* 比 *POP* 更不保守——可能因为在 *POP* 中，被试必须累积资料，而在 *PIP* 中，被试分别判断每一条资料的判断性影响，而由贝叶斯定理来完成累积工作。

368 还要注意的是，图 25—3 中纵轴与横轴的坐标都是取了对数的。如果把比率的差异转换成为几率，*PIP* 和 *POP* 之间的重要差异就会变得明显了。例如，根据回归直线计算，如果某个场景，*PIP* 以 99:1 的比率支持某种战争对和平的假设，那么，支持同样的战争对和平的假设，*POP* 给出的比率是 4:1。

无论如何，误解假设不能够解释 *PIP* 和 *POP* 之间的这种不一致。*PIP* 中，被试分别估计每一条资料的判断性影响；而 *POP* 中，被试必须依靠自己来累积——并且相当保守。既然不存在任何资料—生产程序的模型可以使用，那么就不可能判断出正确的后验几率是多少。但是很明显，*PIP* 和 *POP* 之间的差异的一个原因在于累积过程的差异。

实验的一个合作者菲利普斯关注这一事实：不存在任何资料—生产程

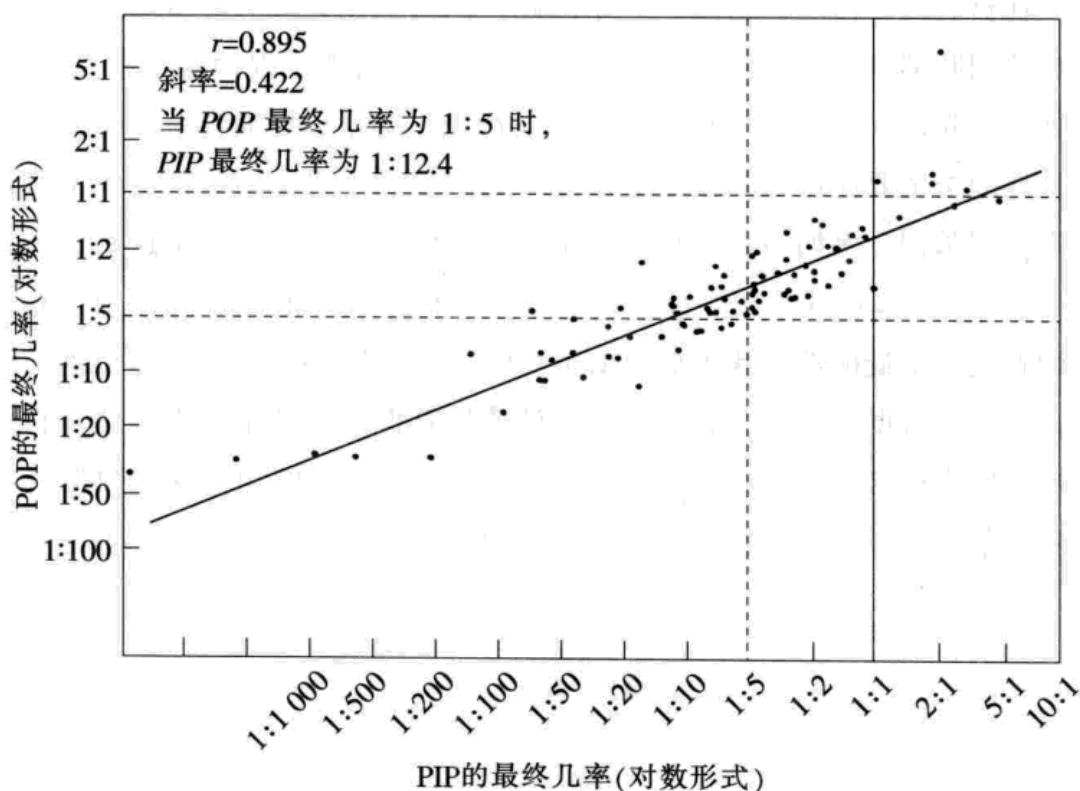


图 25—3 POP 与 PIP 支持战争假设的最终几率，以对数形式表示。

序的模型可以使用，因此就不能够确定是 PIP 还是 POP 更加接近于正确。所以在他的博士论文中，他在一种情境下比较 PIP 和 POP，在这种情境中可以使用一种资料—生产程序的模型，因此对于每一条资料，要求估计可能性比率是有意义的，结果 POP 程序得到了保守的估计。他的被试是密歇根大学学生报纸的编辑。他选取每名编辑一学期的社论文章，数出每篇社论中每个单词开头两个字母和结尾两个字母，这样，他为每名编辑准备了两个书包，一个装满开头双字母，一个装满结尾双字母。在 PIP 任务中，他取出一定的双字母，然后让每名编辑（只对他自己的书包）估计可能性比率，以该双字母是取自开头双字母书包的假设作分子，以取自结尾双字母书包的假设作分母。在 POP 任务中，他准备了一个双字母序列，是从一个书包中抽取出来的，然后当编辑完成了这个序列时，让他估计这个序列是从开头双字母书包，而不是从结尾双字母书包中抽取出来的后验几率。大部分注意力放在编辑的初始训练上，两次收集可能性比率的估计值，一次在后验估计之前，一次在之后。

369

在资料的分析中,出现了一个问题,因为无论是 *PIP* 中还是 *POP* 中,所有的判断都倾向于支持开头双字母的书包。这可能是因为这样做更容易,例如,想起以 *re* 开头的单词要比想起以 *re* 结尾的单词更加容易,尽管 *re* 作为单词的结尾要比作为开头更加常见;我们习惯于用单词的开头而不是结尾来标记单词,例如当我们在字典中查找它们时。然而,修正这种偏差是有可能的。图 25—4 给出了修正后的结果。根据实际的双字母总数计算得到的真实几率最极端。其次是根据第二组可能性比率估计值计算得到的几率,然后是根据第一组可能性几率估计值计算得到的几率,直接估计得到的后验几率最接近中间,因此也是最保守的。如果我们相信这些数据(我确实相信),那么,尽管与 *POP* 相比, *PIP* 已经相当不保守了,然而它还是

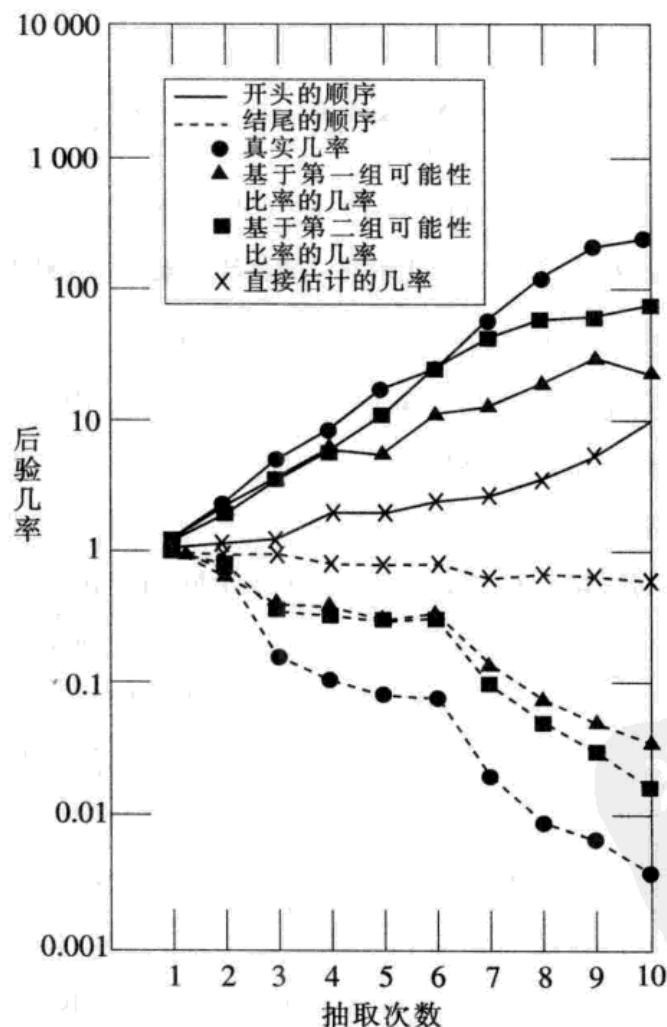


图 25—4 所有被试支持开头双字母书包的后验几率中位数,它是取次数的函数。

太保守了——但是，*PIP* 的估计是随着实践而改善的。

本章节选自 B. Kleinmuntz 主编的《人类判断的形式表征》 (*Formal Representation of Human Judgment*, New York: John Wiley and Sons, Inc., 1968)。允许重印。



第 26 章

多阶段推理中的最佳推测假设

盖提斯等 (Charles F. Gettys, Chinton Kelly III, and Cameron R. Peterson)

370 多阶段推理 (multistage inference), 包括一系列的单阶段推理 (Single-stage inference), 其中, 前一阶段的输出成为后一阶段的输入。在单阶段推理中, 人们从数据资料 (data) 或者明确观察到的证据 (evidence) 中推理出一组假设 (hypotheses)。多阶段推理的第一阶段, 以同样明确的资料或者证据开始, 然而, 下一阶段的输入是前一阶段的输出, 因此, 下一阶段的推理, 就是基于事件的概率 (probabilities of events), 而不是一个特定真实事件的确定知识 (Gettys & Willke, 1969)。

例如, 假设你想预测一场大型游园会成功或者失败。如果因为下雨人们拥挤在室内, 那么游园会成功举行的可能性就很小。你所拥有的资料, 就是天边出现的乌云。推理的第一阶段, 是乌云和游园会期间是否下雨的关系。假设你估计下雨的概率是 0.70。这一估计将成为你下一阶段推理的输入。如果你确切地知

道将会下雨，那么你可以推断游园会成功的概率。但是，由于你不能完全地确定天会下雨；你拥有的资料表明下雨的概率是 0.70，那么你怎样继续你的推理呢？

修正后的贝叶斯定理 (Modified Bayes Theorem, MBT)，为这种多阶段推理提供了一种最优模型 (optimal model) (Dodson, 1961; Gettys & Willke, 1969)。许多研究表明，在多阶段推理任务中，人们的直觉性表现 (intuitive performance)，使根据资料推断出来的结果相对于 MBT 模型的预计结果，具有更多的确定性。例如，在一个几率评估 (odds estimate) 任务中，被试估计的几率，显著性大于 MBT 模型计算出来的结果。这个结果是相当令人惊奇的，因为证据显示，在单阶段推理中，人类行为几乎总是保守的，也就是说，人们的推断相对于资料可保证的结果，总是更少确定性 (例如，爱德华兹，1966)。当然，悖论在于，多阶段推理是系列的单阶段推理。如果在单阶段推理中，人们的推断相对于资料保证的结果更少确定性，那么多阶段情境中，我们可能预期随着推理的每一阶段的进展，被试会变得越来越保守，因为其推断与最优之间的偏差，在每个阶段上，会逐渐累积。事实上，情况正好相反。在两阶段推理结束的时候，被试比最优模型 (MBT) 所保证的更为确信。这意味着在两个单一阶段的“交汇处”，某种过程发生了，它们的效果如此强烈，以至于单阶段推理中的任何保守主义倾向都被克服了。

单阶段推理的任务，总是基于确切的已知资料。然而，尽管一个多阶段推理的任务，开始于确定资料，但是后继的推理阶段，处理的却是不确定资料。在多阶段推理中，几个模型已被构造出来，用以解释如何处理资料的概率，而不是解释确定资料会产生过分确定性的情况。“as-if”模型，是具有预测过分确定性的性能的一种非最优模型 (non-optimal model) (Gettys & Willke, 1969; Howel et al., 1971)。这个模型设计的情境是：人们可以选择收集更多的资料，如果他们感到这是必要的话。模型假定，在推理的第一阶段，资料的收集持续到决策者对于事件的状态有了充足的确定性。一旦他的确定性超出了某种阈值，那么他将开始进入下一个阶段的推理，其行动好像 (as-if) 他已经完全有把握确定下一阶段推理所需的输

入。让我们回到游园会的例子。在看到乌云之后，决策者有可能会得到当前的天气预报资料。假设天气预报警告将有一场暴风雨，那么现在他对可能下雨的确定性，就会超出他的阈值，他将进入到下一个阶段的推理，其行动就好像他确定会下雨一样。因此，推理的第二阶段的结果，将是他基于下雨的 as-if 假设所做出的游园会失败或者成功的概率估计。现在，因为通过做出 as-if 假设，他忽略了可能不下雨的概率，所以他对于失败的概率的估计，将会超过真实（veridical）失败的概率。事实上，如果他的 as-if 假设不正确，即没有下雨，那么游园会很有可能成功。而最优模型，在分配成功或者失败概率时，都考虑了下雨和不下雨两种情况。而 as-if 模型仅仅考虑了下雨的概率，因为这个原因，导致了游园会将会失败的过分确定性。

372 那么，如果一个人对于推理的第二阶段所需要的输入的确定性没有达到 as-if 假设所要求的阈值，并且他也没有希望通过更多的资料来提高确定性，他的行为会怎样呢？一个与先前的研究所发现的过度确定性现象一致的可能假设是：首先，他将作出一个最多是推测的 as-if 假设。这个模型——我们称之为“最佳推测模型”（the Best-Guess model）——在实行中是一个合格的 as-if 模型，共享了 as-if 模型的理念：在推理的第二阶段，由于几乎仅专注于最可能事件，决策者忽略了或者倾向于忽略其他较不可能事件的含义。就前面的例子来说，如果你拥有的唯一信息就是天际的乌云，你可能不会愿意做出一个不合要求（unqualified）的 as-if 假设；但是，你可能首先假定天将会下雨，基于这个假定，你得出游园会成功的主观几率。那么，因为你没有完全确定天将会下雨，考虑到这种情况，你可能降低你的主观几率。主要是因为你明显的没有考虑不下雨的含义，这些主观几率，很有可能和根据 MBT 模型计算出来的结果不同。

斯奈泊等（Snapper & Fryback, 1971）在一个关于资料信度的实验中，报告了与上述解释一致的实验结果。然而，他们的研究设计，并不是直接检验最佳推测模型（Best-Guess model）。而这正是现在这个实验的目的。

一、方法

实验的目标，要求构造至少三个层次（three levels of variables）的变

量, 其中, 中介变量包含两个以上的可能事件 (event); 它还要求操纵最有可能的中介事件之外的所有其他可能事件的概率分布——这一操纵, 作为较低层次 (the lower level) 事件发生的结果, 对于较高层次 (the upper level) 可能事件的最优概率修正的量值 (magnitude of optimal probability revision) 产生影响。

三个层次的变量, 因此采取下列形式。较高层次的变量, 分别由标签为 I 和 II 的袋子 (bag) 组成。每个袋子作为一个容器, 里面装满了小一点的容器, 这些小一点的容器代表中介水平的事件。具体说, 每个袋子装有 18 个小罐 (can) (35mm, 胶卷盒), 每个罐子分别标为字母 A、B、C 或者 D。最后, 每一个罐子装有 100 张小的有色磁盘 (small coloured disc); 每一张磁盘分别是红色、绿色、黄色或者蓝色。

表 26—1 给出每个容器的构成状况。表的 A 部分描述了袋子中的罐子的构成情况, B 部分描述了罐子中磁盘的构成情况。例如, 在袋子 I 中有 8 个标签为 A 的罐子, 而在袋子 II 中仅有一个标签为 A 的罐子。如表的 B 部分显示的, 罐子 A 中有 80 张磁盘, 罐子 B 有 1 张, 罐子 C 有 1 张, 罐子 D 有 18 张。

表 26—1 书包和罐子内部构成的数值表示

罐子的标签	A. 书包的构成			
	书包 I		书包 II	
A	8		1	
B	3		6	
C	6		3	
D	1		8	

磁盘颜色	B. 罐子的构成			
	罐 A	罐 B	罐 C	罐 D
红	80	1	1	18
绿	1	80	18	1
黄	18	1	80	1
蓝	1	18	1	80

实验按照下述程序进行。从两个袋子中随机挑出一个袋子, 再从这个袋子中随机挑出一个罐子, 最后从这个罐子中随机取出一张磁盘。因此,

379 抽出一张红色的磁盘就为支持罐子 A 提供了证据, 同样地也提供了支持袋子 I 的证据。注意这仅仅是一个可以直接观察到的最低层次 (bottom-level) 的可能事件——一张磁盘。这一观察仅仅提供了有关中间层次的可能事件——罐子——的部分证据, 同样地也提供了有关较高层次的可能事件被选择的部分证据。因而, 推理的第一阶段是磁盘颜色 (红、黄、蓝、绿) 和罐子标签的字母 (A、B、C、D) 之间的关系, 第二阶段是罐子的字母和袋子标签的数字 (I、II) 之间的关系。

在一个层次上, 最可能事件为真的行动策略, 将会在下一个较高的层次上导致可能事件的概率分布极端化。因此, 这一策略和下述经验结果是相一致的: 在多阶段推理任务中, 人们会过分修正较高层次的概率。

而从最佳猜测策略, 我们可以推出另一个可检验的假设。在任何中间层次上, 最可能事件为真的行动策略, 会使人忽略了这一层次上所有其他可能事件的概率分布。因此, 在较高层次上他所做出的概率修正, 就会忽视除中间层次上最可能事件之外的所有概率分布的变异。这个实验的设计, 就是为了检验这一假设。

1. 实验设计

我们构造了表 26—1 中所显示的 3 种类型的推理任务。表 26—1 中 A 部分显示的频次, 在所有三个任务中都被使用。表 26—1 的 B 部分显示的矩阵, 只在一个任务中被使用; 在其他两个任务中, 矩阵中的数值 80 分别被换成 70 或者 90。相应地, 数值 18 分别被换成 28 或者 8。为了以后讨论的方便, 这三个任务分别指定为 70—28 任务、80—18 任务和 90—8 任务。在所有三个任务中, 给定从罐子中抽出的一张磁盘的颜色, 让被试来估计袋子的几率。

374 2. 被试

25 个被试, 是来自于密歇根大学 (University of Michigan) 的学生。在此之前, 他们参加了另外一个持续约两个小时的多阶段推理实验。在这个实验中, 被试依照所要求的反应模式进行训练, 并在一个多阶段推理任务中进行了系列的几率估计。然而, 没有涉及最优模式, 也没有使用任何类型的反馈。

3. 指导语

因为被试参加过先前的实验，指导语很简短。向被试解释了任务细节后，要求他们去设想这样一种情况：根据抛掷一枚好硬币的标准随机挑选一只袋子，再随机地从这个袋子中取出一个罐子，接着从这个罐子中随机地取出一张磁盘。然后，要求被试推测事实上是根据这种随机程序抽取出来的一张特定磁盘的颜色，并且基于这张磁盘的颜色来估计袋子的几率。

4. 程序

依照上述指导，在 3 个任务中，让被试估计袋子的几率。用表 26—1 中的矩阵，来告知被试罐子和磁盘之间的相对频次。所有的任务，按照随机的顺序呈现给被试组，每组 4 到 6 人。在每一个任务中，4 种可能的颜色，按照随机的顺序采用。在进入下一个任务之前，基于所有可能呈现的颜色，被试来估计袋子的几率。当被试完成 12 次（每个任务 4 种颜色，共 3 个任务）估计后，三个任务按照不同的随机顺序重复一遍。一共有 24 次判断，在每个任务中，每种颜色采用 2 次。

二、结果与讨论

对数据的检查，显示了被试的几率反应（odds responses）有一种极端的双峰分布（an extreme bimodality）现象。对一些被试而言，蓝色与红色之间理论上的差别，以及黄色与绿色之间的差别，在他们的几率估计中，没有产生任何差别。而其他一些被试，基于蓝色的几率估计，比基于红色的几率估计，更为极端；基于黄色的比基于绿色的，也更为极端。至少在顺序意义上，后面的这些被试的反应，与 MBT 模型是一致的。一些被试似乎没有觉知到蓝—红之间和黄—绿之间的差别，而其他被试，至少就他们能够依照正确的方向做出反应而言，他们觉知到了这些差别。考虑到这些，我们把所有那些基于蓝色的几率评估，至少有一次其评估的几率比红色情况大致少 2%，或者对黄色情况的评估比绿色情况时大致少 2% 的被试归类为“有意识的被试”（aware Ss）。这些被试至少有点意识，因为在一次或更多次的判断中，对于蓝—红和黄—绿的几率评估，他们与 MBT 模型所指示的方向是同向变化的。根据这种保守的标准，25 个被试中的 10 个被归类

为“有意识的被试”。

对于较不可能事件的概率变化,其他 15 个“非觉知”的被试,并没有不同反应。他们完全地忽略了可能性较小的罐子的含义,他们的反应,同最佳推测模型的极端形式是一致的。图 26—1 给出“非觉知的”被试的几率反应的中位数。因为几率评估所偏好的袋子,在形式上是不相关的,所以数据被绘制成绝对对数几率表。在图形上方,对于 3 个层次的数据的不确定性,对红色和蓝色反应的几率对数中位数,用实线连接;在图形下方,类似地,是对黄色和绿色反应的数据。图 26—1 还显示了 MBT 模型的预测值(图形中的主对角线)、as-if 模型的预测值(两条水平线)和极端形式的最佳推测模型的预测值——在图形中标记为模型 I (Model I)。

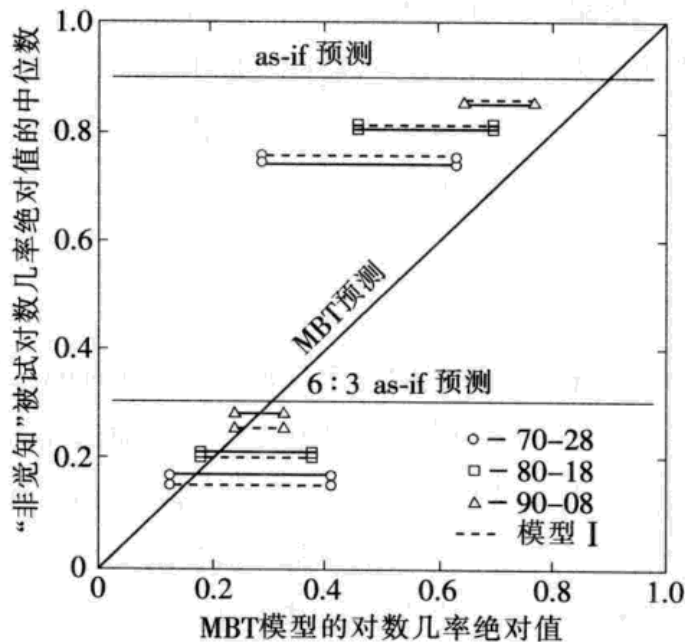


图 26—1 最优模型、as-if 模型和 MBT 模型对“非觉知”被试行为的预测

用最可能事件的概率乘以在这一事件为真的条件下得到的后验几率 (posterior odds), 我们可以得到模型 I 的预测值。例如, 假设在 80-18 类型的任务中, 一张红色的磁盘被抽选出来, 那么支持罐子 A 的概率是 0.8, 并且如果磁盘确实是从罐子 A 中取出的, 那么袋子的几率是 8:1。在关于罐子类型的临界确定值 (the threshold certainty for can type) 被超过的条件下, 对于蓝色和红色磁盘的情况, as-if 模型预测的几率是 8:1, 黄色和绿

色磁盘的情况是 6:3。对于 MBT 模型, 根据下面的后验几率的公式, 我们可以计算出在红色磁盘情况时, 最优的几率是 2.86:1 (Gettys & Willke, 1969):

$$\frac{P(B I | \text{color})}{P(B II | \text{color})} = \frac{P(B I)}{P(B II)} \times \frac{\sum_i P(\text{color} | \text{can}_i) P(\text{can}_i | B I)}{\sum_i P(\text{color} | \text{can}_i) P(\text{can}_i | B II)} \quad (1)$$

公式中字母 B 代表袋子, 根据类似于表 26—1 中显示的条件概率, 可以计算出其他几个数值。

显然, 图 26—1 中的资料既不与 as-if 模型, 也不和 MBT 模型的预测值相吻合。对于图形上方的蓝色—红色的情况, as-if 模型的预测的几率的值是 8:1, 而被试则没有那么极端, 类似地, 在图形的下方, 相对于 as-if 模型的预测值 6:3, 被试也较少有极端性。然而, 最佳推测模型的极端形式 (即模型 I) 与图形中的中位数非常吻合。图 26—1 中的水平的虚线代表模型 I 的预测值。在所有的任务中, 对于图 26—1 的下方的黄色磁盘的情况, 模型 I 的预测值正与 MBT 对角线 (are to the right of the MBT diagonal) 吻合。因此, as-if 模型和模型 I 并不必然的做出比 MBT 模型的几率估计值更为极端的预测值。例如, 在 80-8 的任务中, 当罐子 C 成为最可能事件 ($P=0.80$) 而罐子 A 是较不可能事件 ($P=0.18$) 时候, 这种情况发生。如果最可能事件 (即罐子 C) 为真的话, 它分配给袋子 I 的几率是 6:3, 而较不可能事件分配给袋子 I 的几率是 8:1。在这个例子中, 相对于 MBT 模型, 任何忽略了较不可能事件情况下的比率 8:1 的模型的预测都将是保守的。

如果假定在严重偏离于他们对于确定性的主观感受时, 被试将不会采用一个非最优模型, 那么, 可能最重要的结果是被试使用模型 I, 因为对于模型 I 他们没有发现任何错误。模型 I 决定他们的几率估计的量值, 但是在另外的情形中, 他们可能会使用其他的复合的规则 (combination rule)。更为重要的是, 模型 I 的预测值确实和数据吻合这一事实暗示: 被试倾向于专注最可能的选项, 而忽略了较不可能选项的推论。

在图 26—2 中显示的是 10 个“有觉知的被试”的资料。和图 26—1 中

一样，图形描绘了 as-if 模型和 MBT 模型的预测值，但是因为模型 I 的预测值明显的和资料不符合，所以省略了模型 I。

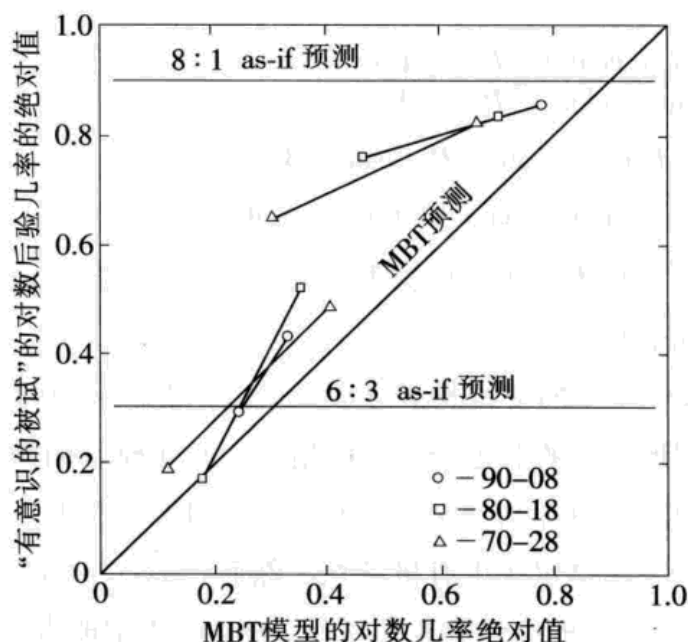


图 26—2 “有意识的被试” 的表现与 MBT 和 as-if 模型的预测比较

一般地说，“有觉知的被试”似乎和 MBT 模型一样对相同的变量做出反应，但是它对于 MBT 模型的数量吻合程度 (quantitative fit) 较差。如同在先前的研究中所发现的，与 MBT 模型相比，被试的特征具有更大的确定性。对于蓝色和红色磁盘的情况，和 MBT 模型一样，被试比 as-if 模型所暗示的具有更少的确定性；也和 MBT 模型中一样，他们对于黄色磁盘情况的判断超过 as-if 模型的预测。当然，这种情况的出现是因为最可能事件具有的几率是 6:3，而较不可能事件的几率是 8:1。如果被试觉知到多阶推理情形中的细微差异的话，他们将会认识到几率必然大于 6:3。对于这种一般状态的唯一例外是 80-18 任务中黄色和绿色磁盘的情况。在黄色磁盘的情况下，当最可能事件的概率减少时，基于这一事件的后验几率应该随之增加，而在绿色磁盘情况下，当最可能事件的概率增加时，基于这一事件的几率应该随之减少。但是，在 80:18 条件下被试的反应并不遵循这种模式。一般地说，“有觉知的被试”似乎使用的是一种复合的规则，既有些像 MBT 模型，但是相对于 MBT 模型又稍稍过度。

显然，在多阶段推理中，“非觉知的被试”的反应，支持最佳推测倾向的假设。很明显，可能因为情境的复杂性，在推理的后续阶段中，一些被试倾向于唯一地关注最可能事件。正如单阶段推理中的保守性，多阶段推理中的最佳推测效应，似乎也是人类处理复杂信息普遍性无能的另一例证。既然大部分的人类信息处理，在本质上是多阶段的和或然性的，那么为了运用贝叶斯定理，下一个合适的步骤，就是去发现种种方式，以防止人们出错：即仅注意最有可能的中介事件，而忽视所有其他事件。

本章原载《组织行为和人类绩效》杂志 (*Organizational Behavior and Human Performance*, 1973, 10: 364-373)。允许重印。



第 27 章

根据回忆信息推论个人特征

特罗普 (Yaacov Trope)

378 社会诊断，经常是以对他人的行为不精确的回忆为基础的。我们经常需要推断他人的个人特质，甚至在我们不能够确定自己是否记得他人真实行为的时候。在这种情况下，我们不得不根据不确定的行为证据进行归因。目前的这个研究，在于探讨人们是否以及怎样把这种不确定性的源材料 (source of uncertainty) 纳入到他们对他人的诊断中去……

一、一个根据回忆的行为进行推论的贝叶斯模型

在目前这个研究中，被试根据他们自己对于行动者行为不可靠的提取，推断行动者拥有某种特质的概率。为了检验研究假设即这些主观的概率会不合理地偏高，并且对于可靠性不够敏感，我们把它们与最优概率作比较。最优概率根据贝叶斯模型推导出来，多步骤概率推论的研究者发展了这个模型 (见 Peterson,

1973)。以一个从实际行为 B_j 到某种个人特质 A_i 的单步骤推论开始, 贝叶斯定理有如下熟悉形式:

$$P(A_i/B_j) = P(A_i) P(B_j/A_i) P(B_j) \quad (1)$$

这里 $P(A_i/B_j)$ 是给定 B_j 时 A_i 的后验概率; $P(A_i)$ 是 A_i 的先验概率; $P(B_j)$ 是 B_j 的总概率; $P(B_j/A_i)$ 是给定 A_i 时 B_j 的条件概率。后一种概率代表关于行动者的特质, 行为所具有的诊断价值 (diagnostic value)。从行动者可能选择的每一种替代行为能够推导出后验概率 $P(A_i/B_j)$ 。然而, 观察者并不知道实际上选择了哪一种行为, 他们只能够根据他们回忆的行为 B^* 推断行为的概率。关于这个推论步骤, 贝叶斯定理表示为:

$$P(B_j/B^*) = P(B_j) P(B^*/B_j) P(B^*) \quad (2)$$

这里, $P(B_j/B^*)$ 是在行为 B^* 被回忆起时, 实际的行为 B_j 的后验概率; $P(B_j)$ 是 B_j 的先验概率; $P(B^*)$ 是回忆的行为 B^* 的总概率; $P(B^*/B_j)$ 是在行动者的实际行为是 B_j 的条件下, 回忆行为 B^* 的条件概率。概率 $P(B^*/B_j)$ 反映了可靠性 (reliability) —— 回忆起一实际发生的行为的概率越高, 则可靠性越高。

经由方程 2 推导得到每种行为的后验概率 $P(B_j/B^*)$, 以此作为权数, 贝叶斯二步骤推论模型把 $P(A_i/B_j)$ 值纳入到一个加权平均数中, 每一种行为的 $P(A_i/B_j)$ 值可以根据方程 1 推导得到。因而, 在给定某种回忆的行为时, 能够用下面的公式表示某种特质的后验概率:

$$P(A_i/B^*) = \sum_j P(A_i/B_j) P(B_j/B^*) \quad (3)$$

方程 3 量化地表达了前面讨论过的规范因素 (normative consideration)。特别地, 既然 $P(A_i/B_j)$ 是 $P(B_j/A_i)$ (诊断性) 的一个递增函数, $P(B_j/B^*)$ 是 $P(B^*/B_j)$ (可靠性) 的一个递增函数, 那么 $P(A_i/B^*)$ 应该伴随着可靠性和诊断性的增长而增长, 一个变量的作用依赖于另一个变量的水平 (level)。用安德森 (Anderson, 1974) 的信息整合理论的话来说, 就是对于特质的概率的诊断, 每一种回忆的行为都有一定的影响力。反过来, 我们假定这种影响力是某一行为的诊断性以及对该行为的回忆的可靠性的一个乘法函数。方程 3 确定的整合规则既能够应用于最优的单步骤推论 [即分别从方程 1 和方程 2 获得的 $P(A_i/B_j)$ 和 $P(B_j/B^*)$ 的值], 也

能够应用于主观的单步骤推论 [即对于 $P(A_i/B_j)$ 和 $P(B_j/B^*)$], 被试自己的估计]。在前一种情况中, 单步骤推论以及对推论的整合都是最优的 (即贝叶斯式的), 而在后一种情况下, 只有整合过程是最优的。目前的这个研究用观察概率与这两种预测概率作比较。被试推断学生被研究生院录取的情况以及学生的政治态度。前一个推论基于被试对学生的成绩的回忆, 而后一个推论以被试对学生的行为的回忆为基础。

380 二、实验 1

1. 方法

(1) 概要

首先, 向被试出示学生一门本科课程的成绩; 之后, 被试必须回忆每个学生的成绩, 并且估计每个学生被工商管理研究生院录取的概率。这个任务包括两个推论步骤: (a) 从所提取的成绩 G^* 到学生的实际成绩 G ; (b) 从实际成绩到学生是否被研究生院录取 A 。改变 (vary) 工商管理学院录取成绩的诊断价值, 我们可以控制下一个推论步骤中的不确定性 (uncertainty)。低诊断性的成绩, 是人类学导论课程的期末成绩, 而经济学导论课程的期末成绩, 则是具有高诊断性的成绩。变动成绩要被识记的学生数目, 来操纵记忆信度。

(2) 程序

告知被试某一年, 工商管理学院的申请者中 50% 会被录取。因而, 被录取或者被拒绝的先验概率 $P(A_y)$ 和 $P(A_n)$ 都是 0.50。告知被试, 他们要估计一些随机挑选出来的学生被录取的概率, 并且这些诊断基于给定的一门本科课程成绩是低于 75 (G_l) 还是高于 75 (G_h)。成绩按 0~100 的标准度量。

(3) 诊断价值的操纵

两种概率分布 (用条形图给出) 显示出成绩与录取的关系。一个条形图显示了被录取的学生的比例 (proportions), 他们的成绩为 G_l 或者 G_h 。这些比例代表条件概率 $P(G_l/A_y)$ 和 $P(G_h/A_y)$ 。另一个条形图显示了被拒绝的学生的比例, 即 $P(G_l/A_n)$ 和 $P(G_h/A_n)$ 。高诊断性成绩 (经济学课

程) 的条形图显示了以下比例: $P(G_l/A_n) = P(G_h/A_y) = 0.85$ 和 $P(G_l/A_y) = P(G_h/A_n) = 0.15$ 。低诊断性成绩(人类学课程)的条形图显示了如下比例:

$$P(G_l/A_n) = P(G_h/A_y) = 0.55 \text{ 和}$$

$P(G_l/A_y) = P(G_h/A_n) = 0.45$ 。每个被试诊断两种学生样本: 一个样本中给出的是高诊断性的成绩, 另一个给出的是低诊断性的成绩。对于所有的被试, 样本出示的顺序保持平衡。

(4) 信度的操纵

广泛的预先测试或前测(pretest)指出, 变动其成绩必须被识记的学生数目, 能让我们获得不同的信度水平。告知被试, 样本中有一半的学生得到低成绩, 另一半得到高成绩, 并且会给出一张得到高成绩的学生的名单。为了使这张刺激名单(stimulus list)中各项目之间的差异最小化, 用两位数的身份号码代表这些学生。用磁带录音机给出这些号码, 以听得见的声音, 每秒 1 个的速度。每个刺激名单结束时, 给出一个高音。两秒钟之后, 磁带录音机给出一个测试名单(test list)。测试名单包括 8 个两位数的身份号码, 其中有 4 个在刺激名单中出现过(即 G_h 号码), 4 个没有在刺激名单中出现过(即 G_l 号码)。测试名单中出现的 4 个 G_h 号码, 是从刺激名单的全部系列位置中取出的。每个测试名单中, G_h 和 G_l 号码的顺序是随机决定的。测试号码以每 10 秒钟 1 个的速度给出。在间隔期间内, 被试必须指出该学生的成绩更有可能是 G_h 还是 G_l , 并且估计该学生被研究生院录取的概率, $P(A_y/G^*)$ 和被拒绝的概率, $P(A_n/G^*)$ 。前测显示, 做出这些判断, 10 秒钟的间隔是足够的。

使用三种名单长度: 4 个项目的名单, 7 个项目的名单和 14 个项目的名单。前测的数据表明, 正确回忆的百分比的范围是从最长名单的大约 60% (低可靠性条件), 到中长名单的大约 75% (中等可靠性条件), 到最短名单的大约 90% (高可靠性条件)。每名被试对应一个高诊断性的名单和一个低诊断性的名单。这些名单中, 号码的组合对比诊断性保持平衡。三种可靠性条件, 每一种分配 20 名被试。

(5) 完全可靠性条件

在这种单步骤推论的情况中, 一组 46 名被试根据确切知道的成绩, 推

断被录取和被拒绝的概率 $P(A_y/G)$ 和 $P(A_n/G)$ 。被试得到一个小册子,上面显示了条形图(即诊断性信息)和学生的成绩。阅读完一个学生的成绩后,被试简单地说明他们的概率判断。每名被试对应一个有低诊断性成绩的学生和一个有高诊断性成绩的学生。对于一半的被试,学生的成绩低,而另一半,成绩高。

(6) 被试

一共 106 名被试(42 名男性和 64 名女性),他们是耶路撒冷希伯来大学心理学导论课程的学生,他们独立地参加实验。参加实验是课程要求完成的部分任务。

2. 结果与讨论

正确回忆低成绩 $P(G_l^*/G_l)$ 和高成绩 $P(G_h^*/G_h)$ 的比例,作为每个被试的可靠性的量度。 $P(G_l^*/G_l)$ 和 $P(G_h^*/G_h)$ 的均值表明,在低可靠性(14 个项目的名单)条件下,被试的表现近似于运气(near chance)(均值分别为 $M=0.54$ 和 0.60),在高可靠性(4 个项目的名单)条件下,表现出色(excellence)(均值分别为 $M=0.93$ 和 0.89),在中等可靠性(7 个项目的名单)条件下,表现中等(intermediate)(均值分别为 $M=0.73$ 和 0.77)。对这些比例的方差分析(可靠性 $\times$ 诊断性)显示,可靠性的主效应(main effect)是高度显著的, $P(G_l^*/G_l)$ 和 $P(G_h^*/G_h)$ 的 F 检验分别为, $F(2, 57) = 37.06$, $p < 0.001$ 和 $F(2, 57) = 20.17$, $p < 0.001$ 。没有其他的变差来源(source of variation)显著。

完全可靠性条件下的数据显示,在从确切知道的成绩到录取的概率 $P(A/G)$ 这一单步骤推论过程中,对诊断性的控制具有明显作用。与根据低诊断性的成绩(均值分别为 $M=0.57$ 和 0.55)进行的推断相比,从高诊断性的成绩(均值分别为 $M=0.81$ 和 0.78),被试推断的 $P(A_n/G_l)$ 和 $P(A_y/G_h)$ 更高;对 $P(A_n/G_l)$ 和 $P(A_y/G_h)$, 差异都是显著的, t 检验分别为 $t(22) = 8.08$, $p < 0.001$ 和 $t(22) = 8.59$, $p < 0.001$ 。

现在,我们来考虑给定回忆成绩时,录取的最优后验概率 $P(A/G^*)$ 。根据方程 3,对于每名被试,给定回忆成绩的情况下,实际成绩的概率 $P(G/G^*)$ 和给定实际成绩的情况下,录取的概率 $P(A/G)$ 二者结合决定

了最优后验概率。经由方程 2, 每名被试的 $P(G/G^*)$ 值根据他自己对回忆行为的可靠性 $P(G^*/G)$ 计算。 $P(A/G)$ 的值是完全可靠性条件下得到的诊断的均值。因而, 在给定回忆成绩的情况下, 录取的最优后验概率 $P(A/G^*)$ 基于一个主观的单步骤推论——从实际成绩到录取的概率——和一个最优的单步骤推论(方程 2)——从回忆成绩到实际的成绩。图 27—1 中左边部分的虚线代表了根据回忆, 学生的成绩为低的情况下, 该学生被拒绝的最优概率 $P(A_n/G_l^*)$ 的均值; 图 27—1 右边部分的虚线代表了, 根据回忆, 学生的成绩为高的情况下, 该学生被录取的最优概率 $P(A_y/G_h^*)$ 的均值。^[1] 显然, 贝叶斯模型预测, 这些概率将伴随着可靠性和诊断性的增加而增加, 当诊断性高时, 可靠性的作用比诊断性低的时候更加显著。

SSS

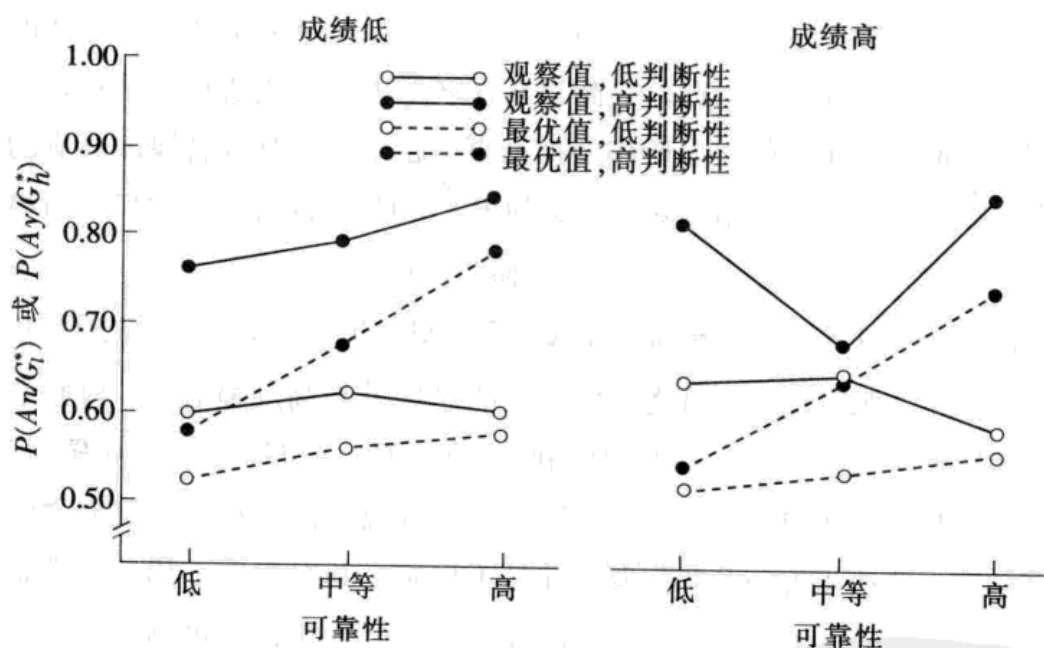


图 27—1 $P(A_n/G_l^*)$ 与 $P(A_y/G_h^*)$ 的观察值和最优值,
作为可靠性和判断性的函数 (实验 1)

每个被试需要确定两个观察值: 第一个是他估计的一些学生的 $P(A_n/G_l^*)$ 的一个平均值, 根据他的记忆, 这些学生的成绩为低; 第二个是 $P(A_y/G_h^*)$ 的一个平均值, 根据他的记忆, 这些学生的成绩为高。图 27—1 中的实线表明, 这些观察概率在性质上和数量上都偏离最优概率。

真实成绩的诊断价值是唯一对观察值 $P(A_n/G_l^*)$ 和 $P(A_y/G_h^*)$ 具有持续作用的因素, F 检验分别为 $F(1, 57) = 96.97, p < 0.001$ 和 $F(1, 57) = 114.82, p < 0.001$ 。真实的录取成绩的诊断价值越高, 后验概率越极端。真实成绩被回忆出的可靠性对于推论只有非常弱而且不稳定的作用 ($F < 1$)。甚至于对每一种诊断性层次, 可靠性的两个极端层次上的推论都没有显著的差异。换句话说, 当正确回忆的比率处于近似运气的水平时, 推论并不比处于几乎完美水平时候的推论更不极端。被试的推论也不能显示出贝叶斯模型所预测的诊断性 $\times$ 可靠性的交互作用。当可靠性减少时, 真实成绩的诊断性的作用并没有减弱。这些变量的交互作用也在关于学生个体的推论的层次上得到检验。规范地来说 (Normatively), 与成绩具有低诊断价值的情况相比, 当成绩具有高诊断价值时, 对于不同学生的成绩, 被试在信心方面的变化 (variation in confidence) 会对关于学生录取情况的推论产生更大的作用。因此, 对高诊断性名单中的学生所作的推论, 应该比对低诊断性名单中的学生所作的推论更不稳定。计算每名被试对应的两个名单中每一个名单包括的 8 名学生的 $P(A/G^*)$ 的方差。与贝叶斯模型相反, 这些方差没有显著的差异, t 检验为 $t(59) = 1.63$ 。

图 27—1 也表明, 观察概率比最优概率更加极端。对于 $P(A_n/G_l^*)$ 和 $P(A_y/G_h^*)$, 观察概率与最优概率之间的总体差异高度显著, F 检验分别为 $F(1, 57) = 33.38, p < 0.001$ 和 $F(1, 57) = 84.67, p < 0.001$ 。最后, 对于 6 种条件中的每一种, 计算所有被试观察概率和最优概率之间的相关系数。这些相关关系非常弱, 并且不稳定, 其中有 5 个相关关系是正的, 7 个是负的。有 2 个相关关系显著, 但是为负 [在低可靠性—高诊断性条件下, 对于 $P(A_n/G_l^*)$ 和 $P(A_y/G_h^*)$, 相关系数分别为 $r = -0.50, p < 0.05$ 和 $r = -0.55, p < 0.01$]。

这些结果证明, 被试不能把可靠性因素整合到他们的概率推论之中。看来, 被试仅仅根据真实成绩的诊断价值进行推论。在实验条件下推论得到的概率, 稍稍低于在完全可靠性条件下推论得到的概率, 这表明被试并没有完全采取一种 “as-if” 策略 (即在进行推论时, 真实成绩好像是确切地知道一样)。但是, 信心方面的减少是非常小的, 并且没有伴随可靠性一

致变化。结果，与贝叶斯概率相比，被试的推论是极端的，并且未经保证。注意，对于从真实成绩到录取的概率 $P(A/G)$ 的单步骤推论，情况并不是这样。在完全可靠性条件下推论得到的概率和根据贝叶斯式单步骤推论模型（方程 1）推导出来的最优概率非常接近。这个结果证明，在二步骤推论情况中，非最优性（或者过度自信）是因为引入了额外的不确定性来源——对真实成绩的记忆的不完善……

三、实验 4

1. 方法

在这个实验中，被试推断男学生支持 (A_p) 或者反对 (A_n) 归还阿拉伯人西岸地区（1967 年战争中被以色列占领的地区）的概率。投票赞成工党（一个相对鸽派的政党）或者利库德集团（一个相对鹰派的政党）作为具有高诊断性的信息。学生的头发长度，长对短，作为具有低诊断性的信息。前测显示，短头发和投票支持利库德集团 (B_n) 意味着对归还西岸地区持否定的态度，而长头发和投票支持工党 (B_p) 意味着一种肯定的态度。前测也表明，学生的投票行为比他的头发长度更有诊断价值。

再现记忆的任务以及设计和程序的其他方面都与实验 1 相同。3 种可靠性条件中的每一种分配 20 名被试。

一个有 57 名被试的附加组被分配到完全可靠性条件。在这种情况下，被试确切地知道学生的头发长度或者他投票支持的党派。每名被试根据短头发、长头发，支持利库德集团和支持工党来推断学生的态度概率。

一共 117 名被试（51 名男生和 66 名女生）参加了实验 4，他们是耶路撒冷希伯来大学心理学导论课程的学生。

2. 结果与讨论

可靠性控制对于正确回忆 B_n （即短头发或者投票支持利库德集团）和 B_p （即长头发或者投票支持工党）的比率有强大作用， F 检测分别为 $F(2, 57) = 55.45$, $p < 0.001$ 和 $F(2, 57) = 22.04$, $p < 0.001$ 。对于低、中等和高可靠性条件，概率 $P(B_n^*/B_n)$ 的均值分别为 0.53、0.76 和 0.96；概率 $P(B_p^*/B_p)$ 相应的均值分别为 0.61、0.82 和 0.91。诊断性以

及它与可靠性的交互作用的影响都不显著。

和预期的一样,被试在完全可靠性条件下推断的概率 $P(A/B)$ 值表明,投票行为的主观诊断价值大大高于头发长度的诊断价值。因而,与根据学生留短发 ($M=0.52$) 推断得到的概率值相比,根据学生投票支持利库德集团 ($M=0.78$) 这一事实推断得到的态度概率 $P(A_n/B_n)$ 更加极端, t 检验为 $t(56) = 13.61, p < 0.001$; 类似地,与根据学生留长发 ($M=0.53$) 推断得到的概率值相比,根据学生投票支持工党 ($M=0.75$) 这一事实推断得到的态度概率 $P(A_p/B_p)$ 更加极端, t 检验为 $t(56) = 9.89, p < 0.001$ 。

图 27—2 显示了最优概率和观察概率。根据 $P(B/B^*)$ 的值 [经由方程 2, 根据被试回忆学生行为的可靠性 $P(B^*/B)$ 计算] 和完全可靠性条件下得到的 $P(A/B)$ 的均值,经由方程 3, 推导出每个被试的最优概率。可以看到,行为的诊断性是引导态度概率推论的唯一因素,对 $P(A_n/B_n^*)$ 和 $P(A_p/B_p^*)$, F 检验分别为 $F(1, 57) = 68.94, p < 0.001$ 和 $F(1, 57) = 90.40, p < 0.001$ 。^[2] 与贝叶斯模型相背,可靠性以及它与诊断性的交互作用的主要作用都不显著。此外,高诊断性名单中,观察概率的方差并不大于低诊断性名单的方差, t 检验为 $t(59) = 0.59$ 。和在先前的实验中一样,观察概率比最优概率更加极端,对 $P(A_n/B_n^*)$ 和 $P(A_p/B_p^*)$, F 检验分别为 $F(1, 57) = 5.68, p < 0.05$ 和 $F(1, 57) = 27.63, p < 0.001$ 。最后,在各种条件下,所有被试的观察概率与最优概率之间的 12 个相关关系中,8 个是正的,但是只有 2 个显著:在高诊断性—高可靠性条件下,对 $P(A_n/B_n^*)$, $r = 0.436, p < 0.05$; 在低诊断性—高可靠性条件下,对 $P(A_p/B_p^*)$, $r = 0.402, p < 0.05$ 。

就我们研究发现的概括性而言,实验 4 提供了额外的支持。它表明,在先前的实验中观察到的偏差,对于根据回忆行为进行的态度推论也有影响。行动者的实际行为的诊断性,是推论的唯一决定因素,甚至在诊断性不得不基于被试自己的信念而没有任何外部的精确证据时。

四、一般讨论

我们的被试根据他们自己并不可靠的记忆做出的推论,系统地违背了

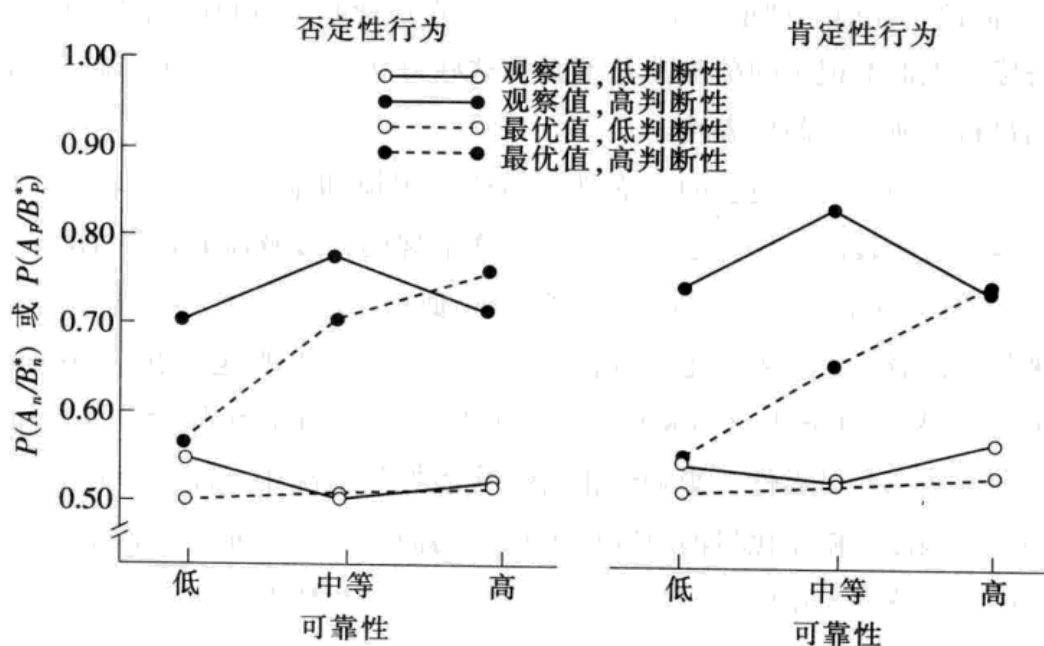


图 27—2 $P(A_n/B_n^*)$ 与 $P(A_p/B_p^*)$ 的观察值和最优值,
作为可靠性和判断性的函数 (实验 4)

概率推论的规范准则。被试倾向于完全依赖信息的诊断性, 尽管觉知到自己的记忆并不可靠, 他们仍然对可靠性不敏感。事实上, 根据正确回忆的信息做出的推论 (例如, 根据回忆, 被试认为学生的成绩低, 并且成绩在事实上低的情况下的推论) 与根据错误回忆的信息做出的推论 (例如, 根据回忆, 被试认为成绩低, 而事实上成绩高的情况下的推论) 之间的比较表明, 全部 4 个实验中的差异不显著 (insignificant), 可以忽略不计。也就是说, 后一种概率推论和前一种一样极端。这些结果与这种观点一致, 它认为人们倾向于运用一些简单的启发式规则, 从而减低任务的复杂性。对信息的可靠性的忽略, 可能反映人们使用了一种诊断启发式策略即代表性启发式策略 (Kahneman & Tversky, 1973, 4)。例如, 被试回忆行动者投票所支持的政党, 根据投票支持该政党这一行动代表持有某种态度的行动者刻板印象 (stereotype) 的程度, 推断该行动者支持归还阿拉伯人西岸地区的概率。与一个留长发的行动者相比, 一个投票支持鸽派政党的行动者更多地体现了这种刻板印象的特征 (在前 3 个实验中, 经济学课程的高成绩, 比人类学课程的高成绩更多地代表了那些被工商管理学院录取的学生的特

征)。因此,在前一种情况中,可以更加确定地推断出对归还该地区持积极的态度。人们的记忆的精确性并不影响这些概率,因为它并不影响人们判断回忆行为所具有的代表性。

在这方面,一个报告的可靠性,与样本规模和基率(base rate)之类的因素是可比较的。作为证据的基础的样本规模,以及需要推断的特质的基率,并不影响证据本身与特质之间的相似性。因此,代表性概念暗示,在进行推论时,人们不考虑这些因素,研究也已经表明这一点(Nisbett & Borgida, 1975; Tversky & Kahneman, 1974, 1)。基于这个概念,人们可能会期望,在那些可靠性会影响到报告的代表性的情况中,能够考虑到可靠性因素。因而,使用比目前研究中所运用的简单的、二元性事件细节更为丰富的行为描述,微弱的记忆可能会影响回忆信息的明晰性、完备性和形象性。这种模糊的信息可能更少代表了该种特质,因此,会产生更具回归倾向或者更加温和的概率诊断……

本章节选自《人格和社会心理学杂志》(*The Journal of Personality and Social Psychology*, 1978, 36, 93-106)。允许重印。

注释

[1] 经由方程1推导出最优概率 $P(A/G)$,据此计算出最优概率 $P(A/G^*)$ 的另一组值。我们不再单独讨论这些概率,因为它们与从主观估计的 $P(A/G)$ 中推导出的最优概率 $P(A/G^*)$ 的值非常接近。

[2] 应该注意,这些结果不能归于诊断性是作为被试内(within-subjects)变量,而可靠性是作为被试间(between-subjects)变量。关于第一个名单中的学生的推论,在每一种可靠性条件下,有一半的被试,学生行为的诊断价值低,而另一半诊断价值高。对这些推论作方差分析,在分析中,诊断性当然是一个被试间因素。分析得到,诊断性的作用高度显著,可靠性以及诊断性 $\times$ 可靠性的作用不显著。对前面3个实验的数据的分析得到了类似的结果。

第八部分
PART EIGHT

校正程序



第 28 章

决策中非常规线性模型的优美之处

道斯 (Robyn M. Dawes)

391

米尔 (Paul Meehl, 1954) 的《临床预测与统计预测：理论分析与证据评论》一书 25 年前发表。该书回顾了一些研究，这些研究表明，根据数值型的预测变量（例如，研究生入学考试的成绩、平均绩点、推荐信的等级）去预测心理数值型的标准变量（例如，刚刚获得哲学博士学位的研究生的能力等级），与那些被认为是擅长这种预测的人的临床直觉相比，常规线性模型表现得更好。本章的目的是为了评论证据，以表明甚至是非常规线性模型也可能优于临床预测。

常规线性模型，是指在模型中，赋予预测变量某种权数，从而使预测与标准之间的关系最优化。简单回归分析，是常规线性模型最普通的例子；它加权预测变量以使得加权后的混合式与实际标准之间的相关最大化。判别函数分析，是另一个常规线性模型的例子；预测变量权数的赋予，使结果得到的线性混合式

能最大地显明两个或者更多组之间的差异。另一个例子是脊状 (ridge) 回归分析 (Darlington, 1978; Marquardt & Snee, 1975), 它试图分配权数从而使得线性混合式与一组新数据的兴趣标准之间相关最大。

因而, 有许多种类型的常规线性模型, 它们在多种情境中被使用。一个例子 (Dawes, 1971) 涉及的是教员预测研究生的能力等级。1969 年春天, 俄勒冈 (Oregon) 大学心理学系评定该系 1964 年秋至 1967 年秋之间入学的所有研究生, 他们没有因为非学术的原因 (例如, 精神病或者结婚) 而中途退学; 各位教员仅仅评定他们为等级相当不错的学生。评定使用下面的等级尺度: 5, 杰出; 4, 中上; 3, 平均; 2, 中下; 1, 由于学习困难而被迫退出。这种完全等级构成了一个心理学上令人感兴趣的标准, 因为系里教员的主观印象, 是一个学生离开研究生院之后能获得 (如果有的话) 工作的主要决定因素。样本一共 111 名学生。评定每个学生的教员数目从 1 到 20 不等, 平均数值为 5.67, 中位数为 5。评定等级是可靠的。为了判定信度, 对等级进行过一次单向 (one-way) 方差分析, 其中每个学生的评定被视作一次处理; 结果得到的处理间的方差比 (n^2) 为 0.67, 它在 0.001 水平之上都显著。这些能力评定, 也通过常规线性模型被预测; 而这个常规线性模型, 则基于学生的研究生入学考试 (GRE) 成绩, 学生的本科平均绩点 (GPA) 和学生本科院校的可选择性的测量。^[1] 能力评定等级与预测变量之间的交叉一可靠性多元相关 (the cross-validated multiple correlation) 为 0.38。与米尔的结果一致, 这些后来的能力评定等级与选择学生的招生委员会给出的平均等级之间的相关关系为 0.19^[2]; 也就是说, 它解释了四分之一那么多的方差。这个例子, 是这一领域心理学研究发现的典型, 因为: (a) 与模型的预测的相关关系高于与临床判断的相关关系, (b) 两种相关关系都低。这些特征, 被心理学家用来解释这一发现: 它意味着, 尽管模型的低相关表明线性模型作为一种方法是不完善的, 但是判断者比之更低的相关关系, 却仅仅表明使用了错误的判断者。

非常规线性模型, 是指模型中权数的选择根据一些非最优的方法。权数, 可以选择为等同的, 也可以根据预测者的直觉来选择, 或者随机地选择。但是, 非常规模型可能具有重要的效用 (utility)。例如, 当对前一个

例子中标准化的 GREs、GPAs 和选择性指数平均加权时，结果得到的线形混合式与后来的教员评定等级之间的相关系数为 0.48。线性混合式的相关关系，不仅高于与入学委员会的临床判断的相关关系（0.19），而且也高于对分半样本的权数所进行的交叉一可靠性检验的相关关系。

一个可能更令人感兴趣——至少是对一般大众而言——的非常规模型的例子，是由一位物理学家促成的，那时他和我一起在一个研究小组研究预测系统。后来有一次，当他和他的妻子及我在酒吧时，他说我的论文可能会令我的同事有点兴趣，但是在心理学系研究生院的成功并不能引起更广泛的兴趣：“例如，你能不能运用你的一个非常规线性模型预测我和我的妻子相处的情况呢？”他问道。我意识到我能够——或者说可能。那时，俄勒冈大学心理学系正在从事性方面的研究，其中大多数是行为主义导向的，因此当该研究的被试做爱时、争吵时、当他们进行社交活动（例如，与姻亲）时等，他们都处在监控之下。这些被试也主观评价他们的婚姻或者夫妻生活的幸福程度。我立刻想到一个非常规线形模型来预测对婚姻幸福的自我评价：用做爱的频率减去争吵的频率。正好我的同事荷华德（John Howard）在堪萨斯城（Kansas City）密苏里（Missouri）大学读本科时，曾经收集过这类与夫妻生活有关的数据，那时他与亚历山大（Alexander）一起工作（1971）。确定夫妻间关于做爱和争吵的判断的可靠性之后，亚历山大从 42 对夫妇的每一对中，找到一个合作者来监控这些事件，她允许我们分析她的资料，得到下列结果：“在 30 对婚姻幸福的夫妇中（如同监控合作者报告的），只有两对夫妇争吵的频率多于他们做爱，所有 12 对婚姻不幸福的夫妻都争吵得更频繁。”（Howard & Dawes, 1976, p. 478）后来，我们在俄勒冈大学重复了这一发现。这一次 27 位监控者根据一种 7 - 项目量表评定幸福等级，从“非常不幸福”到“非常幸福”，有一个中性的中间指标。做爱与争吵的频率之差和这些婚姻幸福程度的等级评定之间的相关关系为 0.40 ($p < 0.05$)；没有一个变量单独显著。爱德华兹等（1977）在密苏里，邵登（Thornton, 1977a）在得克萨斯（Texas），也重复得到了这一发现。后者发现，在 28 对新婚夫妇中，做爱—争吵的差异与对婚姻幸福的自我评定之间的相关关系为 0.81 ($p < 0.01$)。这一相关关系如此高的原

因可能在于邵登是在被试监控他们的做爱和争吵之后，而不是之前得到关于婚姻幸福的等级评价；事实上，一位被试意识到她争吵超过做爱之后，就会决定离婚（Thornton, 1977b）。这些研究的结论是，如果我们的爱超过我们的恨，我们就是幸福的；如果我们的恨超过我们的爱，我们就是不幸的。结论并不深奥，无论是在心理学还是在统计学上。关键在于，这样一个非常粗糙的非常规线性模型，预测了一个非常重要的变量：对婚姻幸福的判断。

自从米尔（1954）的著作发表以来，大部分（事实上是全部）的文献，支持他关于常规模型和直觉的临床判断的一般结论。有学者（Sawyer, 1966）评论了这方面的大量研究，其中的一些研究范围非常广泛（见 L. R. Goldberg, 1965）。但是，在他的书发表 10 年之后，米尔（1965）得出结论，仅有一个简单的案例表明临床判断优于常规线性模型，并且这一结论，立刻受到他人（Goldberg, 1968a）的质疑，后者发现，甚至这一案例也没有表明临床判断的优越性。霍尔特（Holt, 1970）评论了几项研究的细节。他甚至提出，与理解不同，预测可能并不是临床判断的一个非常重要的部分。但是对文献的搜索，并不能发现任何一项研究，当二者都是基于同样的可编码输入变量时，临床判断曾经被证明优于统计预测。并且，尽管大多数非实证主义者会同意理解并不等同于预测，但是很少有人会同意理解并不需要一些预测能力。

为什么呢？因为人们——尤其是某个领域的专家——更擅长于选择和编码信息，而不是整合信息。

但是，人是重要的。统计模型可能以最优的方式整合信息，但是选择变量的总是个体的人（判断者、临床医生、被试）。并且，了解预测变量与兴趣标准之间的方向关系，或者能够以这样一种方式编码变量从而使得变量之间具有明确的方向关系的，也是人类判断者。也恰恰是在那种预测变量有效并且它们与标准之间存在条件性单调关系的情境中，常规线性模型才充分地起作用。^[3]

在决定“寻找什么”之类的事情时，线性模型并不能取代专家，但是恰恰是这种关于在做出决定时应该寻找什么的知识，是人所拥有的特殊的

专家知识。甚至在其复杂程度有如决定下一步棋的走法的判断中，也是以一种合适的方式编码棋盘从而“看到”合适走法的能力，区分了大师、专家和新手（deGroot, 1965; Simon & Chose, 1973）。人胜出，不是因为整合信息的能力（Slovic, 1972b）。其次，棋艺大师考虑的步数，并不比专家更多；他只是知道去考虑哪一步。爱恩荷恩（Einhorn）的一项研究，可能最出色地说明了知道去寻找什么与整合信息能力之间的区别。专家医生编码霍奇金氏病（Hodgkin's disease）病人的活组织检查，然后对病情的严重程度做出一个总体评定。总体评定，并没有预测 193 个病人的存活时间，他们全部死了。评定与存活时间的相关关系，实际上全部为 0，一些相关关系甚至方向错误。但是，医生编码的变量，当被运用到一个多元回归模型中时，确实预测了存活时间。

总体上说，常规线性模型工作的原理非常简单。人们擅长于挑选出正确的预测变量，擅长于以某种方式对它们编码，从而使它们与标准之间存在条件性单调关系。但人们拙于整合来自多个方面且不可比较的来源的信息。当预测与标准之间存在一种条件性单调关系时，常规线性模型优于这种整合。

例如，考虑这个问题，比较一个 GRE 成绩 750、本科平均绩点（GPA）为 3.3 和另一个 GRE 成绩 680、本科平均绩点为 3.7 的研究生申请人。大多数判断者会同意，这些能力指标和先前成绩应该以某种补偿方式结合起来考虑，但是问题是怎么进行补偿。许多尝试这件工作的判断者，对 GREs 和平均绩点的分布特征了解很少，而大多数人，也根本不了解许多研究的知识，而这些研究，主要关注这些分数作为研究生学业成功的预测指标的有效性。此外，如果不具备这些知识，这些数字本身是不可比的。有希望的申请人，其 GREs 范围是从 500 到 800，而平均绩点是从 3.0 到 4.0。在这种情况下，一种统计的加权方案会优于人类判断者，这奇怪吗？

现在来假设，在某些情境下，不可能构造一种常规线性模型。我们之所以不能这样做的一个原因，是我们的样本规模不足。例如在多元回归中，权数 b 是出名地不稳定；权数 b ——最优权数——要能在交叉-可靠性检验中优于简单的单位权数，观察资料与预测指标的比率，应该达到 15 比 1 甚至是 20 比 1。许多学者（Schmit, 1971; Goldberg, 1972; Claudy, 1972），

利用计算机模型理论上证明了这一要求，而爱恩荷恩等 (Einhorn & Hogarth, 1975; Srinivisan, 1977)，则采取分析方法处理了这一问题。一般的解决方法，依赖于大量参数，如总体中的多元相关和预测变量之间的共变模式。但是，应用含义很明显。而在一些情境中，如果没有“合适” (decent) 的观察资料与预测指标的比率，标准回归分析，就不能用。

另一种不能运用常规线性模型的情境，不存在可测量的标准变量。但是，我们可能知道重要的预测变量会是什么，并且如果我们能够测量标准的话，我们可能知道变量与标准之间关系的方向。例如，当决定哪些学生可进研究生院的时候，我们希望预测一些未来的长期变量，它们可以被称作“职业的自我实现”。我们知道，使用这个概念，我们所指的意思是什么，但是仍然没有有效的、精确的定义。即使我们有一个定义，也不可能用目前学生的记录进行这项研究。因为至少要在学生已经完成他们的博士学习之后 20 年，这个变量才能被评估。但是，我们确实知道这一标准与智力、过去的成绩和超过同学的能力之间很可能存在有明确的关系。在我们的申请人档案中，GRE 成绩评估了第一个变量；本科平均绩点成绩评估了第二个变量；推荐信评估了第三个变量。那么，为了评估我们的申请人的潜力，我们能不能希望构造关于这些变量的某种类型的线性组合吗？考虑到我们不能进行标准的回归分析，当我们评估我们的申请人时，难道除了对这些变量进行毫无帮助的直觉整合，就没有其他办法了吗？

构造非常规线性模型的一种可能方法，是通过使用自助法 (bootstrapping) (Dawes & Corrigan, 1974; L. R. Goldberg, 1970)。其过程，是先构造一个专家关于结果标准的判断的常规线性模型，然后用这个线性模型来替代判断。有人 (Hammond, 1955; Hoffman, 1960) 指出，在心理学的文献中，这种线性模型能够精确地预测专家的判断。32 年前，罗斯福的前任副总统 (Henry Wallace) 华莱士就已经这样做了。在 1923 年一篇关于农业的文章中，他建议使用线性模型来分析“玉米专家脑子里在想些什么”。在他的一篇颇有影响力的文章中，霍夫曼把线性模型的使用称作是对判断者的同质异晶描述 (paramorphic)，这个词语的意思是指，判断者的心理过程，并不包括计算输入变量的一种隐含的或者外在的加权平均值，但是能

396

够通过这样一种加权来模拟他的心理过程。在预测变量与标准变量之间存在条件性单调关系的情境中，同质异晶描述已经非常成功（评论见 Dawes & Corrigan, 1974; Slovic & Lichtenstein, 1971）。

自助法模型，利用了根据判断者推导出的权数；因为这些权数，不是源于预测指标和标准变量本身之间的关系，所以结果产生的线性模型是非常规的。然而，这些同质异晶描述的表现，始终优于它们由之导出的判断者（至少是在根据预测值和实际值之间的相关关系评估有效性时）。结果已经表明，自助法具有普遍性。例如，在维京斯等人（Wiggins & Kohn, 1971）主持的一项研究中，伊利诺伊（Illinois）大学心理系的研究生，被呈现描述伊利诺伊大学心理学系其他（真实的）研究生的 10 种背景（background）、能力（aptitude）和人格（personality）指标，他们被要求预测这些研究生第一年的平均绩点。在预测实际的平均绩点方面，每位伊利诺伊大学判断者的线性模型，都要优于判断者自己的表现。在维京斯等（引用见 Dawes & Corrigan, 1974）合作主持的一项研究中，这一结果得到了验证。而戈登伯格（Goldberg, 1970）也证明了这一结果：29 位临床心理学判断者中有 26 位根据明尼苏达多项人格测验（MMPI）的剖面图预测神经症或者精神病的诊断。道斯（1971）在俄勒冈大学评估研究生申请人时也发现了这一结果。关于自助法的效果，我所知道的一个曾经发表的例外，是利比（Libby）主持的一项研究（1976）。他的被试是 16 名来自较小银行（资产在 300 万美元到 5 600 万美元之间）的贷款部门的职员和 27 名来自大银行（资产在 6 亿美元到 44 亿美元之间）贷款部门的职员。他们要判断在公布他们的财务报表之后的 3 年内，60 家公司中的哪 30 家公司会倒闭。贷款部门的职员要求根据 5 种财务比率（例如，现期资产与总资产的比率）做出判断。平均而言，贷款部门的职员，正确归类了 44.4 家企业（74%），它们或者有偿付能力或者将来破产；但是平均而言，贷款部门的职员的两质异晶描述，只能正确归类 43.3 家企业（72%）。结果证明这一差别在统计上显著。因此，利比得出结论：这是一个自助法不起作用的案例——可能是因为他的判断者是受过高度训练的专家，他们试图预测一个高度可靠的标准。但是，戈登伯格（1976）指出，许多比率分布属于严重偏斜分布，

因此，他重新分析了利比的数据，在构造贷款职员模型前先标准化比率。利比发现，他的研究中 77% 的职员表现优于他们的同质异晶描述，但是戈登伯格使用调节过的预测变量发现相反的情况；72% 的模型表现优于它们所推导出的判断者。^[4]

为什么自助法会起作用呢？有些学者（Bowman, 1967; Goldberg, 1970; Dawes, 1971）坚持认为，自助法成功的原因，在于线性模型从其他可变行为（例如，受情境效应或者外部变量影响的判断）中提取出潜在的策略（在隐含的权数中）。

对于自助法的效能的信念，是基于判断者的线性模型的有效性与判断者判断自身的有效性的综合。从逻辑上来说，这仅仅是两种可能的比较中的一种。另一种是判断者的线性模型的有效性与一般线性模型的有效性的比较；也就是说，为了证明自助法起作用是因为线性模型在剔除了不可靠因素的同时吸取了判断者有效的专家知识的精髓，那么必须证明，根据分析判断者的行为得到的权数优于可能用其他方法得到的权数，例如随机方法。因为判断者模型和随机方法得到的模型二者都是完全可靠的，所以随机模型与判断者模型之间的比较，使得对于判断者的潜在线性描述，或者说策略的评估成为可能。如果随机模型的表现同样令人满意，那么判断者就不是在“遵循某些有效的原理，而是拙劣地遵循它们”（Dawes, 1971, p. 182），至少不是比其他任何在正确的方向上加权变量的原理更加有效的原理。

表 28—1 给出了道斯等总结的 5 项研究，它比较了通过各种方法获得的有效性（也就是相关系数）。在第一项研究中，一批 861 名精神病人在各个医院进行明尼苏达多项人格测验（MMPI）；之后，根据更加详尽的信息，把他们归类为神经症患者或者精神病患者。明尼苏达多项人格测验的剖面图包括 11 项得分，每一种代表了调查对象回答问题的方式与患有明确界定的精神病理学类型病人的类似程度。因而每一名病人与一组 11 项分数相连，而问题是预测以后的诊断，病人是属于精神病（编码为 1）还是神经症（编码为 0）。29 名“富有不同经验、受过不同训练的”临床心理学家（Goldberg, 1970, p. 425）被要求根据一个 11 级的选择式正态分布（a 11-step forced-normal distribution）进行预测。接下来的两项研究，涉及的是伊

利诺伊大学心理系的 90 名一年级研究生，要根据 10 种预测学术成果的变量来评估他们。这些变量，包括能力测验成绩、大学平均绩点、各种同龄人评定的等级（例如，外向性）和各种自我评定（例如，责任心）。计算所有这些学生第一年的平均绩点。而问题是根据这 10 种变量预测他们第一年的 GPA。第二项研究，是由伊利诺伊大学的 80 名（其他专业的）研究生进行预测。第三项研究，则是由俄勒冈大学的 41 名研究生进行预测。第四项研究的细节，已经在前面叙述过了；它是一项关于预测以后的能力评定等级的研究，在俄勒冈进行。而在最后一项研究（Yntema & Torgerson, 1961）中，给被试出示一些椭圆，实验者根据椭圆的大小、离心率和灰度为椭圆赋值，使用的公式是 $ij + kj + ik$ ，其中 i 、 j 和 k 指刚刚提到的三种维度的数值。在这个实验中，被试被要求估计每个椭圆的值，并且在每一回合结束时给出结果反馈。问题是根据椭圆的大小、离心率和灰度预测每个椭圆的真实值（即实验者赋予的数值）。

表 28—1 预测与标准值之间的相关

案例	判断者的 平均效度	判断者模 型的平均 效度	随机模型 的 平 均 效度	平均加权 模型 的 效度	回归分析 的 交 叉 效度	最优线 性模型 的效度
预测神经症患者或精神病患者	0.28	0.31	0.30	0.34	0.46	0.46
伊利诺伊大学学生 GPA 的预测	0.33	0.50	0.51	0.60	0.57	0.69
俄勒冈大学学生 GPA 的预测	0.37	0.43	0.51	0.60	0.57	0.69
预测俄勒冈大学学生日后的能力等级	0.19	0.25	0.39	0.48	0.38	0.54
Yntema 和 Torgerson (1961) 的实验	0.84	0.89	0.84	0.97	—	0.97

注：GPA = 平均绩点。

表 28—1 的第一栏，给出了这些研究中判断者的平均效度，第二栏给出了这些判断者的同质异品描述的平均效度。在所有个案中，自助法都起作用。但是那时，同事和我构造的是随机线性模型，也就是说，在模型中，除了符号，权数是随机选择的，然后应用于标准化的变量。^[5]

每个变量的符码 (sign)，根据一种先验的基础来决定，因此变量与标准之间存在正关系。然后，根据一个单位方差的正态分布随机选择一个常规偏差，这个偏差的绝对值用来作为该变量的一个权数。每个案例构造 10 000 个这种模型。(Dawes & Corrigan, 1974, p. 102)

平均而言，这些随机线性模型的表现，大约与判断者的同质异品模型一样令人满意。表的第三栏，预测了这些平均值。第四栏给出的平均加权模型，甚至表现得更好。平均加权模型之所以必然会优于平均随机模型，有一个数学方面的原因。^[6]而最后两栏，给出了标准回归模型的交叉—有效性检验的效度和最优线性模型的效度。

当根据矩形分布选择权数时，得到了本质上一样的结果。为什么？因为线性模型的强检验性克服了对于最优加权的偏离。换句话说，至少在这些研究中，自助法的研究发现，仅仅是重新证实了先前的研究发现，即常规线性模型优于人类判断——根据判断者的行为推导出的权数几乎接近于最优权数，从而模型的结果也高度相似。解决获得最优权数问题的一个方法是——用温特菲尔特 (Von Winterfeldt) 和爱德华兹 (1973) 的话来说——使用“固定最大值” (flat maximum)。接近于最优水平的权数，产生与最优的第二位权数 (beta weights) 几乎同样的结果。因为专家判断者至少了解变量的方向，所以他或她的判断得出的权数几乎是最优的（但是注意，在所有个案中，平均加权模型优于基于判断者行为的模型）。

40 年前，威尔克斯 (Wilks, 1938) 首次指出，不同的线性组合之间，存在相互高度相关。他仅仅考虑了在预测因素之间存在正相关关系的情境。这一结果看来是普遍成立的，只要这些相互的相关关系不是负的；例如， $X + 2Y$ 和 $2X + Y$ 之间的相关关系为 0.80，当 X 和 Y 之间不存在相关关系

时。最近，许多学者（如 Green, 1977; Wainer, 1976; Wainer & Thissen, 1976; Edwards, 1978; Gardiner & Edwards, 1975）调查研究了结果对与系数的变化（假设不涉及符码的变化）相对不敏感的方式。

道斯等人（1974, p. 105）得出结论：“全部的诀窍在于知道寻找哪些变量，然后知道怎样相加。”下述研究充分地证明了这一原则，该研究是在道斯等人的文章发表之后进行的。在这项研究中，哈蒙德等人（Hammond & Adelman, 1976），既调查也影响了丹佛（Denver）城警方应该使用哪一种类型的子弹的决定，这一决定，与上述讨论的大多数决定相比，具有更明显的社会影响。下面引自哈蒙德等人的文章：

1974 年，丹佛警察局（Denver Police Department, DPD）和全国的其他警局，决定更改它们的手枪弹药。警方提供的主要理由是说，传统的圆形弹头的子弹，不能提供充分的“阻挡力量”（也就是说伤残对方因而不能向警员或者其他人回击的能力）。DPD 主管人员（和其他警方主管人员一样）建议用一种尖端凹陷的子弹代替传统的子弹。他们认为这种子弹在撞击时会变平，因而减少了穿透力，增强了阻挡效果，因此减少了潜在的跳弹。警方提议的子弹更改，遭到了美国民权联合会、少数族群和其他群体的抗议。反对改变子弹的人声称，新子弹和非法的达姆弹一样，它们导致比圆形弹头子弹更大的伤害，因此应该被禁止使用。和惯常一样，对这个问题的判断，私下里形成，然后予以热情地、顽固地公开辩护，并召开了公众听证会。双方都求助于弹道专家寻求科学的资料和支持。（p. 392）

争论聚焦于评价特殊子弹的性能，混淆了子弹的物理效力与其对社会政策的意义；也就是说，支持者不是去区分什么是子弹应该达到的（社会政策问题）这一问题和特殊子弹的弹道性能问题，他们仅仅为赞成这一种或者另一种子弹争论不休。因而，如同哈蒙德等人指出的，社会政策的制定者，漫不经心地扮演了（差劲的）弹道专家的角色，或者相反。哈蒙德等人所做的，是从政策制定者那里发现主要的政策维度，然后他们让弹道

专家根据这个维度评价各种子弹。结果证明，这些维度，被证明是阻挡效果（身体被击中的人不能回击的概率）、严重伤害的概率和伤及旁观者的概率。当弹道专家根据这些维度对各种子弹做出评价之后，结果表明，最后两种维度几乎完全被混淆了。但是它们没有完全与第一种混淆。子弹并不会沿着一个混淆了有效力和致死性的单一维度变化。严重伤害和伤及旁观者的概率，与子弹的穿透力高度相关，然而子弹有效地阻止对方还击的可能性，与子弹着体时伤口的范围高度相关。既然政策制定者不同意指定给三个维度的权数，哈蒙德等人建议赋予它们相等的权数。结合平均加权与弹道专家的（独立的）判断，哈蒙德等人发现有一种子弹，它与 DPD 那时使用的标准子弹相比既拥有较大的阻挡效力，也较不易于造成伤残（也比较不易于伤及旁观者）（Hammond & Adelman, 1976, p. 395）。这种子弹，也比 DPD 先前推荐的子弹更不容易导致伤害。“市议会和所有其他相关团体都接受了这种子弹。现在 DPD 正在使用这种子弹。”（Hammond & Adelman, 1976, p. 395）。^[7]再一次，“全部的诀窍，在于决定寻找哪些变量，然后知道怎样相加”（Hammond & Adelman, 1974, p. 105）。402

这样的话，为什么人们没有更经常地这样做呢？我知道有 4 所大学（伊利诺伊大学、纽约大学、俄勒冈大学、加州大学圣巴巴拉分校——可能还有更多）使用线性模型选择申请人，但是甚至是它们也是把线性模型用作一种最初的甄别设备，在对超过分数线的申请人进行最后选择时，也还是使用临床判断代替线性模型。戈登伯格（1965）根据明尼苏达多项人格测验（MMPI）的剖面图用以诊断神经症和精神病的精确方案，也已经证明，在试图完成这种任务方面，它优于临床判断者（根据我和戈登伯格所知，还没有研究者得到判断者表现得更好的情况）。然而，我使用该方案的一次经历（在安亚伯老兵管理医院）是，它因为犯了明显的错误（一个有趣的原因，最后详细讨论）而被停止。1970 年，根据一个基于明确定义的指标、院系资产和需要的拟线性分数系统，我建议我们在俄勒冈大学的研究委员会，按比例削减给予系里的国家科学基金和国防教育法案委员会基金；我被告知说“你不能系统化人类的判断”。仅仅在 6 个月之后，当我们的委员会意识到在直觉判断基础削减基金而面临的政治上和伦理上的困境 403

之后，这一系统才被采用，等等。

在过去的3年里，我不断撰写和讨论在主要的社会决策中使用线性模型的效用（和伦理上的优越性，根据我的观点）。不同的读者和听众反复提出许多相同的反对意见。我愿意归类这些反对意见并给出回答，以此结束本章。

反对使用线性模型的意见

这些反对意见，可以归入3种一般的类别或范畴：技术的、心理的和伦理的。我将按顺序讨论每种类别。

1. 技术的

技术方面最常见的反对意见，是关于相关系数的使用。例如，莱姆斯等（Remus & Jenicke, 1978）写道：

很明显，道斯等人选择相关系数来确立随机的单位规则的效用，这是不合适的（原文如此，为什么不合适呢？）。在道斯等人引用的实验中，同样需要一个标准函数。毫无疑问，存在一个成本函数，它在于错误地分类精神病和神经症，或者是允许勉强合格的学生进入研究生院的同时而拒绝了合格的学生。（p. 221）

首先考虑研究生入学的例子。大多数学校有 k 个空缺位置和 N 位申请人。问题是从 N 中选出最优秀的 k 个人（反过来，他们也愿意选择该学校）。还有比使用一种合适的等级更好的方法吗？没有。莱姆斯等人写作时，好像问题不是关于相对的选择，而是一个绝对选择的问题。然而大多数的社会选择，涉及的是从一组选项中选择更好的或者最好的：更好的学生、最好的子弹、可能的最好的飞机场地址等。很明显，因为相关系数非常好地反映了等级，它适宜于评估这种选择。

神经症—精神病问题，更加微妙，甚至更加不支持他们的论点。“毫无疑问”，他们说，“存在一个成本函数”，但是他们没有指明任何候选的方法。意思很明显：如果他们能够发现，那么临床判断会被证明是优于线性

模型的。为什么？就他们而言，由于缺乏这样的发现，他们的论点根本毫无疑问。但是，这种毫无根据的论点，却非常能够激发人们的兴趣（例如，对于失败的将军或者失败的足球教练而言，他们知道他们的计划“确切无疑地”会有效，“如果……”——然而实际上，无论如何他们的计划注定失败）。

技术方面第二个相关的反对意见，是关于判断者的平均相关系数与线性模型的平均相关系数的比较。可能由于取平均，一些实际上非常出色的判断者的表现被掩盖了。数据表明不是那样的。例如，在戈登伯格（1970）⁴⁰⁴的研究中，29 名受过训练的临床医生中只有 5 人表现得比单位加权模型好，而没有任何人的表现优于常规模型。在维京斯等人（1971）的研究中，没有一个判断者比单位加权模型出色。我们在俄勒冈重复得到了这个结果。在利比（1976）的研究中，在预测破产时，43 名判断者中只有 9 人表现得比资产—可靠性比率更好（有 3 个人表现得与比率一样好）。如果承认临床医生应该能够预测神经症或者精神病症状，研究生应该能够预测研究生的成功，银行贷款职员应该能够预测破产，他们的问题就变为这样的问题：很可能在这些研究中所使用的专家，并不是合适的专家。而这又是毫无根据的论证：如果使用其他的专家，那么结果就会不同。但是，再一次，没有出现这种专家。再一次，合适的答案，是询问为什么这些假设的其他人应该会有不同的理由（如同一所大学的副校长所告诉我的，“你的研究仅仅证明你使用的专家不行；毫无疑问，我们可以找到更好的专家，可以做得更好”——显然，不是从心理学系找到的）。

技术方面的最后一个反对意见，是关于标准变量的性质。无可否认，它们是短期变量，不那么深刻（例如，平均绩点或症状）；否则，大多数研究就不可能进行。那么，问题就提出来了。如果是要预测一种真正长时段的、重要的标准，是否研究会有所不同。答案是，结果当然可能不同，但是我们没有任何理由假定它们会不同。首先，一般来说，久远的将来比最近的将来更加不可预测，原因很简单，更多的不可预见的、外部的因素，或者自我强化的因素，会影响单个的结果（注意，我们不是在讨论累计的结果。例如在中西部地区，一个非常寒冷的冬天，通常会持续超过三个

月)。既然临床预测本来就劣于线性模型，那么，除非线性预测随着时间会变得比临床预测更差，否则该假设不会成立。没有什么先验的理由，相信预测中会出现这种差别退化，也从没有人曾经向我证明过。当然，没有任何证据。再一次，反对意见仅仅是毫无根据的论点。

特别引起人们兴趣的事实是，那些认为不同的标准、不同的判断者、不同的变量，或者不同的时间框架会导致不同的结果的人士，他们制造这种个案已经有 25 年了。然而，他们失败了。

2. 心理的

抵制使用线性模型的一种心理的原因，在于我们关于临床预测的选择性记忆。我们对临床预测的信念，因为成功的临床预测的例证的便利性 (Tversky & Kahneman, 1974)，而被加强了——尤其是那些例外与一些方案的例证：“我知道，曾经有人仅仅……他……”（例如，“我听说有一个智力测验只有 130 的人得到了心理学的高级学位”）。如同尼斯比特等人 (1976, 7) 所指出的，比起基于大量案例的更加有效的统计汇编，这种单个案例对于判断具有更大的影响力。而对于临床心理学家，一种预防根据这种例子而拒绝精确预测的方法，是仔细记录他们自己对自己病人的预测——不受事后认识支配的预期记录。这种记录，能够使所有成功或者失败的预测案例发挥同样的影响力。另外，它们能够运用于另外一项临床与统计相比较的研究，这项研究使用可能是最优秀的判断者——临床医生他或者她自己。

此外，在一些情境中，对积极或者消极结果的预测，存在自我实现效应。由于在这种情境中的选择 (Einhorn & Hogarth, 1978)，有效判断的错觉会被强化。例如，招生部的官员，他们断定某位申请人特别适合一个研究生项目，当那位申请人表现优秀时，他们可能会觉得自己的判断尽了责任，尽管申请人的成功，很大程度上是因为该项目的积极效应（与此相反，在一组经过挑选的申请人中，对选择的线性模型的评估，是通过这个模型预测申请人表现的情况来进行的）。或者一位侍者，他认为桌边的某个人很吝啬，他可能没有平时那么殷勤，结果只得到很少的小费，因而使他的临床判断免除了责任。^[8]

第二种抵制使用线性模型的心理原因，源于它们所公认的低效度。在此，有一种隐含的（与外在的相对）毫无根据的论点，因为它既没有提出评估程序方面的改变，也没有提出改变判断者或者标准。这些隐含的假设是，这些具有心理学兴趣的标准，事实上，是高度可预测的，因此可以得出；如果一种预测方法（线性模型）表现得并不太好，那么另一种方法可能会做得更好（或合情合理），然后就被转换成为相信另一种方法会做得更好（这并不是一个合情合理的推论）——除非这种方法能被发现。一位考虑研究生入学的教务主任，最好地表达了这种反对意见，他写道，“线性组合式与未来的能力等级的相关关系只有 0.4，而招生委员会的判断的相关系数是 0.2。0 的 2 倍还是 0。” 1976 年，我回答如下（Dawes, 1976, pp. 6-7）：

作为回答，我仅仅指出方差的 16% 比 14% 更好。然而，对于我，这个论点的最迷人之处在于隐含的假设，方差的其余 84% 是可以预测的，并且我们能够以某种方式预测它。

现在我们正在处理什么？我们正在处理一群 20 岁左右的（全部都那么聪明机智）人的人格和智识特征……到底为什么我们如此确定我们能够做出这种预测？无疑，不必每晚阅读传道书（Ecclesiastes）才能理解运气的作用。此外，很明显，在职业发展中，存在积极的反馈效应，它会夸大门槛现象。例如，一旦某些人被认为足够“出色”，他们就会被邀请到杰出的机构，他们有最杰出的同事互相影响——因此优点被强化了。同样的问题也出现在那些没有完全达到这样一个门槛的人身上。所有这些因素，不仅抑制了对成功的长期预测，而且对这种有关成功的预测研究，也必然被限制在那些可被接受的预测问题上，这些问题，有范围的限制，并且其预测指标间，有负的方差结构。（Dawes, 1975）

最后，职业的成功还存在各种类型的非智力因素，所有这些因素，在允许进入研究生院学习之前是不可能被评估的。例如，成功地构造一种令

人满意的或者振奋人心的性关系，尚未显现的毒品或者酒精上瘾的遗传倾向，不幸地加入了一个破产的研究团队，诸如此类。

从智力方面来说，值得注意的是，我们能够预测甚至 16% 的方差。但是我相信，我自己的情感反应也代表了我的那些同事。他们单纯地假定，未来因此是更可预测的。我希望它是可以预测的，尤其是当我想预测的未来的某个方面，对我而言是很重要的时候。我认为，这种欲望就转换为一种隐含的假定，未来事实上是高度可预测的，那么逻辑上将会得到；如果某种东西不是一个非常好的预测指标，其他的什么可能会做得更好（尽管认为它必然会如此，从来都是错误的）。

因为已确切地证明我们的预测只能达到可怜的程度（通常是一个低的相关系数），统计预测，就因此直截了当地用事实打击我们：生活并不是可以预测的。与此相反，非系统的临床预测（或者“后测”）许诺给我们一个安慰的幻象：生活事实上是可以预测的，并且我们能够预测它。

3. 伦理的

407 去年，当我在洛杉矶复兴博览会时，我无意中听到一位年轻的女性抱怨“非常不公平”，因为加州大学圣巴巴拉分校心理系仅仅根据分数，甚至没有给她一次面试的机会就拒绝了她。“他们怎么可能知道我到底如何呢？”回答是他们不能。即使有一次面试，他们也不能做到（Kelly, 1954）。然而，许多人坚持认为，不经过面试就做出一个至关重要的社会选择，很没有人性。我认为，人们是否被公平地对待的问题，更多地是与他们是否被非人性地对待有关，而不在于这种对待是不是面对面的问题（一些最糟糕的医生花费了大量的时间与他们的病人交谈，却从来不去阅读任何的医学期刊，很少或者根本不做任何测试，最后只能在葬礼上哀悼病人）。就申请人来说，平均绩点代表了他 3 年半的表现。（无疑，并不是所有的教授都对他或者她的独特的创造力类型有偏见）。GRE 是一种更加精心设计的测试。实际上，我们相信比起仅仅使用这两个数字，一次 10 分钟的文件评估或者一次半小时的面试能让我们把这份工作做得更好或者更公平吗？这种认知上的自负（Dawes, 1976, p. 7）是不道德的，尤其是考虑到没有任何证据表明我们能够比线性方程做得更好（如果要合乎伦理，那么进行例外的

选择也必须极端地仔细，因为如果我们根据他或者她拥有的某些特殊潜力就允许他或者她以低的成绩入学，那么我们就自动地拒绝了一些成绩更高的人，他们很可能具有同样惊人的潜力，如果我们不辞辛劳去评价它的话)。

无论我们多么希望本章所回顾的种种研究的这个或者那个方面发生改变，无论我们发现它们的结果在心理上会是多么令人不情愿或者令人讨厌，无论“把人还原成仅仅是数字”让我们在伦理上感到多么不舒服，事实依然是，我们的客户是人，他们应该得到最好的对待。如果这意味着——如同目前显现的那样——选择、诊断和预测，应该仅仅根据一些代表了重要特质的价值的数字的叠加，而不是其他任何东西，那么，就那样吧。否则就是欺骗我们所服务的所有人。

本章原刊于 *American Psychologist*, 1979, 34, 571-582。允许重印。

注释

[1] 这个指标，是基于卡斯等人 (Cass & Birnbaum, 1968) 在《美国大学比较指南》一书最后给出的可选择性等级。可选择性的类别，根据下列规则被赋予数字数值：最具选择力的，6；高度值得选择的，5；非常值得选择的 (+)，4；非常值得选择的，3；值得选择，2；不值一提，1。

[2] 不幸的是，111 名同学中只有 23 人能够用在这个比较中，因为招生委员会使用的等级尺度，年年发生细微的变化。

[3] 当所测量的变量其较高值能够预测标准的较高值时，关系是条件性单调的。这一条件是两种更加根本的测量条件的综合：(1) 独立性，即每个变量与标准之间的关系独立于剩余变量的值；(2) 单调性，即序数关系是一种单调关系 (见 Krantz, 1972; Krantz et al., 1971)。为了使线性模型起作用，真实关系并不需要是线性的；它们仅仅必须与线性模型接近。“为了计算两个变量之间的相关系数，它们之间的关系必须是线性的。”在一本统计学导论课本中发现的这种建议，是不正确的。最为重要的是，总有可能计算某些东西。

[4] 应该指出，一个常规线性模型既优于贷款部门的职员，也优于他们的同质异晶描述。使用同样的任务，有人 (Beaver, 1966; Deacon, 1972) 发现，线性模型

预测以 78% 的精确性通过交叉—有效性检验。但是我必须指出, 整体而言, 最简单的可能非常规模型表现最出色。在利比所研究的个案中, 资产与可靠性比率 (!) 正确归类了其中的 48 (80%) 家企业。

[5] 不幸的是, 道斯等人没有具体清楚地说明这些变量首先必须被标准化, 并且结果是一个标准化的因变量。如果没有先进行标准化, 对不可比的变量的等同或者随机加权——例如, GRE 分数或者 GPA——就会毫无意义。

[6] 考虑一组标准化的变量 $S_1, X_2, \dots, X_m$, 每一个都与标准化变量 Y 正相关。 X_s 的平均值与 Y 的相关关系等于 X_s 的总和与 Y 的相关关系。这一总和与 Y 的协方差等于

$$\begin{aligned} & \left(\frac{1}{n}\right) \sum_i y_i (x_{i1} + x_{i2} + \dots + x_{im}) \\ &= \left(\frac{1}{n}\right) \sum_i y_i x_{i1} + \left(\frac{1}{n}\right) \sum_i y_i x_{i2} + \dots + \left(\frac{1}{n}\right) \sum_i y_i x_{im} \\ &= r_1 + r_2 + \dots + r_m \text{ (相关系数之和)} \end{aligned}$$

y 的方差等于 1, X_s 的总和的方差是 $M + M(M-1)\bar{r}$, 此处 $\bar{r}$ 是各 X_s 之间的平均交互相关关系。因此, X_s 的平均值与 Y 的相关等于 $(\sum r_i) / (M + M(M-1)\bar{r})^{1/2}$; 它大于 $(\sum r_i) / (M + M^2 - M)^{1/2} = r_i$ 的平均值。因为每一个随机模型都与标准正相关, 所以平均值的相关关系——它是单位加权模型——高于相关关系的平均值。

[7] 应该指出, 仅有 8 种子弹达到帕雷托最优 (Pareto frontier)。也就是说, 仅有 8 种子弹在阻挡效力和伤害概率两方面都不低于某些其他的特殊子弹 (或者在一个变量上低, 而在另一个变量上相等)。因此, 不管是任何加权规则, 都会是在这 8 种子弹中进行选择。

[8] 这个例子由爱恩荷恩 (1979) 提供。



第 29 章

虚构数字的生命力

辛格 (Max Singer)

一般认定，纽约市的海洛因吸食者，每年偷窃的财产价值大约 20 亿美元到 50 亿美元，并且在所有的财产犯罪案件中，瘾君子占了大约一半。如兰德 (RAND) 之类的组织、赛弥尔 (Howard Samuels) 这样的政治人物，甚至美国的司法部长，也使用关于瘾君子犯罪的这种估计。半数的财产犯罪是由瘾君子犯下的，这个估计最初来自于一个警察官员，而它被如此频繁地使用，以致这一估计，现在已经成为公共知识的一部分。

通常以类似下述的方式，估计瘾君子所偷窃的财产的数额：

有 100 000 名瘾君子，平均每人每天花费 30.00 美元维持毒瘾，这意味着为了支付他们的海洛因，这些瘾君子，每年必须获得约 11 亿美元 ($100\,000 \times 365 \times 30.00$ 美元)。因为瘾君子必须在黑市上以仅仅 $1/4$ 的价格出售他所偷窃的财产，或者更低，所以为

了支付他们的海洛因，他们每年必须偷窃大约 40 亿美元到 50 亿美元。

409 这些计算，或多或少能够做得更加细致一些。可以考虑这样的情况：一些依靠非法手段谋生的瘾君子的典型特征是，在监狱中度过他们人生的 1/4 以上，这会减少犯罪数量的 1/4。《纽约时报》最近报道了莱利（William “Donkey” Reilly）的死讯。74 岁的他，是一个超级瘾君子，吸毒上瘾已经 54 年了，他在监狱中度过了其中的 30 年。瘾君子偷窃的东西中，有一些是现金，不必到黑市交易。海洛因费用的大部分，是通过从事海洛因交易，而不是从社会中偷窃来支付的，并且另外一大部分是通过卖淫活动来偿付，包括依靠卖淫为生的男性瘾君子。但是，无论怎样仔细地削减，如果人们在尝试估计瘾君子所偷窃的财产价值时，假定存在 100 000 名瘾君子，并且估计他们为了维持生活以及毒资，必须偷窃的最小数额（在对他们的合法收入给予了慷慨的估计之后），那么，对于纽约市，人们会给出的一个数值是，每年 10 亿美元左右。

但是，如果你从另一方面去入手这个问题，会发生什么呢？假设我们问：“有多少财产是被瘾君子或者其他人偷窃的呢？”瘾君子的偷窃，必定小于总的偷窃。在纽约市，任意一年被偷窃的财产的价值是多少呢？当我第一次提问时，多少让我感到吃惊的是，这是一个很难回答的问题，即使是近似的回答。没有人有过任何估计，即使是具有最微弱可信度的估计，并且这个问题，似乎也没有被经常提问过。官方报道纽约市的偷窃数额是每年大约 3 亿美元，其中约 1 亿是汽车被盗（一种很少由瘾君子从事的犯罪）的价值。但是，很明显，有非常大量的犯罪没有被报告。例如，正常情况下，商店偷窃是不会向警方报告的（许多因偷窃损失的财产，也没有报告给保险公司，保险行业对总的偷窃也没有一个很好的估计）。

然而，事实证明，如果人们仅仅询问一个问题，如“在纽约市，去年瘾君子偷窃了价值 10 亿美元的财产，可能吗？”那么，估计被盗财产的数额是相对容易的。很明显，瘾君子偷窃最常用的两种方式，是商店偷窃和入室盗窃（shoplifting and burglary）。瘾君子们商店偷窃的财产的价值会是多少呢？在纽约，全部零售一年的订购量是 150 亿美元，这包括机车、地毯、钻戒和其他一些通常是商店偷窃者不能偷窃的物品。零售企业存货损

失的一个合理数字是 2%，这个数字包括了管理贪污、职员偷窃、运输部门和卡车等（特别地，百货商店报告，最近一些年，商店偷窃大幅增长，但是，它们属于最容易受损失的零售企业，因此不足以使总的比率大大超过 2%）。实际上，人们一般同意，零售企业超过一半的财产损失，是由雇员造成的，剩下的损失，才是由于外来的商店偷窃者。但是，让我们认为，商店偷窃者偷窃了纽约全部零售财产的 1%——这大概是一年 1.5 亿美元。

入室盗窃的情况如何呢？纽约市大约有 250 万住户。假设平均来说，每年有 $1/5$ 的住户被抢劫或者被盗（robbed or burglarized）。这考虑到了在一些地区，入室盗窃非常普通，以及一些住户一年被盗超过一次。这意味着每年有 500 000 起入室盗窃。每一起案件损失的财产的价值，平均可能是在 200 美元左右。当然，一些案件中，损失了更多的财产，但是这种情况是非常少见的，并且经常是那些非瘾君子的职业窃贼所为。如果我们使用数值 $200 \text{ 美元} \times 500\,000$ 起，那么我们可以得到，在纽约，一年中人们家庭被偷窃的财产为 1 亿美元。

显然，这些估计的数值，没有一个是不能违背的，或者是被证实有根据的。你可以做出你自己的估计。此处的估计有一个特点：如果它们有 $1/10$ 的可能是错误的，那么这会非常令人吃惊；如果它们有 $1/2$ 的可能是错误的，那么这对于结论无足轻重（对于估计者，这是一个很好的位置）。

很明显，并非所有的瘾君子偷窃，都是从店铺或者从人们家中偷窃财产。一种最可怕的瘾君子犯罪类型，是行凶抢劫或者其他形式的抢劫，从纽约市民那里夺走财产。我们也可以估计这种情况。假设平均而言，每一年 10 个人中，有 1 个被行凶抢劫或者抢劫，那么将会有 800 000 起这种劫案。如果平均一起导致损失 100 美元财产（这非常不可能），那么一年内，这种形式的偷窃将取走 800 万美元。

所以，我们可以看到，如果认为瘾君子犯下了所有的商店偷窃、所有的入户偷窃和所有的抢劫，那么在纽约，瘾君子一年内所偷窃的全部财产，总额达到约 3.30 亿美元。你可以加进去所有你想要的“附加因素”（fudge factors），增加瘾君子犯下的所有其他各种各样的罪行，但是无论如何，我

410

们很难为这种估计找到一个根据，即瘾君子一年的偷窃超过5亿美元。2.5亿美元看来是个更好的估计，尽管可能仍然高了。毕竟，肯定有一些贼不是瘾君子。

因而，我相信我们已经证明了，虽然实际的数字仅是1/10，但是人们普遍假定，在纽约，瘾君子们一年偷窃了20亿美元到50亿美元。关于这一点，5分钟的思考就可以证明。^[1]那么是多少？价值2.5亿美元的财产仍然是很大一笔财产，超过了每年花费在戒毒以及其他阻止、控制吸毒的项目上的金钱总数。并且，瘾君子所偷窃的财产的价值，仅仅是他们给社会造成的全部成本中的一小部分，更大的成本在于恐惧、改变居住环境、预防成本以及其他针对偷窃及其危害的回应和再回应措施。

411 估计瘾君子所偷窃的财产的价值，这个练习的一个目的，是为了澄清人们对于数字的态度。人们感觉有许多的瘾君子犯罪，并且20亿美元是一个很大的数字，所以他们倾向于相信瘾君子偷窃的财产价值20亿美元，但是2.5亿美元也是一个很大的数字，如果我们的洞察力（our sense of perspective）没有因为每天关注联邦的支出而扭曲，那么多数人会满足于接受一年2.5亿美元是很多的偷窃。

同样地，这个练习还提醒我们：甚至是负责任的官员、负责任的报纸和负责任的研究团体，偶尔也会得到一些事实上没有任何现实根据的数字，并且审察后把它们作为绝对正确的数据。这个经验提醒我们，人们不能够因为一个估计已经被各种各样应该知道自己在说什么的人物广泛使用，就假定这个估计可能更加正确。

但是，几乎不可能存在如人们所相信的那么多的瘾君子偷窃，这一事实还有更加重要的内涵（implication）。这个意义就是，很可能并不存在如许多人所相信的那么多瘾君子。以零售方式购买海洛因需要支付的钱财大部分来自于偷窃，而大部分的瘾君子是以零售方式购买海洛因。因此，偷窃的数额从根本上限制了瘾君子的数量，尽管是不精确的（Hudson Institute的一项研究得到的估计是，在消费掉的海洛因中，接近一半数量的海洛因由这样一些人使用的，他们并不以零售方式购买，也不是用偷窃到的财产来支付，而是用他们在分配体系中所提供的“服务”）。^[2]但是，尽管

从事交易的人（较低等级的）消费了近半数的海洛因，然而他们仅仅占了瘾君子全部数量的约 $1/6$ 或者 $1/7$ ，他们是那些能够支付得起大量毒瘾的人。

关于纽约市的瘾君子的数量，最流行的、非正式的估计是 100 000 +（通常强调 +）。华盛顿的联邦花名册列出约 30 000 名瘾君子在纽约，纽约市卫生部的花名册列出了约 70 000 名瘾君子的姓名。尽管不是所有在名单上的瘾君子，都仍然在吸毒——他们中的许多人已经死了，或者在狱中——多数人相信还有很多瘾君子，他们没有出现在任何的花名册上。认为纽约市有 100 000 名瘾君子这个估计非常保守，这很普通。1970 年初，邓森-盖宝（Judianne Denson-Gerber）博士的估计被广泛引用，她估计到那一年夏天为止，会有超过 100 000 名的青少年瘾君子（Teenage Addicts）。很明显，还有很多瘾君子，他们 20 岁或者年龄更大。^[3]

在本章中讨论瘾君子的数目时，我们是在谈论当“瘾君子”一词被使用时，人们所想到的那类人。^[4] 一个更好的术语可能是“街头瘾君子”，一个通常每天服用海洛因的人。他是那种人，外形和行为都像一个瘾君子的标准形象。在此，我们排除了那些在医学职业上经常使用海洛因或其他的麻醉剂或者对它们上瘾的人，偶尔使用海洛因的学生，上瘾但是不必偷窃也不经常光顾通常的瘾君子聚会场所的有钱人，等等。当我们论及“吸毒问题”时，包括这些案例不是很重要；毫无疑问，尽管在不同程度上它们是问题，但是它们是一种和典型的街头瘾君子造成的问题非常不同的类型。

瘾君子偷窃的财产数额证明，纽约市街头瘾君子的数目更可能是 70 000，而不是 100 000，并且几乎可以确定，不可能是被经常使用的数字 200 000。其他几种估计街头瘾君子数目的简单方法得到了近似的结论。

与瘾君子人群接触的经验使观察者们估计，平均每一位街头瘾君子在监狱中度过他们一生的 $1/4$ 到 $1/3$ （这方面的一些研究者，如 Edward Preble 和 John J. Casey, Jr. 认为平均超过 40%）。这将意味着，在任何时刻， $1/4$ 到 $1/3$ 的瘾君子是在监狱中，因此，通过用 3 或者 4 来乘在监狱中的瘾君子数目，能够估计出瘾君子的总数。当然，在狱中的瘾君子的

数目，也不是一个已知的数值（并且，事实上，如上面我们已经指出的，甚至不是一个非常精确的概念）。然而，人们能够合理地估计在狱中的瘾君子的数目（为此，我们可以包括在各种志愿治疗中心的瘾君子）。这个数字是 14 000 ~ 17 000，这与纽约市街头瘾君子总数为 70 000 的估计相当一致。

418 估计纽约市街头瘾君子总数的另一种方法，是使用人口统计资料，而关于瘾君子人群的资料是可以得到的。例如，我们能够合理地确定，在纽约市，街头瘾君子人群中约 25% 是波多黎各人（Puerto Rican），约 50% 是黑人。我们知道，大约 5/6 的街头瘾君子是男性，并且 50% 的街头瘾君子年龄在 16 岁到 25 岁之间。这意味着，瘾君子总数中的 20% 是年龄在 16 岁到 25 岁之间的黑人男性。如果有 70 000 名瘾君子，那么这意味着大概有 14 000 名年龄在 16 到 25 岁之间的黑人男性是瘾君子。但是，在纽约，年龄在 16 到 25 岁之间的黑人只有 140 000——一半可能住在贫民区。这意味着，如果这个城市有 70 000 名瘾君子，那么 10 个黑人年轻人中就有 1 个是瘾君子。如果有 100 000 名瘾君子，差不多 6 个人中有 1 个，如果有 200 000 名瘾君子，那么 3 个人中有 1 个。你可以自己断定，这些对年轻黑人男性群体的各种程度的洞察中哪一种最可信，但是相当清楚，200 000 名瘾君子这个数字是难以置信的。类似地，总数 70 000 名街头瘾君子将意味着大概 7 000 名年轻的波多黎各男人是瘾君子，而在纽约市，年龄在 16 到 25 岁之间的波多黎各男孩的总数是大约 70 000。

从任何方面来看，上文的分析都不是为了贬低海洛因吸食问题的重要性。海洛因是个可怕的灾难（curse）。当你想象一下牵涉其中的个人悲剧，70 000 名瘾君子是一个非常庞大的数目。如果你必须靠工作来维持生活，25 亿美元是一笔非常庞大的财富，它们从市民那里被偷走，经过瘾君子和收赃者之手到了那些输入和分销海洛因，那些收受贿赂或者为海洛因行业提供其他服务的人的口袋中。

这篇文章的主要目的，可以说是为了阐明，在界定一个问题的时候，通过认真对待数字，了解它们的意义，在各种意义之间以及根据常识（例如城市中人口或者住户的数目）权衡各种意义，对人们能够有多大程度的

帮助？当普通人以及负责任的官员与各种类型的专家打交道时，这方面的细微努力，就能够给予他们很大的帮助。

本章原载于 *The Public Interest*, 1971, 23, 3-9。允许重印。

注释：

[1] 虚构数字在犯罪领域可能比在大多数领域中更加神秘，更具有生命力。20 世纪 50 年代早期，凯福夫（Kefauver）委员会公布了一个估计值，在美国，赌博每年“取”走 200 亿美元。事实上，这个数字是随便挑出来的。一个成员说：“关于花费的钱财，我们一无所知。加利福尼亚犯罪委员会说是 120 亿美元，Virgil Peterson of Chicago 说是 300 亿美元，我们选择 200 亿美元作为二者的平衡。”

在 1971 年 1 月 13 日的《纽约客》（*New Yorker*）上，爱泼斯坦（Edward Jay Epstein）给出了一个不同寻常的例子——宣称 28 名黑豹党人被警察所杀的精细纪录，这个例子表明虚构数字的生命力的强大（结果证明是 19 名黑豹党人被杀，其中有 10 人被警察所杀，而这其中又有 8 名看来可能是在他们先动手的情况下被杀的）。

[2] 在后来一项对 St. Luke 医院的 81 名瘾君子——他们平均 34 岁——的研究中得到了类似的数据。这些瘾君子在一年内消费的海洛因中，超过一半是通过销售海洛因而支付的。顺便提一句，在前一年中，这 81 名瘾君子偷窃的财产平均价值 9 000 美元。

[3] 在最近的其他估计者中，我们可能注意到一个马克思主义者尤里克（Sol Yurick），他给出了我们“500 000 名吸毒者”（*Monthly Review*, December 1970），还有克尔森（William R. Corson），他在 1970 年 12 月的《阁楼》（*Penthouse*）上坚决主张，“今天，至少有 2 500 000 名黑人美国人依赖海洛因”。

[4] 关于“瘾君子”一词，有一种有趣的反常现象。如果缺乏一个关于“瘾君子”的定义，大多数人会说，“瘾君子”是一个经常服用海洛因（或者一些类似的毒品）的人，如果不能得到他通常的剂量，他会出现不舒服或者痛苦等停止服药的症状。但是，对于大部分一般被认为是“瘾君子人口”的人，这个定义都是不适用的。事实上，它也不适用于多数被证明是瘾君子的人。一个已经戒毒或者被关在监狱中而远离毒品大约一个星期的瘾君子，不适用“瘾君子”的标准定义。他不再因为没有服用海洛因而出现任何身体症状。可以肯定，“驴子”莱利，会符合

不确定状况下的判断

大多数人关于瘾君子的看法。但是在作为一个“瘾君子”的54年中，有30年他是待在监狱中的。可以肯定，待在狱中的大部分时间里，他没有吸食海洛因，而在狱中的时间超过了他“瘾君子”生涯的一半（尽管在狱中，一定数量的海洛因是可以得到的）。



第 30 章

直觉预测：偏差和校正程序

卡尼曼和特沃斯基 (Daniel Kahneman and Amos Tversky)

一、引言

任何有意义的预测活动，都包含大量的判断、直觉以及训练有素的猜想。事实上，专家的观点，是诸多技术的、政治的和社会的预测依据。甚至在一些通过数学模型或者模拟而获得的预测中，观点和直觉仍然扮演着重要的角色。在这些模型中，直觉判断渗透进所考虑的变量的选择，这些变量所负荷的影响因子以及这些变量所假定的初始值。鉴于直觉在所有这些预测中的关键作用，有必要对限制专家进行精确判断的因素进行分析，并且发展出一套旨在改进判断质量的程序。

二、单一的和分布的资料

专家通常被要求对不确定的数值提供最好的推

415 断、估计或预测，如某一天的道-琼斯指数、某种产品在未来的销售状况或者一次选举的结果。预测者可资利用的信息分为两类：即单一性的和分布性的。单一信息或个案数据，由有待考察的特定案例的证据所构成。而分布信息或者基率资料（base-rate data），则包括在类似情景中有关结果分布的知识。例如，在预测一本新小说的销售量时，有关书作者、写作风格以及情节的资料，就是单一信息；而有关各种小说销售量的资料，则是分布信息。类似地，在预测一个病人的寿命时，单一信息包括他的年龄、健康状况、个人病史；而分布性信息，则由相关的人口统计信息所组成。单一性信息，描述那些借以使此问题与其他问题相区别的特定属性。而分布性信息则刻画出以往观察到的结果在同类案例中的发生状况。在这里，分布性数据的概念，不能等同于贝叶斯先验概率分布的概念。前者定义的是数据的性质，而后者则是定义信息获得的序列。

可获得的相关的分布性信息总是稀缺的，从这个意义上来说，许多预测问题，从本质上来说都是独一无二的。例如，预测 2000 年对于核能的需求量，或是人们发现治疗白血病的有效手段的日期。在诸如此类的问题上，专家只能依赖单一性信息。然而，有证据显示，即便当分布性信息是可以获得的时候，人们对于这些信息仍然不够敏感。事实上，最近的研究显示，人们主要还是依赖单一性信息，甚至当这些信息不充分、不可靠时也是这样，而对于分布性信息则往往重视不够（Kahneman & Tversky, 1973, 4; Tversky & Kahneman, 第 10 章）。

计划制定的语境，就提供了许多实例，它们表明在过去经验中的不同结果的分布状况是被忽略的。例如，科学家和作家总是倾向于低估完成一项任务所必需的时间，甚至当他们拥有相当多的过去教训来引导他们做出正确的计划时也是如此。类似的谬误，也曾见于对工程师错误估计修理发电站的完成期限的记载中（Kidd, 1970）。尽管这种计划失误有时候归结于主观动机的因素，比方说，心怀某种希望的想法，甚至当对时间和花费的低估会遭到事实上的惩罚时，它仍然会频繁地发生。

计划失误，是由于倾向于忽略分布性数据并且采取（或许可以被定义为）预测的内隐方式的结果。这种方式下，人们关注的是特定问题的组成

要素，而不是在类似案例中结果的分布状况。在计划评估时，内隐方式往往容易造成估计不足。举个例子，一座建筑只能在没有材料运输的延误、没有罢工、没有不正常的天气状况等情况下，才能按时完工。尽管这些阻碍不太可能发生，但它们中间只要有一个发生，后果都将是严重的。然而在人们的直觉中，这种综合性的考虑是很不充分的（Bar-Hillel, 1973）。试图用加上一个可变（slippage）因素来抵消这一误差，是远远不够的，因为调整值将会与作为参照值（anchor）的初始值，太过接近（Tversky & Kahneman, 1974, 1）。而把一个特定的问题视为类别中的个体的外向方式，将有助于克服这种错误。以这种路径，人们不会试图去神化那个可能使计划失败的特定方式。相反，他会把手边的问题与相类似项目完成时间的分布状况联系起来。它表明通过考察外部性问题，诸如此类项目通常需要多长时间，而不仅仅是内部性问题，比如在操作某个特定问题时的具体因素和困难是什么，可能会获得更多的合理的估计。

416

特定因素会增强主要依赖单一信息而忽视分布信息的倾向，只要这些因素能提高有关问题可感知的独特性。分布性数据的重要性，可能由于对特定案例的细致了解或是高度卷入而被掩盖。专家被要求回答问题的形式，也影响了一个问题的被感知到的唯一性。例如，开发一个新产品需要多少钱的问题引出的内在方法是，整个花费被分解为不同的组成部分。与此相对应的，关于实际花销可能超过目前预算的比率的问题，倾向于去考虑在开发类似项目时花费状况的分布。这样，改变分析单位，比方说从成本转换到超出预算的花费，就可以改变思考问题的方式。

这种普遍的低估或者无视分布性信息的倾向，或许就是直觉预测的主要错误。考虑分布性信息固然不能确保预测的准确性。但它确实提供了防范措施，以免人们进行完全不真实的预测。因而分析者应该尽可能地将预测的问题纳入这样的框架，即，更便利地利用一切可能获得的分布性信息。

三、回归与直觉预测

在大多数预测问题中，专家都同时掌握特定案例的单一性信息和涉及类似案例的结果的分布性信息。类似的例子，如预测学生成绩的教练，评

估一场交易的赢利能力的银行家，一个估计教科书销量的出版商，或是预测经济增长的某项指标的经济学家。

在这些情况下人们怎样进行预测呢？心理学的研究（Kahneman & Tversky, 1973, 4; Ross, 1977）表明，直觉预测通常遵循一个简单的匹配原则：人们会选择一个预测值，使得这个案例在诸种结果分布中的位置与它在人们对不同结果的印象分布的位置相匹配。下面的例子形象地说明了这个规则。一个编辑评论一部小说的手稿，并留下了非常良好的印象。他说：
417 “这本书读起来像是一本最畅销的书。在近些年内出版的类似读物中，我敢说只有 1/20 的书比这本书给我留下更为深刻的印象。”如果当时，这位编辑被要求对这部小说的销量做一个评估的话，他极有可能会预言此书会位于销量最佳的 5% 的图书之列。

有充分的证据表明，人们通常通过把预测与印象相匹配来进行预测。然而，这种预测法则是不合理的，因为它没能把不确定因素考虑在内。我们上例中的编辑无疑也会承认图书的销量是极难预测的。在这样一种高度不确定的情况下，对于一本图书销量的最好预测，应该是介乎个人的主观印象和同类图书的平均销售量之间的某处取值。

统计预测的一个基本原则，也是运用直觉最少的一个原则就是，预测的极值（extremeness），必须由对可预测性因素的考虑来调节。比方说，试想，那个出版商从以往的经验中得知，图书的销量与他的最初印象关系甚微。无论是给他留下良好印象的书稿还是他不喜欢的书稿，都同样地可以卖得很好或是很糟糕。在这种可预测性为零的情况下，出版者对于销量的最好的假定，应该是所有的图书都是一样的——比如，相关类别的均值——而不去计较他对这本书的个人印象。在间接的情况下，当然这种情况是更为普遍的，预测应该是回归的，也就是说，它应当落在类别平均值和最能表征个人对手边这个案例的印象的数值之间。可预测性越低，预测与类别均值就应当越近。直觉预测，是典型的非回归性的：人们往往在可靠性和预测的有效性程度都很低的信息的基础之上做出极端的预测……

四、预测的校正程序

怎样引导专家去做正确的回归预测呢？怎样能够让他遵循统计预测的原则去使用可以获得的单一性和分布性信息呢？在这一部分将提出一个为了达到这些目标而设计的五步骤程序。

1. 步骤 1：选择一个参照类别

这一阶段的目标，是辨识一个类别，使之作为手中案例的有意义的参考，并且这个类别的结果的分布应该是已知的，或者能够有把握作出合理的估定。

比如，在预测一本书的销量或者一部影片的毛收入时，参考类别的选择是简单明了的。在这些情况下，因为书的销售量或是收入的分布是已知的，要定义书籍或是电影的合适的参照类别是相对比较容易的。

在一些预测的问题中，例如对开发一个小说类产品成本或面市时间的预测。由于不同种类的差异如此巨大，以至于他们之间很难做出有意义的比较，因而很难甄别参照类别。然而，如前所述，这个问题，有时候可以通过重新定义预测量而加以克服。比如，比较在不同的科技领域中开发项目时，也许以超出成本（overruns）的百分比做比较，要比比较绝对的花费更容易。相比而言，预测超出成本，突出那些在许多不同的项目中都是相同的计划中现实方面的决定因素。因此，在后一种形式中，定义参照类别，要比前者来得容易一些。

通常，专家会为特定的问题想出好几个参照类别，而从中做出一个选择是必要的。例如，在对一本书的销量做预测时，其参考类别，可以是同一作者的其他图书、同一主题的图书、同一大类的图书如精装本等。选择一个参考类别，往往需要在相互冲突的标准之间作出平衡。这样，最有包容性的类别可以容括对于结果分布的最佳估计，但它也可能因为异质性太强而妨碍了与手中的图书的可比性。另一方面，同作者的图书这一类别，可以为比较提供一个最天然的基础，但有待考察的这本书，却有可能落在以往观测到的结果的范围之外。在此例中，同一主题的图书这一参考类别，将会是最恰当的。

2. 步骤2：对参考群体的分布进行评估

对一些问题而言，比方说卖书，其结果分布的统计值是可获得的。而另外的一些问题，其相关的分布估计，必须建立于各种渠道的信息之上。特别是，专家应该提供对类别均值的估计，以及一些反映结果分布范围的其他估计值。问题样本，诸如此类书籍的平均销售量是多少？这类书籍中，销售量超过 15 000 本的比例有多大？

许多预测性问题的特点，是缺乏直接的相关分布的数据。在远期预测中更是如此。在这类预测中，相关分布往往是关于遥远的未来才会发生的结果。如考虑这样的预测：2000 年，英国在世界个人城市运输市场中的所占份额。也许将这个问题改换成如下的提法会更有用：“2000 年，英国在世界各种高科技领域市场上所占份额的可能分布是怎样的？相比其他的科技状况，你如何预测运输系统这个特定的领域？”需要引起注意的是，在这个问题中，结果的分布是未知的。但是，所需分布的估计，可基于英国目前在世界各种科技市场上所占份额的分布值，并通过英国在世界市场上位置的长期变化趋势的估计对其进行调节。

3. 步骤3：直觉估计

专家对于某个问题的部分信息，通过参考类别的结果分布来进行归整。除此之外，专家通常会拥有相当数量的关于特定案例的单一性信息。这些信息将它与同类别的其他成员相区别。专家此时应当被要求在这些单一性信息的基础之上做出直觉式估计。如上文所述。这种直觉式估计，很可能是非回归性的。这个程序的接下来两步的目的，就在于纠正这个错误以获取更加充分的估计。

4. 步骤4：可预测性的评估

专家现在应该来评估，在此案例中，可获得的这类信息，在多大程度上保证了预测结果的正确性。在线性预测的背景下，对可预测性的适当测量是在预测值与结果值之间的积矩（product-moment）相关。在那些过去的预测及结果的记录尚在的情况下，所需的值可以从这些记录中推知。如果这些数据是缺失的，人们就只能依靠对可预测性的主观评估。一个统计经验丰富的专家或许能够以他的经验直接估计出 ρ 值。当统计经验缺失的时

候，分析者就应当求助于一些非直接的程序。

其中的一个程序，要求专家来比较他所关注变量的可预测性和其他变量的可预测性。例如，某位专家，可能有相当的自信相信自己预测书籍销量的能力，要超过运动解说员对足球赛中得分状况的预测能力，但认为自己不如一个天气预报员对两天后的气温的预测能力。一个技巧熟练并且勤奋的分析员，可以建立起一个粗略的可预测性的量表。这个量表，基于一系列现象的预测及结果之间的数学关系，这些现象的范围，涵盖了高度可预测的现象如温度，到高度不可预测的现象如股票价格。果真如此，（统计）分析员，可要求专家在这个量表中对目标量进行定位，并由此提供一个估计数值。

另外一种对可预测性进行评估的方法，包含如下类似的问题：如果你要考察你们即将要出版的两部小说，你有多少次在预测哪两本将会卖得更多时结果是正确的？关于预测和结果之间的序列相关性的估计，可以通过如下的步骤获得：如果 ρ 是结果被正确预测的案例占整个被预测案例的比例，则 $\tau = 2\rho - 1$ 提供了一个预测准确性的指标。其范围从 0（预测处于随机水平）到 1（预测的完美无缺）。在许多情况下， τ 可以视为 ρ 的粗略估计。

420

对可预测性进行估计并不容易，而且他们需要认真的检验。专家往往事后才发现错误（Fischhoff, 1975）。这样就导致对结果的可预测性给予过高估计。专家也容易犯便利性偏差（availability bias），并且会回想那些在头脑中留下深刻印象和引起震惊而后来才被证实的碎片。

5. 步骤 5：对直觉估计的校正

要对非回归性（nonregressiveness）进行校正，直觉估计应当向参考类别的平均值方向进行调整。如果直觉估计是非回归性的，则在相当一般的情况下，直觉估计与类别均值之间的差距，应该用因素 ρ 来削减。这里的 ρ 是积矩相关系数。这个程序提供了一个人们希望减少的非回归性错误的估计量。

例如，假设专家对给定图书销售量的直觉估计是 12 000 册，并且平均而言此类书的销量是 4 000 册。进一步假设专家相信他能够有 80% 的把握

正确地预测这些书未来的销量。在这个案例中， $\tau = 1.6 - 1 = 0.6$ ，销量的回归估计，就等于 $4\,000 + 0.6(12\,000 - 4\,000) = 8\,800$ 。

在直觉估计相对比较极端，而可预测性中等或者偏低的情况下，这种修正的效果是很明显的。计算的基本原理，应当认真地对专家做出说明，以便他能够决定是否坚持他原有的假设，或是采用计算出来的估计，还是修正他的评估值取得某个中间值。

上文概述的程序，对于在分析者与专家互动过程中产生的反对意见，是开放的。首先，专家可能会对认为他最初的直觉估计是非回归性的假设提出质疑。幸运的是，这个假设可通过如下途径加以证实：（1）要求专家对给他留下强烈印象的案例，例如书稿，在参考类别中所占的比例进行估计。（2）让他对于实际结果超过了他的直觉预测的案例在参照类别中所占的比例进行估计。例如，前例中销售量超过 12 000 册的图书。如果这两个比例大致上相等，则这个预测就应该是非回归性的。

一个更为普遍的反对意见，或许会对“预测应该是回归的”这一基本的观念提出质疑。专家也许会正确地指出，目前的程序通常会产生出保守的不会太偏离群体均值的预测，并且不太可能预测出一个外在于以往所有观察值的异常值。对于这一攻击的回答是，一个会犯错误的预测者，保留正确的预测一些异常值的机会，仅仅以错误地认定许多其他的案例也是以异常值为代价。非回归预测，是过度的预测：它们与真实概率的相关表现在，高的预测就是过度估计，而低的预测就是估计不足。在大多数情况下，这种错误的代价是高昂的，而且应当被排除……

五、结论性评语

这里所陈述的路径，基于下列的一些关于预测的基本主张。第一，大多数的预测和预报，都包含有不可或缺的直觉因素。第二，知识丰富的人的直觉预测，有很多有用的信息。第三，这些直觉预测，通常以一种可预测的方式有偏差地表示出来。这样，问题就不在于是否去在表面价值上接受或拒绝直觉假设，而是这些结果如何被修正或是改进。

对人们判断的分析表明，直觉带来的许多偏差，源于人们对特定类型

的信息的轻视。例如，结果发生的基率频率和他们的可预测性。本章中陈述的校正误差的策略，试图从专家的有关信息中引出通常会被他们忽视的信息，并帮助他将这些信息与他的直觉印象以一种符合统计预测的基本规律的方式整合在一起……

本章原载于《公共利益》（*The Public Interest*, 1971, 23: 3-9）。允许重印。



第 31 章

去偏差过程

费希荷夫 (Baruch Fischhoff)

422

一旦在实验室条件下观测到某种行为现象,那么着手研究它在统计检验上的强弱 (robust) 是合适的。一种流行的并且富有成效的提问策略,就是所谓的解构性检验 (destructive testing) ——一种源自工程学的技术。我们提出一种实验设计,使其受到一些条件的影响而达到并且超出它的可行性极限。处于控制之下的解构 (destruction), 可以使我们明白我们应该相信什么以及它起作用的原理。具体到一种行为现象,同样的逻辑,使我们的研究可以控制观测的条件以及必须被唤起或者控制的心理过程,以消除这种现象。当这种现象是判断偏差 (judgmental bias) 时,解构性检验,也就成为某种形式的去偏差或消除偏差 (debiasing) 的努力。解构性检验指出了实验设计的缺陷;一旦消除了偏差,其结果就是判断的改进 (improved judgment)。

关于启发式和偏差 (heuristics and bias) 的研究本身,就可以看作是解构性检验的一种应用。它所检

验的前期假设是，人类是能干的直觉统计学家。因果观测证明，一般来说，人们的判断“足够好”，使得他们的一生不会因为做出这一判断而陷入太大的麻烦之中。早期研究（Peterson & Beach, 1967）支持这一观点，并且指出，近似地，人类可以被描述为诚实的观测者和规范的判断者。本书所述及的后续研究，通过考察人们的明显成功（apparent success）的局限性，来检验这种近似（approximation）的准确性。更好的判断可以使人们更富有、更健康吗？他们所取得的成功，可以归功于一种宽松的环境，而不需要假定某种明智的行为吗？悲剧性的错误，提供了洞察人们的决策过程的本质和质量的重要机会。所幸这种悲剧性错误少得可怜，以至于我们所掌握的极少的数据不可能清理出那些把人引入歧途的因素。判断研究（judgment research）使用了解构性检验策略，以得到刻意设计的情境中的有偏判断（biased judgment）。理论家们希望得到一种失败和成功的模式，这种模式本身几乎不需要什么合理的解释。因此，关于偏差的研究，澄清了智慧的来源及其局限，就像关于消除偏差的研究，澄清了明显的愚蠢行为一样。这两者对于判断研究都是至关重要的。

429

尽管一些判断研究，主要是为了证明一些或许是人为制造出来的条件下可能出现的某种偏差，但是也有许多其他研究，致力于采取预先措施消除偏差。其中一些明显属于消除偏差的研究，它们希望在实验室中证明有效的程序，同样可以改善实际生活中的表现。其他研究有着更多的理论层面的目标，它们试图弄清楚引发次优判断的情境。本章的核心内容是评论一些研究，这些研究可以看作是减轻两种相似的偏差——后知之后明（hindsight）和过分的信心（overconfidence）——的努力。本章以如下信念看待成功和失败：（a）失败有助于澄清问题的严重程度以及人们对纠正或者预防措施的需要；（b）研究的总体模式，是揭示心理维度的关键所在，而这种心理维度，对于现实生活情境的界定以及预防这种情境下偏差行为的扩展十分重要。

本章的评论，力求穷尽相关的研究，其选择遵循以下三个标准：

- （1）只考虑同行匿名审稿（peer review）杂志中的研究。因此，质量控制的责任被外化。

(2) 逸闻趣事性质的证据（只有很少的例外）不予考虑。尽管这种报告是有关某些类型的消除偏差努力（例如，专家的使用）的主要资料，但是，它们受一些解释和选择上的偏差的影响，超出本章的讨论范围（见第23章）。

(3) 一些经验性质的证据将被考虑。那些还没有被检验过的证明，以及不可能被检验的理论论证（例如，关于实验的生态学效度），排除在外。

在进行文献评论之前，我们会先提供一个关于消除偏差的努力的框架，界定可能的研究思路以及隐含的假定。当应用于各种判断偏差时，这样一个框架可能揭示出一些复发性模式。

一、消除偏差的方法

一旦问题出现，自然首先寻找导致问题的原因。根据它们本身对于问题原因的隐含指控，可以非常清楚地对消除偏差的程序进行分类。最重要的一个区分，在于偏差是源于判断者，还是任务，或者是二者之间的不匹配。偏差是意味着实验设计人为的不合格和解释的不明确，还是明确的判断错误的个案，或者是需要但却错误运用认知技能而导致的不幸结果？正如表31—1中加以总结和下面将要加以描述的，这些类别中的每一个都可以根据所谓的问题深度进一步细分。难度到底起到什么根本性的作用？需要技术或者结构的改变吗？由于因果类型的不同，用于发展消除偏差的技术的策略也就截然不同。

表 31—1 根据隐含的假定消除偏见的方法

假定	策略
错误的任务	
不公平的任务	提高奖励 澄清指导语/刺激 阻止事后预测 使用更好的回答模式 少问一些问题

续前表

假定	策略
被误解的任务	证明备择目标 证明语义的不一致 证明任务不可能 证明区分被忽视
错误的判断者	
可完善的判断者	警告问题 描述问题 提供个人化的反馈 充分训练
不可救药的判断者	替换人员 重新校准他们的回答 基于错误制订计划
判断者与任务之间的不匹配	
重构	让知道的东西显现出来 寻找不一致的信息 分解问题 考虑替代情境 提供替代方式
教育	依赖真正的专家 从小训练

1. 错误的任务

(1) 不公平的任务

实验者对他们自己以及其他人的工作提出了一些标准。只有通过（评论者和编辑的）信度审查，使得他们相信研究已经排除了比较明显的人为因素的影响，他们的研究才会被发表。但是，由于我们不能控制所有条件，也不可能让所有人满意，不论是初始研究还是系列研究，因此当试图推翻某种效应时，对一般人为的方法论上的人为因素的确定，就成了首先工作。在这些主张中，可以举出如下几种：（a）被试不在乎实验——因此应该提高良好表现的奖励；（b）实验任务使被试感到迷惑——因此需要使用更加细致的指导语以及更加熟悉的刺激；（c）被试并不相信实验者所宣称的实验的性质，或者被试所理解的奖惩结构与实验者所想要的不同——因此要

让被试相信，他们对答案的选择才是实验者的兴趣所在；(d) 被试不能表达出他们所知道的东西——因此应该使用更加熟悉的或者更加容易适应的回答模式；(e) 被试被询问了太多的问题，以至于他们形成一种固定的回答模式——因此应该少问一些问题（或者干脆把研究的旨趣设定为研究固定回答）。

处理这种问题本身就可以使研究更具有科学性。但是这种努力通常缺乏理论内容。由于它的目标是创造更好的实验环境，因此，关于人为因素的研究，可能甚至不能提供什么情境方面的资料，使得我们可以有把握地推广观测到的结果。“成功的”人为因素研究提供的主要是反面的资料，质疑某种效应是不是在“公平”的实验条件之下观测到的。至于当提出问题时，现实生活是不是同样的“公平”，则是另一个问题了。

(2) 被误解的任务

关于人为因素的研究本身，就隐含了对实验操作不当的批评。最初的研究人员本应该了解得更多，更加仔细。但是，这种指控，对于第二种类型的任务缺陷就不十分合适了：实验者不能理解被试的行为逻辑或者概念结构。重新阐述任务以澄清被试究竟在做些什么，这种方法已经被启发式—偏差研究路径的批评者和支持者所使用。使用这种方法，研究者可以试着指出存在于显然的偏差行为中的智慧，这种方法包括：(a) 证明某些备择目标是以牺牲所研究任务中的最优性得到的（例如，通过做出错误的诊断来了解一个系统的属性）；(b) 证明被试对关键术语的定义与实验者所持有的或者假定的不同；(c) 证明除非决定做出其他假设，否则被试就不能完成任务，而这些假设只有在非常偶然的情况下才会与实验者给出的假设一致；(d) 证明被试做出了一个合理的区分，而实验者却没有意识到。

426 要真正有所作为，这种任务的重新阐述，就应当有经验证据的支持，而不只是简单地说一句“被试可能是这样想的”。最坏的情况是，这种宣称可以有其特色，使得弄虚作假成为不可能；实际上，相互矛盾的说法，可以用来解释消除不同的偏差。最好的情况是，它们可以做出强有力的关于认知表征的理论陈述（Fischhoff, in press-a）。

2. 错误的判断者

(1) 可完善的判断者

如果任务已经被改进，但是偏差依然存在，那么问题就在于被试了。为了排除多余的行为，研究人员可能会采取一种递进设计（escalation design），每一步都反映了实验者不断增加的对于轻易完善人类表现的悲观情绪：（a）警告可能存在偏差，但却不指明偏差的性质（这一策略不同于通过暗示潜在的错误是系统性的，因此被试需要指导而不只是公平的机会来鼓励被试更加努力工作）；（b）描述典型观测下的偏差的方向（或许还有偏差的程度）；（c）提供一份反馈信息，使警告的含义个人化；（d）提供一个充分的训练计划，包括反馈、指导以及其他任何让被试从认知上掌握任务的東西。

通过假设这种解决方案不会自动出现，或者不会仅仅因为仔细地修改问题的措辞就会出现，这样一步一步地就可以找出判断者的缺点，而非任务的不足。尽管具有重大的实践意义，但是训练练习仍然可能限制了理论上的影响力。试图发现什么起作用的因素的努力，可能创造出系列作用因素不明的操作策略。可能还需要更多的系统性实验来确定这些因素。最终的目标在于理解被训练计划创造出的人为经验与现实生活所提供的、自然而然形成的经验如何不同。为什么一种技术可以消除偏差，而另一种则不行呢？

(2) 不可救药的判断者

有时候，自以为是的训练者可能认定成功是不可能的，或者只有迫使被试采取最优反应才可以得到。由于实质上是给予被试正确答案，或者是制造出某种不可避免的需求特征，“成功”既失去了理论上的意义，也缺乏实践意义。人们会按别人所说的去做，这已经不是什么新鲜事；如果他们不得不时刻被告知该如何回答，那么谁还会需要这样的被试呢？

有三种可能的选择：（a）用一些更高级的回答装置取代人；（b）重新校准不可靠的判断为更加合理的值，假设误差的总量和方向可以预期；（c）当人们根据这些判断计划行动时，承认判断的不准确。已经在上述三种方法中的任何一种上放弃对人的信任的决策者和决策分析人员，仍然有

427 可能通过估计那些不能消除的偏差的大小、范围以及弹性，帮助我们理解人们的判断。但是，因为改善判断并不是这些校正行为的意图，所以我们在此只粗略地介绍。

3. 判断者与任务之间的不匹配

(1) 重构

或许最为温和的心理学视角，是不单纯地指责判断者或者任务。相反，假设问题已经被合格地提出，并且被试拥有所必需的技能，但是不知为何，这些技能并没有被使用。秉承人类工程学的理念，这种研究方法认为，合适的观测单位应该是一个一个的被试—任务（person-task）系统。成功正是因为可以使这二者尽可能地相容。正如一架机械上完好无损的飞机需要良好的仪器设计才可以飞行一样，诚实的（即不是误导的）判断任务，也只有当它被重构成为一种使得被试能够有效地使用现有的认知技能形式时，才是可以处理的。

尽管这种认知工程学倾向于根据具体的任务采取策略，但是仍然有一些策略是共通的：（a）迫使被试把他们所知道的东西表达出来，而不是让他们留在脑袋里；（b）鼓励被试寻找不一致的证据，而非收集支持所选答案的证据细节；（c）向被试提供把重大问题分解为更加熟悉的、更加容易处理的组成部分的方法；（d）建议被试考虑他们可能经历过的所有情境，使得他们可以更好地理解面对的特定情境；（e）提供其他表述问题的方式（例如，使用不同的术语、具体化以及给出类比）。

(2) 教育

被试—任务“系统”研究方法的一个变种认为，人们可以完成这个任务，但不是现有的这些被试。可以使用其他人选：（a）专家，他们除了拥有实质性的知识，还掌握了在不确定性条件下处理信息的能力；（b）“新生代”，他们从小就受概率思维的熏陶。从某种意义上说，这种观点认为，即使从原则上讲人不是不可救药的，但是目前的大多数人做不到这一点。教育与训练（前一个类别）的区别，就在于前者培养通用的能力，而后者只培训特定的技能。

二、后知之明的偏差：一个消除偏差努力的实例

任何负责任的工作的一个重要方面，就是从经验中学习。一旦我们了解了某件事情是怎样出现的，我们会试着去了解它为什么会发生以及评价我们——或者其他人——的相应计划。尽管这种从结果中得到的知识被认为赋予我们的判断以后知之明的智慧，但是它的优点可能还是被夸大了。在后知之明中，人们总是会夸大那些本应该事先预料到的事情。他们不仅倾向于把已经发生的视为不可避免的，而且倾向于把它视为在发生之前就已经看起来是“相对不可避免的”。人们相信其他人本应该可以比实际情况更好地预期某一事件。他们甚至不记得他们自己的预期，以至于在事后夸大他们事前所做的预测（Fischhoff, 1975）。尽管相信我们始终知道那些只有在后知之明中我们才可能知道的东西会令人高兴，但是这种信条，几乎不能提供什么来帮助我们公正地评价意外与失败在什么程度上是不可避免的。单纯地指责在难免出错的系统中犯错误的决策者，而不承认这种错误的不可避免性以及改进系统，这既不公平，也是在自欺欺人。通过鼓励我们夸大我们的知识范围，这种偏差使得我们对于自己的预测能力过度自信。一个没有感觉到任何意外的过去，可能意味着一个充满意外的未来。

对这种偏差的研究，已经包括了对前一部分提到过的大多数可能的消除偏差策略的研究。但是几乎没有什么技术成功地减轻了后知之明偏差；更没有人曾经消除过这种偏差。这些技术列在表 31—2 中，下面将会加以讨论。

1. 错误的任务

(1) 不公平的任务

在最初一个证明后知之明偏差的实验中，被试阅读一段描述一个历史事件的文字，然后估计如果没有被告知真实结果，他们会赋予每一个可能结果的概率。无论被报告的结果是真是假（即是否和历史事实一致），他们相信，与那些没有被告知结果的被试相比，他们赋予结果的概率更高。这一研究被列入消除偏差的各种努力之中，因为通过把注意力放到一些小故事上，它回应了一个方法论上的批评，即“提出了太多问题”，而这一

批评可能影响到后续的研究。其他一些问题较少的研究,并没有涉及消除后知之明偏差的问题,包括斯洛维奇等人的研究 (Slovic & Fischhoff, 1977),他们要求被试分析几项科学实验的可能结果的概率;米希尔等人的研究 (Mitchell & Kalb, in press),要求一些护士估计出院后发病的概率;和其他学者的研究 (Pennington et al., 1980),他们要求一些妇女估计每个人在一次怀孕测试中结果呈阳性的概率。

429 其他一些试图证明后知之明偏差的人为来源的努力都失败了,它们包括:使用“惊奇”判断等级量表代替概率估计 (Slovic & Fischhoff, 1977);使用更具有同质性的问题使得被试能够更加全面地唤起一类知识,而不是使用涉及多个领域内容的常识性问题,结果没有一个领域得到深入的思考 (Fischhoff & Beyth, 1975);恳请被试更加努力地工作 (Fischhoff, 1977b);试图消除被试对实验性质的疑问 (Wood, 1978);使用当代事件让判断者先予考虑,然后给出他们的事后估计 (Fischhoff & Beyth, 1975)。

(2) 被误解的任务

后知之明偏差的一个可能吸引人的地方,在于它把人们描述为他们始终知道将会发生什么,而这可能令人相当高兴。人们要为这种不应有的自吹自擂付出代价,只要 (a) 人们的先知之明 (foresight) 导致了事后看来很愚蠢的行为,或者 (b) 系统性地夸大人们的所知导致人们在现有的知识水平上自负,这很可能导致行为的反复无常,或者不能寻找到所需要的信息。由于这种长期后果在典型实验中并不重要,我们可能会担心被试被引诱着给自己涂上光辉的色彩。尽管大多数实验已经被用来证明被试能够重构事前性质的知识,而不是检验知识有多大的扩展余地,但是夸大的诱惑仍然存在。如果真是如此,应该反思被试和研究人员对任务解释的差异。一种旨在消除这种可能性的巧妙处理,需要被试首先回答问题,然后记住自己的答案,而他们的记忆的敏感性正是问题所在 (Fischhoff, 1977b; Fischhoff & Beyth, 1975; Pennington et al., 1980; Wood, 1978)。另一种巧妙处理,要求被试在事后估计他们的同伴的事前回答,假设他们没有理由要夸大其他被试的知识范围 (Fischhoff, 1975; Wood, 1978)。这两种处理没有一种被证明是成功的。被试记得他们自己曾经比实际情形中更有见识。就

夸大其他人事前原本（或者应该）知道多少而言，他们是毫不留情的“事后诸葛亮”。

2. 错误的判断者

学会避免囿于自己当前的视角而产生偏差，构成了历史学家的训练的一个——或许是唯一——核心所在（见本书第 23 章）。但是这种努力尚未见成功的先例。历史学家对于第一手资料的强调，以及他们化石似的对过去的观念的记录，可能反映了这样一种感觉，即人们的思想是不可救药的，需要某种类型的文件训练。尽管使用了极不严格的程序，但是这项关于训练的实验研究，并没有给出任何理由使我们可以保持乐观：费希荷夫（1977b）向被试明确描述了这种偏差，要求他们在判断中避免它——但毫无用处。

3. 判断者与任务之间的不匹配

（1）重构

为了使这些任务与判断者运用的认知技能和倾向更加相容，我们采用了三种策略重构后知之明的任务（hindsight tasks）。其中的一种策略，是及时地把被试和对事件的报道分隔开来，希望能够减轻报道对于被试的感知的支配趋势（Fischhoff & Beyth, 1975; Wood, 1978）；这种策略并不见效。第二种策略，是被试估计被报道的事件再次发生，而不是首先发生的概率，以希望在这种前瞻性的视角中，不确定性可以更加容易地被利用（Mitchell & Kalb, in press; Slovic & Fischhoff, 1977）；这种策略也以失败告终。最后一种策略，要求被试简要说明他们如何解释出现了事实上并未发生的结果（Slovic & Fischhoff, 1977）。这种相反证据的使用，显然减轻了被报道的事件在判断上的必然性。显然，被试的记忆或者想象中有这种相互矛盾的证据可以利用，但是如果没有重构问题，这些证据就不可能被真正使用。

（2）教育

几乎没有什么实验证据证明，职业教育带来的对于某一主题的深入了解，可以减少后知之明的偏差。有学者（Detmer et al., 1978）在外科医生（专家医生和住院医师）的判断中发现了后知之明的偏差，在实验中，这些医生评估一段涉及可能性腹部动脉瘤溃散的影片。而另一些学者（Arkes

et al., 1981) 在医生关于一名酒保的急性膝痛病症的临床描述中, 也发现了这种偏差。同样, 米希尔他们 (Mitchell & Kalb, in press) 在护士对属下行为结果的评估中也发现了这种偏差。如果对自己生活中的事件进行判断的人可以被看作是实质上的专家的话, 那么彭宁顿等 (Pennington et al., 1980) 关于判断自己怀孕测试的结果的妇女的研究, 也就可以看作是又一个证明专家偏差的实例。甚至对于更为严格意义上的专家, 伍德 (Wood, 1978) 发现, 当涉及常识性问题时, 那些富于学识的专家被试并不会比普通人更少受偏差影响。关于专家沦为后知之明的偏差的受害者的逸闻趣事, 在本书第 23 章中已简要地述及, 它们包括因果观测和广泛的研究, 如有人 (Wohlstetter, 1962) 把由国会珍珠港事件调查委员会成员组成的情绪高昂的 39 名专家的努力, 描述成一大堆后知之明的偏差的集合。

4. 小结

491

尽管后知之明的偏差是一个较少被研究的判断问题, 但它也已经有了足够的研究可以给出一些检验性的一般结论: 后知之明的偏差是一种统计检验上相当强并且广泛存在的现象。若想减轻这种偏差, 需要对人类的认知过程有一定程度的了解以及一些假设。其中一个假设是, 人们处理后知之明的偏差的通常方式, 并不是最有效地使用他们的知识或者推理上的技巧。利用相互矛盾的证据, 似乎可以部分地解决这一问题, 帮助他们更好地运用自己的心智 (Slovic & Fischhoff, 1977)。

但是, 在认可这一解决方案之前, 还有许多经验问题需要加以说明: (a) 还需要什么步骤才可以消除而非减轻这种偏差? (b) 这种程序对于结构不那么明显的任务适用吗? (c) 处理很少一些示范性任务的实践足够改变其他任务中的行为吗? 在这些任务中没有给出具体的指导。如果消除偏差的程序, 增加了人们对于他们的判断能力的信心, 而不是提升这种能力本身, 那么, 这种程序的使用, 很可能会陷入更多的麻烦, 以至于不值得被采用。

三、过分的信心: 消除偏差的努力

“不确定状况下的决策”, 意味着知识的不完备。因此, 这种决策制定

的一个重要组成部分，就是估计所有可以利用的知识的质量。尽管统计方法可以指导这种估计，但是在一些或者其他情况下，判断需要估计人们赋予他们关于事件状态的最优猜测的信心水平。因为通过引致过度的或者不充分的警告，信心过度或者信心不足的估计，可以导致不良的决策，所以判断研究长期以来的一个兴趣所在，就是识别对信心产生不适当作用的各种因素。就其引导人们夸大他们所掌握的知识完备性而言，从结果中得到的知识，是其中的一个因素。尽管人们怀疑结果知识是否使得人们对于他们的所知过度自信，但是可以相信，人们受到特有的某种类型自卑倾向的影响，而后知之明的偏差，则有效地抵消了这种自卑倾向。澄清这种可能性，需要有关评估置信判断的绝对有效性的研究。

由于难以估计任何单一置信判断的绝对有效性，因此这一领域中的大多数研究，着眼于成组判断的质量或者校准。每一个判断，代表了被试认为某件事实的陈述正确的主观概率（见本书第 22 章）。比如说对于完美校准的个体，估计为 0.70 就对应着有 70% 的时候陈述是正确的。

显然，过分的信心，是最常观测到的结果。一项典型研究的结果可能是：0.75 的概率对应着仅为 60% 的“正确率”（hit rate），而关于确定性（ $p = 1.00$ ）的陈述，只有 85% 的时候是正确的。当人们估计他们对于数值型变量的量值的了解情况时 [例如，“我有 98% 的把握认为莱恩县（Lane County）注册共和党投票人数在 12 000 到 30 000 之间”]，人们发现，真实答案落在他们的 98% 置信区间之外的情况，竟然达到了 20% 到 40%，而这并不是非常不同寻常的事情。这种结果，对于那些必须依赖置信估计以及那些（直接或者间接的）把问题归咎于夸大其所知的人而言，都是令人烦恼的。试图证伪、质疑、支持或者限制关于过分的信心现象的大量研究发现，可根据如下的消除偏差的努力的视角来进行分类。几篇全面评论这方面的文献的文章的出现，极大地支持了对现有研究进行的二次分析，尽管它们之中，有一些评论，其目的并不在于此。它们包括下面的这些研究：Henrion, 1980; Hogarth, 1975; Lichtenstein et al., 22; Wallsten & Budescu, 1980。这种二次分析之所以变得复杂，正是因为被引用的许多研究并非为此目的而进行的。因此，我们不能总是简单地把它们划入一个单一的消除

偏差的类别。这种温和的不匹配，可能反映了目前的分类表的局限（由于不甚明确的区分）或者研究的局限（由于混淆了各种消除偏差的处理技巧）。

1. 错误的任务

(1) 不公平的任务

过分的信心的现象的实践意义，在于它引发了技术层面大量消除自负的努力，其中绝大多数已经被证明无效。它们中很多技术涉及“回答模式”的处理，例如比较信心的概率和几率表达（Ludke et al., 1977），或者改变创建主观概率分布时的估计置信区间（Selvidge, 1980）。在基于一些实质性的理论免除了创造和证明他们的实验操纵的必要性之后，采取“工程学”研究方法的实验者，频繁地表现出他们在乐于尝试的实验程序方面的创造性。但是，理论的缺乏，也使得解释和推广他们的成功和失败变得更加困难。例如，西福等（Seaver et al., 1978）发现，当使用“定值”方法（即实验者选择数值，被试估计它们的概率）而非“固定概率”方法（即实验者给出概率而被试给出相应的数值）推导置信区间时，可以观测到更少的过分的信心。这一成功可能反应了定值方法与被试的心理过程之间存在某种类型的更大的兼容性，或者反映了实验者选择的定值所传递的信息。特沃斯基和卡尼曼（1974, 1）的研究，基于所假设的初始化—调适判断启发式，得到了类似的结果，尽管他们在研究中也向被试提供了定值。

除了对置信的正确回答模式进行相当深入的研究之外，表 31—1 的上部，还列出了其他一些零散的试图消除另一种威胁任务公平性的因素的研究。例如，许多校准研究，为了得到统计上可靠的个体结果，需要大量的回答，而如果每名被试使用少于 10 个，甚至只有 1 个问题的研究，就没有观测到过分的信心的现象，那么回答数目的多少，可能就是值得关注的问题了（如 Hynes & Vanmarcke, 1976; Lichtenstein & Fischhoff, 1977）。一些研究中使用的指导语简短精练，而如果使用了冗长烦琐到似乎会令被试无法忍受的指导语的研究，没有发现类似的结果，那么指导语的长度可能就是问题的所在了（例如，本书第 21 章；Lichtenstein & Fischhoff, 1980b）。这种详尽无遗甚至有些迂腐的指导语，似乎可以看作是一种缓解被试任何试图猜测实验者的意图的方法。至于所使用的刺激的明确性，当多样的常

识性问题组被换成同质的问题（例如，Fischhoff & Slovic, 1980; Oskamp, 1962）或者非言语的“知觉”问题（例如 Dawes, 1980; Lichtenstein & Fischhoff, 1980b）时，没有发现过分的信心有任何变化。

当奖励被提高并且判断者对“现实”（即不是仅对实验）情况进行判断时，我们就可以放心地相信过分的信心现象将会消失。但是不幸的是，被用于研究这一假设的研究策略，很可能面临解释上的困难。一种明显的方法，是监测执行习惯任务的专家们的信心表达。但是，评估专家表达的标准很可能与校准冲突，这种可能性击毁了上述方法的合理性；也就是说，很有可能会奖励谨慎流露出的过度自信或者夸张的过分小心。例如，当医生过高估计了某种病症时（例如，Christensen-Szalanski & Bushyhead, 1981; Lusted, 1977），这很可能是因为他们无法确定自己到底了解多少或者是因为担心治疗失当，或者是贪图患者额外检查可能多付的金钱报酬，或者是其他与当前目标不相干的考虑。由于这些复杂因素，本节所列举的置于表 31—2 底部的专家研究专用于此，而不是作为提高奖励的努力。

表 31—2 去误差的经验

策略	关于事后偏见的研究	关于过分信心的研究
错误的任务		
不公平的任务		
提高奖励	4	1, 30
澄清指导语/刺激	6	3, 10, 13, 14, 21
阻止事后预测	11	13, 21
使用更好的回答模式	9	13, 14, 20, 22, 23, 32, 34, 35?, 36, 40?
少问一些问题	3, 7, 8, 9	16
被误解的任务*		
证明备择目标	3, 4, 6, 8, 9	14
证明语义的不一致	—	3, 14, 19, 30?
证明任务不可能	—	13
证明区分被忽视	—	15?
错误的判断者		
可完善的判断者		
警告问题	—	13
描述问题	4	3

434

不确定状况下的判断

续前表

策略	关于事后偏见的研究	关于过分信心的研究
提供个人化的反馈	—	21
充分训练	5?	1, 2, 4, 17, 21, 26, 27, 31, 34
不可救药的判断者		
替换人员	—	—
重新校准他们的回答	—	2, 5, 24
基于错误制订计划	—	—
判断者与任务之间的不匹配		
重构		
让知道的东西显现出来	—	18
寻找不一致的信息	9	18
分解问题	6, 11	—
考虑替代情境	—	—
提供替代方式	7, 9	35?
教育		
依赖真正的专家	1, 2, 7, 8, 10, 11	11, 16, 20, 24, 29, 33, 38, 39/8, 9, 23, 28, 31, 32 ^a
从小训练	—	6, 7

注：研究凡例见注后，黑体字表明实验操作被证明至少部分成功，而仍有待经验检验或者证据不明显的研究以问号标记。

a. 斜线前的条目表示研究使用的专家没有接受过校准训练；斜线后的条目表示研究使用可变的难度层次。

研究凡例

事后偏见

1. Arkes, Wortmann, Saville, & Harkness (1981)
2. Detmer, Fryback, & Gassner (1978)
3. Fischhoff (1975)
4. Fischhoff (1977b)
5. Fischhoff (1980)
6. Fischhoff & Beyth (1975)
7. Mitchell & Kalb (in press) (付印)
8. Pennington, Rutter, McKenna, & Morley (1980)
9. Slovic & Fischhoff (1977)
10. Wohlstetter (1962)
11. G. Wood (1978)

过分的信心

1. Adams & Adams (1958)
2. Adams & Adams (1961)
3. Alpert & Raiffa (1969, 21)
4. Armelius (1979)
5. Becker & Greenberg (1978)
6. Beyth-Marom & Dekel (in press) (付印)
7. Cavanaugh & Borkowski (1980)
8. Clarke (1960)
9. Coccozza & Steadman (1978)
10. Dawes (1980)
11. Dowie (1976)
12. Ferrell & McGoe (1980)
13. Fischhoff & Slovic (1980)
14. Fischhoff, Slovic, & Lichtenstein (1977)
15. Howell & Burnett (1978)
16. Hynes & Vanmarcke (1976)
17. King, Zechmeister, & Shaughnessy (in press) (付印)
18. Koriath, Lichtenstein, & Fischhoff (1980)
19. Larson & Reenan (1979)
20. Lichtenstein & Fischhoff (1977)
21. Lichtenstein & Fischhoff (1980b)
22. Lichtenstein, Fischhoff, & Phillips (Chap. 22)
23. Ludke, Stauss, & Gustafson (1977)
24. Moore (1977)
25. Morris (1974)
26. Murphy & Winkler (1974)
27. Murphy & Winkler (1977a)
28. Nickerson & McGoldrick (1965)
29. Oskamp (1962)
30. Phillips & Wright (1977)
31. Pickhardt & Wallace (1974)
32. Pitz (1974)
33. Root (1962)
34. Schaefer & Borcharding (1973)
35. Seaver, von Winterfeldt, & Edwards (1978)
36. Selvidge (1980)
37. Sieber (1974)
38. Staël von Holstein (1971a)
39. Staël von Holstein (1972)
40. Tversky & Kahneman (1974)

另一种提高奖励的方式,是对一些本身就很重要的研究附加置信估计,而这些估计不具有任何行为上的暗示。有人(Sieber, 1974)根据学生自己的测试答案推导他们的信心水平,做到了这一点。结果(现在广为人知)

435 发现了过分的信心现象，这可能是因为校准对于涉及的奖励的不敏感，也可能是因为这种方法本身就不能有效提高奖励。理论上完美的奖励控制策略，是根据适当的评分规则奖励被试，同时惩罚不坦诚地表达不确定性的被试。但是，这种规则相当不对称，也就是说过度地惩罚了过分的信心，而较轻地惩罚了信心不足。因此，那些掌握了规则要领但是又对其详细情况不感兴趣的被试，可能把评分规则理解为是在迂回地指导他们绝不要表现出较高的信心水平。在这种情况下，人们可能是机械地减低他们的信心水平，而没有提高他们的理解能力。总之，或许让被试努力工作的最好方法，就是训练研究人员的基本技巧，以提高任务的内在吸引力和被试的参与精神。

(2) 被误解的任务

436 无论如何仔细地向被试描述一个任务，仍然可能存在这样的一些疑问，即被试是否确实了解了任务并且接受实验者预期的奖励结构。检验一种设计是否可行的一个基本策略，就是在一个新的任务中，使用这些回答并且保持原有的奖励结构，看一看被试是否忠于他们自己已经做出的回答。费希荷夫等人（1977）采用了上述策略，他们询问被试是否愿意参加一场基于他们刚刚做出的置信估计的赌局。在这场赌局中，如果被试是诚恳地或者倾向于低估他们自己的信心水平就会受到奖励，而如果他们夸大了自己的所知，则会受到惩罚，无论什么原因。例如，谨慎地夸大其词可能会作为替代的目标，从而他们表现得比真实情况下更具有学识。这样的被试，更加愿意参与赌局，而不顾他们和其他地方被观测到的被试一样自负。

另一种主张被试对任务的理解与实验者不同的观点，来源于这样一个观测结果，即“确定性的程度经常被用于日常话语之中（就像提及温度一样），但是这种程度很少用数字的方式表达出来，而且也很难有机会去确定它们的有效性，……人们不能恰当地估计 0.80 的概率，这并不更令人吃惊，正如人们很难评价蜡烛的亮度和华氏温度一样。”（Fischhoff et al., 1977, p. 553）对这种可能性的一种反应，就是把注意力限定在概率的极端取值上，相信“大多数人更容易接受我 100% 地确定某个表述为真这种表述，也更容易评价这种表述的恰当性”（Fischhoff et al., 1977, p. 553）。另

一种反应，是给数值型的概率贴上一个文字标签以使它们更容易被理解（见本书第 21 章；Lichtenstein & Reeman, 1979）。不过没有一种处理技巧，被证明是明显有效的。一种更为深入的关于实验者与被试之间的语义学分歧的观点主张，“不确定性”这一术语本身可能有多种解释，并且并不是所有这些解释，对每一个人都有意义（Howell & Burnett, 1978；Phillips & Wright, 1977）。基于这种思想的消除偏差的经验努力，可能被证明是卓有成效的。

对于一些任务，被试根本毫无所知。在这种情况下，我们可以观察到一些最极端的过分的信心现象。尽管一般来说，实验者不会指点被试应当抱有什么水平的信心，但是仍然有可能存在一个隐含的假设，即被试认为“实验者不会给我一个不可能完成的任务”。如果被试真的这么想，那么想要他们抱有适当的信心水平就是不可能的了。费希荷夫等人（Fischhoff & Slovic, 1980）使用一系列任务检验了这种可能性，这些任务的内容（例如，诊断溃疡、预测不知名的股票的价格）以及指导语的设计，都是为了使任务不论是看起来还是本质上都是不可能完成的。但是，只有当被试被提醒“很有可能不能做出这种类型的区分。试着尽最大努力完成任务。但是在极端情况下，如果你感到（对于自己的答案）完全没有把握，那么请别犹豫，每一个都回答 0.5（表明这是猜测）好了”（p. 752）时，过分的信心才被减轻（但是也只是部分地）。任何更强一些的指导语，都可能具有研究者自身的需求特征的嫌疑。

2. 错误的判断者

(1) 可完善的个体

适当地修改解释性假设，则使前一节中最后提到的研究，在本节中将变得抢眼起来。让被试相信他们可以把每一个回答仅仅看作一种猜测，这种做法可以看作是一种消除任何残留的对于任务的误解的方法，或者是朝向纠正那些理解了任务但却不了解其自身的被试迈出了一步。这种方法有一个隐含的警告，即不接受给出的猜测性答案，可能成为一个问题。阿尔泼特等人（见本书第 21 章）的指导语，就把这种警告明显化了。在这项研究中，为了避免过分的信心，他们指导被试延长他们的主观概率分布的双

437

尾。这种巧妙处理的部分成功，是否反映了人们对于命令的理解力和敏感性，仍不清楚。但这种模棱两可，可能解释了采用这种方法的研究为何如此之少。

随着“实验者效应”成为审慎的训练研究中的常见现象，对于要求特征的担心逐渐消失。正如表 31—2 所指出的，人们已经进行了一系列训练方面的努力，并且取得了极好的成绩——尽管人们可能会担心，学术期刊缺乏对于反面结论研究的热情，可能会削减人们对于失败的可能性的关注。为了获得某种效应，训练者们甘愿做任何事情，这可能使训练努力变成相当复杂的技术处理，而起作用的因素在某种程度上变得不再重要。就学习而言，一些更为必要的条件似乎是：得到大样本回答的反馈，被告知自己的表现（并且不只是关于一般性的问题），有机会讨论人们的主观不确定性感受与数值型概率回答之间的关系。让他们感到吃惊的是，利希坦斯坦和费希荷夫（1980b）发现，加上精细的个性化的反馈，一个回合的训练就可能像一系列的试验那样有效。就狭义上的掌握一种特定的任务（例如，学习实验者所要求的回答分布）或者获得更为通用的技能而言，这些成功，在多大程度上反映了训练的效果，仍不清楚。

（2）不可救药的个体

对训练研究的不耐烦或者对训练研究的普遍性的怀疑，使得许多研究人员认为，错误的置信估计是不可避免的，因此他们把注意力集中在帮助决策者处理这些估计上。一些人建议用专家小组替代个人，专家的估计可以同方向相加或者机械地整合（例如，Bechker & Greenberg, 1978; Morris, 1974）。其他人则要求自由地使用敏感性分析，分析决策分析中随时出现的置信估计（例如，Jennergren & Keeney, in press）。还有一些人提出使用校正因子重新校准估计，而修正因子则表明信心评估应该是他们当前的信心水平的函数（Lichtenstein & Fischhoff, 1977）。例如，过分的信心现象的普遍存在可能意味着，当某个人宣称确定性时，我们可以认为它的正确率为 85%。不幸的是，就这一策略而言，当人们被错误地校准时，他们的自负程度依赖于他们面对的特定任务的难度（Lichtenstein & Fischhoff, 1977）。因此，只有在了解手头任务的难度，并且能够观测到类似难度的任务中被

试的（过度）信心水平，或者至少能够推测出观测到的和预期的过分信心之间的关系时，才能决定所需要的校准总量（Ferrell & McGoey, 1980）。

3. 判断者与任务之间的不匹配

(1) 重构

关于校准的研究，和判断研究领域的其他主题一样，仍然相对隔绝于认知研究的主流，它从心理学文献中借鉴的更多是方法论而不是理念。无论这种联系上的缺乏是否反映了判断问题研究者本身的褊狭固执，或者是当前的认知过程模型对信心的不充分表征，它很可能已经妨碍了减轻过分信心的方法的发展。过程模型不仅应当提出更有效力的处理技巧，而且应该指出工程学的方法为什么起作用或者不起作用（以及它们的效应可以推广的范围）。当前关于目击证人的证词、可知感（feeling of knowing）和元记忆的研究，最终可能为我们提供这二者的契合点（例如，Gruneberg et al., 1978）。

一种帮助人们以一种与置信估计的要求更为兼容的方式使用他们现有的认知技能的方法，可能是要求被试列出为什么他们选择的答案可能错误的原因，从而减少过分的信心，这是一些学者的贡献（Koriat et al., 1980）。列出为什么某人可能正确的原因，或者给出一个支持和一个反驳某人所选答案的原因，并没有产生什么效果，这也表明，关键的因素并不仅仅是更加努力工作或者诚实坦率，而是要让被试表达出他们的记忆，而这与通常的置信估计任务不同。如果没有明确提示“请想一想为什么你可能是错的”，那么人们似乎不会对他们最初的答案持批评态度，更不会去想证明这些答案。或许类似地，有人（Markman, 1979）发现，除非被告知要去寻找，否则他那些 9 到 12 岁的被试不会发现文本材料中的前后不一致。

尽管在实践而非心理学领域取得了进步，但是西福等（Seaver et al., 1978）的定值技术（fixed-value technique），可以看作是另一种重构被试处理任务的方式。与“传统的”固定概率法把注意力集中于被试对正确答案的最优猜测相比，定值法围绕一组假定为错误的数值而组织一个人的知识，这可能导致对于“人们知道什么”这一问题更为全面的评价。

(2) 教育

过分信心的消除，是不是专家在其专业领域中所接受的实质性教育的

间接结果呢？探索这一问题的一种明显方法，就是考察真实任务表现中自信的表达。但是正如前面所提到的，由于现实压力可能限制专家的自由表达，这一方法被复杂化了。例如，人们可以在那些没有充分证明有效性的情况下就做出自信的判断的专业中〔例如，预测股票价格的走势（Dreman, 1979; Slovic, 1972c）；诊断精神病的危险程度（Cocozza & Steadman, 1978）〕发现过分信心现象的证据。当然，如果作为他们所激起的信心的函数，咨询（并付费）这种专家，那么他们可能会被诱惑而歪曲他们的所知。

毫无疑问，确保诚实的最大努力，是关于天气预报员的研究，他们的训练要求对良好的校准经常给予明确的奖励。他们的表现非常出色（例如，Murphy & Winkler, 1974, 1977a）。这种成功是否应归功于校准训练，或者是作为他们的一般职业教育的副产品，至今仍不清楚。其他一些研究中也使用了专家，但是专家没有接受过校准训练。对这些研究的评论证明，这种训练而非实质性的教育，才是起作用的因素。有关心理学系研究生（Lichtenstein & Fischhoff, 1977）、银行家（Staël von Holstein, 1972）、临床心理医生（Oskamp, 1962）、行政官员（Moore, 1977）、文职工程技术人员（Hynes & Vanmarcke, 1976）以及未受训练的职业天气预报员（Root, 1962; Staël von Holstein, 1971a）的实验研究，分别根据被试所在的专业领域出题，并且排除了现实世界的压力，结果都发现了过分信心的现象。

杜威（Dowie, 1976）在报纸上赛马预测栏目的撰稿人身上发现了良好的校准。尽管这些专家既没有采用什么明确的支付函数，也没有得到任何正规的反馈，但是人们可能仍然怀疑他们是自我服务，并时刻关注自己的表现，有良好的校准就自我奖励。也许我们从小就应该被训练学会这种类型的自我监控。最近有一些提议（例如，Beyth-Marom & Dekel, in press; Cavanaugh & Borkowski, 1980）主张判断应该成为学校课程的一部分，就体现了这种观点。这些提议的前景如何还有待检验。

最后，有一种相当狭义形式的专家知识，被证明是最有力的消除过分信心的方法（也是最无趣的）。被试的平均信心水平的变动，相对于他们面临的任务的难度的变动更缓慢，这一事实反映了人们对于他们的所知不

敏感 (Lichtenstein & Fischhoff, 1977)。正确答案比例与平均信心水平构成的典型配对是: 0.51, 0.65; 0.62, 0.74; 0.80, 0.78 和 0.92, 0.86。当准确率的变动范围为 0.41 时, 信心水平的变动只有 0.23。与这些概括性的统计数据一致, 校准曲线在某种意义上同样不佳 (或者说非常平坦); 但是, 他们的过分信心程度的变化却非常显著。前两对数值代表了信心过分, 第三对体现了适当的总体信心水平, 而第四对则说明信心不足。这些例子, 来自于利希坦斯坦等人 (Lichtenstein & Fischhoff, 1977) 的研究, 但是其他的研究 (Clarke, 1960; Nickerson & McGoldrick, 1965; Pickhardt & Wallace, 1974; Pitz, 1974), 也揭示了同样的模式。实际上, 对不同条件下的过分信心的任何比较, 都必须考虑所使用的任务的难度。从这方面来看, 文献中过分信心现象的普遍性, 部分反映了一种没有给予被试非常简单的问题 (或许是自然的) 的趋势。

4. 小结

我们已经采取了各种各样的方法从各种人群那里获取置信估计, 这些估计几乎总是表现出人们对于他们的所知非常不敏感。尽管并不需要关闭方法论操纵的大门, 但是迄今为止, 这些技巧都被证明是相对无效的, 并且它们的结论也难以推广。它们所做到的事情就是证明, 相对来说, 过分信心能够抵抗多种思维形式 (除了任务的难度水平的变化)。更多地依赖心理学理论似乎是得到更加有效的、可预计的处理技巧的关键。校准训练的有效性证明, 仔细分析这种训练而非专业化教育提供了什么独特的经验, 既可以指导我们消除偏差, 又能丰富我们的心理学理论。

四、讨论

假设此处所评论的这些研究都已经得到准确的描述, 并且穷尽了 (或者至少在相当程度上代表了) 相关的研究, 那么它们所传递的总体信息, 似乎可以让认知心理学家们相当放心了。后知之明和过分信心这两种偏差, 都已经被证明在统计检验上是相当强的, 它们抗拒那种把它们解释为人为因素, 并且通过机械的处理技巧 (例如, 使被试更加努力工作) 就能消除它们的努力。有效地消除偏差的方法, 通常需要改变任务的心理学性质

(以及被试处理任务的方法)。在这种情况下,我们必须至少部分地相信心理学理论。例如,一个关于人们在估计信心水平之前如何提取记忆信息的假设,指导了有些学者(Koriat et al., 1980)对提取过程的操纵。即使是
441 “把所有一切都丢给被试”的训练方案,也是基于经过良好检验的、可以广泛应用的基本学习理论。

本章的总结,还需要涉及几个概念上的限制(caveat)(除了它所接受的方法论上的限制)。第一个限制,关于人为的和心理学的处理技巧之间的区别,它可能比这里已经指出的更加模糊。例如,当根植于非正式的指导语不会推导出“真实的行为”这样一种观点时,劝说被试更加努力的工作,就是一种人为的处理技巧。但是,如果研究者能够提出一种关于不同的指导语如何影响判断过程的实质性假设,那么这种人为的因素,就会成为一种主效应,在有或者没有明显的劝说情景下,它能够独立地预测现实世界的行为。

第二个概念上的限制在于,质疑偏差的性质,可以反映出一种有限的、无效的心理学研究视角。继续前一段中的例子,生活中既有偶然的情况,也有努力工作的情况;其中哪一个都不比另一个在本质上更加“真实”。类似地,关于偶然和努力工作两种情况的实验室实验的相对效度,依赖于它们的结果能够外推到现实世界中的情境。每一种情况都有其位置。理解实验室和现实世界之间的匹配,需要描述这两种情况的良好判断。例如,努力工作的情况并不必然等于重要的情况。对于某些重要问题,人们可能并不努力工作,除非他们既意识到一个判断对于问题的结果的核心重要性,又意识到该判断可能出错的潜在必然性。

利用消除偏差的研究以揭示观测偏差的边界条件,引出了第三个概念上的限制。在本章中,总结表格和讨论隐含地赋予了各种研究同样的重要性,这些研究之所以被选中,可能是根据某种关于每一项研究的可靠性的观念(即由它们的研究能力和深广程度等标准来定)。这种对统计上显著的和并不显著的结果的记录程序,从方法论上来说就是不可靠的。当人们怀疑是否对所有可能的研究充分抽样了的时候,这在概念层面上就已成问题。在这种情况下,这些收集的数据构成了概念上相互依赖的观察,并且需要

赋予相同的权重。任何关于人类如何行为的总结，都需要仔细指明行为情境的子集，因为研究就是从这一子集中抽样的。例如，一些批评家指责关于判断启发式的早期研究是在“自寻烦恼”，因为这些研究需要寻找（掌握）人们反常行为的情境。如果这种说法正确，那么就不必把每一种对有偏行为的论证都看作是人类总体判断能力的不利因素；它们的重要性，受到这些实验所研究（或者过度解读的）的情境类型的限制。由于集中关注偏差估计的边界条件，大部分近期研究，都受到它们自身的抽样偏差的影响。在推广它们的结论时，需要考虑到这一点。

442

进一步的讨论

类似的模式是否会出现其他类别的偏差中，这需要类比的文献评论。表 31—3 给出了偏差领域的特征描述，其中可以寻找到反复出现的模式。表 31—3 也区分了本书所描述的偏差和困扰心理学家的其他偏差。

这些文献评论始终不能摆脱的一个元问题是，人到底有多优秀？他们是认知上的残疾者还是博学家呢？一个简单的答案，就需要回答无数关于生命的本质和人类经验与实验室情境之间的总体相似性的问题。现有的文献评论得出了一个令人困惑的结论，即人类判断技能的宝库是半空半满的。他们能够娴熟地应付现实生活，但是他们又会笨拙地犯下一些可以预见的重大错误；他们足够聪明，可以设计出许多应用广泛并且简便易行的启发式有效地帮助他们，但是他们又不够聪明，以至于不能意识到这些启发式的局限。因此，只有在具体的判断任务情境中，才能对人的能力做出比较明确的评价。

表 31—3

话语域：偏差和去偏差努力

1. 推理得以获得的隐含过程是概率性的。也就是说，判断是在不确定性的条件下进行的，而偏差就源于决定论的心智和概率性的环境之间的遭遇。
2. 问题源于证据的整合而非证据的发现。尽管刺激是完备的并且是尽可能明确的，但它们还是很少显示任务是如何被结构化的。而被试的工作，就是解释和利用被提供的这些片段信息。
3. 偏差是非实体性的。认知过程的操作，在给定信息结构的任何内容领域中，都应该是类似的。这就使错误不能归咎于由于有意欺骗而导致的错误信息（misinformation）和“错误观念”（misconceptions）。

续前表

4. 有些规范理论有助于合适判断。这条标准规则，就使问题脱离偏好王国（如不一致的态度），在其中，没有什么反应可看成是最优的。
5. 没有提供或允许计算的辅助（不用纸笔）。这种对直觉判断的关注，排除了种种辅助工具如计算器、统计参考书或互动计算机。
6. 对次优行为显然没有明显的诱导。也就是说，偏差就其本质是认知性的，而非动机性的。偏差研究的“核心”，当然在于人没有理由去进行次优行动，而错误则意味着他们仅仅是不知道更好的方案。

448

这种关于“人”的总体判断（或者说是借口），反映了大多数判断研究的一个共同特征——缺乏对个体差别的关注。尽管这种对群体效应的偏爱可能只是一种研究口味上的差别，但是由于主张判断研究中的主效应如此重要并且研究得如此不充分，因此关于个体差异的研究应该被暂时搁置，其结果可能从理论上合法化了这种偏好。而关于特征已知的群体的研究提供的洞察力相当贫乏，这也从经验上支持了这种说法。特别令人吃惊的是，对于已经知道的最为重要的群体——那些在专业领域内进行判断的专家——进行的实验室研究中，也缺乏对差异性的关注。道斯等（Dawes, 1976; Eddy, Chap. 18; Fischer, 1970）和其他人所收集的逸闻趣事以及个案研究证据也表明，充分的训练和高额的奖励并不能确保判断的优秀。尽管如此，我们仍然需要进一步的研究，这既是因为很多人顽固地相信专家会做得更好，也是因为有效使用专家判断在应用层面上的重要性。

对于充分利用专家以避免偏差这一直接的实践目标，了解专家是否比普通人更优秀，或者至少他们比普通人更能意识到自己的判断的局限，这就足够了。至于消除所有的判断偏差这一最终的实践目标，最重要的是了解专家可以做到什么程度以及他们为什么不能做得更好。下面列出了有助于学习的一般条件。对于其中的每一条，依据其所处的环境，我们可以看一看专家在哪些方面可能占有优势，或者处于劣势：

（1）练习大量的同质性任务。专家应该有过这样的经验。他们可以利用这种经验磨炼自己的判断技巧，或者发展出一套针对特定情境的习惯性解决办法，从而使自己从分析（和思考）需要之中解放出来。

(2) 有明确标准的事件。尽管专家经常被要求给出相当明确的判断，但是这些判断的对象一般是某个极其复杂的（自然、社会或者生物）系统的组成部分，因此很难评估判断者自身对对象的了解程度。不准确的判断，可能被归咎于不可预测的偶然性；而准确的判断，很可能是因为错误的原因。

(3) 特定任务的强化。从原则上说，专家为其表现付出代价。但是，即使在他们的判断意见可以被识别出来的时候，他们也可能因为其他理由受到奖励（例如，他们带来好消息了吗？他们打乱了计划吗？事情结果会变好吗？）。

(4) 明确承认对学习的需要。参加教授专业知识的培训，无疑是一种谦虚的表现。然而，在这个过程以及接下来的职业生涯的每一个阶段中，对于那些表现优秀以及流露出竞争能力的人，特定的优势会不断积累。

这些都是有关学习的纯粹操作性原理，是在假设不了解潜在的认知过程的情况下操纵行为的原则。澄清和利用这些认知过程，是消除偏差研究的一个主要的理论和实践任务，尤其考虑到这种操作在某种程度上优于更为机械的努力。尽管对于从基础研究到实地应用这一长段道路，关于偏差和消除偏差的研究，已经走完了相当一部分，但是它还没有在任何一端扎下根。现在看来，要想到达一端需要同时到达另一端。好的实践需要更好的关于人们的思想如何起作用的理论。而好的理论，也需要更好的实践来澄清和把握人们的思想真正起作用的条件。

感谢 Ruth Beyth-Marom、Don MacGregor 和 Paul Slovic 对于本章初稿的建设性意见。本项研究得到了海军研究室的资助，项目号 N00014 - 80 - C - 0150, Perceptronics, Inc.。

第 32 章

改善归纳推理

尼斯比特等 (Richard E. Nisbett, David H. Krantz, Christopher Jenson, and Geoffrey T. Fong)

445 本章，我们讨论改善人们日常生活中的推理的可能性。尼斯比特和罗斯（1980）提出，以推理格言的形式，使普通人能够运用科学家的推理工具，可能改善推理，如“这是一个经验问题”，或者“你总是能够为例外情况辩解”。关于一些最重要的非正式推理指南，我们会涉及“统计启发式”，如“把证据作为一个样本考虑，然后考虑一下样本规模”。我们相信，扎根于统计学和概率理论的知识，以具体问题中运用启发式策略的经验为依据，可以证明，这种粗略的推理指南，对于减少日常生活中的推理错误具有很大价值。

系统地考察这种方案能够做到的事情以及在哪里可能会遇到困难，是可能的。我们相信，目前，有三个此类的困难；我们也相信，就拓展关于人们怎样推理以及他们应该怎样推理的知识而言，解决每一个问题的努力都会有所助益。

(1) 给定一个模棱两可的推理, 确定它事实上是错误的可能非常困难。在卡尼曼和特沃斯基的大部分研究中, 作为规定推理 (prescribed inference) 基础的概率模型, 是标准的统计模型, 而把它应用于所述事件是不会引起争议的。然而, 当同一性质的社会心理学研究展开时, 越来越清楚的是, 对于事件, 准确地知道哪一种模型是正确的, 因而知道需要哪一种推理程序, 这是非常困难的。

(2) 甚至在能够相当确信错误已经发生的时候, 想知道错误是怎样发生的也是非常困难的。很难知道, 错误是应归于错误的推理, 也就是说, 拙劣的推理程序, 还是由于错误的模型, 即对于该事件性质的错误的先验信念。

(3) 甚至当我们能够相当确信合适的模型的时候, 我们可能仍然缺乏清楚有效的推理指南。目前, 我们并不清楚如何把一些最根本的统计考虑转换为日常生活中的推理指南。这在考虑样本偏差因素的情况中, 尤其明显。人们易于忽视证据被偏差影响的可能性, 但是, 远不明确的是, 对于大多数现实世界的问题, 哪一种“统计启发式”才适宜使用。

一、归纳推理中的模型和启发式

让我们通过分析一个相对缺乏上述困难的案例, 开始我们关于错误的归纳推理的讨论。本书中的许多案例可以满足要求, 但是卡尼曼和特沃斯基 (1972b, 3) 的产房问题 (the maternity-ward problem) 尤其明确, 并且帮助我们给出一些定义。被试被要求判断, 是大医院还是小医院在一年中拥有更多这样的天数, 该天出生的婴儿中, 超过 60% 是男孩。多数被试选择了“大约一样”; 其他的人, 一半选择了“大医院”, 一半选择“小医院”。换句话说, 作为一个群体, 被试相信, 这种男婴出生比超过 60% 的异常天数在大医院和小医院出现的可能性, 是相同的。

根据这个实验以及许多其他实验, 卡尼曼等人进行理论概括, 导致这种结果的主导的判断模式是代表性启发式 (representativeness heuristic) 的运用。一个被试使用这种启发式策略, 就会集中注意异常结果 (60%) 与假定的典型结果 (在这个案例中是大约 50%) 之间的不一致。既然近似的

程度或者“代表性”不受医院的规模影响，那么被试就会断定，异常结果对于大小医院的可能性是一样的。

然而，处理这个问题的正确方法是，给定任何一天，人们把一所医院该天实际的“男”和“女”出生结果这个子集，看作是类似于50—50总体的一个随机样本。然后依据二项公式或者大数定理，样本的规模越大，一个异常的样本百分比越不可能。根据这样运用的一种概率模型，人们得出结论，大医院有60%或者更多男婴出生的天数的可能性更小。

447 在这个例子中，直觉启发式（代表性启发式）所指导的推理，与运用数学模型的推理相对比。我们做了一个规范判断，批评直觉的结论，认为它是错误的，因为根据一个我们相信是新生婴儿性别恰当代表的模型，得到了不同的结论。在此，根据模型作出规范判断，我们的程序阐明了一种一般性原则：直觉推理的合理性，必须根据该事件的潜在模型的恰当性予以证明。

数学模型仅仅是模型的一种，而且，是人类推理中很少被使用的模型：它需要数学的训练和充足的时间思考。但是，我们将用模型这个术语指代任何对于现实的某些方面的部分代表。我们包括了物理模型（例如，一个根据原型按比例缩制的飞机模型）、数学模型（例如，描述飞机机翼上方空气流的方程）和更为普遍的直觉的概念模型。这些模型，是人们对飞机或者气流或者新生婴儿性别的决定程序或者现实的任何其他方面的心智表征（mental representation）。对于这些模型，一个极其重要的区分就是那些纯粹决定性的模型（即它们没有包括对不可预测性的任何描述）和那些概率性的模型（即它们假定，在给定的一般信息条件下，事件并不是完全可以预测的）之间的区分。概率模型可以是物理模型（掷骰子或者从壶中抽取弹子）或者数学模型（随机变量），但是它们经常是更不确定的。例如，田径运动会上，某个观看跳远比赛的人可能相信，下一个选手会跳到大约8米远，但是跳得稍远或者稍近一些不是稀奇的，甚至7.5米或者8.5米也是可能的。这个人对于跳远的心智表征，是一种直觉的概率模型，它包括一个心智上的“典型距离”和一种类型的“误差分布”，“误差分布”产生了对典型值的偏离。

一个人的事件模型的性质，对于将要使用的推理工具的选择至关重要，这些工具包括各种启发式策略。启发式，是任何一种转换信息从而解决问题或者形成判断的指导原则。上面我们谈到了代表性启发式，但是实际上，它包括两种不同的启发式：一种用来构建模型（根据代表性启发式策略，潜在的模型应该与观察资料的结构特征密切吻合），一种用来判断结果的可能性（如果某个结果的结构更类似于假定的潜在模型的结构，那么这种结果就更可能）。统计的启发式鼓励人们根据资料的性质考虑信息，如可靠性和有效性，使用概率性的概念操作信息，如基率（base rate）。如果对于一种特殊类型的事件，人们有一个潜在的概率模型，那么当考虑那些事件时，人们就可以运用统计的启发式策略。

让我们更具体一点，看看在产房问题中，这些模型和启发式策略概念是如何运用的。对于产房问题，一种完全令人满意的解决方法是使用一个数学模型，一个从 50—50 总体中随机抽样的模型。大多数的被试，可能使用一种更为简单的新生婴儿性别模型——该模型指定了一个 50—50 的划分，但是没有包含任何关于抽样过程的随机性与可能的抽样误差之间的关系。毫无疑问，大多数被试，确实了解一些样本规模与抽样误差之间的关系，但是这些知识并没有被纳入到他们为解决该问题而确立的心智的直觉模型中。相反，被试使用代表性启发式策略，补充简单的 50—50 模型，而这引导他们得出结论：60% 的男婴出生比不太可能，但是对于大、小医院，可能性是等同的。最后注意，至少从定性的方面来说，被试不必像统计学家一样解决这个问题，如果对于观察到的性别比例，他们的模型中包含有直觉的抽样概念，那么他们可能已经注意到了一种统计的启发式策略。在这个案例中，需要的启发式策略，是大样本比小样本更可能包含有代表性或者典型结构这样一种观念。 448

简单地说，推理是以模型为基础的。依据他的模型，一个人可以运用各种算法或者启发式策略。一个数学家可能从一个数学模型演绎得到定量的推论；一个使用直觉的抽样模型的人，可能成功地使用了一种统计的启发式策略；而一个使用过于简单的模型的人，可能使用了代表性启发式策略，并且至少是在这个问题中，他被误导了。

二、概率模型在日常推理中的应用

概率模型或者统计的启发式策略运用的失败，并不仅仅在回答产房问题之类的难题中导致错误，它也影响到日常生活情境中人们的推理，尤其是在社会领域。罗斯（1977）提出，人们会犯基本归因错误：他们倾向于把他人的行为更多地归咎于个人品性，而忽略了情境因素或者多变的环境对行为的影响。我们将给出两个关于这种错误的例子，并且指出考虑概率模型，如何帮助我们限定并且纠正这种类型的错误。

尼斯比特等（1975）指出，被试经常不能利用“共识”信息：对于一个特定个体的行为，甚至当他们被告知在同样的情境下大多数人也会有同样行为的时候，他们仍然推断是特质、个人品性的原因。例如，在一个研究中，被试认为“Greg R.”是一个缺乏感情、残忍的人，因为他没有去帮助一个他认为正被抢劫的“受害者”。对于实验组和控制组的被试，给予Greg R. 这样一种否定评价的倾向是同样高的。实验组的被试，被给予了共识信息，即在同样的情境中大多数人不会去帮助受害者；而控制组的被试，没有被给予这种共识信息，因此相信大多数人会帮助受害者。

449

人们怎样才能利用共识信息缓和他们关于个人特质或者品性的推论呢？在我们看来，一些强调行为的情境决定因素的社会科学家，运用一种直觉的（有时候甚至是形式的）概率模型，在该模型中，某些情境使得几乎所有人都会采取某种特定的行为，而一些个体几乎在所有情境中都会采取该种行为。在这种模型提供的结构中，共识信息，很容易被解释为特定情境属于高概率类型的证据。因此，该种行为的出现，就不是给定个体在所有情境中都会采取该种行为的有效证据。

我们认为，这种类型的概率模型很好地描述了社会行为，它们为批判那些不能利用共识信息的推理提供了根据。在尼斯比特他们的研究中，被试可能完全依赖代表性启发式策略，构建一个与资料匹配的目标个体的模型：Greg R. 是残忍的，缺乏感情，因为他的行为看起来是残忍的并且缺乏感情。

在这个例子中，和在产房问题的例子中一样，并不存在太多的引诱，

以至于不能为被试的推理辩护。一种社会行为的纯粹特质模型，没有情境对概率的影响，不是非常合理的，被试自己最有可能承认这点。但是，我们的第二个认为是基本归因错误的例子，导致了更多的争论。

假设一个重要的管理职位的候选人，由于以前雇主的强烈推荐，得到一个面试机会。面试从一个团体午餐会开始，在午餐会上，候选人表现得紧张不安。之后，几位参加午餐会的人认为，候选人缺乏该工作必需的人际交往能力。

一个社会心理学家可能会说，雇主关于“人际交往能力”的推理意义不明。关于冒犯行为的情境解释（包括午餐面试场景的要求），或者一种概率性的解释（他今天有点失常），可能是合理的。事实上，考虑到以前雇主的推荐，这种解释，看来比根据候选人的能力的解释更加合理。但是，雇主可能会回答说，任何不能在午餐会面试情境中表现得一贯泰然自若的人，非常可能在其他需要同等自信的相关工作情境中失败。

这种论点很难成功。雇主们根据一个模型来证明他们的推理，而该模型要求一种情境中的失败和其他情境中的失败之间存在一种高的统计相关，这些情境在某些方面与第一种情境类似，尽管不是在其他方面。很难知道最初的推理所潜在的实际认知机制是否包括了这样一种模型的使用，或者相反，仅仅依靠代表性启发式策略（考查候选人午餐会上的行为与一个刻板的理想候选人在午餐会上的典型行为之间的符合程度）。并且，也很难确切地知道雇主们提出的模型是错误的。社会心理学家根据经验知道，当进行经验检验时，这种相关关系通常是非常弱的。但是，这种特定的相关关系，并没有被直接检验，要检验它，运用现有的心理学技术，将会涉及众多细致的、成本过分高昂的研究。

这些例子，阐明了归纳推理可能发生错误的两种不同方式，它们也指明了可能需要的两种不同类型的规范建议或者训练：（a）有时候，推理的程序是错误的。有时候人们可能同意（至少基于反思）哪一种模型是合理的，但是可能在实践中，使用了其他更加简单的模型和相应简易的启发式策略。规范建议，强调避免过度简化以及代表性启发式策略过于简易的运用。训练的目标，包括教授具有广泛使用价值的模型，如二项抽样模型，

450

教授这些模型的具体应用，强调那些有时候能够替代根据模型的严格演绎推理的统计启发式策略。(b) 有时候，对于给定的模型，推理是正确的，但是模型可能有问题，或者甚至可以证明是错误的。规范建议，强调已知的事实（例如，关于某种类型的行为相关关系为弱），而根据这些事实，某个特定的模型看来更可取。

分析也指明了规范建议的可用性方面，存在两种非常严重的空白。在一些情况下，我们可能缺乏形成适当模型所需要的知识。在其他一些情况下，我们可能缺乏足够简单和可用的统计启发式策略。共识信息的例子，说明了目前社会科学的模型中存在的缺口。我们可以批评那些根本不能利用共识信息的被试，但是假设被试回答说，“是的，你是对的。情境因素至关重要；现在告诉我们，根据大多数人会以同样的方式行为这一事实，我们应该在多大程度上修改我们关于 Greg R. 是缺乏感情的看法？”回答这个问题需要一个特质/情境交互作用的理论，而目前，人格与社会心理学家还没有提供这样一种理论。

我们缺乏某些基本的统计的启发式策略，对这一事实的论证需要独立的一段。

三、针对样本偏差调节推理：统计启发式的需要

在推理过程中，考虑到样本规模的正确程序，已经得到了很好的发展，因为许多统计理论致力于这种类型的问题。作为其部分结果，我们预期，教授人们要更多注意日常生活提出的推理问题中的样本规模因素，不会存在很大的困难。但是，当我们考虑样本偏差问题时，事情就很不一样了。
457 很不幸，因为正如尼斯比特和罗斯（1980）已经论证的，看来推理错误更有可能来自于样本偏差而非小样本规模。

在讨论为了解决样本偏差，人们尝试开发启发式策略中所遭遇到的问题时，区分一个观察样本的两种主要类型的信息，将是有所帮助的：关于抽样程序的信息和关于共变特征的典型性的信息。对于第一种类型的信息，即与程序相关的，我们相当粗略地区分出偶然样本（haphazard sample）、统计样本（statistical）和例证样本（illustrative）。这些术语，立刻将我们拉

离了理论统计学的领域——它仅仅关注这些范畴中的一个——把我们推入偶然观察与观测事实的沟通（communication）这一领域中去。

偶然样本，是以任何可能的或者最便利的方法从所有可能的观察中选出的。大多数的推理以偶然样本为基础，不仅是在日常生活中，甚至是在细致的科学工作中。例如，在卡尼曼等人的研究中，所使用的被试之所以被观察，因为他们是当时能够得到的。统计样本，涉及一个已知的概率结构，用来从一些可能观察的总体中选择观察资料，它们是唯一一种可以进行严格的统计计算的样本。在民意测验或者质量控制测验中，普遍使用这种类型的样本。然而，通过构造一个合理的总体以及一个能够满足实际抽样程序的概率抽样模型，一个聪明的统计建模者，可以把一个偶然样本转换成为一个统计样本。最后，例证样本是一种用来提升一个基于大量数据的报告的正确性或者形象性的沟通工具（communication device）。电视报告和杂志文章，有时候使用生动的单个案例来说明猜想的总体特征。但是，甚至是在科学领域中，许多资料描述，也是例证性的。例如，在电生理学中，一些结论是根据对大量神经元的不完全观察得到的，对它们的说明，经常是使用那些来自更好地展示典型特征的神经元的记录样本。

关于样本，另一种类型的信息，是关于共变特征的典型性。这种信息，涉及多个变量而不仅仅是主要的兴趣变量的量值，它可能与主要的兴趣变量相关，并且使其与已知的总体值比较可以评估样本的典型性。考虑一个日常生活的例子，其中主要的兴趣变量是收入，但是每个个体的年龄和性别也都被记录了。假设我们想知道，以写诗作为业余爱好的群体和弹奏乐器的群体，哪一个群体的收入更高。碰巧我们有一个朋友杰克，50岁，他有很多朋友（多数是中年人）写作诗歌；我们还有一个朋友珍，25岁，她有很多朋友，大部分是年轻女性，她们演奏乐器。这些诗人的收入分布，远远高于音乐家的收入分布。所有人都会同意，对于所讨论的这两个群体，这不能证明什么：在与收入（在当下）高度相关的变量的水平上，它们是总体的非典型样本。很明显，在多大程度上样本的特征是非典型的，以及对于预测目标变量的值，这些特征的有用程度，非常重要。卡尼曼和特沃斯基的许多实验，使用以色列高中学生作为被试。然而我们和他们还是根

据结果作了更广泛的推论，因为尽管大多数成年人既不是以色列人也不在高中就读，但是，我们认为这些特征与他们的研究中的目标变量即使有相关关系，也是弱相关。应该注意的是，甚至在一个仔细构造的统计样本中，也必须考虑非典型性：如果一项对注册民主党人收入的调查使用一个小规模的随机样本，样本中白人中年男性占多数，那么我们不会认为这种标准类型的推理有效，尽管它是以随机抽样为基础。

为了调节非典型性，一个统计建模者，可能试图更加精确地构造目标变量和共变因素之间的关系，并利用这种关系来校正推理。例如，人们可能猜测或者估计出年龄、性别和收入之间的关系，然后比较调节后的诗人和音乐家的收入。这种调节需要经验、反思，并且经常需要定量的资料和足够大的样本（对模型检测和参数估计而言）。

我们已经注意到，有时候可以使用精密的模型来处理偶然抽样或者非典型样本，但是期望不是专家的普通人采取同样的方法来解决样本偏差，这是不合理的。那么，外行人使用哪种直觉程序修正样本偏差呢？我们能够提供哪些源于统计学家的拇指法则的建议，来改善他们的推理呢？

人们如何针对样本偏差进行调节，哈密尔等（Hamill et al., 1980）的两项研究，着手处理这个问题。总体的结论是人们根本不调节：根据偶然样本与根据相对的典型样本做出的推理中，人们没有表现出系统的差异，甚至对于特别指明为非典型的（可以说是反例证性的）样本，他们也得到了同样类型的推论。

如果读者想认真考虑一下人们应该使用哪种统计启发式策略，那么这些研究的一些细节，可能是有价值的。在一项研究中，被试观看一位据称是狱警的人的采访录像，一半被试观看到的狱警显得很仁慈，而另一半，狱警显得非常残酷。一些被试被引导而相信，给出的这个案例说明了更大规模的狱警样本；而其他被试被告知这个样本是反例证性的，他们被明确地告知，被采访的狱警是作为一个最极端的例子挑选出来的（对于那些观看到仁慈的狱警的被试，是一个最人道的狱警；而对于那些观看到残酷的狱警的被试，是一个最不人道的狱警）。第三组没有提供任何有关例证性的任何信息，实际上，所呈现的是一个样本规模为 1 的偶然样本。在另一项

研究中,被试阅读一篇杂志文章,文章生动地刻画了一位福利救济领取者的消极形象。一些被试被引导相信,关于人们依赖福利的时间长度,他们所阅读的个案是非常典型的:他们被告知(错误地),和大多数中年的福利救济领取者一样,他们读到的这位妇女作为成人的大部分时间依靠救济度日。关于这方面的特征,其他的被试被告知这个妇女的非典型性的真实情况:大多数福利救济领取者,只有几年是依赖救济生活的。

根据各种量度评估被试对一般的狱警(在第一项研究中)和一般的福利救济领取者(在第二项研究中)的态度。比起观看残酷狱警的小组,观看仁慈狱警的小组,对狱警的态度更加积极。这表明,人们是根据单个的案例推理一般的狱警。阅读了杂志文章的被试,对于福利救济领取者的态度比控制组更加消极,这再一次表明,人们是根据关于一位领取者的文章推理一般的福利救济领取者。

在狱警研究中,被告知该狱警是典型情况的被试,与那些没有得到任何抽样程序信息的被试根据录像做出的推理之间,并没有系统的差异。甚至那些相信该狱警是非典型的被试,也和其他组一样朝同一方向推理。事实上,他们与其他组之间的差异,在统计上是不可靠的。同样地,在福利救济领取者研究中,“典型”组与“非典型”组做出类似的推理。事实上,在他们的平均态度分数方面观察到的差异,可以忽略不计。

对于我们后面的规范分析,重要的是考虑怎样才能根据认知过程解释这些结果。如果假定被试根据过去的经验和信息形成了明确的关于福利领取者或者狱警的先验态度,那么结果相当令人吃惊。对于那些被提供了一个例证性或者典型性案例的被试,一种对他们关于该总体的先验态度的强烈的影响,可能是合情合理的,因为他们很有可能相信对于更大的群体,这个案例是个典型;但是,对于“偶然”组的被试,不存在这种辩护。并且,为什么那些被提供了一个非例证性或者非典型案例的被试应该完全依照样本指示的方向推理,而不是坚持他们原来的态度,或者甚至可能是沿着与提供的案例相反的方向推理呢?

当前关于态度形成和态度变化的理论,提供了一些试验性的解释(Bem, 1967; Nisbett & Ross, 1980; Nisbett & Wilson, 1977)。我们可以推测,

454 在回答关于他们的态度的问题时，被试根本不是从记忆中检索态度。相反，他们是根据手边的材料构建态度，包括他们当前对于客体的情感反应，他们关于客体的语义联想，观察客体最近的行为。大部分的构建过程，是快速的，没有被意识到的。一个生动的个案，可能引起对于它所代表的整个类别的客体的情感反应，尽管存在对典型性的补偿保证，但却是苍白无力的（似乎情感反应可能“逃离认知的牢笼”。无论多少保证，即使伤害过我们的巴黎出租车司机是个例外，也不可能恢复我们从前对于巴黎出租车司机的放心态度）。生动的个案，也可能作为一种从记忆中恢复类似信息的刺激。因而，被试可能因为仁慈的狱警而不自觉地回想起“霍根的英雄”中和善的纳粹警卫，或者在夏日重新放映的童年影片中由安迪·葛里费兹扮演的善良的南方治安官。当被试开始构建他们对于一般狱警的态度时，他们可能提醒自己不要忘记刚刚看到的只是偶然的或者非例证性的证据，他们甚至很可能成功地做到这一点，但是毫无用处。情感反应以及受到现在的偏差影响的记忆内容，足够导致一种与控制组被试完全不同的表现态度（expressed attitude）。

记住上述被试是如何被观察个案所影响，让我们转向规范问题：根据关于抽样程序以及相对于重要的共变特征而言的样本的典型性的有关知识，人们应该怎样调节他们的推理？形式理论和应用统计学的实践，提供给我们的，要比表面看到的还要少，更少于我们所需要的。

一条精明而讲究实际，富于坚定的科学特色的规则指出，除非根据合适的随机抽样，不应该作出关于一个总体的任何结论，否则就违背了统计程序的假定，并且也不可能知道多大程度上应该相信该结论。根据这条严格的规则，那些没有得到任何关于抽样程序或者录像中的狱警的典型性信息的被试，应该中止所有的推理，因而他们对于狱警的态度应该与他们看到的狱警是仁慈的还是残酷的完全一致。

根据三种不同的理由，可以拒绝这条规则：认识的、实用主义的和逻辑的。

（1）第一个论据来自于哥德曼（Goldman）的观察：“认识上的建议或者规则，必须能够被遵循”（1978，p. 513）。我们关于态度构建的可能机

制的讨论证明，就人们的日常认知而言，“中止推理”是一条不太可能的规则。一条更加具体、指明如何去除联想式记忆的内容偏差以及如何补偿情感反应的规则，可能是所需要的。

455

(2) 即使我们能够应用高度禁止性的规则，它也会产生令人遗憾的作用，妨碍我们学习大部分确实是我们学习得到的东西，不论是在日常生活中还是在科学中。在获取大学教务信息或者高速公路上的交通状况时，人们几乎不能摆脱他的偶然样本，亲历的或者间接遇到的。本章引用的那些关于人类推理的科学研究所使用的被试、材料和环境，都不是经过统计抽样得到的，而是方便使用或者易于构造的。

(3) 最后，随机性并不是抽样过程的一种可以直接观察或者不需证明的性质：它是数学模型的一个性质，数学模型可能也可能没有有效地描述现实世界的过程。一个偶然样本，可能在效果上是“随机的”，因为我们相信某些模型是对这个世界的有效描述，并且，正如我们所指出的，一个聪明的统计建模者，能够用适当的数学公式容纳这种直觉的洞察力。

我们怎样解决与共变特征的典型性相关的问题呢？表面上看来，可能值得一试的是制定一条规则：对于那些可以想象得到的重要特征，如果样本与目标总体存在实质的差异，那么禁止从中作一般性的推论。根据这条规则，就某个重要特征（依赖福利的时间长度）而言，被试知道关于福利的样本个案的非典型性之后，应该中止关于一般的福利个案的其他特征的所有推理。但是，这条规则与随机性规则一样，受到同种类型的批评。认识方面的批评完全相同：一个正确的规则，必须考虑在日常推理中所使用的实际认知机制，必须能够被遵循。实用主义方面的批评，类似于对前一规则作出的批评。在大多数科学研究中，人们可以发现，样本的一些特征是非典型性的，并且被认为是相关的。人们在多大程度上相信一个宣称这些特征非常重要的模型，这至关重要。最后，即使我们知道，或者强烈地怀疑一种非典型的共变特征很重要，我们仍然可能在一个模型中容纳它的重要作用，或者引入合适的修正。在业余音乐家和诗人收入的例子中，我们能够尝试粗略地估计年龄和性别的作用，然后比较两个群体调节后的均值。当然，在那个例子中，我们能够得到任何可靠程度的推理，这几乎都

是不可想象的。但是，这主要是因为我们缺少一个关于年龄和性别对收入的作用的有效模型。

456 可能存在能够运用的、更宽松并且更少强制性的原则吗？我们可以尝试，例如，“知识是试验性的”。就其本身而言，它完全正确，但是它太过含糊以至于没有太大的用处。我们需要更加具体的统计启发式策略。在多大程度上，我们应该相信从偶然样本中得出的一般性结论？根据非典型样本，人们能够得到什么类型的试验性推论？在某种程度上，这些问题，是向统计学家和哲学家提出的。存在一种普遍的模型建构方法吗，使这些模型能够作为直觉推理的统计启发式策略的来源？但是，如同哥德曼（1978）极力主张的，我们关于态度形成的机制以及影响推理的无意识过程的讨论表明，认识论的原则，必须通过可用性和有效性的检验。并且，在此我们还面临另外一个问题。我们想要修正的推理程序，有一部分是自动的和无意识的（Bem & McConnell, 1970; Goethals & Reckman, 1973; Nisbett & Wilson, 1977）。以什么代价战胜自动性呢？去理解指导、故事、动机诸如此类的东西，需要每一个清醒的时刻都进行大量的推理。对于自动性，可以合理地怀疑能够发现与知觉几乎同样程度的这些功能。有效的统计启发式策略应该是可以学会的，能达到容易运用甚至自动地使用的程度。

四、概率模型能够在日常推理中运用吗？

我们设想了一个方案，其中，概率模型和统计启发式策略，最终被纳入到多数人的日常推理中。我们已经讨论了这个方案的两个主要困难：无法知道哪一种模型是正确的；在一些重要的领域缺乏便于使用的统计启发式策略。第三个问题，涉及人们在日常思考中纳入概率模型和启发式策略的能力。这个问题同样困难，当然，它在逻辑上是最先的。尽管存在这些困难，我们仍然有理由保持乐观，并且一些问题本身可能有助于一种类型的实验研究，我们会在后面概述。

保持乐观的一个理由是，人类的推理能力伴随新的文化创造而改变。看来，推理方面的真正进步，只是在最近的历史时期内才发生的。现代的概率概念的历史，刚刚超过 300 年（Hacking, 1975）。在大约 1 660 年之

前, 几乎任何意义上的概率概念, 都是用于理解随机装置产生的事件, 如骰子和纸牌。而今天, 事实上每一个受过教育的人, 都会在一些领域内运用基本的统计推理, 例如运动和天气, 而他们对于如骰子和纸牌之类的随机装置的统计意义的理解, 也完全不同于前现代的概念 (Hacking 甚至说, 任何具有现代的概率概念的人在古代玩骰子, 都能在很短的时间内赢得整个高卢地区)。

除了纯粹的统计概念之外, 许多普遍的归纳原则和准则, 也是相当晚近时代的产物。相关关系不足以确立因果关系这一观念, 在休谟之前还没有出现过, 并且直到边沁 (1824/1952) 的《政治谬误手册》一书才得到明确的和全面的说明。直到密尔 (1843/1974) 的《逻辑体系》发表, 一种对“相似性准则” (Nisbett & Ross, 1980) 作为原因—结果关系推理基础的普遍控诉才出现。事实上, 直到 18 世纪末, 物理学家们还在遵循完全相反的规则, 他们习惯于根据“印迹学说” (doctrine of signatures) 推理。这个学说认为, 我们能够期望每一种天然药剂明显的外在性质, 指明它自身可以医治的疾病, 因而坚硬的、砾石般的物体, 对于治疗胆结石是有用的, 而黄色的物体对战胜黄疸病有帮助, 诸如此类。幸运的是, 物理学家证明他们能够抛弃这种归纳启发式策略, 支持其他更优越的策略。

保持乐观的第二个理由是, 有时候, 有效的推理可能比错误的推理更迅速、更简单。我们已经指出, 错误, 有时候源于过度简化以及正确的模型可能很复杂。但是, 反过来也同样正确。错误可能是由于过于复杂的因果推理, 而缺乏一种有效的、简单的解决问题的方法, 可能导致人们创造几种复杂的拙劣的方法。在我们当前的一个实验中, 我们进行关于体育的电话访谈, 我们在访谈中包括了一些适合于概率回答的问题。我们经常得到一些简单的、迅速的 (并且依我们的观点是正确的) 概率回答。其他的被访谈者, 不是根据一种概率的方法考虑问题, 他们犹豫不决地给出几个冗长的决定论式的解释, 这些解释看来甚至不能使被访谈者自己相信是正确的。

历史和逸事方面的证据, 都反映了一个同样的潜在观点。即使是迅速的、自动的人类推理过程, 也是基于反复学习获得的概念和模型。因此,

457

人类的推理以及偏爱的论证形式的变化，不仅仅是个体成熟的作用，也是语言、文化和教育方面的变迁的作用。

最近，我们开始一项研究计划，考查人们怎样推理以及他们怎样学习推理那些我们认为应该运用一种概率观点解决的问题。我们使用各种不同结构的问题，以及内容上或多或少易于引起概率性思考的问题。在此，甚至是试图对结果作初步的报告也是不合适的，但是，为了指出我们认为值得探索的问题，我们使用的材料可能有所帮助。为了说明这些材料，我们给出下面的问题。

大学的选择

458

戴维是东海岸一所中学的高年级学生，他正准备读大学。在中学阶段，他取得了优异的成绩，而且他被允许享有两个最高选择：一所小型的文学院和一所常春藤盟校的大学。两所学校在声誉上不相上下，并且费用也相当。两所学校都位于迷人的东岸城市，离他的家乡距离相同。戴维有几个好朋友在文学院读书，也有几个好朋友在常春藤盟校就读。他们和他一样成绩优异，兴趣也与他相近。他在文学院的朋友都对他说，他们非常喜欢那个地方，发现那里非常促人上进。而在常春藤盟校的朋友则抱怨在人际、社会背景和教育背景方面的不适。开始戴维认为他会去文学院，然而他决定某天亲自去看看这两所学校。他对自己在私立文学院看到的很不满意：他遇见的几个人看起来都很冷淡、很令人讨厌；与他短暂交谈的一位教授看来粗鲁无礼，对他也不感兴趣；他也不喜欢对校园的感觉。他确实很满意自己在常春藤盟校的见闻：他遇见的几个人很有活力、热情而且令人愉快；他碰到了两个不同的教授，他们对他表示了个人的欣赏；对于校园，他离开时留下了非常愉快的印象。

戴维应该选择哪一所学校呢，为什么？试着分析双方的论据，并解释哪一方的论据更有力？

这个问题特别困难，因为它包括两个不同的概率成分：（a）一个基率

论据，大致是说戴维不可能根据分布的尾端进行判断（从他的朋友的意见，他能够猜出对于文学院，他的典型反应是非常肯定的，而且变异性低）；(b) 一个样本规模/样本偏差论据，关于他某一天的见闻的适当性。根据预先研究或先导研究（pilot work），我们猜测，每一个概率成分都会受到一堆常识信念的阻碍，也就是说，每个个体的喜爱和厌恶情况都是独一无二的，仅能根据该个体的经验和人格加以预测，并且个体的情感，无论在开始是多么不可思议，但至少是可以信赖的，只需要很少的关于客体的信息就足够导致一种喜爱和厌恶情感的公开表达，这为未来的反应提供了值得信赖的向导。大多数对该问题的随意回答反应了这种信念（如，“他是为自己选择，而不是他的朋友”）。但不是全部如此。这里有一个本科生给出的回答，他没有学习过任何概率或者统计：

我想他应该去文学院。他在那儿的负面经验是因为和那个学校短暂的、非常肤浅的接触。他的朋友已经在那儿一段时间了（推测起来），他们所有人都和他非常类似，他们熟悉那个地方，并且喜欢那儿，而关于常春藤盟校的相反陈述也是真的。但是，他可能会跟从他自己对这两个地方的感觉来证明。经常，这种直觉是一种更高级的感觉能力，我们不能加以分析，他这样做可能是对的。尽管如此，我认为我提到的第一个选择更加可靠，因为他的经验太局限于这两个学校。

我们认为这个回答是一个非常令人满意的统计式回答，当然，尽管它 459
不是用概率模型的语言来表述的。

在开始这些研究时，我们猜测，在大多数情况下，统计学家能够有效地运用一种形式的概率模型，在日常生活中也存在类似的情况，经常有一些聪明的外行人直觉地运用类似的概率思维。例如，早些时候我们注意到，通常人们不能构造出这种模型，在模型中，一个个体的行为被视为情境变量和人格变量的共同作用，一些情境使得大多数个体采取该种行为。然而我们认为，当行为在检测中表现良好时，人们确实使用了一些和那种模型类似的东西。在我们的文化中，“容易检测”（easy test）的概念广为人知，

并且被广泛使用，它认为大多数人有可能表现良好。我们推测在日常推理中，为解决其他类型的问题，可能也使用了其他类型的概率模型。我们的实验试图证明这一点，人们确实能够有效地解决一些问题，虽然在其他具有类似形式结构的问题上失败了，因为对于后一种类型的问题，他们的潜在模型缺乏一个概率成分。

如果这个推测被证实，我们就被引向大量我们可能尝试用实验来回答的问题，在多大程度上，一个有效的概率推理，应该归因于应用了某种类型抽象或者形式的模型？为什么在一些类型的具体问题中，可以容易地运用形式模型，而在其他问题中却很难？如果成功的推理不是因为运用内容随意的形式模型，那么应该怎样使用理论术语来描述它？其他的问题，涉及模型建构和规则使用方面的个体差异。在多大程度上，概率推理的使用属于一种稳定的个人特征？它与智力相关吗？它与教育背景有关吗？

最后，还有一组重要的问题，涉及概率模型的可教授问题。基本的问题，已经在我们的研究中得到回答：通过传统的统计学课程和更加迅速的方法，人们能够学会以更加具有概率性的方式回答至少是上述类型的“文字问题”。至于哪一种教授方法更有效，需要教授的最佳归纳原理和方法是什么？这些重要的问题，我们仅仅是开始提出，它在很大程度上有待于日常生活中的实际判断。我们希望我们会得到多方面的帮助。

本章以及其中报告的一些研究的写作部分由国家科学基金 BNS 79-14094 资助。感谢 Lee Ross、Saul Sternberg 和 Paul Thagard 对于本章初稿的意见。



第九部分
PART NINE

风险知觉



第 33 章

事实对恐惧：理解可觉知的风险

斯洛维奇等 (Paul Slovic, Baruch Fischhoff,
and Sarah Lichtenstein)

463

人们会对他们所感知到的风险作出反应。如果他们的感知就是错误的，那么有关个人、公众以及环境保护上所作出的努力，就很可能受到误导。对于一些如机动车事故那样的风险来说，已经有大量的统计数据可以被采用。但对另外一些我们熟悉的活动，如抽烟、酗酒等，对其风险的估计，需要进行复杂的流行病学研究以及实验研究。然而，即使统计数据很充足，这样的一些“硬性”事实，对于我们形成应对措施的作用，也是极其有限的。在很多时候，人们作判断时需要解释这些事实，并判断它们的相关性。

而且还有其他一些新的风险，例如与 DNA 重组研究或核能相关的风险。正由于它们是新的，所以对其风险的估计必须建立在像错误树 (fault trees) (见图 33—1) 这样复杂的理论分析之上，而不是以直接的体验为基础。不过，尽管这种理论分析有一个客观的表象，但其中仍然包含着很大的个人判断成分。一

个人必然会依赖于他通过接受教育而产生的直觉，去判断所分析的问题的结构，考虑其结果，以及分析错误树的各种分支的重要性。一旦这些分析得以完成，将被传达给那些实际处理风险的人，他们中间包括企业家、环保主义者、政府调节员、立法者以及选举人。如果这些人不理解或者不相信传达给他们的这些数据，那么就可能产生不信任、冲突以及无效的风险处理。

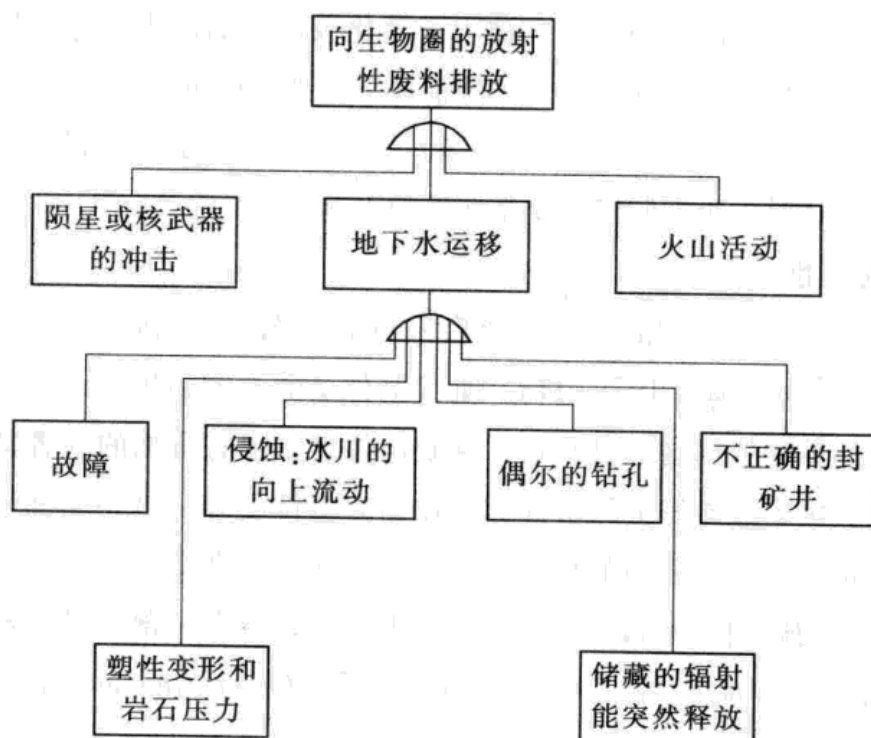


图 33—1 该错误树表示掩埋在盐积层下的核废料会意外泄露放射性物质的各种途径。位于底部两行的每种可能的初始事件都会导致地下水对放射能的传送。这种传送会逐渐导致放射性物质泄漏到生物圈。另外如图中的第二行所示，巨大的陨石、核武器或者火山爆发都能直接导致放射性物质的泄漏，而不需要地下水的帮助。

本章要探索的是危机估计过程中的一些心理因素。其中最基本的前提是公众和专家都是这个过程中必然的参与者，其估计从本质上讲是主观的，而且理解这种判断的局限性，对于进行有效的决策是非常关键的。

一、风险感知中的判断偏差

当人们普遍被要求对风险进行评估的时候，他们很少握有统计证据。大多数情况下，他们必须依据记忆中所听到或看到的关于要考察风险的情况去进行推论。心理学研究，已经识别出人们在此种情况下会采用的一些非常普遍的推论规则，这些研究中的很大一部分，在本书前面已经讲述过。这些被称为启发式的判断规则，被用于简化极困难的脑力任务。虽然它们在一些情境下很有效，但在其他多种情境下，如金融分析（Slovic, 1972c）以及自然灾害管理（Slovic et al., 1974）等，它们会导致较大的并且是持续性的偏差，这对于决策制定将产生严重的影响。

1. 便利性

与风险感知特别相关的一个启发式，被称作便利性（availability）（Tversky & Kahneman, 1973, 11）。如果某件事非常容易想象或者易于被回忆起，人们就会依据这个启发式，判断这件事是很有可能发生的或者其发生是很频繁的。因为经常发生的事件通常比稀有事件更容易被想象和回忆，便利性就经常成为一个适当的提示。然而，便利性经常受到许多与发生频率无关的因素的影响。例如，一个最近发生的灾难或者一部生动的电影，比如《大白鲨》（*Jaws*）或《中国综合症》（*The China Syndrome*），都能严重歪曲危机判断。

便利性偏差，有助于解释人们关于某种自然灾害的错误感知和错误的决断。如在讨论受洪水困扰的平原的居民时，凯茨（Kates, 1962）写道：

对人们采用先进的洪水灾害信息的能力的一个主要限制，在于他们对经验的基本依赖。生活在为洪水困扰的平原的居民，在很大程度上受制于其自身体验……最近经历过的一次洪水，似乎让管理者们认为他们自己应该为关注的损失规模设置一个上限。（p. 140）

凯茨认为在改善洪水控制中的许多困难，都应该归因于“个人不能对从未发生过的洪水形成明确的概念”（Kates, 1962, p. 92）。他观察到人们

对洪水发生可能性的预测，严重地受到刚过去的体验的调节，他们把将来视作刚结束的过去的一面镜子，从而将自己的推断局限于一些已被简化的概念。同样地，在一场地震发生之后，对地震保险的购买将会大幅度上升，而当对这场灾难的记忆逐渐淡化，其购买则会直线下降（Steinbrugge et al., 1969）。

便利性启发式的一个相当重要的意义在于，对一个发生概率较低的风险的讨论，可以增加它的易记忆性和可想象性，并由此提高它的可感知的风险性，而不管这种讨论中的证据表明的是什么。比如，在 DNA 重组研究中的领头人物，很快就后悔自己曾将重组创造的生物所带来的污染这样一个还算遥远的风险，带到了公众的注意范围内。罗森伯格（Rosenberg, 1978）将揭示这种假定的风险之后的反应概括如下：

起初，反应仅仅是对此研究中的科学家们所展现的社会责任感的某种称赞……然而，逐渐地并且是可以预料地，争论开始升温。各种推测层出不穷，整个场面越来越让人恐慌，而且得知此风险的公众范围也越来越大。许多关于此问题的讨论，完全忽略掉了这个风险首先只是假设的这样一个事实，并且假想在 DNA 重组实验室里装满了各种狂暴的动物。最终，那些因为自我约束而使整个过程进入骚动状态的科学家们，受到了严重中伤。（p. 29）

（1）对致命事件的频率判断

便利性偏差得到了一些研究的例证，在这些研究中，大学生以及妇女选举人联盟的成员，对 41 个死亡原因的发生频率进行了判断（Lichtenstein et al., 1978）。在其中一项研究中，研究者首先告诉这些人在美国的某一个死亡原因（机动车事故）所造成的每年的死亡数量（50 000），然后要求他们估计其余 40 个死亡原因的发生频率。在另一个研究中，参加者被要求判断两个死亡原因中哪一个的发生频率更高。在这两个研究中，从全局的意义上看，被试的判断都是基本准确的：人们通常知道哪一个是最常发生以及最少发生的致命事件。然而，在这种全局性的图景中，人们仍会犯严重的判断错误，其中的许多错误，显然反映了便利性的影响。

图 33—2 比较的是被试所判断的每年的死亡人数和公众健康统计资料所发布的数字。如果其频率判断是准确的，那么它们将与统计的比率相等，所有的数据点都将落在图中的线上。尽管更易发生的风险，通常会唤起更高的频率估计，但这些点显然散落在一条曲线的周围，它们时而位于代表准确判断的直线之上，时而又位于其下。大体而言，被试对那些较少有的死亡原因，进行了过高的估计，而对那些常见的死亡原因则估计不足。

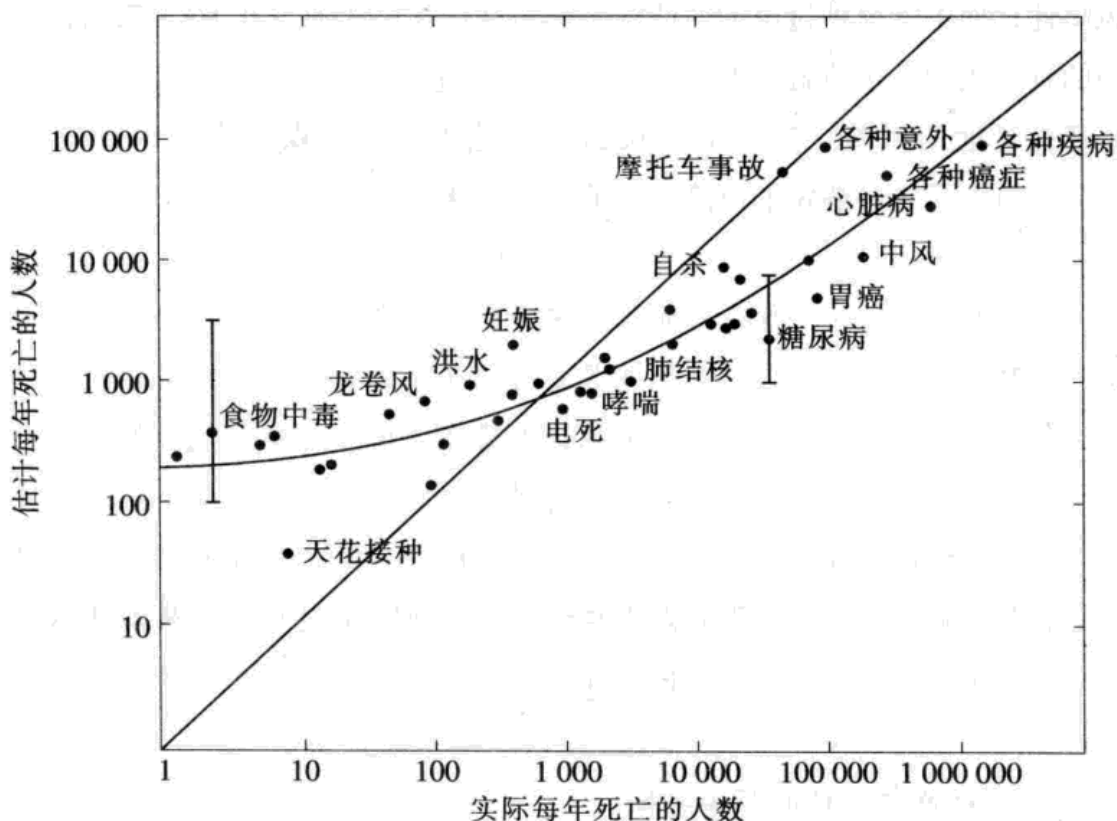


图 33—2 41 种死亡原因所造成的实际死亡人数与估计的频次之间的关系。如果估计频次与实际的值相等的话，所有的数据将会落在中间的直线上。用曲线拟合的那些点表示许多人的回答的平均值。图中的垂直线段两端分别表示 1/4 和 3/4 的受调查者对食物中毒、糖尿病以及各种意外所引发的死亡人数的估计值，两线段本身反映出受调查者之间回答的差异度。在对其他 37 种死亡原因的估计中也存在着类似的差异。

除开这些一般性的偏差，相当多特殊的偏差也很明显。比如，交通事故被认为是和疾病一样，导致了同样多的死亡，然而疾病实际上夺走了 16

倍于此的生命。谋杀被错误地判断为比糖尿病和胃癌死亡的发生更为频繁。谋杀还被认为与因中风导致的死亡一样经常发生，而后者实际上导致的是 11 倍于前者的死亡数。因食物中毒、龙卷风和妊娠（包括分娩以及流产）所造成的死亡频率，也同样被极大地过度估计。表 33—1 列出了在我们的各种研究中其频率被过低估计的一些致命事件。与我们对于便利性的考虑相一致，那些被过高估计的死亡原因，是引人注目的或者耸人听闻的，而那些被过低估计的原因，则倾向于是一些不引人注意的事件，它们只是在某个时候引发了一例命案，而通常是处于一种非致命的状态。

表 33—1 死亡频率判断中的偏差

被高估最多的	被低估最多的
各种意外事故	天花接种
摩托车事故	糖尿病
妊娠，生产和流产	胃癌
龙卷风	闪电
洪水	中风
食物中毒	肺结核
各种癌症	哮喘
火灾	肺气肿
毒虫叮咬	
自杀	

资料来源：Slovic, Fischhoff, & Lichtenstein (1979)。

(2) 偏差性的报纸报道和偏差性的判断

便利性启发式，突出了经验作为一个风险感知的决定因素的重要作用。如果某人的体验是偏差性的，那么它的感知也很可能是不准确的。不幸的是，人们所接触到的许多信息所提供的，都是关于灾难世界的一个歪曲的图景。我们来看一下作家巴赫（Richard Bach）对于一对首次乘坐飞机的夫妇所显示出的恐惧的观察：

当地面在风以及飞机引擎所引发的气流中渐渐倾斜并且在我们下面变得越来越小时，我观察到威斯康星与他的女伴，注意他们的变化。尽管他们在欢笑，他们对飞机仍充满了恐惧。他们对于飞行的知识，

仅仅来源于报纸的标题，其中多是撞机、坠毁和事故死亡的消息。他们从未读到过哪怕一篇对于一架小型飞机起飞、飞过气流并且重新安全降落的报道。他们只能够相信尽管所有这些报纸都在进行灾难报道，但安全飞行应该是可能的，因此他们依靠此种信念，拿3美元以及自己的生命作为赌注。(Bach, 1973, p. 37)

在关于上述研究的一个追踪研究中，有学者(Combs & Slovic, 1979)考察了美国海岸对面的两份报纸关于死亡原因的报道。各种关于报纸报道的指标，在一年期间的交替月份，对两份报纸进行了记录。结果显示这两份报纸在他们关于生命威胁事件的报道中都存在着类似的偏差。例如，表33—2的检查显示出，许多在统计上发生频率很高的死亡原因(如糖尿病、肺气肿、各种形式的癌症)，在研究进行的这段时间内，却很少为任一家报纸所报道。除此之外，那些暴力的、经常是灾难性的事件如龙卷风、火灾、溺死、谋杀和机动车事故等，比那些不怎么引人注目但具有同样的(或者更高的)统计频率的死亡原因，得到更多的报道。比如，疾病造成的死亡人数是交通事故的16倍，但是关于交通事故的报道文章，却是关于疾病的文章的3倍，记录了几近7倍于后者的死亡人数。在那些最常发生的事件中，与实际的频率相比，谋杀是被最大量地加以报道的一类。尽管疾病夺走了100倍于谋杀的生命，但关于谋杀的报道文章，却是关于疾病死亡报道的3倍。而且，关于谋杀的文章的篇幅，几乎要比关于疾病和交通事故死亡的文章长2倍。

表 33—2 Eugene, Oregon, Register Guard 和 New Bedford, Massachusetts, Standard Times 两地关于 41 种死因的报纸报道数以及统计上的频次。

死因	统计上每 2.05×10^8 名美国居民的 死亡比率	被试的 估计	报道的死亡		报道的事件		文章	
			R-G	S-T	R-G	S-T	R-G	S-T
1. 天花	0	57	0	0	0	0	0	0
2. 维生素中毒	1	102	0	0	0	0	0	0
3. 食物中毒	2	183	0	0	0	0	0	0
4. 麻疹	5	168	0	0	0	0	0	0
5. 烟火	6	160	0	0	0	0	0	0

续前表

死因	统计每 2.05×10^8 名美国居民的 死亡比率	被试的 估计	报道的死亡		报道的事件		文章	
			R-G	S-T	R-G	S-T	R-G	S-T
6. 天花接种	8	23	0	0	0	0	0	0
7. 剧烈咳嗽	15	93	0	0	0	0	0	0
8. 小儿麻痹症	17	97	0	0	0	0	0	0
9. 毒虫叮咬	48	350	0	0	0	0	0	0
10. 龙卷风	90	564	36	25	10	6	14	7
11. 闪电	107	91	1	0	1	0	1	0
12. 无毒动物	129	174	4	2	4	2	4	2
13. 洪水	205	736	4	10	2	2	2	2
14. 严寒	334	314	0	0	0	0	0	0
15. 梅毒	410	492	0	0	0	0	0	0
16. 妊娠, 生产 和流产	451	1 344	0	0	0	0	0	0
17. 传染性肝炎	677	545	0	0	0	0	0	0
18. 阑尾炎	902	605	0	0	0	0	0	0
19. 电击	1 025	766	5	0	5	0	6	0
20. 撞车	1 517	689	0	1	0	1	0	1
21. 哮喘	1 886	506	1	0	1	0	1	0
22. 枪击	2 255	1 345	8	1	8	1	9	1
23. 中毒	2 563	1 013	3	3	1	1	1	1
24. 肺结核	3 690	658	0	0	0	0	0	0
25. 火灾	7 380	3 336	94	46	33	9	38	10
26. 溺死	7 380	1 684	47	60	44	24	45	37
27. 白血病	14 555	2 496	1	0	1	0	1	0
28. 摔死	17 425	2 675	15	7	15	6	16	9
29. 谋杀	18 860	5 582	278	208	167	122	329	199
30. 肺气肿	21 730	2 848	1	0	1	0	1	0
31. 自杀	24 600	4 679	29	19	28	18	36	20
32. 乳腺癌	31 160	2 964	0	0	0	0	0	0
33. 糖尿病	38 950	1 476	0	1	0	1	0	1
34. 摩托车事故	55 350	41 161	298	83	245	69	180	73
35. 肺癌	75 850	9 764	3	2	3	2	4	2
36. 胃癌	95 120	3 283	0	1	0	1	0	1
37. 各种意外	112 750	88 879	715	596	421	152	374	177

续前表

死因	统计每 2.05×10^8 名美国居民的 死亡比率	被试的 估计	报道的死亡		报道的事件		文章	
			R-G	S-T	R-G	S-T	R-G	S-T
38. 中风	209 100	7 109	12	4	12	4	13	4
39. 各种癌症	328 000	45 609	25	12	25	12	26	15
40. 心脏病	738 000	23 599	49	30	45	25	46	25
41. 各种疾病	1 740 450	88 838	111	87	100	76	104	78
报道总数 (第 10、11、13、29、31、37 和 41 个 死亡原因)			1 174	945	729	376	860	483
相关系数 (R-G vs S-T)			$r = 0.97$		$r = 0.94$		$r = 0.98$	

注: R-G = *Register Guard*; S-T = *Standard Times*。

资料来源: Combs & Slovic (1979)。

而且,在报纸报道中的偏差与人们判断中的偏差非常类似。人们所判断的死亡频率与报纸所报道的死亡数量的相关性达到了 0.70。而这种高相关,并不是由于判断的以及报道的死亡与统计频率具有共同的联系。如果我们将统计频率处理为一个常量,对于这两份报纸而言,人们的判断与报道的死亡数之间的偏相关分别为 0.89 和 0.85。虽然我们试图从这种相关中推导出媒体的报道会使人们的危机感知偏差化,但实际情况是人们关于什么是重要的观点也会影响到媒体的报道。新闻业著作中充满了大量的关于两个方向上的影响的实例 (Brucker, 1973)。

(3) 它不会发生在我身上

考虑到人们既不是关于所考察危机的专家,也没能接触到有代表性的关于风险信息样本,人们对于死亡原因的判断,应该是有如所期待的那样好。对令人误解的信息样本的正确感知,应该被看成是另一个明显的判断偏差的基础,这个偏差即是人们有将自己视作是个人对灾难具有免疫力的偏好。大多数的个人,都相信他们自己会比那些普通的司机技术更好 (Näätänen & Summala, 1975; Svenson, 1981), 比一般人更可能活过 80 岁 (Weinstein, 1980), 而且与一般人相比更不可能被自己所使用的产品伤害到 (Rethans, 1979), 如此等等。虽然这样的一种感知显然并不现实,但从每个个体的体验这个角度来看,风险就显得非常之小。比如汽车驾驶,

尽管一个蹩脚的司机开得很快，车距很小，等等，但他跌跌撞撞地前进却没有遇到什么灾祸。这种个人的经验证明给他们的，便是他们超人的技能以及安全感。而且，个人的经历经由新闻媒介展示给他们的是，当事故发生时，事故是发生在其他人的身上。由于这种令人误解的经验，人们在拒绝采取如系安全带这样的保护措施时，便可能会觉得非常合理（Slovic, Fishhoff & Lichtenstein, 1978）。

（4）眼不见，心不烦

在许多情境下，觉察不到“便利性”信息的局限性，会使人们受到误导，进入一种自我满足的状态。在由费希荷夫等人（Fischhoff et al., 1978）所做的一项研究中，三组大学生被试，被要求对一个显示汽车发动时的风险的错误树的完整性进行评估。其中一个组看到的是完整的错误树。另外两个组的被试看到的则是被不同地修剪过的错误树。一种形式是去掉了启动、点火和恶作剧损坏这几个枝节；另一种形式则是除去了对电池、燃料以及其他引擎问题的详细说明。

实验中对于此任务的说明如下（括号中的数字，是提供给那些只看到修剪过的错误树的被试）：

每一天，在全美国有数百万的司机在进行这样的操作，即钻进一辆汽车，把钥匙插进点火开关里面，然后试图去发动引擎。很多时候，引擎无法发动起来，于是整个行程就被延误了。我们希望你们思考一下与此相关的各种问题，它们严重到足以导致一辆汽车不能被发动起来，从而使驾驶者的行程至少被延误 1 分钟。

导致汽车引擎无法发动的有 6 个（3 个）主要缺陷。这些主要的原因类别很可能无法包含所有的可能情况，因为我们加进了第 7 个（第 4 个）类别，它包括所有其他的问题。

请回答下面的问题：

请估计，在每 100 次因“发动失败”引起的行程延误中，由这 7 种（4 种）因素各自引发的有多少次。将你的估计写在下面所命名的各种因素旁边的空白线上。你的估计必须加起来总和为 100。

472

如果那些只看到修剪后的错误树的被试能够正确地觉察到哪些被遗漏了的话，他们归因于“其他”类的问题所占的比例，就应该等同于那些看到整个错误树的被试归因于被修剪掉的原因类和“其他”类的问题所占有的比例之和。表 33—3 的结果显示，那些被试没有看到的原因，完全处于被试的考虑之外。比如，在浏览修剪过的错误树的第一组中，“其他”项本应该上升到原来的 6 倍（从 0.078 到 0.468），以反映那些在图中被删掉的发动、点火问题以及恶作剧损坏所导致的失败所占的比例。然而，“其他”项仅仅是翻了一倍，而被提到过的整个错误树体系的重要性都被大大提升。另一个研究不仅复制了上述的这样一个结果，而且显示出相对于那些看到整个错误树的被试而言，观察修剪后的错误树的被试，认为发动失败（出于所有的原因）更不可能出现。

表 33—3 看到未经删减以及有所删减的错误树的被试对发动失败的归因

组	n	几种发动失败原因所占比例						
		电池	发动系统	燃料系统	点火系统	引擎	损坏	其他
未经删减的错误树	93	0.264	0.195	0.193	0.144	0.076	0.051	0.078
有所删减的错误树 1	29	0.432	—	0.309	—	0.116	—	0.140 ^a
有所删减的错误树 2	26	—	0.357	—	0.343	—	0.073	0.227 ^b

注：表中的横线表示该种原因被删掉了。

a 应为 0.468；

b 应为 0.611。

资料来源：Fischhoff, Slovic and Lichtenstein (1978)。

2. 过分的信心 (overconfidence)

(1) 确定自己所知

启发式的一个相当有害的方面在于，人们对以它为基础的判断持有高度信心。在另一个死亡原因研究的后继研究中，研究者要求被试指出自己在两个致命事件中的哪一个发生更为频繁的选择中，正确的比率是多少 (Fischhoff, Slovic & Lichtenstein, 1977)。表 33—4 显示的是，对应于每一个最常用的比率类型，正确回答所占的百分比。在实验 1 中，当被试给出的比率为 1:1、1.5:1、2:1 和 3:1 的时候，他们的回答得到了相当好的校正。也就是说他们的正确回答所占的百分比，接近于与这些比率相应的正确百分比。然而，当比率从 3:1 上升到 100:1 的时候，正确率仅有少量

的提高或者几乎不变。在给定比率为 100 : 1 时，只有 73% 的回答是正确的（而不是 99.1%）。在比率为 1 000 : 1 的时候，正确率“跳跃”至 81%，并在比率为 10 000 : 1 的时候进一步上升至 87%。而在给定比率为 1 000 000 : 1 或更高的时候，正确率为 90%；那么其适当的信心度应该是 9 : 1 的比率。有 12% 的回答没有列入表 33—4，是因为它们处于最普通的几个比率类型之间，显示出的是同样的过度的信心形式。总的来讲，即使是在最高的比率水平上，被试也经常会出现错误。而且，他们给出了许多非常极端的比率回答。他们半数以上的判断超过了 50 : 1。且几乎有 1/4 的回答高于 100 : 1。

表 33—4 一些主要比率类型中正确回答的比例

比率	相应的 正确率 ^a (%)	致使事件						常识问题		
		实验 1 ^b			实验 2 ^b			实验 3 ^b		
		N	% N	正确率	N	% N	正确率	N	% N	正确率
1 : 1	50	644	9	53	339	8	54	861	19	53
1.5 : 1	60	68	1	57	108	2.5	59	210	5	56
2 : 1	67	575	8	64	434	10	65	455	1	63
3 : 1	75	189	2	71	252	6	65	157	3.5	76
5 : 1	83	250	4	70	322	8	71	194	4	76
10 : 1	91	1 167	17	66	390	9	76	376	8	74
20 : 1	95	126	2	72	163	4	81	66	1.5	85
50 : 1	98	258	4	68	227	5	74	69	1.5	83
100 : 1	99	1 180	17	73	319	8	87	376	8	80
1 000 : 1	99.9	862	13	81	219	5	84	334	7	88
10 000 : 1	100	459	7	87	138	3	92	263	6	89
100 000 : 1	100	163	2	85	23	0.5	96	134	3	92
1 000 000 : 1	100	157	2	90	47	1	96	360	8	94
总数		6 098	88		2 981	70		3 855	75	
总的正确率			71.0			72.5			73.1	

注：%N 表示落入每个主要比率类型中的比率判断的比例。有 66 个被试参加实验 1，40 个被试参加实验 2，42 个被试参加实验 3。

a 表示适用于那些已受到很好校正的被试

b 实验 1，2，3 在最初的报告中是表示为实验 2，3，4

资料来源：Fischhoff, Slovic and Lichtenstein (1977)。

第二个实验，力图通过给被试更多的指导以改进他们在实验中的表现。这个实验过程以一个 20 分钟的讲座开始，该讲座对概率和比率这两个概率

进行了细致的解释。也讨论了将某人不确定的感觉用数字比率判断的形式表达出来的精妙之处，而且着重强调了当一个人对正确答案非常不确定的时候如何使用小比率（介于1:1和2:1之间）。研究者还给出一个图表，提供了各种比率与相应的概率之间的联系。最后，他们还给被试讲解了校准（calibration）的概念（参看本书第22章），并要求他们用一种会导致他们被较好地得以校准的方式进行比率判断。虽然由此被试的表现有某种程度的改进，但他们仍然表现出了一种毫无根据的确定性（见表33—4）。他们对大约1/3的选项都给出了等于或者大于50:1的比率。与这些比率相关的回答中，只有83%是正确的。

在第三个实验中，被试们在回答一些常识性的问题〔如：下面哪一本杂志在1970年的发行量最大？（a）《花花公子》（b）《时代周刊》〕时，仍然表现出了与之前在回答有关致命事件问题时的同样的过度的信心（见表33—4）。费希荷夫等人在此研究后作了一个附加的研究，试图通过询问被试是否愿意打下述的这个赌，投注于他们之前的比率估计，来检测被试对于自己估计的信任程度。

“琐碎问题快速决断游戏”介绍

现在实验已经结束了。你们过一会儿就能够得到2.50美元的回报。但是在你们拿到这笔钱并且离开之前，我希望你们能考虑一下是否愿意和我们玩一个游戏，它有可能增加你们的报酬。这个游戏的规则是这样的：

（1）请看你们的答案表。找出你们估计你们回答正确的比率为50:1或更高的那些问题。然后将这样的问题的数目写在这里。

（2）我会告诉你们那些你估计正确比率为50:1或更高的问题的答案。我们来数一下这些问题中，你有多少个是回答错误的。由于在那样一个如此高的确定率下出错是非常让人惊讶的，因此我们将这些错误的回答称为“你的惊奇”（your surprises）。

（3）我面前有一袋筹码扑克。袋子里面有100个白色的筹码和2个红色的筹码。如果我伸手进去，随机抽取一个筹码，那么我将选到

一个白色筹码的比率是 100 : 2 或者 50 : 1，此比率跟你回答正确的比率是一样的。

(4) 每一次你给出一个比率为 50 : 1 或更高的回答时，我将从袋子里面抽取一个筹码（如果你想的话，也可以来替我抽取）。在我再一次抽取之前，我会将这个筹码放回去，因此比率是不会改变的。我抽到筹码为红色的概率为 $1/51$ 。由于抽到红色几乎是不可能的，所以每次我抽到红色时也可以被称为“我的惊奇”（my surprise）。

(5) 在你每次因自己在正确比率为 50 : 1 或更高的问题前出错而感到惊讶之余，请付给我 1 美元（在某些情况下上涨为 2.50 美元）。而在我抽到红色筹码时，也会付给你 1 美元。

(6) 如果你的回答被良好校准，那么这个游戏对于你是非常有利的。因为平均而言，在我 51 次的抽取中，就会有一次输掉 1 美元给你。但是由于你的比率有时候是高于 50 : 1 的，因此你有望输的机会更小。

(7) 你愿意玩这个游戏吗？

打这个赌，对那些其回答被较好地校准、对自己的回答信心度略有不足的被试，是很有利的，而对于那些信心过度的被试而言则非常不利。在我们的研究中，大部分被试都愿意加入这个游戏。由于他们的信心毫不合理，因此他们在游戏中损失了许多钱（我们在实验结束后退还他们了）。

尽管此种毫无根据的确定性的心理学基础非常复杂，但显然其中的一个关键因素在于，人们没有意识到自己的知识往往是建立在一个通常相当脆弱的假设之上。比如，在实验 1 中，30% 的被试对于谋杀比自杀更常发生这个错误的断言给出了比 50 : 1 更高的比率。谋杀事件通常更易于被回忆，被试有可能是受此误导，而没有意识到易记忆性在这样的推论中并不是一个有效的推论基础。 475

(2) 高度精确 (hyperprecision)

过分的信心，也会以其他方式表现出来。一个典型的对错误率这样的不确定量进行估计的任务，实际上就是设置一个较高和较低的界限，从而使正确值有 98% 的机会落入其间。在由不同人群参加的实验中，被试们

作了许多不同的判断，这些判断显示，并不是只有 2% 的真实值落在 98% 的确定域之外，20%~50% 的判断值都未进入确定域（本书第 21~22 章）。因此人们所认为的自己估计这种值的准确度，要高于他们实际展现出来的。特沃斯基和卡尼曼（1974，1），把这种高度精确性归功于对初始化和调适启发式（anchoring and adjustment heuristic）的依赖。

过度自信的专家。不幸的是，专家们一旦被迫放弃他们的数据，而依赖自己的判断，就会跟普通人一样倾向于过度自信。费希荷夫等人在 1978 年对专业的汽车机械师重复了前述的错误树研究，这些机械师平均都有 15 年的修理经验。结果发现这些专家几乎跟普通人一样感觉不到错误树中的被删除部分（见表 33—5）。有人（Hynes & Vanmarcke, 1976）让 7 名“国际知名”的地球技术工程师预测能够引起一个土质地基倒塌的堤坝的高度，并围绕这个估计确定一个足以有 50% 的机会将正确的倒塌高度包含其中的置信范围。结果这些人所设定的置信范围中，没有一个实际上包含了正确的倒塌高度。图 33—3 显示的是他们的估计结果。

表 33—5 专家对汽车无法发动错误树中的遗漏的不敏感性

组	n	几类发动失败原因所占比例						
		电池	发动系统	燃料系统	点火系统	引擎	损伤	其他
未经删减的错误树普通被试	93	0.264	0.195	0.193	0.144	0.076	0.051	0.078
未经删减的错误树专家	13	0.410	0.108	0.096	0.248	0.051	0.025	0.060
经过删减的错误树专家	16	0.483	—	0.229	—	0.073	—	0.215 ^a

a 应该是 0.441。

资料来源：Fischhoff, Slovic, & Lichtenstein（1978）。

一个耗资数百万美元的由美国核能管理委员会在 1975 年进行的反应堆安全研究，是为了评定在一个核反应堆中其活性区溶解的概率。它们采用了在本书第 21 章和第 22 章中所讲的设定置信范围的方式，以营造出高度
476 的过分的信心。实际上，“刘易斯委员会”在总结他们对反应堆安全研究的审查时提到，尽管这个研究取得了很大的进展，但“我们相信误差范围
477 被低估了。只是不知道被低估了多少。产生这种结果的原因包括数据库不
充分、统计处理太差以及在整个计算过程中对于偏差的不一致的宣传”
（美国核能管理委员会，1978，p. vi）。

在许多其他的技术风险估计中，将会发现过分的信心的另一些具体证

据 (Fischhoff, 1977a)。下面列出的是一些致使专家忽略或错误判断灾难来源的基本原因。

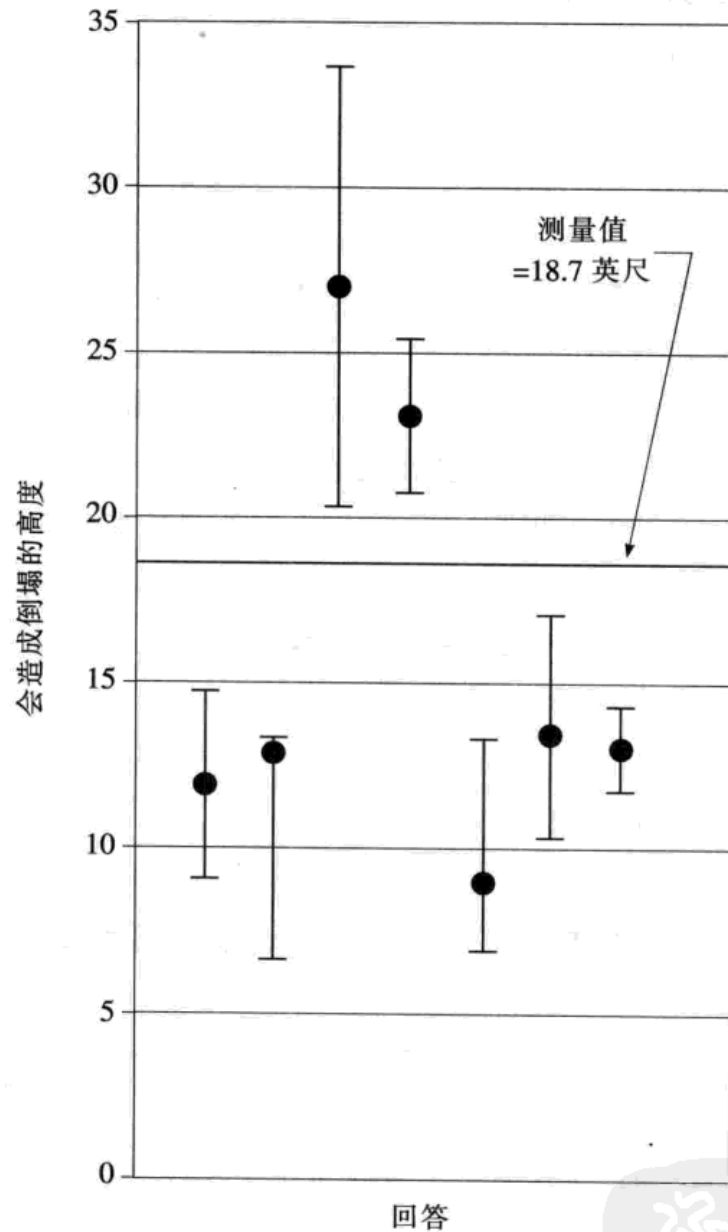


图 33—3 图中的竖直线段表示专家给出的置信范围，但它们无一包含正确值，这无疑是专家们在判断中过度自信的一个例证。图中的数据表示的是 7 位国际知名的地球技术工程师对堤坝倒塌的高度的估计。

未考虑人为失误影响技术体系的方式

例子：在三英里岛 (Three Mile Island) 上的研究者们，由于训练

不足以及在控制间设计上的失误，以致多次错误判断反应堆出现的问题，并采取了不恰当的措施（Sheridan, 1980；总统调查团, 1979）。

对当前的科学知识有过分的信心

例子：DDT 的使用，在科学家们考虑其产生副作用的可能性之前，就已经非常广泛并且无法控制了，而正是它的这些副作用，使得它在今天看来成为一个混乱的、不可逆转的恩赐（blessing）（Dunlap, 1978）。

未认识技术系统如何作为一个整体在运作

例子：DC—10 在早期的几次飞行中都失败了，因为它的设计者没有意识到货舱的减压会破坏重要的控制系统（Hohenemser, 1975）。

对慢性的、累积性的效应反应迟钝

例子：尽管很久以前人们就已经认识到煤矿事故是运营矿物燃料厂的成本之一，但酸雨对生态系统的影响，却是很晚才被发现的。

未预料到人们对于安全措施的反应

例子：水库和防洪堤所提供的不完全的保护，使人们产生一种虚幻的安全感，并促进了受洪水威胁的平原地区的发展。因此，尽管洪水减少了，但每一次洪水所造成的破坏增大了很多，使得平均每年所损失的美元数目大大超过堤坝修筑以前的水平（Burton et al., 1978）。

未预计到“共态破坏”（common mode）能同时损坏本应独立存在的系统

例子：位于阿拉巴马布朗斯渡口（Browns Ferry）的反应堆的多重安全系统，是由一些电缆控制的，而它们没有的空间上分开，因此在一场大火中，五个紧急活性区冷却系统全被破坏掉了（美国众议院, 1975；Jennergren & Keeney, in press）。

1976 年发生的泰顿（Teton）水库倒塌事件，提供了另一个有关专家的过分信心的悲惨例证。政府执行委员会将这场惨案的发生，归因于工程师们毫无根据的自信，因为他们自己完全确信已经解决了在修建过程中出现的许多严重问题（政府执行委员会, 1976）。甚至当水库首次蓄水时就出现

了大约 1/300 的塌陷时，他们也没有计算新水库的倒塌概率。

478

二、提醒人们风险的存在

清楚地预测风险是非常困难的，但不幸的是，这又非常必要。辐射风险、药物副作用、职业病、食品污染、有毒的化学制剂以及机械故障等风险，日益充盈于我们的报纸和我们日常的思索中。由于对这些风险的处理对于个人以及整个社会的福祉都非常重要，因此目前人们坚持自己在决策制定过程中扮演一种积极角色的权利。结果，风险性较高的企业的发起人和管理者，面临着越来越大的压力，得去告知人们可能面对的风险（见图 33—4）。例如，在最近几年：

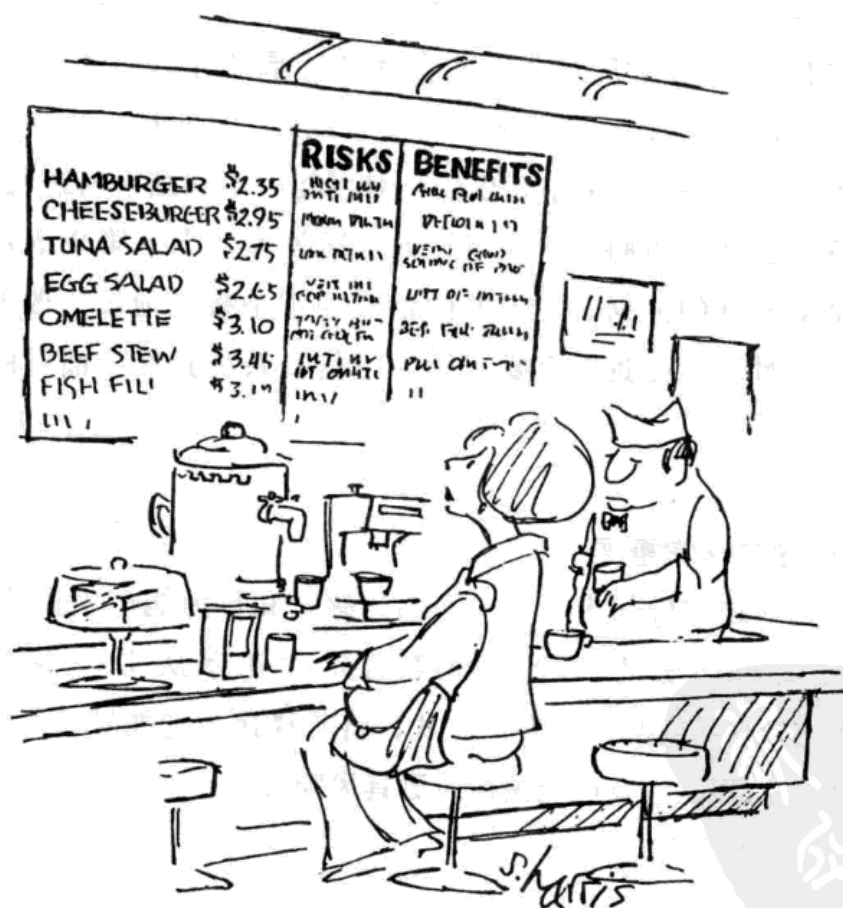


图 33—4 S. Harris 绘，引自《纽约客》杂志。1979

食品及药物管理局，向患者颁布了关于数量日益增长的处方药品的信息。

住房及城市发展局，开始要求出售建于1950年前的房屋的售房者向买者介绍房屋中含铅涂料的使用情况。

被提议的联邦产品质量责任法，越来越注重让消费者及工人充分了解他们可能遇到的风险。

白宫要求卫生、教育和社会福利部部长制定一项关于辐射接触对健康的影响的公众信息系统。

尽管这些意图是非常好的，但要建立一个有效的信息系统应该是相当困难的。因为那意味着要找到一些令人信服的解释复杂的并且是技术性的资料的方式，这些资料通常为不确定性所遮蔽。在完成这项工作的过程中，不仅所分配到的时间经常是非常有限的，而且提供给人们的信息必然要对抗接受者对所讨论的风险及其后果的先入之见（也许是误解）。比如，在某些情况下，误导性的个人经验，有可能引出一种虚幻的安全感，而在另外一些情形中，仅仅是对可能的不利结果所进行的讨论，也会提升它们表面的威胁性。而且，如罗斯等人（本书第9章）已经证明的，在面对对立的证据时，人们的信念通常是缓慢地发生改变，显示出非凡的持久性。下面我们简要回顾一下信息系统必然会面对的其他一些挑战。

1. 表达形式非常重要

风险得以表述的确切方式，对人们的感知和行为能产生显著的影响。

479 比如，某种活动能够将一个人每年的死亡几率从 $1/10\ 000$ 提高到 $1.3/10\ 000$ ，如果我们换一种表达方式，将其描述成它能导致每年的死亡风险增长30%，那么它就可能被视为更具风险性。下面介绍一个说明字面表达形式的影响的例子。

(1) 错误树

错误树的设计者，肯定要作许多随意的决定，以判断如何组织并且表现那些不同的故障原因。一个与此显然无多大差别的决定，是我们应提供多少细节描述。费希荷夫等人（1978）发现对于不同的详细程度，人们所

获得的感知是相同的。仅仅提到错误树的一个分支，能让人们正确地估计若它被详细地描述时，会显得十分麻烦。但是合并分支（比如将发动系统和点火系统合并到一个范围更广的类别里面），或者分裂分支〔如将点火系统分成点火系统（螺旋管故障和火花塞失效）和配电系统（配电器和配线故障）〕就会产生差异。一组给定的问题，若被表达为两部分而不是以一个整体来表述，那么其被误解的几率会上升 30% 以上。

480

（2）坐椅安全带

表述形式的重要性的第二个例子，来自于一个关于汽车坐椅安全带使用态度的研究（Slovic et al., 1978）。依据前面的研究，遭遇不幸的概率对于激发保护措施采用，具有非常重要的作用（Slovic et al., 1977），因此斯洛维奇等人认为人们不愿主动地系安全带，是由于他们在个别的一次开车路途中遇到致命性事故的概率相当小。由于每 350 万人在路途中，才会遇到一例致命性事故，每 10 万人中，才会有一人遭遇致残性损伤，因此拒绝使用安全带，应该被视作是非常合理的。然而，如果我们从多次旅行的角度来考虑在这许多次旅行中出现一次事故的实际概率，就会发现不系安全带是不太明智的。在 50 余年的驾车生涯中（大约有 40 000 次旅行），遇到致命性事故的几率上升至 0.01，而受到至少一次致残性损伤的概率为 0.33。斯洛维奇等人在一次飞行员研究中，要求一部分被试从一生的角度来考虑是否使用安全带，另一部分被试则从一次次单独旅行的角度来判断，结果发现前者对安全带（以及气囊）的使用更持欢迎态度。不过，由前述的长时期角度出发所导出的对于安全带使用的赞同态度，是否会得以保持并且转换成实际行为，还需要作进一步研究。

（3）伪确定性（pseudocertainty）

按照“预期理论”（prospect theory）（Kahneman & Tversky, 1979b），仅仅是具有可能性的结果与那些确定能实现的结果相比，其影响力是不够的。因此，能将受伤害的概率从大约 0.01 减小到 0 的保护措施，会被给予更高的评价，尽管另一项保护措施能将遭受同样伤害的概率从 0.02 降低到 0.01。

特沃斯基和卡尼曼（1981）认为，保护措施在人们头脑中的印象很容

易被调整，使得其能够预防风险的外在确定性也随之发生改变。例如，一项适用于火灾但并不适用于洪水的保险政策，既能被描述成对于特定的风险即火灾具有完全的防护作用，也能被解释为可以减小财产损失的全部可能性。期望理论认为，用前一个角度（标示为“伪确定性”）来看，这项政策会更具吸引力，因为它提供了一个对于有限的几个问题的绝对保护。

481 我们用一个特别的保护措施——接种疫苗来检验这个推测。我们设计了两种“种痘问卷”。卷1（可能的保护）先告诉被试，某疾病可能会危害到20%的人口，如果接种某种疫苗，那么其接受者中的一半就会对此疾病产生免疫，然后询问被试是否会自愿接种这种疫苗。而在卷2（伪确定性）中，研究者告诉被试，这种疾病有两个类别，它们之间互相排斥并且发生概率相等，其中每一种都会使10%的人口受感染；我们的疫苗能完全抵御其中的一种疾病，但对另外一种却没有丝毫作用。这项研究中的被试，都是由一个发布在俄勒冈大学的学生报纸上的广告招募而来的。其中一半的被试拿到的是卷1，另一半填写的是卷2。在阅读问卷的说明之后，他们要对自己在问卷所述情境下接种疫苗的可能性进行评定，其评定的等级从1（几乎一定不会接种该疫苗）上升到7（几乎一定会接种疫苗）。

尽管这两份问卷都暗示着该疫苗能够将某人总的患病风险从20%降低到10%，我们还是预计相对于阅读卷1（可能的保护）的被试，回答卷2（伪确定性）的被试会更希望接种该疫苗。其实验结果证实了我们这种推测：回答问卷2的被试中，有57%的人指出他们会接种疫苗，而回答问卷1的被试中，只有40%的人作此回答。

伪确定性效应，突出了减小风险和消除风险之间的比较。如特沃斯基和卡尼曼已经指出的，这种区分，在任何规范性的基础上，都很难证明其合理。而且，对确定性进行操作化，似乎对于其他保护形式（如医疗、保险、洪水和地震防护措施）的设计和描述，具有重要的意义。

（4）初始化（anchoring）

最常见的一种表述假象，是我们的判断倾向于停留在最初展示给我们的信息上（Poulton, 1968; Tversky & Kahneman, 1974, 1）。在图33—2所介绍的实验后，利希坦斯坦等人（1978）对原有条件进行了一定修改，他

们让第二组被试估计在美国 40 种不同原因中的每一种所导致的死亡频率。不过，他们事先告诉被试，不再是每年有大约 50 000 人死于机动车事故，而是告诉他们每年有大约 1 000 人被电死。尽管这两个数据都是准确的，但因后者提供的是一个较小的数字，而使被试减少了对于最大频率的估计。这种对最初数字的初始化，使得这两组被试的估计，在某些项目上相差了差不多 5 倍。

费希荷夫等人（1980）在实验中让被试判断几种潜在的死亡原因的致命性时，对四组被试采用的是形式不同但实际上等价的询问方式（例如，相对于每一个垂死的受害人而言，有多少人幸存下来了？在每 100 000 个受害人中，有多少人最后死亡了？）。表 33—6 用一种普通形式展示了这些被

表 33—6 用不同的回答模式，不同的数学方法对死亡可能性的估计

疾病	100 000 受灾祸者的死亡率				
	估计的死亡率	估计的死亡数	估计的存活数	估计的存活数	实际的死亡率
流行性感胃	393	6	26	511	1
腮腺炎	44	0.114	19	4	12
哮喘	155	12	14	599	33
性病	91	63	8	111	50
高血压	535	89	17	538	76
支气管炎	162	19	43	2 111	85
怀孕	67	24	13	787	250
糖尿病	487	101	52	5 666	800
肺结核	852	1 783	188	8 520	1 535
车祸	6 195	3 272	31	6 813	2 500
中风	11 011	4 648	181	24 758	11 765
心脏病发作	13 011	3 666	131	27 477	16 250
癌症	10 889	10 475	160	21 749	37 500

注：4 个实验组按照下列指令：（a）估计死亡率：在 1 000 个受灾祸者中，有多少人死亡？

（b）估计死亡数：X 人受灾祸，有多少人死亡？

（c）估计存活率：在受灾祸者中，存活下来的和死亡的人数的比率？

（d）估计存活数：Y 人死了，受灾祸但是没死亡的有多少人？

对问题（b）、（c）、（d）的回答被转换成每 100 000 人中的死亡数以利于比较。

资料来源：Fischhoff & Mac Gregor, 1980。

试所作的判断，显示出了提问时的措词对于被试表现出的风险感知的更强的影响力。比如，当人们直接估计流感所导致的死亡率时（第一列），他们的平均答案是每 100 000 例中会有 393 人死亡。若告诉他们平均一年里有 80 000 000 人感染流感，让他们估计其中死亡的人数（第二列），被试的平均回答是有 4 800 人会死亡，若用死亡率来转换，这仅仅表示每 100 000 例中有 6 人死亡。在问题中作的小小的一点改变，就使被试估计的比率改变了 60 多倍。同样的差异，也出现在其他的问题和其他的风险种类中。

（5）其他的效应

在有关冒险行为的文献中，记载了大量其他的格式效应（format effect）。例如，我们已经发现人们在和同伴一起时对赌博的评价明显不同于他们单独所作的判断（Grether & Plott, 1979; Lichtenstein & Slovic, 1971; 1973）。费希荷夫等人的研究（Fischhoff et al., 1980; Hershey & Schoemaker, 1980; Schoemaker & Kunreuther, 1979）都指出，如果将是否购买保险这个问题重新表述为在进行一场赌博和接受少量钱财损失之间进行选择，那么被试的决定通常是与先前相反的。同样的一种风险，若以幸存生命数的方式来描述，而不是以其造成的生命损失来界定，那么人们对它的评价将是完全不同的（Tversky & Kahneman, 1981）。一些其他的表述格式及内容的影响效应，可以在大量研究中看到（Fischhoff et al., 1980; Kahneman & Tversky, 1979b; Slovic et al., in press-b; Tversky & Kahneman, 1981）。

在表述风险的方式上的细微差别能严重影响到该风险如何被感知，而这意味着那些负责建立信息系统的人有相当大的调控人们的感知的能力。而且，由于这些效应现在还没有被广泛了解，因此人们在作如何组织他们的认识这样随意的决定时，就会毫无觉察地调整他们自己的感知。

2. 风险的交叉比较具有误导性

最普遍采用的一种深化人们的认识的方式，是将对许多种风险所进行的定量风险估计全部展示出来。在分析这些数据时所获得的精确性，或许对于个人的以及社会的决策制定非常有用。威尔森（Wilson, 1979）观察到：“我们都力图从数量上测量我们所面对的风险……因此我们可以对各种风险进行比较并且判断哪些可以接受或者哪些必须戒绝。”（p. 43）另一位

学者 (Lord Rothschild, 1979) 补充说：“在你还没有将威胁到你的风险与那些现在没有但可能会影响你的风险进行比较之前，是没有必要陷入对生命风险的恐慌的。”

有意思的是，这些告诫得到了人们的重视，各种风险被精心制作成表格甚至是“风险目录”，在其中列出了一系列生命风险致死或致残的不同指标。而且，索比 (Sowby) 在 1965 年提供了大量关于每一小时的外出的风险性的数据，比如，它显示摩托车一小时的骑行与一个 75 岁老人再存活一小时同样风险。威尔森在 1979 年设计了一张活动表格（如乘坐喷气机飞行 1 000 英里，在煤矿工作 3 小时），其中的每一项被预计能使一个人每年的死亡几率上升百万分之一。威尔森宣称“这些比较能帮助我估计风险，而且我猜测它们也能帮助其他人完成这项工作。不过这种比较最重要的作用，在于帮助整个国家制定决策以改善我们的健康状况，并减少我们的事故发生率” (p. 45)。类似地，科恩等人在 1979 年，按照所预计的各种风险对预期寿命的减少数目对这些风险进行了排序，并设想“这些顺序（此表中的顺序）应近似社会对它们的考虑次序。然而，我们可以看到一些非常重要的问题，没有受到足够的重视……而一些在此表中接近最末端的风险类别，尤其是那些涉及辐射的风险，却受到了过多的关注” (p. 720)。

正确地讲，对风险进行比较并不是一个决策制定过程。比如，在摩托车骑行和高龄的风险比较中，并不需要我们进行任何特别的推论 (Fischhoff et al., 1981)。而且，即使这种风险的交叉比较能作为对人的直觉的一种辅助，它还是具有许多内在的局限性。比如，虽然一些人在获悉仅仅是一架商务客机的起飞或者降落就能使某人的预期寿命平均减少 15 分钟后觉得豁然开朗了，但另一些人却觉得自己被这些信息完全弄糊涂了。在飞机降落的时候，某人或许会过早地死亡（几乎是确定无疑地要提早 15 分钟以上），又或许不会。对许多人来讲，这种平均值并不足以表现这些风险的特性。实际上，有学者 (McNeil et al., 1978) 发现，面临肺癌手术未知结果的病人，对手术过程中即将发生的死亡可能性和他对预期寿命的影响，都同样担心。

另一个局限在于，这种概括性的统计表，会掩盖关于各种风险的许多

重要特征。若因这种概括性而导致对于事实的不确定或不一致，那么对点估计 (point estimates) 的介绍，就会引发过度的信心。由于人们非常关心灾难性事故发生的可能性 (Slovic et al., 1980)，因此有必要提供一些关于最大伤害产生可能性及其数量的指标。还有一些在统计简表中被忽略，但仍会影响到人们对于风险的态度特征，其中包括自愿性，可控性，熟悉性，后果产生的即时性，对于后代的威胁性，减少风险的容易性，以及对那些遭遇风险的人给予救济的公平性 (Slovic et al.)。虽然一些失误如遗漏不确定范围 (uncertainty bands)，很容易得以纠正，但确定如何衡量灾难潜在性、公平性和其他一些重要的特征，仍需要作进一步的研究。

3. 小结

无论是采用警戒标、包装上的插图或是大量的媒体节目，给人们提供关于风险的信息，对于帮助他们应对现代生活中的风险及不确定性这一更大问题而言，仅仅是很小的一部分。我们认为我们的学校应该承担一部分责任。公立学校的课程，应该包含一些经过专门设计的内容，以告诉人们他们所生活的世界是具有或然性的，而不是宿命论的，并且帮助他们学会一些应对这个世界的判断技巧和决断策略 (Beyth-Marom & Dekel, in press)。这些策略，对于在一个充满不确定性信息的环境中生存，是非常必要的，就如同几何学和三角学在处理物质实体世界中一样重要。

三、核能：风险感知的一个个案研究

485 核能曾引起了极大的争论，或许再没有其他任何领域能让人们感知到如此显著的风险性或者如此高的风险性。在本节中，我们将依据刚介绍过的研究来考察这些争论。

1. 一般问题

即使是在三英里岛的事故发生之前，核工业也一直处于一种不利的公众舆论之中。大规模的而且是持久性的反对运动，曾导致了美国新核电站的审批和修建延期，损失惨重，并且在欧洲一些国家，此运动还引起了政治上的动乱。

三英里岛的反应堆事故，又激发了反核能浪潮可预见的、迅速的升温。

任何对核能在国家能源前景中的作用的设想，都必须把这些反对派的意见纳入考虑之列，并预见其将来的变化。近年来研究中的一条线索显示，在反对使用核能的公众心目中所形成的对于潜在的核事故的印象，完全不同于由许多技术专家所提出的估计。我们将描绘出这些印象，并且推测它的来源、持久性以及其中暗含的意义。

2. 基本的感知

对那些反对核能的人所作的问卷调查显示，他们认为核能所带来的益处是非常少的，但它的风险性则是难以接受地高 (Fischhoff et al., 1978)。关于核能的益处，这些人并不认为它是满足基本能源需求的一个重要环节 (Pokorny, 1977)；相反，许多人把它视作是对其他能源的补充，而这些能源目前自身就很充足（或者只要稍节约就足够了）。而在核能的风险性这一方面，与其他科技活动相比，核能引发了更强的恐惧感 (Fischhoff et al., 1978)。一些人将人们的这种反应归因于他们对于辐射所产生的无形的并且是无法消除的污染的恐惧，因为这种污染将可能导致癌症和遗传缺陷。然而，虽然用作医疗诊断的 X 射线，也是一种能引起同样危害的放射技术，但人们对它却不那么恐惧，相反，还常常低估射线的风险性 (Slovic et al., 1979)。造成这种感知差异的原因在于，人们经常将核能与核武器联系在一起。由于核能发展于这样一个暴力性的起源，就总是被视作一个有着不可控的、致命性的和潜在毁灭性风险的技术，而这些特征，却与诊断用的 X 射线沾不上边。

那些反对使用核能的人，在描述他们脑海中对核事故及其后果的印象时，显示出他们有一个基本的预测，即在他们的一生中很可能会遇到一次严重的核反应堆事故，它会造成成千上万的甚至是上百万的人死亡 (Slovic et al., 1979; Slovic et al., 1979)。而且，这样一次事故，还可能会引起一大片地域的无法挽回的环境破坏。这些预测与核工业对此的看法俨然不同，核能行业认为，即使在一次严重的事故中许多危害性事件也是不太可能发生的，而且就算发生了，它们拥有的若干安全系统也会限制其造成的破坏。

上述二者之间的“感知差异”所产生的必然后果，是大众的怀疑和不信任态度，他们认为核能的危害要远远大于专家的估计 (Kasper, 1979;

Starr & Whipple, 1980)。而专家们又反过来质疑公众的理性，并谴责他们的“感情主义”阻碍着技术的进步。结果导致两者之间敌意的，有时候还是非常激烈的对抗。

对这种感知差异的认识，使得一些技术专家呼吁必须教导公众了解核能“真正的”风险性。一个大众观点分析 (Pokorny, 1977) 是这样来解释这个问题的：

阻碍人们对核能问题作出成熟、正确的判断的最大障碍，在于他们对相关的知识和事实缺乏基本了解。在目前这种态度背景中，恐怖故事、疑惑和不理性经常会占上风。只有通过核能知识和事实的耐心讲解，才能使人们明白正确的选择是…… (p. 12)

我的个人观点是这种试图缩小感知差异的尝试，面临着一些基本的障碍。这些障碍主要来自关于此问题的两个关键点，一个是技术上的，另一个是心理上的。

3. 技术障碍

实际上，从技术上讲，几乎没有关于严重核反应堆事故发生可能性的任何“铁定证据”。这项技术是如此的新颖，其事故可能性也是那样微乎其微，以至于根本不可能在经验观察的基础上进行准确的风险估计。相反，这些估计，只能来源于复杂的数学模型以及主观的判断。

进行风险估计的困难，已使许多评论家开始质疑它的有效性 (Bryan, 1974; Fischhoff, 1977a; Primack, 1975)。一个主要的质疑，是认为那些引发事故的重要事件或途径有可能被遗漏掉，从而使风险估计偏低。在反应堆设计的安全性评估中，所存在的另一个问题是，我们很难对“共态破坏”作正确估计，那些设计来支持另一个体系的表面上独立的系统，却可能因为同一个未预料到的共态原因而崩溃。核评论家荷德伦 (John Holdren) 对估计罕有的核灾祸的有效性提出怀疑，他的这个质疑，比较简洁地概括了我们所面临的技术问题：

在每一个反应堆使用年中，在从万分之一到百万分之一的范围内 487
估计其事故可能性是否具有切实的可能性？围绕这个问题专家群体出现了分歧。我是站在怀疑派一边的，认为基本上不可能有令人信服的理论证据来支持如此小的数字估计。我持这种观点的原因非常简单：核能系统是如此复杂，在其安全分析中纳入严重错误的几率非常大，致使那些计算出来的微小的事故发生概率变得毫无意义。（Holdren, 1976, p. 21）

4. 心理障碍

公众对于核能的恐惧，并不能被视作毫不理性的。他们也了解核能的问题仍处于争论中，而且一些专家在过去有的做法是错误的，比如他们用射线直接照射肿胀的扁桃体或者允许人们近距离地观看原子弹爆炸（A-bomb）实验。人们有可能质疑的，是人们思考问题的基本方式（比如依赖于便利性启发式）而导致他们的认识发生一定程度的曲解。核能风险肯定是便利性偏差的一个主要表现，因为大量的媒体都在对此进行报道，而且它还与核战争所带来的清晰的、可想见的风险联系在一起。

如前面所提到的，便利性启发式，意味着对核事故的任何讨论都将增加它们的可想象性，从而放大人们感知到的风险性。一位工程师指出，辐射有可能意外泄漏的各种途径，几乎都是不可能的，因此将核工业垃圾倾倒在盐床上很安全（见图 33—1）。但这种说法并没有让听众放心，反而可能导致他们产生这样的想法：“我才不相信真是那样呢，许多事情都可能出现意外。”由此，对易记忆性和易想象性的依赖，往往会模糊将来可能的事情和现在可能的事情之间的区分。如一个核能支持者所感慨的，“当常人谈论可能发生的事情时，他们有时候甚至懒于提‘可能’两个字”（B. L. Cohen, 1974, p. 36）。另一个分析，也详细阐释了存在于对“最差”方案的误解中的类似问题：

由于在一些分析中人们只能指出一些不良后果可能以略大于 0 的概率水平发生，因此说这些分析中的假设是异乎寻常的或者不可能的

通常没有任何差别。反对已处于计划中的某基地的修建的人，只需要发挥他们的想象力，虚构出一系列能导致某些不良结果的情况就可以将这个计划取消掉，即使他们自己也承认那些情况是非常不可能出现的。在这样一种普遍盛行的氛围中，设计一特定的核基地就变得……充满风险……（J. J. Cohen, 1972, p. 55）

5. 结论

488 虽然上述的讨论，指出了支持核能使用和反对核能使用的人之间产生感知差异的一些可能原因，但它并没有明确地指出究竟其中哪一方对核能各种危害的评价是最准确的。易记忆性和易想象性效应，不仅能强化大众的恐惧心理，也同样能导致专家不易觉察到系统可能崩溃的方式。由于永远无法精确地知道实际的风险，新的信息也常常以一种与我们先前的信念相符的方式被解释，这种感知差异将继续伴随我们很长时间。因此，对一些入而言，三英里岛“证实”了发生灾难性的原子反应核熔化的可能性，而对另一些人来讲，它支持了他们对于多种安全体系和防泄露系统的可靠性的信任。

四、究竟谁来决定？

本章所介绍的研究，告诉我们人们在对风险作判断时很容易出错。它还显示出其出错的程度通常很高，但人们却以极大的信心坚持这种错误估计。由于即便是那些见闻广博的普通人也很难准确地判断风险，因此有人试探性地提出公众应该远离社会的风险估计和决策制定过程。但这种观点，显然是在几个方面被误导了。首先，一些比较详尽的研究发现人们对许多东西具有很好的感知力，尽管他们的观察角度通常与技术专家迥异。在误解发生得比较多的情境中，人们的错误通常都根源于他们偏差化的经验，而这可以由教育来弥补。一些情况下，人们强烈的恐惧感，以及他们对于专家解释的抵制，是由于他们敏感于潜在的灾难性事故，并了解到专家不赞同关于此类事故的发生可能性和危害强度，但却因此在过去犯下了许多严重的错误。即使是在一些比较麻烦的问题上，如关于核能问题的争论等，

若能有一种信任的气氛，并意识到无论是专家还是普通外行人都能有所贡献的话，将会促进信息的交流，并进而深化我们的认识。

而且，在许多情况下（即使不是所有的情况都如此），有效的风险处理都需要很多普通人的合作。这些人必须同意不使用某些东西，并且接受其他东西的替代品；必须明智地投票选择合适的投票方式，支持那些将适合成为风险管理代理人的议员；还必须遵守安全规则并合理地利用法律体系。即便那些专家确实比普通人能更好地判断风险，赋予他们独有的进行风险管理的特权，将意味着用短期的功效代替了建立一个见识广博的公民群体所必需的长期努力。

我们所讨论到的这些研究结果，对非专家人士形成了一系列的挑战：要进一步扩展见识，不要信赖那些未经调查或未得到证实的判断，了解可能会使风险判断发生偏差的各种因素，并对新的信息保持更大的接纳度；简而言之，就是要认识到自己可以进一步接受教导的潜力。

而对于专家和政策制定者而言，这些发现则提出了一个更难对付的挑战：要认识到并承认自己的认知局限，要尝试着培养公众而不是仅靠宣传，要意识到公众焦虑的合理性，并采用一些可以让这些焦虑在社会决策过程中得以表达，但又不会进一步激化的措施。 489

这是一篇论文的修改稿，它原载 R. Schwing 和 W. A. Albers, Jr. 主编的《社会风险评估：怎样才足够安全？》（*Societal Risk Assessment: How Safe is Safe Enough?* New York: Plenum Press, 1980）允许重印。

本研究为国家科学基金会的“技术评估和风险分析”项目的子项目。该项目的资助号码是 PRA79-11934，承担者为克拉克大学，而此分项目的承担者为感知功效学公司（Perceptronics, Inc.）。

第十部分
PART TEN

附录



第 34 章

论统计直觉的研究

卡尼曼和特沃斯基 (Daniel Kahneman and Amos Tversky)

493 近年来,许多有关判断和归纳性推理的文献,都把注意力集中在与各种脑力任务有关的错误、偏差和谬误之上(如 Einhorn & Hogarth, 1981; Hammond et al., 1980; Nisbett & Ross, 1980; Shweder, 1980; Slavic et al., 1977; Tversky & Kahneman, 1974, 1)。这种对错误研究的强调,是判断研究的显著特征,但绝不仅限于此领域:如我们会通过错觉去理解正常感知所采用的原则,通过研究遗忘来了解人的记忆。不过,推理错误在两个显著方面与其他认知错误明显不同:一是它会稍稍令人感到困惑,二是它们看起来似乎是可以避免的。我们通常不会因为我们容易感受到垂直线—水平线(vertical-horizontal)的错觉或者不能记住多于8位的一串数字而觉得不安。与之相反,推理失误却常常让我们烦躁——这有可能是因为在我们事后进行反省时,会发现先前没有找到的答案是如此明显,也可能是由于我们所犯的错误在我们知道它是

个失误后，仍对我们保持着相当大的吸引力。目前许多关于判断的研究，都集中关注具有其中一个或另一个特征的问题。

对判断错误的揭示，是通过将人们的回答和某个既定事实（如那两条线的长度是相等的）进行比较，或者与一公认的数学、逻辑或统计规则相比照。但是，并不是每一个看起来与一确定事实或公认规则相抵触的回答，都是一个判断错误。这种抵触，也有可能是来源于被试对问题的错误理解，或者源于研究者对被试回答的误解。因此，在将某特定回答描述成是一个判断错误时，就纳入了一些关于实验者和被试之间的交流的假定（我们将在稍后讨论这个问题）。研究判断的学者，必须避免对被试的回答作过于狭窄的解释，这会导致将合理的回答视为错误，同样也必须避免过于宽泛的解释，因为这会将所有的回答都合理化。 494

虽然判断错误仅仅是一种研究某些认知过程的方法，但这种方法已经成为研究内容的一个重要部分。当越来越多的证据显示那些有才智的人也会违背基本的逻辑或统计规则时，对于所谓的有关判断和决策制定的理性模型的合理性的怀疑也就与日俱增。在第二次世界大战后的 20 年里，一些对实际行为的描述性研究，主要依据的是标准模型：主观期望的效用理论（subjective expected utility theory）用于风险选择分析，信念改变研究中采用贝叶斯运算，信号鉴别理论则用在心理生理学的任务研究中。对这些情境以及相对较少的一些实验结果所作的理论分析，都认为人非常能干，几乎是最佳的决策者。在这样一个大的背景下，观察到人们对于基本的逻辑和统计推理的违背，就非常令人惊讶；而这种惊讶，促生了另一种被某些作者批评为不合理的消极的对人的智商的看法（L. J. Cohen, 1979, 1981; Edwards, 1975; Einhorn & Hogarth, 1981）。

有三个原因致使人们在推理研究中把注意力集中在系统性错误和推论性偏差上。其一，这些错误使我们智力上的许多局限得以暴露，并提供了一些有助于改善我们的思维能力的方法。其二，错误和偏差，通常能揭示我们的心理过程和启发过程，而这些过程是主导我们的判断和推论的。其三，这些错误和谬误，通过指出哪些统计或逻辑规则是非直觉性的或者说反直觉性的，而帮助我们绘制人类的直觉之图。

直觉和直觉的这两个词，常用的有三种意义。一是如果有一个判断是通过不规范的和非系统的推理模式得出来的，也就是没有采用分析的方法或审慎的计算的话，它就会被称为是直觉的判断。比如，许多心理学家按照一种直觉的程序决定其样本的规模，但却采用分析程序去验证其研究结果的统计显著性。二是如果一个正规的规则或者一个天然的事实与我们日常的关于世界的模型一致的话，就会被称为是直觉的。因此，在一次彩票抽奖活动中，赢得奖金的概率会伴随着彩票数目的减少而减小，在直觉上是显然的，但是若认为有超出一般的机会在 23 人中找到两个生日相同的人，则是反直觉的。三是当我们在日常行为中采用一些规则或者程序时，这些规则和程序就被称为我们的直觉库中的一部分。例如，语法规则，对于一个以该语言为母语的说话者而言，就是其直觉的一部分；而平面几何的一部分（尽管不是全部）规则，也被结合到了我们的空间推理之中。

本章所要关注的，是试图描绘人们对于机会和不确定性的直觉时所产生的方法上的一些方法论上和概念上的问题。我们先讨论一些对统计直觉的不同测试；然后我们转向一个对判断研究中采用问卷一回答范式的批评；接下来我们讨论一下一些统计规则中的非直觉特征，并以此结束本章。

一、统计直觉测试

不确定状况下的判断中出现的错误和偏差，是我们描绘人们的统计直觉界限的主要资料来源。在这种情况下，在应用错误和理解错误之间作一个区分是有好处的。若有证据显示人们知道并且承认某个规则，但并没有采用它，那么这个特定问题中的失误，就被称作是应用错误。而如果人们并没有意识到他们所违背的规则合理性，那么这种情况下犯的错误，就被称为是理解错误。

对于应用错误，一个最具有说服力的表现，就是当一个人本能地或者因受到一丝触动而抱住自己的头，然后大声惊呼：“我怎么能忘记这个呢？”这个时候，他就是犯了应用错误。虽然许多读者都记得这样的经历，但仅仅依靠这种情绪表达作为证明是不够的，我们还需要设计其他的程序，来证明人们对那些自己违背了的规则是理解的。

对于某个规则的理解，能用两种方法来检测，一是启发被试，二是让他们要么认可关于某个普遍规则的陈述，要么支持某个赞成或反对一特定结论的论点。将这些方法结合使用就有了四种程序，我们现在将对此进行讨论并举例说明。

我们先来看一个不太正式的例子，其中人们对于某规则的理解是由他们接受或者支持某观点来确证的。我们中的一人给许多壁球选手提出了下面这个问题。

你们都知道，一次壁球比赛可以采取 9 分制，也可以采取 15 分制。假定比赛的其他规则保持不变，如果选手 A 要比选手 B 更优秀，那么哪一种计分制可以让 A 有更大的取胜机会？

虽然我们所有的被调查者都具有一定的统计知识，但他们中的大部分人，都认为采用这两种计分方式并不会造成任何差别。于是我们让他们考虑一下另一个观点，即认为越优秀的运动员越喜欢较长时间的比赛，因为相对于小样本而言，大样本中通常不太可能出现比较反常的结果。之后，除开少数几个例外，被试们马上接受了这个观点并承认他们先前的回答是一个错误。显然，被调查者对于样本规模对抽样误差的影响是有一些了解的，只是他们没能把壁球比赛的进行时间长短转换为样本规模的一种。一旦这种联系得以建立，正确的结论就变得非常令人信服，这说明先前的回答是一个应用错误，而不是理解错误。

一个更系统的判断错误类型的尝试，来自于对一种被称作结合或联结效应（conjunction effect）（见本书第 6 章）的现象的研究。概率理论中最基本的法则也许就是结合规则了，它认为某种结合（A & B）发生的概率，不可能大于 A 或 B 单独发生的概率。但是，就像下面这个例子所显示的，在我们所设计的许多测试中，大多数被试，即使是非常精通统计的那些人，都认为事件共同发生的可能性要大于其中任一单独事件。

为了导出结合效应，我们让被试看一段对下面这类人格的描述：

琳达今年 31 岁，单身，非常坦率，也非常活泼。她主修哲学。作为一个学生，她对歧视和社会公正问题相当关注，并且参加了反核游行。

我们所提问题的一个版本，是让被试判断下面两种对琳达的陈述中哪一个更具可能性：(A) 琳达是一个银行出纳员；(B) 琳达是一个在女权主义运动中相当活跃的银行出纳员。在我们从一群对统计一无所知的大学生中抽取的大样本中，有 86% 的人认为第二个陈述更可能是真实的。在另一个从心理学研究生中抽取的样本中，只有 50% 的人犯这个错误。然而，当这两个判断陈述被置于 8 个关于琳达的比较陈述序列中时，缺乏统计知识的被试和精于此道的被试之间的上述这种差别就消失了。两组被试中都有 80% 以上的人显示出了结合效应。在另一个被试间 (between-subject) 实验设计中，对两个判断陈述作的是间接比较，但还是得出了类似的结果 (见本书第 6 章)。

我们采用了规则支持 (rule-endorsement) 和论点支持 (argument-endorsement) 程序，来考察人们是否理解并接受结合规则。首先，我们让一群无统计知识的大学生看几个类似规则的陈述，并要求他们判断其正确或者错误。“事件 X 的概率总是大于事件 X 和 Y 的概率”这一判断得到了 81% 的被试的赞同。与之相比，仅有 6% 的被试支持“如果事件 A 比事件 B 更有可能发生，那么它们不可能同时发生。”这一陈述。这个结果显示被试对于结合规则还是有一定理解的，虽然并不是每个人都支持它，但那有可能是由于其抽象的和不太熟悉的表达方式造成的。

我们也采用了一个论点支持程序，即先让被试看一段对琳达的描述以及 (A) 和 (B) 两个陈述，然后要求他们回答下面哪一个观点是正确的：

(1) A 要比 B 更具可能性，因为琳达既是银行出纳员又是一个活跃的女权主义者的几率，必然要小于她是一个银行出纳员的概率。

(2) B 比 A 更可能，因为与她仅仅是一个银行出纳员相比，琳达更像是一个在女权主义运动中非常活跃的银行出纳员。

支持结合规则的观点 (1) 得到了 83% 的心理学研究生的赞同, 但只有 43% 的无统计知识的大学生同意其说法。之后与被试们的广泛讨论也证实了这一结果。精通统计的被试立即认识到了结合原则的合理性。而缺乏统计知识的被试, 则没有受到那些非常正规的论点的太大影响, 而且许多人一直坚持他们最初的回答, 尽管那与结合原则不相符合。

让我们非常惊奇的是, 那些缺乏统计知识的被试, 对于结合原则并没有一个可靠的把握; 当它以抽象的方式表达的时候, 他们会支持它, 但若它与一个非常鲜明的对于某代表事物的印象相冲突的时候, 他们就不再认同它。相反, 受过统计训练的被试, 能意识到这个规则的合理性, 并且能够在一个相当明朗的问题中应用到它。但是, 对于同一问题的另一个不太明显的表达, 他们这帮精通统计的人也难以避免犯结合效应的错误。依据这项实验, 可看出结合效应是一个应用错误, 至少对那些受过更多统计训练的被试来讲是如此。关于此问题的进一步讨论请看第 6 章。

为了考察不同统计知识水平的被试所具有的统计直觉, 尼斯比特等 (见本书第 32 章) 采用了一个启发程序 (elicitation procedure), 即由某个简短故事里的许多特征推出一些结论和推断, 然后让被试对这些结论作出判断并证明自己的观点。研究者发现, 在对基本统计原理的理解上, 被试间存在着巨大的个体差异, 而这种差异与他们所接受的统计训练水平是有密切联系的。统计直觉无疑会受到智力、经验以及教育水平的影响。与其他知识领域一样, 对于该领域的专家而言是直觉性的事物, 但对初学者来讲就常常是非直觉性的 (参考 Larkin et al., 1980)。然而, 即使是对那些统计理论专业的学生来说, 一些统计结果 (如同生日的情况或者在掷硬币游戏中头像的变化) 仍然是反直觉的 (Feller, 1968, p. 85)。而且, 有许多证据显示, 缺乏统计知识的被试, 通常所犯的那些错误 (如赌徒谬误) 也能在精通统计知识的被试那里发现, 只是通常出现在一些较复杂的问题中 (Tversky & Kahneman, 1971, 2)。

在著名的利用四卡片问题 (Wason, 1966) 考察逻辑直觉的研究中, 也用到了启发式 (Evans & Wason, 1976; Wason & Evans, 1975)。在这个问题的标准版中, 实验者给被试出示了四张卡片, 分别写着 A、T、4 和 7,

然后让他们选择哪些卡片应该被翻转过来，以检验其是否符合“如果一张卡片的其中一面是元音字母，那么它的另一面就必是一偶数”的规则。正确的结果应该是将写有 A 和 7 的卡片翻转，因为若第一张卡片是奇数或者第二张卡片上是元音字母都会违背这个规则。大多数被试都选择翻看表面写有 A 和 4 的卡片，这显然犯了明显的逻辑推理错误。瓦森等人继续用这同一问题的不同版本进行了实验，并让被试对他们所做的是否查看每一张卡片的隐藏面的判断给出理由或者论证。研究者认为，被试在证明他们的回答时，给出的仅仅是一种合理化的论证，而不是对那些实际引导他们作决定的规则的陈述。

其他的研究 (Wason, 1969; Wason & Johnson-Laird, 1970), 公布了一些其他的有关人们未能理解证明 (verification) 规则的证据。为了给人们提供一些“药方”，这些研究者给被试出示了他们的判断的结果，并让他们注意到其回答的不一致性。但是这个举措，对于被试在其后同样的任务中的表现，并没有多大的影响。总而言之，研究结果认为，人们在完成证明任务中的困难所反映出的是他们理解上的错误，而不是应用错误。

我们已经仔细讨论过的这些例子，既采用了对规则或论点的支持程序，也利用了启发被试提出观点来证实他们的某个回答的程序。我们没有谈到另一个程序，即让被试直接讲出相关的规则，因为这样一个实验通常包含着一个人不合理的期望：我们希望相信人们能理解一些他们不能正确表达的规则。

为了确定一个应用错误，编好的程序，需要将人们对某特定问题的回答与他们对相关规则或论点的判断进行比较 (McClelland & Rohrbaugh, 1978; Slovic & Tversky, 1974)。不过也可以采用其他的研究设计，去确立一个应用错误。比如，哈密尔等 (Hamill et al., 1980) 做过一个实验，他们让被试观看了一段采访录像，其采访对象据称是一个监狱守卫。有一半的被试被告知，这个守卫的看法 (非常人道或非常残忍) 代表了所有监狱管理人员的观点，而另一半被试则被告知，这个守卫的看法不具有代表性，而且他要么比他的同事仁慈很多，要么残忍很多。之后实验者让被试估计

监狱管理人员在许多问题上的代表性意见。这个研究的结果非常让人惊讶，对于被试的普遍化推理而言，不具有代表性的守卫所表达的观点，与该群体中典型代表的观点，具有同样的影响力。在这类判断中显然出现了问题，虽然我们不可能认定任何一个判断都是错误的，而且也不可能有这么多被试会意识到他们没有受到有关守卫代表性的信息的影响（Nisbett & Wilson, 1977）。通过这个研究，以及另一个被试间设计的研究，我们有理由认为如果群体间的比较产生出了一个多数人都会认为是站不住脚的结果，那么这里面必然包含了一个应用错误。

我们已经将应用错误界定为人们的反应（回答）违背了某个合理的规则，而这个规则是人们理解并且接受的。然而，通常我们很难判断一个错误的类别，因为对理解和接受某个规则的不同检验，可能会得出不同的结果。而且，同样的一个规则，在这个问题环境下可能会遭到违背，而在另一个问题中则不会。上述证明任务的实验，就是一个鲜明的例子：那些没有正确地验证“如果一张卡片的一面是元音字母，那么另一面就必是一偶数”这一规则的被试，在证明另一个形式上相同的规则：“如果一封信是密封的，那么它上面贴的必是一张 5 美分的邮票”时，却没有丝毫的困难（见 Johnson-Laird et al., 1972; Johnson-Laird & Wason, 1977; Wason & Shapiro, 1971）。499

这些研究结果，说明了在推理研究中存在着的一种典型情况，即人们似乎并不拥有有效地证明包含“如果—陈述”（if-statement）的普遍规则，否则他们就能够解决前面的卡片问题了。不过，他们对于正确的规则也不是毫无感知，要不然他们在处理邮票问题时就不会成功。严格来讲，如果掌握某个规则就意味着人们通常会遵守它的话，认为人们并不拥有正确的直觉是正确的。然而，从另一个方面来讲，这种说法又是令人误解的，因为它可能暗示着人们的直觉缺陷，比我们实际上观察到的更为普遍。

之前的代表性研究得出的一些结论，似乎也有与之类似的困境。这些研究显示，许多成年人对于大数定理、对于贝叶斯推论中基率的作用或者回归预测的原理，并不具有普遍有效的直觉。但是事实并非如此简单：对与这些规则相关的任何问题，人们并不都会给出错误回答，或者在任何

特定背景中，这些规则都无法引起人们的关注。

让这些形式上等同的问题变得易于或难于解决的因素，似乎跟这些问题所唤起的思维模式或图式有关（Rumelhart, 1979）。例如，要看出“非Q”与“P意味着Q”的含义的相关性，采用性质—控制（quality-control）图式（他们忘记了给这封密封了的信贴上邮票吗？）显然要比采用证实（confirmation）图式（否定结论就意味着对其前提的否定吗？）更容易。由此看出实际的推理过程，似乎是受模式限制或者受内容限制的，因此对不同的问题会采用不同的推理过程或者推理规则（Hayes & Simon, 1977）。由此，人们的推理过程，仅仅用一些与内容无关的形式规则来描述，是远不够的。

还有另一种可能，即通过一系列极具直觉性的步骤很可能得出一些非常意想不到的结论，而这就让我们勾勒人们的统计或逻辑直觉的任务变得更加复杂了。苏格拉底正是通过采用这种多重直觉步骤的方法，非常成功地让他那些还很单纯的弟子相信，他们一直都懂得那些原理，而他只是让他们自己觉察到这一点而已。通过一系列的直觉步骤得出的结论，能被视为是直觉性的吗？布莱恩（Braine, 1978）在演绎推理的框架下讨论了这个问题，他建议用即时性来检验：如果某个陈述的真理性能立即受到人们的关注，而且只需一步就能证明它，那么这个陈述就是直觉性的。

苏格拉底所暗示的问题，并没有明确地在不确定性判断的框架中研究过，而且在对直觉作直接检测与精心设计出谜一般的问题来检验，以及通过苏格拉底式的引导来验证之间，并不存在什么标准来作严格区分。500 例如，我们可以想象苏格拉底会如何启发他的学生对下面的问题给出正确的回答：

“哪一家医院——不论是大医院还是小医院——将会有更多这样的天数记录，即该天所有新生婴儿的60%是男孩？”

对于斯坦福的大学生来说，这是一个很难的问题（Kahneman & Tversky, 1972b, p. 441, 3），但是通过一系列很简单的步骤，就能引导学生得

出正确的回答。这些步骤可能是下面这样的：

“你不赞同将某天在某个医院出生的婴儿视作一个样本吗？”

“非常正确。那你对从大样本或小样本中得出的结果具有同样的信心吗？”

“好的。难道你不赞成对于那种更不容易出现错误的样本，你的信心更大吗？”

“当然你一直都知道这一点。现在你能否告诉我你认为在一群婴儿中男婴所占的比例达到多少才最接近标准的现实？”

“我再次表示赞同。那么，难道这不意味着若在某天出生的婴儿有 60% 都是男孩，就是对这个标准的严重偏离？”

“那么，如果你非常信任某个样本，你难道不希望这个样本揭示的是事实而不是错误吗？”如此等等。

苏格拉底所采用的这套程序不太合理，会引导被试作出实验者所期望的回答，不过要达到上述同样的目的，还有许多更巧妙的方法可以采用。费希荷夫等人在 1979 年所作的实验显示出，当被试面临一连串的问题，且这些问题只是一些关键的变量上有所变化时，他们就会关注到基率以及证据的可靠性。虽然这些研究者即使在一个被试内（within-subjects）设计中也没有发现样本规模的影响，但其他研究（Evans & Dusoir, 1977; Bar-Hillel, 1979），通过更明晰的表达和更显著的样本结果，观察到了其影响。

一系列并行的问题所给出的暗示，会让被试对那些实际上与正确答案无关的变量给予关注：费希荷夫等人（Fischhoff & Bar-Hillel, 1980a）证明了这一点，他们发现如果某个基率信息是在一组问题中唯一有区别的变量，那么被试就会注意到这毫不相关的基率。实际上，被试倾向于相信若某数据存在着系统变化，那它的任何信息都与正确答案有关。被试内设计，在心理学研究的一些领域中，总会产生显著的解释问题（Poulton, 1975）。而在直觉研究中，它们容易引出实验者所希望检验的效果。

二、论问题——回答范式的局限

501

在前面一节里，我们指出被试内设计和苏格拉底的暗示，都可能唤起我们所要研究的直觉。实际上，问题绝非如此简单。大多数关于不确定判断和归纳推理的研究，都是采用一种对话的范式来进行的，实验者提供给被试一些信息，然后要求他们以口头或书面的方式回答问题或是进行价值评估。在这一节我们要讨论这种问题——回答范式的一些困境和局限。

研究中所采用的让一些漫不经心的被试填写简短问卷的方式经常受到批评，批评者认为这些被试如果更认真地对待实验情境，他们的回答将会完全不同。然而，有证据显示，最初通过假设性问题而形成的人们在推理和选择上的错误，在引入真实的动机之后，并不会完全消失（Grether, 1979; Grether & Plott, 1979; Lichtenstein & Slovic, 1971, 1973; Tversky & Kahneman, 1981）。在人们能够预测他们在一个更现实的环境中会如何回答，以及他们没有要撒谎掩饰自己真实回答的动机的时候，使用这种假设性问题是恰当的。这也并不是说报酬和动机不会影响人们的判断。我们仍然坚持即使存在某种回报，推理和选择错误依然不会消失。无论是日常的报纸还是对以前政治和军事决策的研究，都无法支持某些人的乐观看法，即理性总是占上风的，即便其风险很高（Janis, 1972; Janis & Mann, 1977; Jervis, 1975）。

关于问题—回答范式的一个更深的顾虑，也许在于我们不能很肯定地假定，被试获得某种信息，然后回答问题的“实验性对话”，能够使被试得出的判断与他们在和周围环境的日常互动中所进行的推断无差异。虽然日常生活中的许多判断，也是针对一些很明确的问题，但大多数问题比较隐晦。而且，对话实验，在许多方面都不同于日常的社会互动。

在解释被试的回答时，实验者试图假定：（i）这些问题只是引导被试公开表达那些在他们脑海中自然出现的思想；（ii）所有被试可以得到的信息，都包括在实验提供的资料中了。从被试的角度来看，情形完全不同。首先，实验者所问的问题，在该实验所意欲模拟的实验情境中，并不会自然发生。其次，被试通常关注许多实验者从来没有想过要问的问题，

如：“这个问题有正确答案吗？实验者希望我找到正确答案吗？真有一个可能正确的明显答案吗？问题中是否提供了一些关于实验者所期望的答案的提示？是什么让实验者决定选择我拿到的这份资料？在这份资料里面，是否有一些只是用来误导我们的无关信息，或者它们都与答案相关呢？”实验者所看到的公开的单个答案，其实在一定程度上取决于被试对这一连串隐含问题的回答。而且实验中提供的信息，只是被试在作出隐秘的以及公开的回答时所利用的信息源中的一个 (Orne, 1973)。

继格雷斯 1967 年的詹姆斯 (William James) 的讲座 (Grice, 1975) 之后，哲学、语言学和语言心理学领域的大量文献，都在关注合作原则对于话语意义的影响 (可参考 Clerk & Clark, 1977)。利用这一原则，谈话中的倾听者，有理由假定说话者总是力图让自己成为“可提供资讯的、诚实的、切题的以及表达清楚的”对话者 (Clark & Clark, 1977, p. 560)。格雷斯列出了一个合作的讲话人通常会遵循的几条准则。比如，数量准则，它限制说话人讲述那些听者已经知道或者能够轻易从上下文或其他信息中推断出来的事情。正是按照这一准则，“约翰试图打扫房子”这句陈述所传达的意思，是这个企图并没有完成：因为听者能够设想如果该企图已经实现的话，说话人会用“约翰打扫了房子”这个简单的句子来表达。

被试们是带着差不多一生的谈话经历来参加实验的。他们通常都期待遇到一个合作的实验者，虽然这种期盼常常是错误的。这种关于合作性的假设，对被试对于实验中提供的信息的理解，有着许多非常微妙的影响。特别是，它使实验者对“不相关”信息影响的研究变得相当困难。因为引入不相关信息违背了对话原则，被试们习惯于在实验信息中寻求相关性。如泰勒等 (Taylor & Crocker, 1979a) 谈到，被试对某个人的印象，会受到一些对每个人来讲都是真实的陈述的影响。比如，有一个陈述是“马克在他的教授面前很害羞。”但被试推断马克通常并不害羞，因为他们相信一个合作的实验者不会将一个完全是多余的陈述加入到人格描述中。在其他一些关于不相关或无价值信息的影响的研究中，也产生了同样的问题 (如, Kahneman & Tversky, 1973, 4; Nisbett, Zukier, & Lemley, 1981)。

乐伏特斯等人 (Loftus & Palmer, 1974) 的研究, 提供了一个对嵌入在问题中的先在假设作用的例证。这个研究显示, 某事故的目击者在被问到“这辆车撞碎另一辆车时的速度是多少?” 时, 比他们遇到“这辆车撞到另一辆车时的速度是多少?” 这个问题时, 对该车速度的估计要高。问题中撞碎这个词的使用暗示着, 提问者如果确是真诚的并且合作的话, 就会相信那辆车是跑得飞快的。

509 我们能将对这样一种推论的规范分析, 分成两个独立的问题。(i) 目击者会受到包含了关于汽车速度的个人观点的问题的影响吗?(ii) 目击者在构想一个公开的估计时会受到问题的影响吗? 如果问题中传达了新信息的话, 对于问题(i)的回答必然是肯定的。而对问题(ii)的回答则不是那么明确。一方面, 在回答某问题时, 重复问题中所包含的信息显然是不恰当的。另一方面, 一个合作的证人, 在回答有关数量的问题时, 理应给出最合理的估计。那么如果这个估计已受到了问题的影响, 这个目击者会怎么做呢? 他的回答会是“在你问我之前, 我本认为……”吗? 不管这个例子的正规价值是什么, 它至少给出证据, 显示人们通常不能将过去的观点与现在的意见分离开来, 或者不能估计到那些影响他们观点的因素的分量到底有多大 (Fischhoff, 1977b; Goethals & Rechman, 1973; Nisbett & Wilson, 1977; Ross & Lepper, 1980)。

我们对初始化的研究 (Tversky & Kahneman, 1974, 1), 为细微提示的强大作用, 提供了进一步的例证。在一个研究中, 我们要求一组被试估计土耳其的人口数超过 500 万的概率, 让另一组被试估计土耳其的人口数不足 6 500 万的概率。之后, 我们让这两组被试记下他们对土耳其的人口数的最佳估计; 被分别提供了一个高标准和一个低标准的这两组被试, 各自的平均估计是 1 700 万和 3 500 万。这些回答可以用下面这个假设来合理化, 即在概率问题中出现的数值, 不会离正确值太远。

我们已经指出过暗示效应在有的时候能被证明是合理的, 因为在暗示和信息之间并没有明确的界限。但是, 重要的是我们应指出人们并不会接受暗示, 因为这样做是不正确的。首先, 他们通常并不知道自己受到暗示的影响 (Loftus, 1979; Nisbett & Wilson, 1977)。其次, 即使在被试并不

能合理地相信提供给他们的标准能够传递信息的时候,同样的暗示效应也会发生。在那些被要求通过在一个随意选出来的值上上下下调整而得出一个数量估计的被试身上,初始化效应得到了显著的体现 (Tversky & Kahneman, 1974, 1)。并不是任何的暗示性在困惑他们,而是他们明显不能排除那些并不具有价值的信息在捣鬼。

当被试被要求在一个答案列表中选择一个是他们的答案,或者对一组限定的选择对象根据其概率作出排列时,实验者对选择类型的选择,将提供给他们一些信息。乐伏特斯在 1979 年已经发现在询问被试每周头痛次数的时候,在给出的供选数量等级是 1~5、5~10、10~15 等的情况下,被试选择的次数,将多于他们在其供选数量等级为 1~3、3~5、5~7 等时的选择。在这个例子中,数量等级,必然会影响到怎样才被称为头痛这样一个界定。即使不可能进行这种再解释,被试也一定会在他们的数量估计中,选择某个选择范围的中间值,并形成主观的概率分布,其中每一个选择类别,都被赋予一个并不算小的概率 (Olson, 1976; Parducci, 1965)。

504

费希荷夫等人在 1978 年发现,在问卷中隐含着的提示,也会对其回答产生影响。他们让一些对汽车一无所知的被试和一些熟练的汽车机修工,对能导致汽车发动失败的各种故障进行概率估计。结果发现,对“所有其他问题”这一选项的概率估计,与整个选择列表的完整性并无关系,而且在某个主要因素(如整个电气系统)从列表中删除之后,它也几乎没有什么增长。

即使是细微的和间接的线索,也会对被试的回答有所影响。在最近的一项研究中,我们给被试提供了下面的信息:“A 先生是高加索人,33 岁。他的体重为 190 磅。”然后我们让一组被试猜测 A 先生的身高,而让另一组被试在先猜他的腰围尺寸后再猜测他的身高。第一组被试的平均估计值,大大高于第二组,大约高出 1 英寸。我们推测,先猜腰围的被试相比于那些只猜测其身高的被试,会更多地将其体重归功于其腰围大小。

我们认为,在判断研究中采用对话的方式,值得作更细致的考虑,而不是像过去的研究,也包括我们自己的研究所认为的那样简单。我们不能总是假定,人们亲身观察某个事实后得出的推论,会或者应该与他们通过

被告知该事实而得出的推论相同，因为调节人们之间的交流的谈话规则，并不适用于由观察真实事实得来的信息。在问问题的时候，很难避免给出有关正确答案的（有用的或者误导性的）线索，而且也难以避免在其中传达关于所期待答案的某些信息。对证据解释的一个常见相关问题的讨论，可参阅巴-希勒尔等人（Bar-Hillel & Falk, 1980）的研究。

毫无疑问，我们已经提到过的偏差性因素，极有可能在高度不确定的情境下，具有较大的影响力。如果被试对实验者谈话态度的理解与他们关于某问题的正确答案的可靠认识相冲突的话，这种理解就不会有太大影响。然而，在进行大多数判断研究的中间地带（gray area），谈话背景的变化，将影响到推理过程以至其被试的回答。

三、判断错误：积极的和消极的分析

在判断错误的积极解释和消极解释之间做出区分，通常是很有用的。一个积极的分析，聚焦于那些导致某个特定错误回答的因素；而消极分析，
505 则试图解释为什么没能作出正确回答。比如，针对一个孩子在皮亚杰的守恒任务中的错误，积极的分析是试图识别出那些决定这个孩子的回答的因素，如相对高度或两个容器的表面积等；而对这同一问题的消极分析，则会注意使这个孩子难以获知并理解容量守恒的那些障碍。在不确定判断研究中，积极分析关注的，是人们用以进行判断、估计和预测的启发式。而消极分析，则关心人们在理解和运用基本推理规则中的困难。对于理解错误，消极分析把注意力集中在一些妨碍人们独自发现相关规则或接受有关其问题的简单解释的障碍上。对一个应用错误的消极分析，力图找出编码的困难的方式，它使人们已知并接受的一些规则的相关性得以掩盖。

对于一系列不同问题的判断，有时会违背许多常用规则，在同样的一个启发式能够对此都进行解释的情况下，对该错误进行积极分析通常是最有用的。相应地，当人们在不同问题中一直违背某个规则，但这些错误又不能只归因于某一个启发式的时候，消极分析最具有启发性。因此，如果每天对日常事件的观察已经提供了足够的机会，让人们学习那些判断规则，那么是应该问一问人们为什么会认识不到这些规则。而且也该思考一下如

果人们无法被那些简单而且正确的论据说服的话，他们为什么还要拒绝遵从规则呢。许多学者都讨论过人们在日常经验中了解统计规则的困难，比较著名的有下列的这些研究（Einhorn & Hogarth, 1978; L. R. Goldberg, 1968b; Nisbett & Ross, 1980）。这种认识上的失败，通常可归结于没能完成必需的对相关事例的编码，或者是缺乏对错误判断的正确反馈。而拒绝接纳规则的理由，通常是认为其规则具有反直觉的性质。为了举例说明这一点，我们现在来分析一下人们拒绝采用回归预测规则的原因。

对直觉预测的研究提供了大量的证据，它们证明人们在预测中存在着一种普通的倾向，即他们的预测太过极端或者回归率不足（要想回顾最近的文献请看第 15 章）。在早期的文章中，我们对这种现象进行了一种积极的分析，以作为一个例子说明代表性启发式的作用（Kahneman & Tversky, 1973, 4; 1979a, 30）。然而，我们将看到，有理由对这些现象作某种消极的分析，以提供一个更加全面的认识。

人们常把正确的规则视作非直觉的甚至是反直觉的，对于这种理解性错误，进行一种消极的分析是很有意义的。很多基础统计学的老师，经常会发现他们的学生认为回归这个概念难以理解和运用，即使日常的经验都在告诉他们那些极端的预测，大多数时候都太极端了。比如，体育比赛评

506

论员和老师会经常遇到那种向平凡状态回归的例子：异常的成功通常带来的都是失望，而失败却往往导致进步。

而且，当一个标准变量实际上对预测值呈线性回归，而且该标准变量的条件分布（对预测值的每一个固定值而言）是对称的话，回归预测的规则，就能被下面这个引人注目的论点所证明：我们有理由对所有拥有同一个预测值的个案作出完全相同的预测，也有理由选择这样的预测来保证对于所有预测值为 Y 的个案，其标准变量值的平均值和中值都等于 Y 。然而，这一规则与人们所拥有的其他一些直觉相抵触，我们将在下面讨论其中的一部分。

(1) “一个最佳的预测规则，即使不能保证，至少也应该能对所有的案例作出非常准确的预测。” 回归预测的原理，与这个表面上看来非常合理的要求是相违背的。回归预测提出的一系列预测值，比实际相应的标准值

的偏差更小，因此排除了得到相当精确的预测的可能性。实际上，回归规则允许每一组相关预测值都存在误差：我们绝不会发现一个儿子的身高能准确地从他父亲的身高推测出来，而且从他的身高也能准确地推测出他父亲的身高，除非他们两人的身高都处于整个身高分布的平均值上。如果相信一个允许出现误差的预测规则是最佳的，似乎不太合理。

(2) “观察值与在此基础上作出的预测值的关系应该是对称的。”我们似乎有理由期待如果 B 是从关于 A 的信息中推测而来的，那么在已知 B 的情况下，推测 A 也是正确的。回归预测，当然不支持这种对称性，因为这两个反方向的由一个变量推测另一个变量的预测遵从的不是同一个回归等式。在比较回归预测值与标准变量的实际值时，我们会发现相关的非对称性。在回归预测中，若对某特定值 Y 进行预测，那么所有案例的平均标准值将是 Y，从这个意义上讲，回归预测对所有的案例是公平的。然而，如果我们看一下标准值为 Y 的所有案例，会发现它们的预测值的平均值处于 Y 和组平均值之间。这种不对称性对那些聪明但缺乏统计知识的人来讲，是令人费解并且反直觉的。

就像在同一个总体中进行重复抽样或者采用同一测试的不同种对等形式的情况一样，其得到的最初观察值和标准值产生自同一过程，且无法区分先后，在这种情况下，回归预测的不对称性，显得尤其麻烦。在此种情况下，唯一满足对称性的预测模式，就是恒等法则 (identity rule)，即推测第二种形式下的值与最初的观测值是相等的。回归预测的原理，引入了一种毫无理由的区分：两个值都来自同一个总体，一旦知道了其中某个值，如何可能推测到有关这两个值的差别的迹象？

(3) “任何系统的结果必然有一个原因。”在最初的观察值和相应的标准值之间确实存在着区别，我们可以从任何一个散点图 (scatterplot) 中发现这一点。然而，它似乎是一个没有原因的结果。例如，在测验一再测验的情况下，若了解到第一次的值比较高，那么就会得出第二次的值会降低的预测，但第一次的观察值并不一定导致第二次的值低。出现这种无原因的结果，显然是违背有效的直觉的。实际上，对任何回归的例子，尽管人们偶尔会在这上面出错，都可能给出一个因果解释，这样一个事实严重阻

碍了人们对回归的理解。例如,对于那些熟练的操作,从最初的测验向下一次测验的回归,通常都归因于第一次失败后的努力奋斗,或者归因于最初的成功带来的过分的信心。人们很难认识到操作者即使对于结果一无所知,其操作也会出现这种回归,这仅仅是由于他的操作存在无法排除的不可靠性。第一次操作向第二次操作的回归,通常也是让你惊讶的,因为无法对它给出一个简单的因果解释。

我们已经勾勒出了一个对人们在理解和运用回归预测概念中的困难的消极分析。我们假定人们对于统计预测存在着很强的直觉,一些统计上正确的规则,往往被人们视作是反直觉的,仅仅因为它们与人们现有的直觉相抵触。照这种观点,人们所采用的“规则”代表着一些重要的信念,它们不仅仅是一些理性化的解释,还对人们理解正确的规则起到了很大的阻碍作用。而且,这些信念通常是相互矛盾的,因此也是无法实现的。比如,我们不可能建立起一个关于父亲和(第一个)儿子的身高的非衰退性的联合分布,以使父亲的平均身高成为其儿子的身高的无偏估计值,而且儿子的身高也可以作为父亲身高的无偏估计值。

总而言之,我们已经提出对不确定性判断中存在的一些错误和偏差应该采用双重分析:积极的分析用于解释人们依据启发式选择某个特定错误回答的行为,消极的分析则用于说明人们为什么不了解正确的规则。尽管这两种分析并不是互相抵触的,但它们倾向于强调所研究现象的不同方面。试图结合积极和消极解释的努力,应该可以加深对于归纳性推理的理论分析。

四、总结

在本章中,我们提到了在不确定性判断领域内存在的三组方法上和概念上的问题。其一,我们在应用性错误和理解性错误之间进行了区分,并讨论了研究统计直觉的不同方法。其二,我们反观了判断研究中所采用的问题—回答范式的一些局限,并探讨了隐性暗示、苏格拉底式提示以及对话规则的影响。其三,我们讨论了对判断错误进行积极和消极解释的作用。

本章中所引发的一些思考,使我们对不确定性判断的经验和理论分析

更加复杂化了；它们也提出了一些将来研究的新方向。我们希望通过加深对统计直觉研究中的概念上和方法上的问题的理解，使我们能进一步理解人们归纳推理的复杂性、它的微妙之处以及局限性。

本章原载《认知》(*Cognition*, 1982, 11: 123-141)。允许重印。



第 35 章

不确定性的变式

卡尼曼和特沃斯基 (Daniel Kahneman and Amos Tversky)

在哲学、统计学和决策理论中，对不确定性 (uncertainty) 的分析，通常都只按照单一维度的概率或信念程度来处理所有形式的不确定性。近年来，对不确定状况下的判断的心理学研究，也基本承袭了这一传统，把注意力集中在直觉判断与标准概率逻辑的对应性上 (Einhorn & Hogarth, 1981; Nisbett & Ross, 1980; Slovic, Fischhoff, & Lichtenstein, 1977; 本书第 34 章)。然而，关于不确定性的一种全面的心理学视角，揭示出许多不同的过程与经历，它们既涵盖了一些基本的机制，如单个神经元对重复性刺激的适应性，也包含了一些诸如评价科学假设之类的复杂活动。

在本章中，我们的概述，将延伸到有关不确定状况下的判断的心理分析所通常考虑的观察范围之外。我们会讨论到对不确定性的两种反应水平。首先，我们将描述一些感知中存在的预期和意外的基本

过程，它们能被视作人们进行主观概率判断的先兆。然后，我们将转向一种现象学的考察，其中我们会区分不确定性的内在和外在特征，并概述人们在评估不确定性时可能采用的4种判断模式。

一、概率的基本形式

510 不确定性，是一个所有生命形式都必须准备去面对的事实。在生物复杂性的一切层级上，都存在着对于迹象或刺激的意义或是行动的可能结果的不确定性。在所有层次上，都必须在不确定性被化解之前做出行动，并且必须在高水平的对最有可能发生事件的特定准备状态与当未预料事件发生时能做出恰当反应的基本能力之间，达到一种平衡。鉴于目前的处理都是将重点放在信念而非行动上，我们将不讨论低级生物组织根据强化的概率来分配它们的反应精力的令人惊叹的过程（Herrnsteinn, 1970）。在这一部分，我们主要关注感知的不确定性。

1. 感知的预期

在事件发生之前，我们总存有期望或预期（expectations）；而在事件发生之后，则可能产生惊奇（surprise）。对惊奇的研究，主要是采用心理生理学的方法，并通过数种指向反应的指标以及与事件相关的潜能的P300成分（component）来进行测量（Duncan-Johnson & Donchin, 1977; Dochin, Ritter, & McCallum, 1978）。而在许多语境中，则有大量的方法，被用于对期望的研究。

我们关于感知预期的讨论，将依图35—1所示的图式展开，其中区分

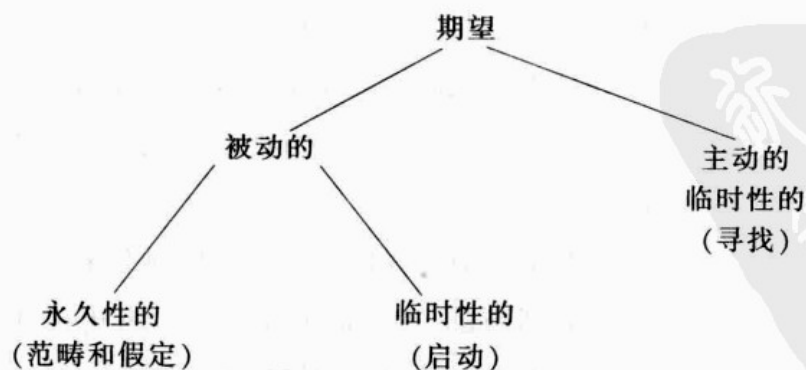


图35—1 感知的期望

了 3 种主要的期望类型。第一个主要的区分,是将主动的和被动的期望区别开来:主动的期望会占据我们的意识,并利用到少量的注意力;相比较而言,被动的期望是自动且毫不费力的,我们与其说它是一种行动,还不如说是一种品性 (disposition) (Posner, 1978)。

一些期望是相对持久的。对于特质之间共变性的长期期望,使得我们用于对经验进行组织和编码的感知范畴得以界定。对于物体的特定期望,例如,房间和窗户可能是矩形的,就作为永久性的假设来帮助我们确定对于模糊刺激的解释 (Ittelson & Kilpatrick, 1951)。经过长时间的累积,相对于另一些事情而言,我们会变得对某些事情更有准备,正如实验中单词的过去频次对于单词再认 (recognition) 阈限所具有的强大影响一样 (Broadbent, 1967; Morton, 1969)。事实上,期望有时会产生虚幻的经历,使得人们无法将它与现实区分开来,就像音位复原效应 (phonemic restoration effect) 所表现的一样。因此,有关单词 “legislature” 中的 “s” 的所有感知信息,都可以从这个单词的录音中抹掉,并代之以一声咳嗽或是某种其他的自然声音。而听此录音的被试,完全相信他们听到了虚构的音位 (Warren, 1970)。

被动的和临时性的期望,调节了背景对于再认的巨大影响 (Foss & Blank, 1980), 以及启动效应的一些变式 (Posner, 1978)。如在某试验的警示标志中引入一个字母,会加速被试在速度匹配任务中对这个字母的反应,即使这个附加信息是被如此有意地加以安排,使得这一警示标志并不传达任何关于此目标的有效信息。波斯纳 (Posner, 1978) 证明了两种期望的一些重要区别,这两种期望分别是由并不包含某种信息的警示标志所建立起来的被动期望,以及当那个标志事实上能够,尽管不完美地,预测目标时所建立起的主动期望。一个被动的期望,在被确定时会带来益处 (如更快的反应),但它不会妨碍对于尚未成型的目标的反应。相比而言,一个能引起被试对某特定目标进行主动准备的标示,也会减缓对于未预期目标的反应。用概率理论的话来讲,主动期望遵循了互补原理:对于特定事件的高度准备是以对其他事件的准备不足为代价的。而被动准备则符合一种非互补性的无成本的利益模式。

511

被动期望可能会与有意识的期望相冲突,并且有证据表明,被动过程对模糊刺激的解释有更大的影响力。有学者(Epstein & Rock, 1960)利用由左视和右视的剪影相接合从而形成的一个可置换的图形背景结构图像,检验两种类型的期望的相互竞争。合成图像的观察者,只能看到其中一个与正常轮廓相符的剪影。他们在建构了两个可加入这一形式的剪影之后,在数次试验中有规律地交替展示这些剪影,从而制造出一种有意识的期望,即任何一个必然伴随着另一个出现。然后实验者第一次将合成的图像展现给被试,并记录下他们在其中看到的头部剪影。根据启动效应(priming effect),观察者通常看到的是前面的测试中出现的剪影,而非他们有意识地期望出现的那一个。

512 在一个已发表的关于不同水平的期望之间的冲突的相关研究中,大脑电位检查(electroencephalogram, EEG)中的P300成分,是其主要的因变量。P300值是EEG中的正向偏移,它在观察者认为是显著的或与任务相关的任何刺激得以呈现之后的300兆秒内出现。许多审慎的研究都显示出,在事件的先验概率和它们所引起的P300偏差值之间,有着紧密的联系(Donchin, Ritter, & McCallum, 1978)。当一个被试面对一个伯努利(Bernoulli)序列时,与低频率事件相比,经常重复出现的事件引出的P300值更小。而且,同样事件的重复出现,往往伴随着P300值的直线下降,它表明人们主观认定的会继续重复的概率在增长。与此相对的是,根据我们熟悉的赌徒谬误,对重复的有意识的期望,长期而言是持续下降的。显然,一个观察者有可能在准备或“启动”某件事的同时,却在有意识地预期另一件事,而且当有意识地预期的事件发生时,他们会显现出惊奇的心理学迹象。因此,我们可以认为个体可能在同一时间对同一事件有着互相冲突的概率判断。这些观察,意味着人的意识似乎是一个科层结构(Dennett, 1979),其中不同的部分接纳不同的数据,赋予不同的权重,并由此对同一情境产生不同的看法。

2. 作为打赌的感知

在毕生的视觉经历中发展起来的期望,对于感知有着深远的影响,同时它们又明显地难以影响到有意识的知识或意向。对这些事实的最著名的

展示,是由有关感知的相互影响主义(transactionalist)的学者所发展出来的(Ittelson & Killpatrick, 1951; Kilpatrick, 1961)。观察著名的扭曲房间和旋转窗户的人,往往被导向一种与他们的日常知识和对该错觉中的物体的特定了解相冲突的视觉经历。这样,当他的朋友沿着扭曲房间的墙壁行走的时候,会被认为其在改变大小,或被看成巨人或被视为侏儒,而且透过旋转的窗户,一张餐巾纸或许会变成薄薄一小片。之所以会产生如此令人惊异的效果,是因为我们有一个主导的假设,即都认为房间和窗子是矩形的。尽管观察者们非常清楚这一假设并不适用于当前的情况,但这种认识对有意识的感知无法产生任何实际的影响。在以往的岁月中已经建立起的对于现实的模式,不可能应某个特殊情况之需而改变。这些观察又一次证实:对于同一事件,观察者能够同时持有相互冲突的观点。

我们曾经提到过,感知的期望决定了我们在一个模糊的刺激中能“看出”什么。事实上,相互影响主义者们曾将感知解释为与现实打赌(Kilpatrick, 1961)。这种感知选择的一个有意义的方面,在于对所选解释的高度信奉。我们的经验中,不包含任何有关刺激的模糊性的印记,即使在感知的解释会如 Necker 立方体一样随时间而波动的时候,它们也力图在任一特定时刻保持确定性。在感知中,对不确定性和模糊性的抑制表明,我们或许在生理上就被安排为要按照感知的最佳博弈来行动,就好像这场赌博没有犯错的风险似的。有意识的感知经历与思维之间的最显著差别在于,后者可以表现出怀疑或不确定性,而前者在正常情况下是不能的。

尽管对不确定性的抑制使得感知博弈与对于不确定事件的有意识判断区别开来了,但这两个层次上的对不确定性的处理,在其他一些方面是相同的。相互影响主义研究的两个引人注目的观察,揭示出似乎适用于有意识信念的一些假设。第一个观察是,重构的环境图像倾向于具有连贯性,并反映出景物或刺激的特征中所包含的自然约束和倚赖条件。因此,当一个物品被放置在其尺寸和距离都被模糊化的环境中时,被人们挑中的感知解释,将选择与按正常方式获得的视觉尺寸相关的尺寸和距离:如果感觉某物品很大,那么它同时也会看上去比那些显得比较小的物品离我们更远。

第二个观察是,感知的建构似乎是一个层级化的过程,其中关于景象

的整体特征的判断，约束并主导着对其中所包含物体的判断。扭曲的房间提供了最好的例子。被试所看见的并不是两种极端观点的中和：一个扭曲房间中有正常大小的人，或是一个正常房间中却有奇怪大小的人。后一种观点压倒了前一种观点，就如同房间的形状在有人在其中行走之前就被计算好一样。是否类似的规则也同样会在如建构未来事件的情节这样的任务中发生作用，是一个非常值得研究的问题。

二、不确定性的现象学

前面的章节力图说明控制感知期望的规则与概率规则是明显不同的。本节将会把这一分析扩展至在主观的概率判断中可能表现出的怀疑和不确定的经历。正如我们将会看到的，在自然语言中，概率的概念可表示好几种不同的意识状态，而对这些状态来讲，标准的概率积分规则，可能已经不能同样适用了。

要理解期望的复杂性，先考虑它们其中的一个表现：当期望未实现时我们会体验到惊讶。设想将一枚硬币抛 40 次，你会期望有多少次是“头像”朝上的？如果你假定硬币是均质的，你将认为更有可能出现 20-20 的结果，而不是其他，然而如果真出现这个结果，你所表现出的惊讶将会超过你在面对 22 枚正面和 18 枚背面的结果时的表现。这两件事的“正确的”
514 主观概率，是由对它们的相对可能性所作的慎重判断所决定的，还是依据它们所引起的自然的惊讶反应呢？

一个可能的解释是，这个例子展示了概率判断的两种路径之间的冲突：对最有可能的结果是 20-20 的判断缘于对机会规则的认识，但是类似 22-18 或 17-23 这样的结果在另一层面上更具可能性，这时概率取决于代表性。一个稍稍不平等的结果，体现了硬币的均质性和抛掷的随机性，而一个完全公正的结果却无法代表这些。依照这种观点，对于期望的更大的心理现实，是建立在体现于惊讶反应中的代表性基础上的。

一个稍稍不同的将注意力集中于对可能结果进行的编码上的解释是可能的。正如我们将看到的，在谈话中将事件 X 的定义延伸为“X 或与此相似的事件”通常是很恰当的。如果对事件的自发编码遵循着类似的规则，

诸如 22-18 或 17-23 这样的结果将会被自发编码为“大致平均的分配”，而 20-20 的结果将被赋予一个特殊的编码，即“完全平均的分配”。一个试图判断事件发生的相对可能性的人，将会考虑如何清楚地陈述结果，并指出 20-20 比 22-18 更有可能发生。但是否出现惊讶的反应可能取决于对事件的自然编码。22-18 的结果相对而言将不会令人觉得意外，因为它被编码为一个大致平均的结果，而且事实上它要比完全平均的结果更有可能发生。

给事件编码所扮演的角色，在对不确定主张的解释中有所体现，这种主张包括“我估计……”或有时候表达成“我认为……”不确定的主张是一类言语行动，这类行为是以特定的真实状态和对有效性的检验为特征的。比如，我们可以看一下下面这个预测：“我认为在 6 个月内黄金价格将比今天的价格上扬 50%。”从字面上看，这是一个分数估计，而且其被证实的概率非常小。然而，人们不会试图对这个预测进行字面理解。分数估计通常被理解为一种相对性的陈述，或是一个对结果可能落入其间的范围的陈述，比如，“我认为黄金价格上扬的幅度将更接近 50%，而不是 X% 或者 Y%。”我们认为演说者和倾听者通常都会认同 X 和 Y 所暗含的意义。例如，如果在接下来的 6 个月中黄金的价格实际上扬了 53%，尽管上文中提到的预言者的预言并不是绝对正确，但人们还是会认为他已经预言得非常准确了。因此，做出一个数字性预测的人，通常诉诸一个范围而非一个点。演说者也同样会赞同这样的主张，即实际值高于或者低于这个估计值都是同样可能的，除非该预测的性质使此种现象不可能发生。因此，一个说“我认为在未来 6 个月中黄金价格会上扬 50%”的人，如果同时认为实际价格更有可能高于而非低于估计值，却不表达出来，通常就会被认为是在有意误导人们。

显然，与预测相关的真实状态，并不要求一个变量的预测值（或预测区间）被视作高度可能的，而需要它比用作比较的值（或区间）的可能性更大。例如，如果一个声称“我认为比利将在下届奥运会上赢得跳高金牌”的人，宁愿在该事件不发生而不是发生上下赌注，他并不会被人们看做骗子，然而，他不许再进行“杰克获胜的可能性更大”这样的补充。因此，自然语言赋予了最佳估计一个极具特权的角色，并且对所偏爱的估计

的认同，也传达了有关与之相比较的其他值的信息。提及某个喜爱的运动员，意味着将把他与其他单个运动员相比，而并不是和可能获胜者的间断分布（disjunction）作比较。这个规则所带来的一个结果是，如果某事件的补充事件被自然编码为一个间断分布（disjunction）的话，有时候是可能对某个被视为比其补充事件更不可能发生的事件进行“预测”的。

一个相关的限制，应用在对信心的表达上。一个有关信心的陈述，表达了某人对自己已经赞同的预测、估计或推论的不确定性。因此，询问“你有多大的信心认为自己是正确的？”是很自然的，而询问“你有多大信心认为自己的推测是错误的？”就不太正常了。信心是我们对自己“认为”会发生的事件的主观概率判断或信任程度。

普通语言同样提供了大量的表达方式，以描述那些可能发生，但我们不必然“认为”它们会发生的事件。因此，人们可以评估候选者获胜的机会，估计不同活动的风险性，在足球赛上下注，并理解气象播报员对下雨可能性的预报。现在我们跟随图 35—2 所示的框架转向一个更详细的对这类陈述中可能表达出的不确定状态的分析。接下来的章节，我们将讨论图表中的两个层次，即对不确定性的归因和不确定性的变式。

1. 对不确定性的归因

图 35—2 所示的主要区分所代表的，是不确定性可以被归因的两种场所：外部世界或我们的知识水平。比如，我们会将与抛掷一枚硬币、从一副牌中抽出一串纸牌、足球比赛的结果以及海伦火山的活动状况等事件相

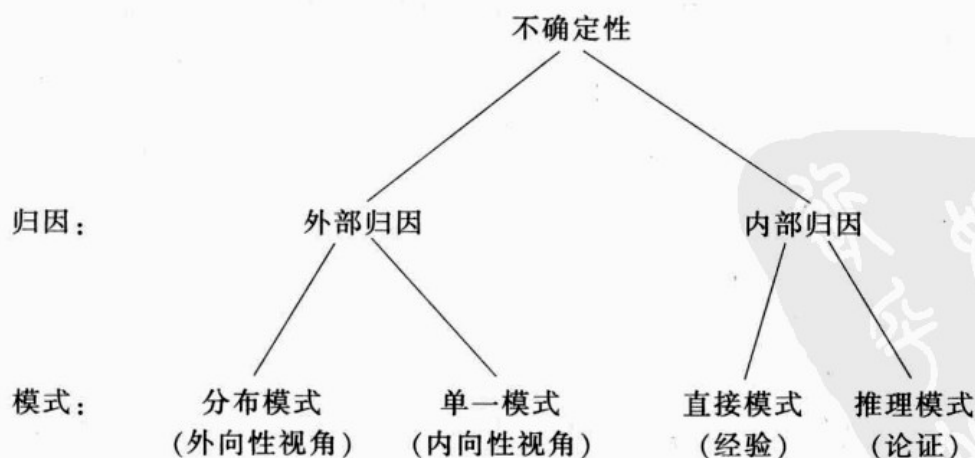


图 35—2 不确定性的变式

关的不确定性，归因到真实世界中存在着的因果关系体系。这些因果关系体系，具有制造不同事件的品性，我们通过评估相互竞争的各种品性的相对强度来判断这些事件发生的可能性。相比而言，“我认为勃朗峰（Mont Blanc）是欧洲最高的山”或者“我希望我能正确拼出她的名字”这样的陈述，反映出的是某人思想中的不确定性，而不是那座山或者那个女孩的模糊性（Howell & Burnett 在 1978 年，就已经用内在不确定性和外在不确定性这两个词来分别表示人们能够或者不能够控制的事情了）。 516

我们对个体无知与外部不确定性之间的区分，与更为广泛的有关经验的内在归因和外在归因之间的差异紧密相关。例如，我们通常将颜色、大小、质地归结为客观物体的外部属性，而将痛苦、情感和记忆归结到具有该体验的主体身上，而不是那些导出这些情绪的客体。

对不确定性的归因，有时候可以从一个简单的语言测试中推导出来：用“这个的概率是……”这样的陈述，来描述对不确定性的评估合适吗？或者我们应该说“我认为是……”贝叶斯派的观点认为，所有的概率都是主观和个人的，与之相对，自然语言却标明了不确定性的内部和外部之分。因此，我们可以说“对下一年内沙特阿拉伯王国会进行政权更替的可能性的最好评估”，但是若说“对尼罗河是世界上最长河流的可能性的最佳评估”就不合理了。对可能性的最佳评估，是属于公众领域内的，至于私人领域中的无知就不算了。

这项测试，并不总能区分这种不确定性到底是外部的还是内部的。例如，一个人可能说 Marlow 写《哈姆雷特》的可能性，尽管他说的这种可能性，更多的是来自于他的无知而不是 Marlow 有多大可能性写剧本。在这个例子中，公众知识体系的存在，可以证明使用“可能性”这种字眼是适当的，它能减少但不是排除关于《哈姆雷特》著作权的不确定性。当然并不是每个人都需要了解这种知识，但是这种估计的可能性，涉及从现有可获得的证据中可得出的合理的或共识性的推论。然而，在尼罗河这个例子中，公共的知识体系，显然包含了正确的答案，无知只能是个人的问题了。 517

在众多因素中，将某事件的不确定性归因为事物的性质还是个人的无知，

得取决于时间。对于过去事件的不确定性很可能是一种无知，尤其是当事件的真实性和人尽皆知的时候；而对未来事件的不确定性，更自然地应归因于相关体系的各种性质。事实上，人们已经注意到对于硬币抛掷结果的不同态度，取决于这枚硬币是否已经被抛掷（Rothbart & Snyder, 1970）。

2. 不确定性的变式

根据评估可能性时我们所需考虑的数据的性质，图 35—2 的第二层次，将不确定性区分为四种原型变式（prototypical variants）。外部不确定性，可以通过以下两种方式来评估：（1）分布模式。在这种模式中，需要研究的案例，可以看成是一系列相似案例中的一件，因为其中所有结果出现的相对频次都是已知的，或者是能被评估的。（2）单一模式。在这种模式中，可能性的评估只能通过手头特定事件的倾向性来评估。接下来，我们通过一个真实的故事来阐释上述两种评估可能性的模式。

一个参与高中课程发展项目的团队，所有的考虑，都面临着不确定性，目前正处于计划制定阶段。他们现在遇到的问题是，完成第一版教科书需要多长时间。参与讨论的每一个参加者，都要求尽可能现实地估计所需的时间；所得出的 7 个评估结果从 18 个月到 3 年不等。然后，团队领导者，转而问其中一个从事教育工作并在课程发展问题上有着丰富经验的参与者如下的问题：“其他那些试图编写教科书，并且在之前从未有过相关课程研究的全新领域内开设新课程的团队，有些什么经验？他们从与我们现在阶段相似的地方开始，完成一本教科书需要多长时间？”所给出的答案中令人沮丧的含义，使所有参与者，包括给出答案本身的专家都感到惊讶：“我所能想到的大多数团队失败了，并且从没有完成一本教科书。而对于那些成功的团队，完成的时间处于 5 年到 9 年这个时间范围内，中位值为 7 年。”

随后进一步的调查显示：通过构建计划和设想情节，并且为未知的突发事件留出一定的安全空间，所有的参与者都是按照单一模式提出他们最初的评估。由于初始化效应（Tversky & Kahneman, 1974, 1），在目前的计划中加入一定的安全空间，这样得出的评估往往是非常乐观的。但是上述故事中一个值得注意的方面是，决策者并没有同时利用那些彼此相关的分布信息，尽管一个专家从其个人知识中提取这些信息是完全可能的，而且

其他参与者还能非常准确地评估这些信息。

上述例子揭示出如何应用单一的和分布的推理模式对连续性变量——完成项目所需时间——进行预测。在该例子中，分布信息包括了有关不同完成时间的相对频次的知识。当然，一个类似的推理过程，也可以用来对离散型结果，例如项目失败的可能性进行评估。这个结果在相关类别中出现的相对频次，就构成了对可能性进行分布模式评估的基础，而对于某特定事件，在单一模式分析中运用到的其他信息，则会导出某种关于该事件会成功或者失败的倾向的印象。还有很多的事例，我们都可以运用单一模式或者分配模式解决同样的问题。

比较下面的例子：

(1) “如果明天早上你给约翰打电话，可能会发现他在家。他说过他喜欢在家工作。”

(2) “如果明天早上你给约翰打电话，可能会发现他在家。我给他打电话时他通常都在那里。”

陈述(1)只允许唯一的一种约翰会在家的可能性判断。而陈述(2)中既包含分布模式又包含了单一模式的估计。有关约翰早晨在家的相对频率，提供了一个明天依然会发现他在那里的可能性的自然估计，但是这个陈述同样赋予约翰一种早上在家的倾向性，如陈述(1)中所述。

我们曾推测 (Kahneman & Tversky, 1979a, 30)，人们通常更喜欢单一模式，通过这种方式，人们采用一种有关能直接导出结果的因果体系的“内向性视角”，而不是那种将当前事件同某个抽样图式联系起来的“外向性视角”。上述课程计划的例子，表明人们对单一模式的偏爱。它同样说明了另外一种我们设想为极为普遍的结果：较之单一模式，判断的分布模式，更有可能导出精确的数值估计以及合理的概率评估。

通过下面几个例子，我们来看内部不确定性的不同评估模式之间的区别：

(3) “我相信纽约在罗马北面，但是不太肯定。”

(4) “我想她的名字叫多利斯，但我不太确定。”

这些陈述所表达出的不确定性，很显然是内部的：陈述（部分地）反应了个人的无知，而不是外部事物的特征。谈到纽约在罗马北面（顺便提及，并非如此）或是将琳达记作多利斯这两种倾向，显然都是很牵强的。

519 两种陈述，在所基于的证据的本质存在差异。陈述（3）反映出一种对证据和论点反复筛选与权衡的过程（如纽约比罗马冷得多；罗马位于意大利中部，等等）。陈述（4）具有完全不同的特点。它所表达出的信心建立于一种对两事件相关强度的内省式判断之上。正如我们对单词拼写的检查经常通过判断“它是否看起来正确”这一方式来完成一样，此时的信心，依赖于一种未加分析的经验。在心理物理学和有关记忆的研究中，与判断相联系的确信度与精确度密切相关：人们在判断正确时通常会比其犯错过时更加自信，尽管他们对自己判断正确的可能性的估计，常常存在偏差（参见本书第22章）。

与外部不确定性类似，与某一给定问题相关的内部不确定性，有时可以通过推理和内省两种模式来评估。例如，关于某影星年龄的问题，我们既可以通过内省的方式寻找一个听起来比较熟悉的答案，也可以运用推理的方式从其他的知识中推导出答案。

我们并不是想说任何不确定性的经验，都可以落实到图35—2中的4种原型变式内。毫无疑问，这里还存在许多混合型的以及不确定的例子。我们已经知道，在一个给定的问题中，不确定性可以归因于外在事物的特征、个人的无知或者这两者的结合，并且可以通过单一模式、分布模式或者这两种模式的结合来进行评估。我们采取任何一种处理方式的目的，就是要突出不确定性经验的变异中的一些重要维度，而非提供一种详尽的并且相互排斥的经验分类。若要试图对各种测量主观概率的实验操作进行分类，可以参见豪厄尔等人的这本著作（Howell & Burnett, 1978）。

3. 讨论

尽管概率的术语，能被用于表达任何形式的不确定性，但概率理论的

原则，并不同等适用于所有的不确定性的变式。当采用分布或者频率模式对外在不确定性进行估计的时候，这些原则，最有可能被直觉判断所采纳和满足。比如，在我们参考气象统计数据以预测明年 4 月 12 号下雨的概率时，主观概率的互补性（complementarity）是非常明显的：过去相应所有 4 月份的日子，显然都可区分成下雨的日子和不下雨的日子。

而在其他变式中，互补性并不这么明显。若不确定性是依照倾向性、论据或确定性来评估的话，即使我们很确定地知道其中的一个选项是正确的，也很难发现这些概率能够互补并整合成一个整体。例如，人们会质疑为什么对纽约在罗马的北部这个论断的信念程度以及对纽约在罗马的南面这一论断的信念程度之和，会等同于对其他任何一组互相补充的陈述的信念程度。事实上，一些作者（如 L. J. Cohen, 1977; Shafer, 1976）认为互补性不应该适用于信念程度。具体而言，萨福（Shafer）指出，既存在两个互相排斥的并且绝无其他可能的假设都能得到坚实的支持的情况，也存在着两种假设都缺乏足够支持的情况，由此他反对信念互补论。类似的有关互补论的性的问题，也出现在对信心的表现和对相互冲突的倾向的评估中。

不确定性的变式，可能在它们被评估的信心上也存在差异。设想一个图钉被抛掷了 4 次，其中有两次钉帽朝下，两次钉针朝下。根据这些数据，大多数的观察者会把下次抛掷图钉时钉帽朝下的概率确定为 0.5。他们同样会认为抛掷硬币时“正面”朝上的概率为 0.5，但他们在对硬币的判断中会比在对图钉的判断中表现得更有信心。正如此例所示，对同样的倾向判断，人们有可能赋予其不同程度的信心度。有关概率的信心度是非常重要的，因为它控制着人们如何做决定。有证据表明（Ellsberg, 1961; Raiffa, 1961），人们更愿意在概率已知的事件上下赌注，比如抛掷硬币，而不愿下注于那些与外部的不确定性和个人的无知相关联的事件，比如抛掷图钉。

在这个领域的不同思想派别所发展起来的关于概率的概念和我们曾讨论过的不确定的模式之间，有着天然的联系。因此，对概率所作的频次的或客观的解释，就将这一概念局限于由抽样过程产生的外部不确定性。相比较而言，贝叶斯学派或个人学派（personal school）将所有的不确定性均

520

视为无知。在贝叶斯学派那里，偏好是信念的基础，而概率就产生自在各个赌注间存在的偏好。然而，从一个心理学的观点来看，这种赌注式的启发式似乎不太现实。由于存在着单个来讲都极具说服力，但彼此又互不相容的各个学派，而且当学派之间互相冲突时，又没有各方都认同的解决规范冲突的标准，这个领域中的论争往往会被激化。通过提供不同观点、不同诉求的各种学派对直觉的充分描述，心理学的分析可能会对规范的讨论有所助益。

本章原载《认知》(Cognition, 1982, 11: 143-157)。允许重印。

