

心理学的研究方法



2

■ 发现的背景

■ 验证的背景：客观性的保障

观察者偏见和操作性定义
实验法：另一个可选择的解释和控制的必要
相关法

阈下的影响？

■ 心理测量

获得信度和效度

自我报告法

行为测量和观察

■ 人类和动物研究中的道德问题

获得赞同

冒险/收获的评估

故意欺骗

事后解说

动物研究中的问题：科学，道德规范，政治

■ 成为明智的研究用户

■ 21 世纪的心理学：心理研究和信息爆炸

■ 要点重述

关键术语



诺贝尔奖得主、物理学家理查德·费曼 (Richard Feynman) 先生 (1985) 在他的自传《别闹了，费曼先生》中向我们描述了一个研究蚂蚁导航能力的非正式的程序：

“在我普林斯顿的家中，有一扇有着 U 型窗槛的凸肚窗。一天，一些蚂蚁爬上了窗槛，围着一小块食物转来转去。我对它们发现目标的方法感到奇怪，很想知道它们如何知道该去那里。它们能够像蜜蜂那样，相互传达食物在哪里的信息吗？它们是否能够感知几何结构？”

这简直太业余了：每个人都知道答案，但我不知道答案……

在（一个）实验中，我摆出很多玻璃显微镜的载片，让蚂蚁在上面走过，同时将一些糖放在载片周围

的窗槛上。然后，通过用新的载片换掉旧的，或者通过重新安排载片的位置等一些方法，我证明蚂蚁并不能感知几何结构：它们弄不明白东西究竟在哪里。如果它们通过一条路来到糖的旁边，但同时有一条更短的路回去，蚂蚁们从来都不能找到这条更短的路。

重新安排蚂蚁留下某种踪迹的玻璃载片是非常容易理解的。用这种手段可以导出很多简单的实验来发现踪迹干掉需要多长的时间，它是否很容易被抹去，等等。我还发现踪迹没有定向作用，如果我捡起一只蚂蚁放在一张纸上，把他转来转去，然后把它放回到踪迹上，在遇见另一只蚂蚁之前，他不知道自己走错了路。”

费曼简单而聪明的实验，使他以一种严格的方式，满足了他对蚂蚁行为的好奇心。

尽管你对蚂蚁交流的方式可能从来没有产生过兴趣，但你可能对人和动物的行为提出一些其他的问题，并迫切渴望得到答案。你可能记得在第 1 章里，我们曾经请你提出一个问题清单，这些问题是你想在读完《心理学与生活》这本书后可以得到回答的。在这一章，我们将描述心理学家是如何回答这些和你息息相关的问题的。我们将集中阐述心理学家应用科学的方法来解决这类问题时所采用的特殊的手段。我们希望你能了解心理学家如何设计他们的研究：可靠的结论是如何从心理学家所研究的（例如你如何思考、感觉和行动）复杂的且经常令人迷惑的现象中得来的？

即使你一生中从来不做科学研究，掌握这一章所阐述的内容仍然是大有裨益的。我们的目的是通过教会你如何提出正确的问题，如何评估心理现象的因果和心理现象的相关来帮助你提高批判性思维的技巧。大众传媒经常以“有研究表明……”这样的开头向我们传达信息，而通过不断锻炼你有理性的怀疑态度，我们将帮助你成为一个能够利用基于研究的结论来解决生活中面临问题的、更为老练的消费者。我们将向你展示如何“通过事实，而不是辩论”来解决争论 (Miller, 1992)。

发现的背景

心理学的研究过程可以分为两个主要的范畴，它们通常是依次发生的：形成想法（发现）然后去检验它（验证）。发现的背景 (context of discovery) 是研究的开始阶段，在这个过程中，通过观察、信念、信息和一般的知识，人们形成一个新的观点或者对于某种现象形成一种不同于以往的思考方法。研究者的问题是从哪里来的？其中一些来源于对环境中的事件、人物和动物的直接观察。另一些则来源于这些领域的传统内容：一些问题被看做是从早期学者中传下来的“伟大的但没有答案的问题”。一般来讲，研究者以一种独特的手段来组合已有的思想，以提出创新的观点。真正有创造力的思想者的标志是发现一个新的真理，它可以朝着更好的方向推进科学和社会的发展。

一般来讲，心理学理论试图揭示大脑、心理、行为和环境的功能以及它们是如何相互联系的。理论 (theory) 是一个组织起来的概念集合，可以用来解释一种现象或一系列现象。大多数心理学理论有一个一般的核心，即决定论

(determinism)的假设。这一假设认为,一切事件,包括物理的、心理的或者是行为的都是特定原因因素的结果,或者说是由其所决定的。这些原因因素被限定在个体环境或个人之内。研究者还假设行为和心理过程以有规律的方式联系,它们的关系和模式可以通过研究被发现和揭示出来。一般来讲,心理学理论总是阐述暗藏在这些有规律模式中的原因的影响。

当一个理论在心理学领域被提出来的时候,人们通常希望它既能解释已有的事实,也能够产生新的想法和假设。假设(hypothesis)是对原因和结果关系的试探性的、可以检验的阐述。一般而言,假设总是被描述为“如果……那么……”的形式,主要指特定的结果是从特定的条件中得来的。例如,我们可以预期,如果儿童在电视中看了大量暴力的场面,那么,他们将对同伴表现出更多的攻击行为。研究就是用来验证“如果”和“那么”之间的联系。理论对形成新假设起着重要的基础作用。当科学数据不能证明一个假设的时候,研究者必须重新思考他们理论的面貌,因此,理论和研究之间进行着连续不断的交互作用。

发现的背景另一个重要部分是研究的参与者必须具备特别的态度和价值观。科学要求对任何结论持有开放的、批评的和怀疑的态度,直到它已经被独立的调查者所接受。保持开放性有两个目的,首先,它使得真理保持一种暂时性特点,随时准备接受新的数据的修正;其次,开放性使得研究者愿意去评估一些重要现象的陈述,而这些现象是他们个人所不能相信和接受的。如超感知觉

科学理论需要经受严格的考验,在理论被认为是可靠的之前,它的结果必须能够被其他的研究者重复。



(extrasensory perception, ESP)(Bem & Honorton, 1994)。

科学实践的基础是尊重那些通过控制的观察和仔细测量而获得的事实。在科学的领域内,当好的数据和专家的观点发生冲突的时候,数据将赢得胜利。保密在研究程序中是被禁止的,因为所有的数据和方法必须最终公布出来以得到公众的验证,也就是说,其他的研究者应该有机会来审查、批评、重复或是驳斥这些数据和方法。对数据、结果和收集数据的方法的描述要和任何推论以及关于证据意义的结论保持分离。在科学出版物中,调查的每一部分都应该被清晰地报告出来,以使读者能够很好地将数据的客观特征从研究者主观的解释中分离出来。最后需要说明的是,我们发表论文,既是对这一领域内知识累积增加的需要,也是允许其他研究者重复这些发现的需要。

小结

发现的背景是研究者在研究阶段,运用观察、信念、信息和一般的知识来形成新的想法和理论。心理学的理论力图理解大脑、心理、行为和环境之间的决定论关系。

验证的背景:客观性的保障

在大多数情况下,研究者进行的工作是从发现到检验理论。验证的背景(context of justification)是把证据拿来验证假设的研究阶段。当试图得到可以产生有效结论、令人信服的证据时,心理学家面临严峻的挑战。他们依赖一个同盟者——科学的方法(scientific method),它使成功成为可能。科学的方法是通过将错误降低到最小,提出可靠的归纳等方法来收集和解释证据的一般的程序集合。心理学被认为是一门在某种程度上遵循科学方法所建立的规则的科学。

因为在科学研究的数据的收集和分析阶段主观性必须被降低到最小,人们用程序的保障措施来增加客观性。其中一种保障措施不言自明:那就是研究者必须以一种别的研究者可以理解和评估的形式来保存观察时和数据分析时的完整记录。在科学方法的其他方面,我们希望强调为什么特别的程序如此重要,因此,在以下两部分中,我们都从客观性的挑战入手,然后来描述被科学方法所规定的补救措施。

观察者偏见和操作性定义

当不同的人观察同样的事件时，他们并不总是“看到”同样的事物。在这一部分，我们将描述观察者偏见的问题以及研究者所采用的补救步骤。

客观性的挑战

观察者偏见(observer bias)是由于观察者个人的动机和预期导致的错误。通常，人们看见的、听到的只是他们所预期的，而不是事实的本来面目。让我们来考察一个有关观察者偏见的非常极端的例子。在20世纪的开始，一位重要的心理学家雨果·蒙斯特博格(Hugo Munsterberg)给大量听众，包括很多记者作过一次关于和平的演讲。事后，他以如下方式对记者的见闻做了总结：

“这些记者坐在离讲台很近的地方。一个人写道，听众对我的演讲感到非常惊讶，会场上鸦雀无声；另一个人写道，我经常被大声的鼓掌打断，在我讲演的最后，鼓掌持续了几分钟的时间；一个人写道，在我的对手讲话的时候，我经常微笑着；另一个人则注意到我的脸保持严肃，没有一丝笑容；一个人说到我的脸由于激动变得紫红，另一个人则发现我的脸变得惨白(1908, p. 35-36)。”

回到原始的报纸上，看看记者们如何带着他们的政治观点来做解释是非常有趣的——然后我们可能会理解为什么记者“见到了”他们所做的。

参与者们以及观众、广播员都受到观察者偏见的支配。你如何才能看出真正发生了什么？

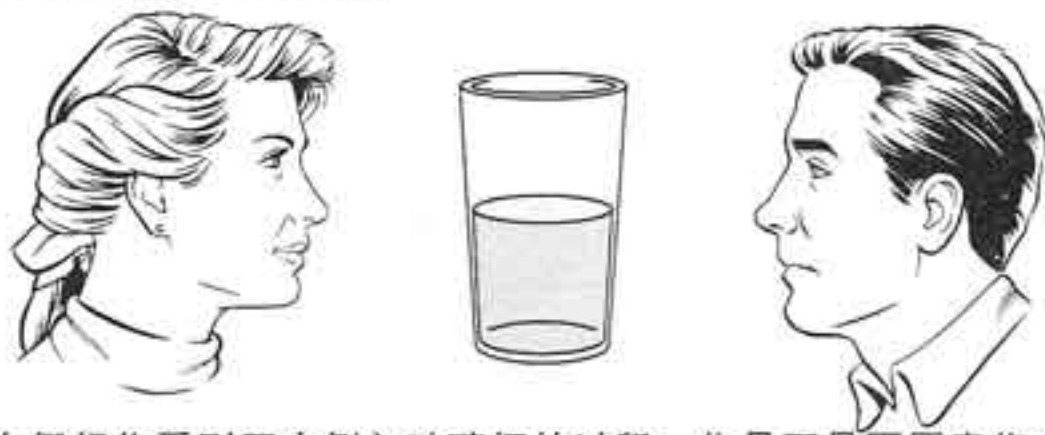


在一个心理实验中，我们不希望发生像蒙斯特博格所报告的那种可笑的观察者之间的差异。然而，这个例子表明，同样的事件如何能够引起不同的观察者得到不同的结论。观察者偏见所起的作用像一个过滤器，一些事情被看成是有关系的和重要的，而另外一些则被忽略为无关的和重要的。

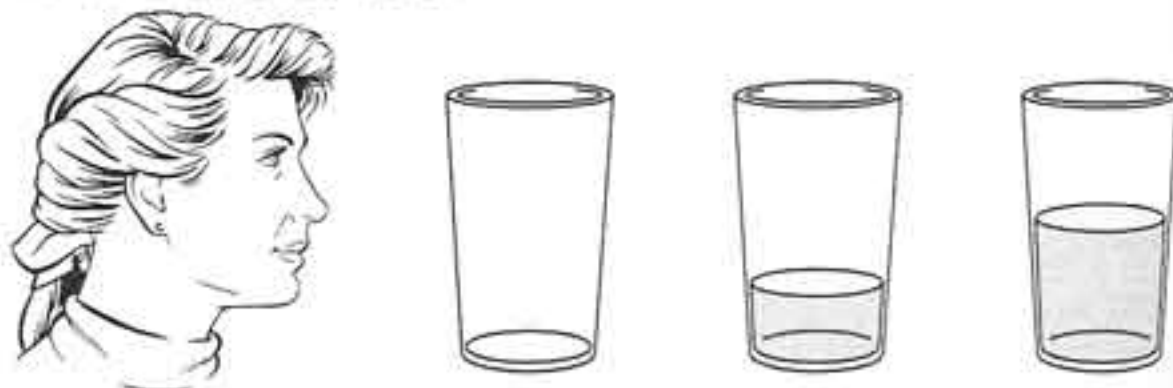
现在，我们希望你试着来描述图2.1，这个例子表明产生观察者偏见是多么容易。你迅速的描述将会使你认识到：你以前进行观察的经验如何影响你对现在看到东西的解释。

让我们把这些经验应用到心理实验的情境中来。研究者经常忙于进行观察。假设每一个研究者都带着不同的先前经验来进行观察——而且通常那些经验包括对特定理论的认同——你将可以看到为什么观察者偏见会引发问题，研究者用什么办法才能确保他们的观察最小程度地受先前预期的影响呢？

请看图示中的玻璃杯。你将如何回答这个经典的问题：这个玻璃杯是半空的还是半满的？



现在假想你看到了水倒入玻璃杯的过程。你是不是更愿意将它描述为：玻璃杯是半满的？



假设你看到一杯水被不断地倒走，那么现在这个杯子是不是看上去是半空的？

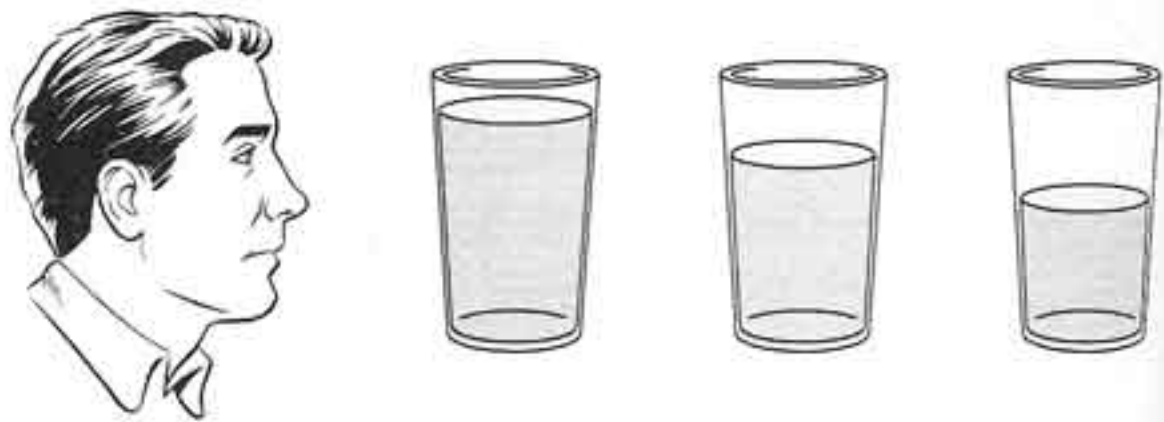


图 2.1 这个玻璃杯是半空的还是半满的呢？

补救措施

为了使观察者误差降到最小，研究者依赖于标准化和操作性定义。标准化(standardization)意味着在数据的收集阶段使用统一的、一致的程序。测验或实验条件的所有特征应该充分标准化以便所有研究的参与者经历完全一样的实验情境。标准化意味着以同样的方式来问问题，以事先建立的规则来量化反应，将结果打印或记录下来以确保在时间、地点，以及不同的被试间和研究者间的可比性。

观察本身也必须标准化：科学家必须解决这样的问题，即如何将他们的理论转化为含义前后一致的概念。对概念含义进行标准化的策略我们称为操作化。操作性定义(operational definition)是以测量它或决定它存在的特定的操作或程序来界定一个概念，在一个实验内使含义标准化。一个实验中的所有变量必须给予操作性定义。变量(variable)是一些在量和质上变化的因素。在实验中，其值相对于情境中其他变量而言独立自由变化的刺激条件被称为自变量(independent variable)。其值是一个或更多自变量变动的结果的变量被称为因变量(dependent variable)——他们依赖于刺激条件的变化。

例如，假设你希望检验那个我们先前曾经提到过的假设：在电视中看到很多暴力镜头的儿童将在对同辈的行为中表现出更多的攻击倾向。你可以设计一个实验来操纵每名被试所看到的暴力镜头的数量（自变量），然后评估他或她表现出来的攻击倾向有多强（因变量）。在你的实验设计中，一个重要的部分是去操作包含在各种电视节目中暴力镜头的数量，以及参与你这项研究的被试所表现出来的攻击性。思考一下，你能设计一个什么样的程序使所有这些测量精确？

操作变量往往非常复杂，这是心理学家经常面临的问题。例如，你如何对孩子和母亲关系的质量进行操作性定义（你可以在第11章中看到答案）？心理学家还必须操作那些无法直接观察到的变量，如你如何对自尊进行操作性定义（见第14章）？如果你是一名心理学研究者，所有其他的研究者可能并不同意你对某一变量所做的操作性定义——他们可能认为，你未能抓住自尊的实质——但如果你给出一个精确的操作性定义，他们就会知道该如何判断和重复你的工作。花一点时间去看看你如何操作图2.2照片中提到的快乐的概念。

实验法：另一个可选择的解释和控制的必要

你知道从日复一日的经验中人们能够对同一个结果提出很多原因。当试图对因果关系进行正确的阐述时，心理学家面临同样的问题。为了克服因果关系中的模棱两可，研究者运用实验法(experimental methods)：它们操作一个自变量来观察其在因变量上产生的效果。这种方法的目的在于明确一种强烈的因果关系，即一个变量对另一个变量有影响。在这一部分，我们将阐述备选解释的问题以及面临这一问题时所采用的一些步骤。

客观性的挑战

当心理学家检验一个假设时，他们经常在头脑中对为什么自变量的变化可以以一种特别的方式影响因变量有一个解释。例如，你可以预期并且实验性地论证，电视中暴力的镜头可以导致高攻击性。但是，你怎么知道恰好是电视中的暴力镜头诱发了暴力呢？为了使这一假设得到最强有力的支持，心理学家必须对可能存在的备选解释非常敏



图 2.2 操作性定义

假设对于这两张照片而言，你只对定量描述每个人有多么高兴感兴趣，那么你应该考虑哪些因素呢？



暴力行为是否由于在电视中看到暴力镜头而引起的？你如何去发现？

感。其他可以导致同样结果的解释越多，最开始假设的可信性就越弱。当一些并不是实验者有意引入到实验情境中的因素确实影响了被试的行为，并对数据的解释增加混乱的时候，我们称这些因素为混淆变量(confounding variable)。当一些被研究的行为效果真正原因被混淆的时候，实验者对数据的解释就会冒风险。例如，假设暴力的电视场景比大多数非暴力的场景更引人注目或者包括更多的动作时，“暴力”就和场景中的表面现象相混淆了，研究者不能确定究竟是哪一个因素导致了攻击行为。

尽管每一个不同的实验方法都潜在着一些特殊的备选解释，我们仍然可以明确两种几乎存在于所有实验中的混淆，即期望效应和安慰剂效应。当研究者或观察者向被试暗示他所预期发现的行为，并因此引导出期望的反应时，非有意的期望效应(expectancy effects)便发生了。在这种情况下，真正诱发所观察到反应的是实验者的预期，而不是自变量。

罗伯特·罗森塔尔(Robert Rosenthal)研究了这种期望偏差的现象以及这种偏差是如何对研究结果产生歪曲的(Rosenthal, 1966)。

我们如何知晓



老鼠将满足你的期望 贯穿《心理学与生活》这本书，我们将和你一起分享心理学的研究成果。一般来讲，我们的结论是由一些分散在世界各地不同的研究小组研究得来的。“我们如何知晓”栏目的目的是定期给你机会来接触一些个别的研究，这些研究都是直接回答特定问题的。每一个“我们如何知晓”栏目都可以使你有机会领悟到实验和认识间精确的关系。在这第一个“我们如何知晓”中，我们将简要地阐述一个研究，它形象地说明了预期能够影响实验结果的途径。

在一个经典的实验中，将几组即将进行走迷宫训练的老鼠分配给12名学生。其中一半学生被告知他们分到的老鼠都是一些擅长走迷宫的种群，而其他的学生则被告知他们的老鼠都不擅长迷宫测验。正如你可能猜想的那样，这些老鼠实际上都是一样的。然而，学生们的实验结果符合他们对其老鼠的预期，那些标记为聪明的老鼠往往比标记为笨的老鼠被认为是更好的学习者(Rosenthal & Fode, 1963)。

你认为学生们将他们的预期传达给他们的老鼠了吗？当实验是在人类实验者和人类被试间进行时，你知道你为什么应当更担心期望效应了呢？因为期望效应歪曲了发现的内容。

当没有任何一种实验操作时，参加实验的被试也改变了他们的行为，这时，安慰剂效应(placebo effect)就发生了。这一概念的来源是医学中的一种现象：当病人接受了化学上无效的药物或无针对性的治疗后，他或她的身体康复了。安慰剂效应指的是由于个体对于治疗会产生良好效果的信念所导致的健康的增强。一些采用无疗效药物的治疗已经表明，在这种情况下，有70%采用过这种药物的病人都有好的或极好的结果(Robert et al., 1993)。

在心理学的研究情境中，当行为反应受到个人对做什么和如何感受的预期的影响，而不受特定的介入或产生某种反应程序的影响时，安慰剂效应就发生了。回忆一下关于看电视导致日后进攻性的实验，假定我们发现根本没有看过任何电视的被试也表现出很高水平的攻击性，我们就可以推测这些被试是由于被放到一个允许他们表现攻击的情景下，他们觉得他们应该表现出攻击行为，因而就这么做了。实验者必须要考虑到被试会仅仅因为意识到他们在被观察或被测验而改变行为。例如，被试可能会对选择他

们来参加这项研究感到很特别,因此所做的跟平常的时候有所不同。这种效应会危及实验的效果。

补救措施

由于人和动物的行为很复杂,往往有多种原因,因此好的研究设计应该能够预期到可能出现的混淆并且采取策略来消除它们。类似于运动中的防护策略,好的研究设计应该能预期对手将做些什么,并制定计划来应付它。研究策略被称为**控制程序**(control procedures)——它是一些力图使所有变量和条件(除了那些与被验证的假设相关的)保持恒定的方法。在一个实验中,指导语、室内温度、任务、研究者行动的方式、时间的安排、记录反应的方式,以及其他一些情境中的细节必须对所有被试都一致,以确保他们的实验是相同的。被试所经历的惟一不同应该是那些由自变量的不同条件。让我们来看一些补救措施,他们是针对一些特定的混淆变量的,如期望和安慰剂效应。

例如,假想你发展了攻击的实验,加入了一个看喜剧节目的处理组。你觉得应该小心避免根据自己的预期,把喜剧组和暴力组的被试用不同的方式来看待。因此,在你的实验中,我们需要研究助手的参与,由他们来迎接被试,并在并不知道被试是看了暴力节目还是喜剧的情况下,来测验他们的攻击性。在最理想的情况下,可以通过保证实验助手和被试都不知道(双盲)哪一名被试进行了哪项处理来消除偏见。这一技术被称为**双盲控制**(double-blind control)。在很多研究设计中,知道实验假设的人是不能被安排来收集被试的数据的。

为了解释安慰剂效应,研究者通常引入一个不进行任何处理的实验条件,我们称为**安慰剂控制**(placebo control)。安慰剂控制属于控制的一般范畴,以使实验者确保他们自己正在进行恰当的比较。设想有一位年轻的女孩,当被问及她是否爱她姐姐的时候,她回答:“和什么比呢?”在你能够真正理解研究结果意味着什么之前,这是一个必须问——而且必须被满意回答的问题。假设你读到一个研究表明“在尼古丁替代物的帮助下,一个群体中超过3/4的人试图戒烟的努力能够成功”(Andrew, 1990)。和什么比呢?控制组怎么样呢?在这项研究中,安慰剂控制组,服用没有任何疗效的替代物,39%的人戒了烟!而且,使用这个并无疗效替代物的时间越长,戒烟的可能性就越大(Abelin et al., 1989)。因此,尼古丁替代物是一种有效的治疗,但是超过一半的疗效归因于预期它会有效的安慰剂效应。控制情况下的数据为我们评估实验效应提供了一个非常重要的基线。

有一些实验设计,我们称其为被试间设计

(between-subjects designs),被试被随机地分配到实验条件(接受一个或多个实验处理)和控制条件(不接受实验处理)下,来接受不同的程序。随机的安排是研究者力图消除潜在被试中和个体差异有关的混淆变量的主要步骤之一。我们所记得的攻击实验就是采用这样的设计。在实验和控制情境下随机安排的两组具有很高的一致性,这是实验开始阶段非常重要的一步。因为每名被试都有同样的可能性被分配在实验情境或控制情境中,我们就不必担心分配在实验组的每个人都喜欢看暴力电视而分配在控制组的每个人都厌恶看暴力电视了。随机的安排可以使每一组中都混有不同种类的人,如果在结果中发现不同情境下存在差异,我们能够更加确信这些差异是由加入的处理引起的,而不是由于原先就有的差异引起的。

研究者还力图在带被试进入实验室的过程中接近随机。一般来讲,心理实验使用20~100名被试——但是,实验者经常希望利用这些样本所表明的情况,从样本(sample)来估计总体(population)。假定你想检验这样一个假设:6岁的儿童比4岁的儿童更爱说谎。你只能从世界上所有4岁和6岁的儿童中取一个非常小的子集,把他们带到你的实验室中。为了使你的样本一般化,你必须相信你所选取的4岁和6岁的儿童类似于任何其他随机选取的儿童组。如果一个样本在诸如男女性别、种族等方面的分布都与总体的特征非常匹配,这个样本就是总体的一个**代表性样本**(representative sample),你能够从这个样本推导出它所代表的总体。如果你有关说谎的研究中只有男孩

当学生们参加万圣节聚会的时候(那也是实验的一部分),在不被认出的情况下,他们比易被认出的时候表现出明显更强的攻击性。在你自己的生活中,你能够想出匿名能够改变行为的事例吗?



做被试, 你将对女孩可能的行为得出错误的结论。

另一种试验设计称为被试内设计 (within-subjects design), 用每一个被试作为他自己的参照。例如, 开始进行实验处理前被试的行为可以与处理后的行为作比较, 在我们已经知道的 A-B-A 设计 (A-B-A design) 中, 被试首先经历基线情境 (A), 然后进行实验处理 (B), 最后再回到基线 (A)。

我们如何知晓



A-B-A 设计中儿童的攻击性 A-B-A 设计是由一位持有这样假设研究者使用的: 只要有环境提供机会, 当儿童感到不会被人知道的时候, 他们的攻击水平会提高。学校的学生被邀请参加万圣节聚会, 在那里他们可以参与各种各样的游戏, 既有诱发攻击性的, 也有不诱发攻击性的。在基线条件下 (A_1), 儿童不穿万圣节的服装进行玩耍; 在处理情境下 (B), 他们穿上服装继续进行他们所选择的的游戏; 最后, 回到基线条件下 (A_2), 他们被告知, 这些服装需要归还, 但是他们可以在没有服装的情况下继续玩。实验结果支持了研究者的假设, 正如你在图 2.3 所看到的, 在其他条件一致的三种情况下, 同样的孩子在匿名的时候 (B) 比可以认出的时候 (A_1 或 A_2) 更具有攻击性。

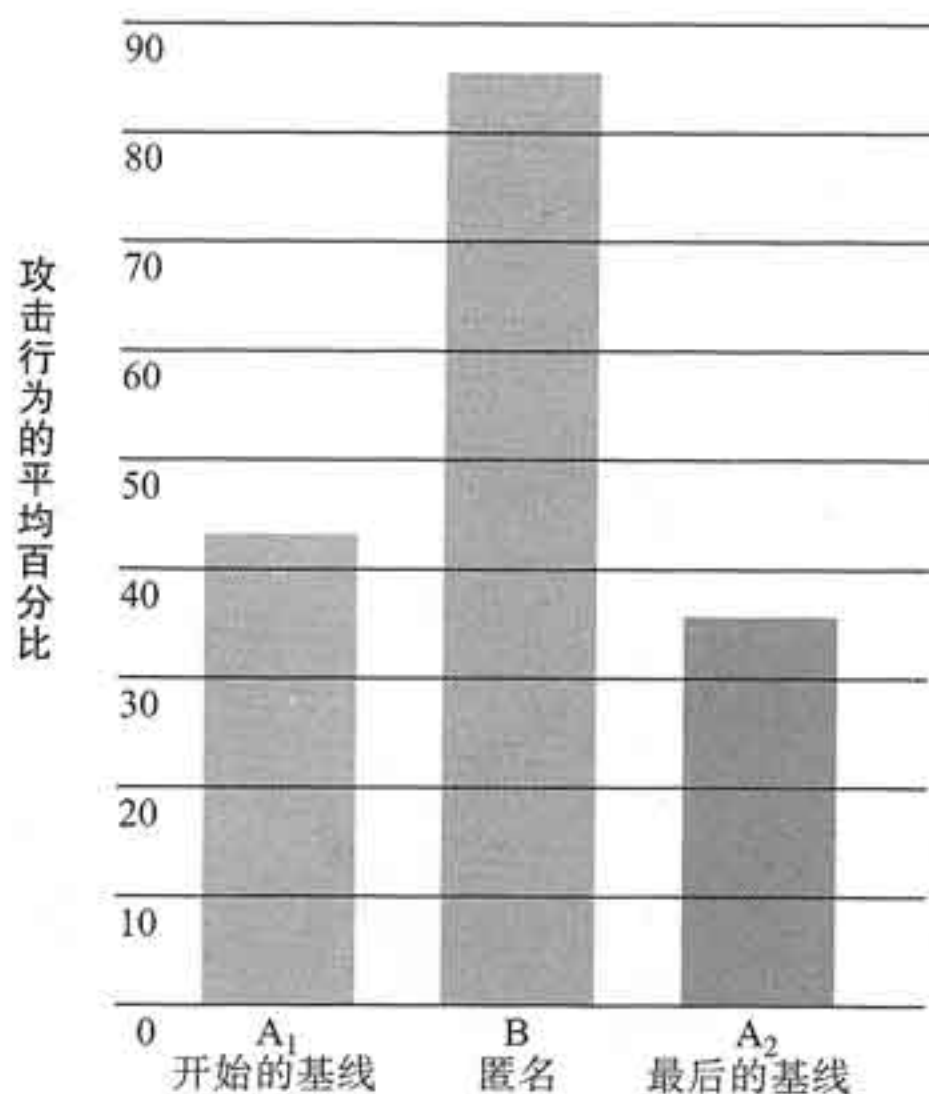


图 2.3 匿名—刺激攻击

匿名的效应很富戏剧性: 同样是这些儿童, 他们在匿名条件下比不穿万圣节服装时表现出了高得多的攻击性行为。

回到基线 (第二个 A 阶段), 使得实验者能够非常确信处理导致的变化不存在混淆的变量, 如时间的流逝等。

我们所描述一系列方法都包括对自变量的操作以观察其对因变量的影响。尽管这些研究方法经常允许研究者对变量间的因果关系进行最强的表述, 但一些条件会使得这些方法不能令人满意。首先, 在一个实验中, 行为经常在一个表面的环境中进行研究, 条件因素被如此强烈地控制, 以至于环境本身就可能歪曲行为, 使它们和自然发生的时候不一样。批评家声称, 大多数自然行为模式的丰富性和复杂性在控制的实验中消失了, 成为简单的、只有一个或少数变量的处理。其次, 研究中的被试往往知道他们在参加实验, 在被测量和检验。他们可能会受这种意识的影响, 表现为力图取悦研究者、努力地去“推测出”研究的目的, 或者改变他们的行为, 使其与未意识到被监控时的行为不一样。第三, 有一些重要的研究问题受到伦理的约束而不可能实施。例如, 我们不能通过设置一个实验组的儿童被虐待, 而设置一个控制组的儿童不被虐待, 以发现虐待儿童是否具有代代传递的倾向。在下面一部分, 我们将谈论人们经常提及的研究方法的类型。

相关法

智力和创造力存在联系吗? 乐观的人是否比悲观的人更加健康? 儿童早期受到虐待和后期的精神疾病有关系吗? 这些问题中设计的一些变量在心理学家看来, 有一些不容易操作, 而另一些操作起来不合道德原则。为了回答这些问题, 正如我们在后面的章节要讨论的, 我们的研究需要基于相关法 (correlational methods)。当力图决定两个变量、特质或者属性关联到什么程度时, 心理学家使用相关的方法进行研究。

为了确定存在于两个变量之间相关的精确程度, 心理学家需要计算一个名为相关系数 (correlation coefficient) (r) 的统计量。这个值在 +1.0 到 -1.0 之间变化, 其中 +1.0 表示完全的正相关, -1.0 表示完全的负相关, 而 0.0 表示根本没有相关。一个正的相关系数意味着当一系列分数增加时, 第二列分数也增加。而负相关正好相反, 第二列分数和第一列分数朝相反的方向变化 (见图 2.4)。一个非常接近于零的相关意味着两个测量分数之间存在很弱的联系, 或是根本不存在联系; 当相关系数越来越大, 直至接近于最大值 +1.0 的时候, 根据一个变量的信息来预测另一个变量将变得越来越精确。

例如, 一位研究者想考察工人的生产效率和压力的相

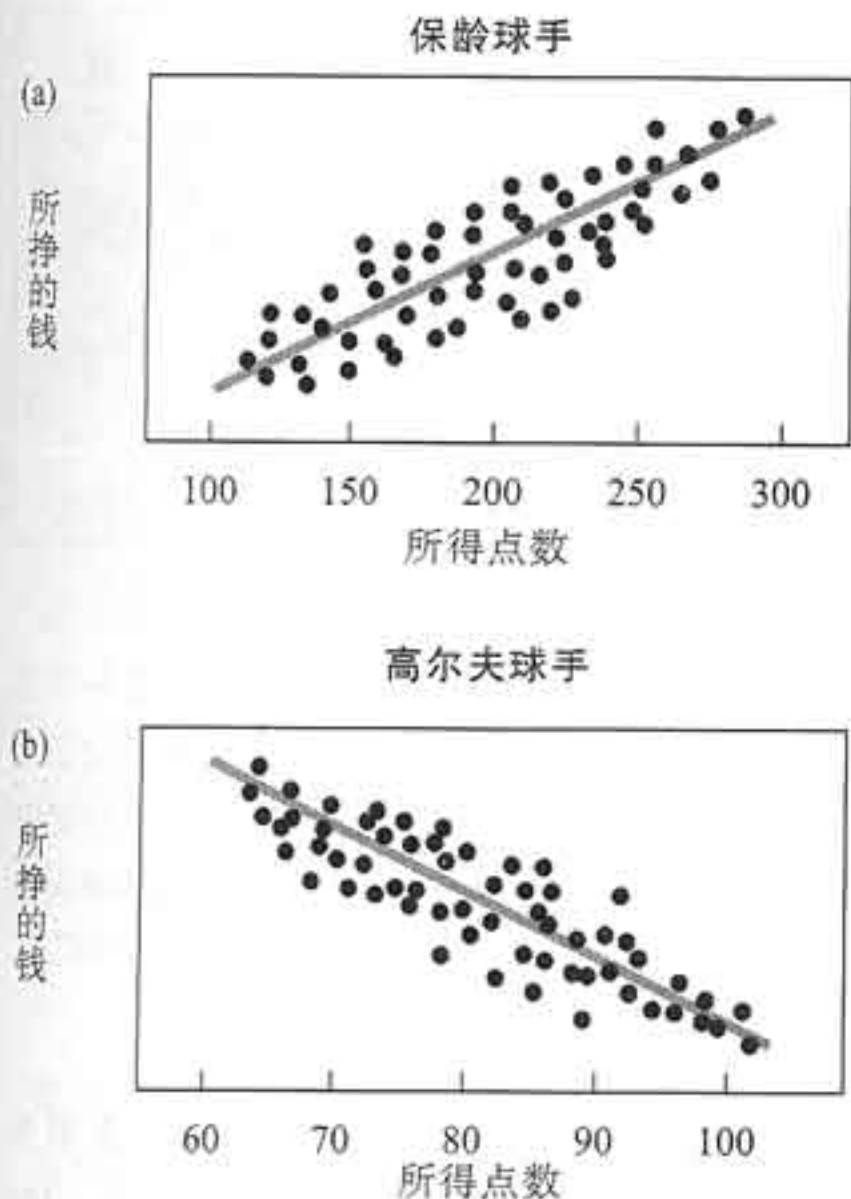


图 2.4 正相关和负相关

这些图示的数据表明正相关和负相关的不同。每一个点代表一名保龄球手或高尔夫球手。(a) 一般来讲, 职业保龄球手所得的点数越多, 他或她挣得就越多。因此, 在这两个变量之间存在正相关。(b) 高尔夫的相关是负的, 因为高尔夫球手得的点数越少所挣得钱越多。

关, 他可以测量人们在生活中经受了多大的压力, 还可以测量人们在工作中表现得多么好。压力可以被操作性地定义为人们在压力量表中的得分, 而工作效率可以被定义为工人每日生产特定产品的数量。然后, 研究者可以选取大量工人, 测量每一个变量的值, 然后计算出它们之间的相关系数。一个很强的负相关意味着当压力增大时, 生产效率下降。知道人们生活压力的分数, 研究者就可以对人们的效率建立合理的预期。

研究者可能想进一步说明: 可以通过减少压力的方法来增加生产效率。这种认识是错误的。高的相关只是表明这两列数据以一种系统的方式建立了联系, 相关并不意味着一个事件可以导致另一个事件的发生, 相关并不意味着因果关系。相关可能反映一部分因果关系, 也可能根本就不反映因果关系。例如, 压力和效率之间的负相关可能意味着: (1) 压力从根本上是人们工作做得差的原因; (2) 不佳的工作生产率使人们经受了更大的压力; (3) 存在一种特定的人格类型, 他们更可能经受压力, 同时也



为了研究压力和工作表现的关系, 研究者必须想出测量这两个因素的方法。如果发现存在正相关, 研究者可以推导出什么来呢?

在工作中表现不佳。注意其中最后一种情况, 第三方变量影响到其他两个的变化。请想象最后一种情境, 假设新技术在一定的 workplace 导致了更多的噪音, 更多的噪音导致了更大的压力 (因为工人们无法思考), 也导致了生产效率的下降 (因为工人们无法交流)。此时, 相关依然存在——高的压力和低的生产效率同时发生——但这根本就不是因果关系。

相关也可能是伪造的, 因为研究者不能通过恰当的控制来进行比较。例如, 1965 年纽约市的能源管制和 9 个月后报告的出生率剧增就存在虚假的联系。“纽约人非常浪漫, 那是因为烛光。”一位新父亲这样说。一位有关的官员解释道: 因为管制, “所有性的替代物——开会、听课、卡片聚会、戏剧、沙龙——在那个晚上都没有了。你还能做什么呢?” (《纽约时报》, 1966 年 8 月 11 日)。在世界的很多地方, 类似的报道在大雪暴或是其他灾难发生之后经常出现。“在海港区, 地震可以导致出生率的剧增”是《旧金山新闻》近期的一个标题 (Chen, 1990)。然而, 当人们花一些时间, 将这些明显急剧上升的出生率和一般的季节变量做比较时就会发现, 相关只是一种伪装成因果的巧合。也就是说, 季节和出生率确实存在相关——它和偶然出现的“灾难”的相关形成控制的对比。很明显, 当我们应用从实验方法中得出的研究结果时, 我们必须对这样的相关结果保持高度警惕。

我们并不想给你留下这样的印象, 即相关的方法不是

一个有价值的科研工具。贯穿整个《心理学与生活》这本书，我们还将看到很多相关的研究，它们都促进了非常重要的认识。我们在此仅提供一个例子来助长你的胃口。

我们如何知晓



年幼男孩中行为问题的来源 什么样的环境因素能够解释，一些男孩，甚至只有五岁，就表现出行为问题，而另一些男孩却没有？一个研究小组试图证明，男孩中的这些差异部分是由于男孩与他们兄弟姐妹相处过程中破坏性的同胞冲突的数量不同而引起的（Garcia et al., 2000）。研究者论证道，和同胞高水平的冲突可能会强化男孩的攻击倾向或对生活情境不恰当的反应。为了测量破坏性的同胞冲突，研究者录制了180个男孩和他们的同胞用不同玩具进行玩耍时的录像。并从以下维度对录像的结果进行评估：孩子间冲突的次数有多少？冲突的强度有多大？相关分析强烈地支持这样的预期，即经历高水平同胞冲突的男孩更有可能表现出进攻性和过失行为。

你能够看到为什么相关设计适合这样的预期吗？你不能随机地安排一些学生和他们的同胞有少的或多的冲突。你只能等着看儿童在某种情境或另外的情境之下有什么样的差异发生。

阈下的影响？

作为这一部分的结束，我们提供一个具体的例证，让我们来看看心理学研究是如何被广告人迫切地用来宣传他们的产品的。你几乎肯定经历过那种录音磁带的商业宣传，宣传中声称这种磁带能够利用意识觉知之外的信息，即潜意识的信息来改变你的生活：它简直是魔幻磁带！一盘磁带保证你拥有更好的性生活，另一盘可以帮你迅速地增加自尊，第三盘帮助你安全有效地减肥。怎么做呢？所有你需要做的只是听——在床上、路上、做作业的时候——“沙滩上舒缓的清涛拍岸声”。

“阈下”的影响有很长的历史。1957年，阈下广告的“发明家”宣称“买爆米花”的信息在电影放映的时候快速地显示在屏幕上，导致爆米花增加了58%的销售量（Rogers, 1993）！尽管几乎可以肯定这是一种欺骗，但它仍然引起了轰动。《华尔街时报》曾经报道，当在Muzak系统中打出“如果偷窃，我将会进监狱”的阈下信

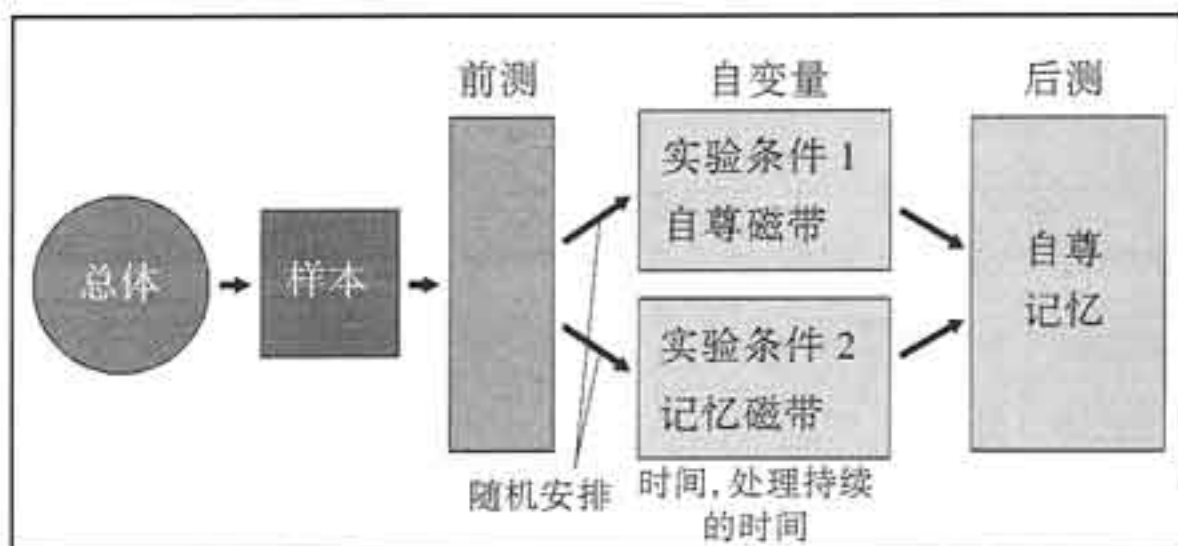


图 2.5 用来检验关于阈下录音磁带效力假设的实验设计

在这个实验的简化版本中，样本是从一个更大的一般的总体中选取的。经过一系列前测之后，他们被随机地分配，以接受记忆或自尊信息的阈下磁带。然后接受后测以客观地评定他们在记忆和自尊这两个因变量上的变化。研究发现阈下劝说并没有明显的效果。

息之后，新奥尔良一个超市的偷窃和资金流失显著减少了。俄亥俄州托莱多（Toledo）的一个电话调查表明，400名被调查者中有75%熟悉阈下广告（Rogers & Smith, 1993）。在那个群体中，又有将近75%的人相信阈下广告在超市中使用得很成功。一般来讲，回答者受教育程度越高，越有可能相信阈下广告的效力。

你现在已经具备了判断这一关键的问题知识：潜意识录音带是否真像他们宣传的那样，影响了大脑的状态和行为？我们的回答来源于我们描述过的实验方法的应用（见图2.5）。

我们如何知晓



对阈下效应的评估 实验小组着手来判断，那些为了提高自尊和记忆力而设计的商业性录音磁带是否真正有效果。237名男女被试参加了实验，他们的年龄范围从18~60岁。首先是前测，通过标准的心理测试和量表测定他们最初的自尊和记忆力，然后，被试被随机地分配到两种情景中，其中的一半被试得到阈下记忆磁带，其他的被试得到阈下自尊磁带，要求他们定期地听这些磁带。五周以后，他们回到实验室进行后测，接受记忆（使用四个记忆测验）和自尊（使用三个自尊量表）测验。研究者并不知道哪名被试接受了哪种处理（Greenwald, 1991）。

这些磁带提高了自尊，增强了记忆吗？这项被控制的实验结果表明：在任何一个客观的测量中，被试无论在自尊还是在记忆上都不存在明显的增加。然

而,一个非常强的效应确实出现了:那就是预期会获得帮助的安慰剂效应。预期到会有安慰剂效应,研究者增加了另一个独立的变量。每一组中的一半被试得到误标记为自尊的记忆磁带,其他被试则得到放在记忆盒子中的自尊磁带。被试相信如果接受了标记为自尊的磁带,他们的自尊就会增加,或者如果接受了标记为记忆的记忆磁带,他们的记忆力就会增强——这种现象甚至在当他们听的根本是另外的磁带时也会发生!

这一严格的实验为我们提供了非常具体的建议:省着你的钱吧,阙下的自助磁带所起的作用并不比安慰剂效应更多。《心理学与生活》这本书的一个重要的目的就是向你提供这种建立在坚实实验方法上的具体的结论。

这个实验还给了你一个心理测量中变量类型的例子,在这个例子中,是被试关于自尊和记忆力提高的信念以及自尊和记忆力的客观测量。在下一部分,我们将一般性地对心理学家测量实验中重要过程和维度的方法进行讨论。

小结

研究进程的第二个阶段是验证。为此,心理学家依赖科学的方法。研究者对数据的收集使用标准化的程序,用操作性定义来界定概念,以此来克服观察者偏见。实验的方法允许研究者宣称其发现的因果关系。研究者必须谨慎地选择解释,包括期望效应和安慰剂效应。为了确保结论的准确性,研究者常使用双盲控制、安慰控制等程序。对于被试间设计,随机分配原则是一种重要的控制方法,它使控制组被试和实验组被试相匹配。被试内设计包括A-B-A设计,被试作为自己的控制组。研究者使用相关研究方法测量两变量的相关程度。相关研究的主要局限是它不能得出因果关系。实验研究方法可用于检验现实世界的一些主张,如阙下录音磁带效应。

心理测量

由于心理过程是多样化的、复杂的,因此,对于测量心理过程的研究者来说这是一项挑战。尽管一些行为和过程很容易看到,但大多数是看不到的,如焦虑、梦等。所以,心理学研究的任务就是使这些看不到的心理过程和行为能够被看到,使内部事件和过程外化,使个人经历公开化。由此你已经了解对于研究者来说,为研究的现象提供

操作定义是多么重要。这些操作定义为实验提供了如何安排变量的数目、定量法、不同水平、大小、强度或数量的程序。许多测量方法是有效的,每种方法都有其独特的优势和弊端。

下面我们回顾一下心理测量,首先讨论两种评估测量精确性方法(信度和效度)的差异。然后,我们来看看几种收集数据的测量技巧。无论心理学家使用哪种方法收集数据,他们都必须使用适当的统计方法验证假设。在本章后的统计附录中记述了心理学家如何分析数据,你可把它与本章内容相结合进行阅读。

获得信度和效度

心理测量的目标是产生可信和有效的发现。信度(reliability)指心理测验或实验研究得到的行为数据具有一致性或可靠性。结果具有可信性指在相似的测验条件下该结果具有可重复性。一种可信的测量手段在重复使用时(所测量的内容不变)将产生可比较的分数。上面我们叙述过一个实验,它表明阙下录音磁带仅能产生安慰剂效应。那些实验使用了237名被试。实验者声称结果可信,这意味着他们能够使用大小相同的另一组被试重复这一实验结果得到相同的数据模式。

效度(validity)指研究或测验得到的信息精确地测量了研究者想要测量的心理变量或品质。例如,对于高兴的有效测量能够预测在特殊情况下你多么高兴。一个有效的实验意味着研究者能把研究结果概括到更大的范围。当我们为你提供基于录音磁带实验的建议时,我们接受研究者提出的研究结果是有效的主张。测验和实验可能是可靠的,但同时可能是无效的。例如,我们把鞋子的大小作为高兴的指标。这是可靠的(我们总能得到相同的答案),但这一指标是无效的(我们不能获得日常的高兴水平)。

当你阅读不同的测量类型时,尽量根据信度和效度评估他们。

自我报告法

研究者往往会对获得那些无法直接观察到的经验的数据感兴趣。有时这些经验是内部的心理状态,如信仰、态度、感觉等。在其他时候,这些经验是外部行为,但如性行为或犯罪活动之类的行为通常是不适合目击的。在这些情况下,研究要依靠自我报告。自我报告法(self-report measures)是口头报告(写或说)研究者提出的问题。研究

者设计可信的方法量化这些自我报告,这使他们能对不同个体的反应进行有意义的比较。

自我报告包括问卷法和访谈法。问卷或调查包括一系列问题,其内容有事实问题(“你是一名登记的投票者吗?”),有过去或当前行为的问题(“你吸多少烟?”),有态度和情感问题(“你对当前的工作满意吗?”)。开放式问题指能自由组织言语回答问题。问题可以有許多固定的备选答案,如是、否、不确定。

访谈指研究者为了获得详细信息和个体进行对话。不像问卷十分标准化,访谈是交互式的。访谈者可以根据回答者说的内容变化问题。好的访谈者除了对社交中发现的信息敏感,对社交过程也是十分敏感的。训练访谈者与回答者建立和善的、积极的社交关系,鼓励回答者信任访谈者,并与访谈者分享个人信息。

尽管研究者依赖各种各样的自我报告法,但自我报告法也有其局限性。很明显,许多形式的自我报告不适用于前语言期儿童、文盲的成人、其他语言使用者、一些心理紊乱的个体、非人类动物。甚至即使自我报告能够使用,它们也可能是不可信的或不可靠的。被试可能不理解问题或不能清楚地记得他们当时的经历。进一步说,自我报告可能受社会期望的影响——为了产生赞许性(或者有时是非赞许性的)印象,人们通常给出错误的或误导性的答案。如果他们报告出自己的真实经历或情感,通常是很尴尬的。如果回答者意识到问卷或访谈的目的,那么为了获得工作,或免除进入精神病院,或完成其他目标,他们可能说谎或选择事实。访谈的环境也能产生个人偏见和成见,这都影响访谈者怎样提出问题和回答者怎样回答问题。

行为测量和观察

作为一个群体,心理学家对许多行为感兴趣。他们可以研究老鼠走迷宫、儿童绘画、学生记忆一首诗或工人重复完成一项任务。行为测量(behavioral measures)是研究外显行为和可观察、可记录的反应的方法。

观察是一种研究人们做什么的主要方法。研究者可有计划、准确和系统地进行观察。观察或者集中在行为的过程或者集中在行为的结果。例如,在学习实验中,研究者可以观察被试复述一系列单词多少次(过程),然后观察被试在最终测验中记住多少单词(结果)。对于直接的观察,研究的行为是清晰可见的、外显的、可记录的。例如,在情绪的实验室实验中,研究者能观察到当看能引起



研究者通过单向玻璃能观察到儿童不受影响或干扰的行为。当你知道你被观察时你曾经改变过你的行为吗?

情绪的刺激时被试的面部表情。

对研究者的直接观察经常存在技术上的争议。例如,当代心理学家经常依赖计算机,因为计算机能精确记录被试完成各种任务时的反应,如阅读句子或问题解决。尽管在计算机时代之前也有许多精确的测量方法,但在收集和分析精确的信息上,计算机具有非凡的灵活性。在第3章,我们将叙述最新的技术,这些技术能使研究者对不寻常的行为进行测量:工作中的脑图。

在自然观察中,不改变或干扰自然环境,研究者能观察到一些自然情况下发生的行为。例如,通过单向玻璃,研究者能观察儿童游戏,而儿童并没有觉知到被观察。一些人类的行为只有通过自然观察才能进行研究,因为在非自然条件下研究是不道德的或不切实际的。例如,研究生命早期的严重剥夺对儿童后期发展影响的实验就是不道德的。

在塑造复杂的行为模式中物种的自然栖息地所具有的长时效应,在实验室的人为环境中是观察不到的。珍妮·古道尔(Jane Goodall)的研究是自然观察法中最有价值的例子(1986, 1990; Peterson & Goodall, 1993)。古道尔花了30多年在非洲贡贝的坦噶尼喀湖研究黑猩猩的行为模式。古道尔知道,如果她只进行10年的观察——这是她最初的计划——她将无法得出正确的结论:

我们观察到黑猩猩的行为和人类有许多相似之处,但黑猩猩比人类更爱好和平给我们留下了深刻印象。由于在最初的10年后我们继续进行观察,我们能证明一个社会团体的分裂以及新分离的小团体间爆发的暴力侵略。我们



珍妮·古道尔花费了她一生中大部分的时间来观察黑猩猩。她发现，如果不在动物的自然栖息地观察动物，她将不能得到任何发现。

发现在一定情况下，黑猩猩可能杀死自己的同类，甚至吃同类的肉。另一方面，我们也观察到在它们家族成员间有着特别的、持久的情感联结……高级的认知能力，[和发展] 文化传统……（Goodall, 1986, pp. 3-4）。

在一项研究的初期，自然观察是特别有用的。它有助于研究者发现某一现象的范围，或者发现一些重要的变量以及变量间的关系是什么。自然观察得到的数据为研究者提供线索，这有助于明确表达假设或研究计划。

在结束心理测量这一话题前，我们必须强调许多研究计划结合使用自我报告法和行为观察法。例如，研究者可能特别寻找人们怎样报告他们的行为与人们实际上的行为表现间的关系（见17章）。另外，一些研究计划并不使用大样本被试，而只是用个案研究（case study）的方法对个体进行各种测量。有时对特殊个体进行透彻分析有助于理解人类经验的普遍特性。例如，在第3章你将知道对单个脑损伤患者进行仔细观察能为语言功能的脑定位这一重要理论提供依据。

我们已经叙述了研究者使用的几种程序和方法。下一节我们考虑使用这些程序和方法的道德标准问题。

人类和动物研究中的道德问题

在测验潜意识信息有效性的研究中，研究者通过错贴

磁带标签来欺骗被试。这么做的目的是为了看被试的预期是否能使他们相信信息是有帮助的，即使对记忆和自尊的客观测量并没有任何改进。从道德上来讲，欺骗总是令人怀疑的，但在这种情况下，其他方法怎能评估被试错误信任的安慰剂效应呢？在一项研究中怎样权衡得与失（包括被试被迫完成冒险的、痛苦的、有压力的或欺骗的程序所付出的代价）呢？心理学家也不断的询问自己这些正在形成的基本原则问题（Rosenthal, 1994）。

尊重人和动物的基本权利是所有研究者的基本义务。为了保证这些权利得到尊重，特别委员会监视每一个研究提议，强制实行美国健康和人类服务部发布的严格的指导方针。每个大学、学院、医院和研究院都有一个评审委员会来对关于人类和动物的研究计划进行评审。美国心理学会（1992）建立了关于研究者道德标准的详细的指导方针。这些方针和道德所关注的究竟有哪些问题呢？

获得赞同

对于以人为被试的实验室研究，一般在实验开始时都对被试解释实验程序、可能的风险以及他们将体验的益处。确保被试的隐私受到保护：严格保密被试的所有行为记录；公开发表的内容要得到被试的同意。要求被试表明他们已经被告知这些事宜，并同意继续实验。提前告诉被试他们可在任何时间停止实验，没有处罚，并且留给他们正式的名字和电话，如果他们有任何抱怨可随时联系。

冒险/收获的评估

大多数心理实验不需要被试冒险，特别是那些只要求被试完成常规任务的实验。然而，大多数人类本性中的个性研究，如情绪反应、自我形象、一致性、压力或攻击性等能令人苦恼或造成心理烦扰。因此，无论研究者什么时候进行这类研究，一定要把风险减低到最小程度，必须把这些风险告知被试，并且必须采取适当的防范来纠正错误的反应。无论涉及到任何风险，研究者都必须依照每条制度规定仔细权衡所冒风险与被试、科学和社会的受益情况。

故意欺骗

对某些类型的研究来说，事先告诉被试实验怎么进行会影响实验结果。例如，如果你研究电视中的暴力对攻击

性的影响，你将不想让被试事先知道研究目的。但是你有充足的假设证明这种欺骗是正当的吗？一些研究者指出任何欺骗与获得赞同的基本权利都是不相容的（Korn, 1987）。美国心理学会（1992）关于研究中欺骗的指导方针清楚地指出：（1）要使研究中的欺骗获得批准，研究必须具有充分科学的价值和教育的价值；（2）研究者必须证明除了欺骗没有任何同等有效的程序；（3）实验中影响被试参与自愿性方面的问题不欺骗被试；（4）根据研究结论必须对被试解释实验中的欺骗。对于有欺骗行为的研究，评审委员会可以施加强制规定，坚决监视最初提出的程序，否则不批准试验（Steininger et al., 1984）。

事后解说

参与心理学研究应该是一种研究者与被试相互之间的信息交换。研究者通过被试的反应获得某种行为现象中的新东西，而被试知道研究目的、假设、预期结果以及预期研究益处。实验结束后，给被试一份详细的事后解说（debriefing），在这份报告中研究者提供尽可能多的有关这个研究的信息，并且确保被试没有疑惑、没有心烦、没有尴尬。如果在实验中的某一阶段必须误导被试，实验者要认真对被试解释欺骗的理由。最后，如果被试觉得他们的数据被误用或他们的权利被滥用，他们有权收回数据。

动物研究中的问题：科学，道德规范，政治

在心理和医学研究中应该使用动物吗？这个问题经常引起两极化的反应。一类研究者指出在一些科学领域，可以使用动物进行非常重要的突破性的研究（Domjan & Purdy, 1995; Petrinovich, 1998）。动物研究有助于得到除了包括药物成瘾的重要知识，还包括治疗焦虑和精神疾病的一些发现和药物测试（Miller, 1985）。动物研究也有利于动物自己。例如，心理研究者已经指出如何减少动物园中的动物所体验的禁闭压力。关于动物学习和社会组织的研究已经导致了动物围栏的设计和设备的改进，这些都促进了动物的健康（Nicoll et al., 1998）。

对于动物权利的维护者来说，上述的成绩并不能减弱他们极端的错误信仰：从其他生物中分离出现代人类在道德上存在分歧（Bowd & Shapiro, 1993, p. 136; 也可见 Shapiro, 1998）。为了纠正这种错误，伦理学家赞成“从基于实验室的入侵性的研究向自然和半自然情景下最小程



使用动物作为被试时，要求研究者提供仁慈的环境。你认为科学上的收获能证明用动物做研究是合理的吗？

度的控制性研究过渡”（Bowd & Shapiro, 1993, p. 140）。每名动物研究者都必须依据详细审查来评估自身工作。

对 1 188 名心理学学生和 3 982 名美国心理学会成员对动物研究的态度调查结果表明，他们支持严格的审查标准（Plous, 1996a, 1996b）：

- ◆ 大约 80% 被调查的人认为自然情境中的观察研究是正确的。少部分人（30% ~ 70%）支持进行笼框或禁闭研究，这取决于动物种类（如，老鼠、鸽子、狗或灵长类动物）。学生和教授都反对造成心理痛苦或死亡的研究。
- ◆ 大多数（大约 60%）学生和教授都支持在大学心理学课上使用动物，但每组中只有 1/3 的人支持应该把动物的实验室研究作为主修大学心理学的必修课程。

与这些人相比你有什么主张呢？你会如何决定动物研究中的付出与收益问题？

小结

心理测量必须得出可信的和有效的发现。自我报告法通过问卷和访谈得以实施。心理学家依据兴趣使用行为测量法如直接观察或自然情境观察。人类研究中的重要道德问题包括获得赞同、研究风险和收益的评估、故意欺骗以及必要的任务报告。研究者必须检查非人类的动物研究的道德重要性。

成为明智的研究用户

本章最后一部分，我们将集中讨论几种关键的思考技巧，这些技巧有助于使你成为一名明智的心理学知识用户。对于任何处于不断变化社会中有责任的人，都有必要仔细推敲这些思考技巧，因为我们的头脑总是被所谓的真理、错误的“常识一样明白的”虚构的故事，以及服务于特别利益集团的带有偏见的结论所填满。一个有判别力的思考者要超越已有的信息，发掘隐藏在信息表面下的真正含义，以理解信息的本质为目标而不被表面的映像和风格所迷惑。

在这个充满复杂心理现象的社会中，心理学家的主张和断言始终是如何能思考、有感情和行动着的个人的日常生活的一部分。不幸的是，许多心理学信息并不来自于合格从业者的书籍、文章和报告，而是来自于报纸和杂志

上的文章、电视和广播中的节目、通俗的心理学自助类书籍。回到潜意识心理控制观点上。尽管这一观点最初被只关心利润的市场顾问詹姆斯（James M. Vicary）的传播所愚弄（Rogers, 1993）——如我们所看到的一样，在实验室里它已经不被信任——潜意识影响外显行为的这种观点持续对人们信仰和他们的钱包施加影响。

学习心理学有助于你依据自己和他人的证据做出英明的决策。你将总是努力把对正式心理研究的见解应用到你周围的非正式心理问题中：对自己和他人的行为产生疑问，根据合理的心理理论寻求答案，并且依据可利用的证据验证这些答案。

- ◆ 避免把相关关系推论为因果关系。
- ◆ 要求关键术语和概念有操作定义，并对其含义达成一致意见。
- ◆ 你很容易在寻求辩解时发现确定的证据，但在寻找确定的证据前，首先要考虑如何反驳某一理论、假设或

21 世纪的心理学

心理研究和信息爆炸

如果你上互联网，就会发现关于心理话题的范围是如此令人惊异。例如，我们写这些段落的那天上午，第一次搜索找到 83 760 页涉及精神分裂（见 15 章）；第二次搜索找到 63 812 页，毫无疑问，在出版社把我们这段文章印成教科书期间，有关这一主题的文章数目会变得更多。这两个数据清楚地表明了为什么许多人把我们生活的时代定义为信息爆炸的时代。所有人都面临着一个挑战，即如何成为明智的信息用户。你怎样确定网上信息来源哪些可信，哪些不可信呢？

在自然图书馆，人们很容易确定信息来源。大多数心理学研究发表在期刊上，这些期刊由一些组织如美国心理协会或美国心理学会出

版。研究原稿投到这些期刊后都要经历预审（a process of peer review）。每个稿件都主要寄给两到五名本领域专家，这些专家对稿件的基本原理、方法以及结果进行详细的分析。只有专家们非常满意的稿件才能在期刊上发表。例如，1998 年美国心理协会（1999）出版的期刊拒绝了 69% 投来的稿件。预审过程并不完美——毫无疑问会忽视一些有价值的研究，其中一些甚至被不公平地漏掉——但是，总的来说，这一过程还是确保了你在大部分期刊上读到的研究都是高质量的。

相比之下，你能很容易看到网上大多数信息的问题：你常常不能断定谁对网页所提供的建议或主张

进行了评估。当你接受网页上的信息时，你必须确定它的来源是可信的。一个好的方法是寻找那些在图书馆中可得到的杂志在线版本。同时，寻找研究者经常用来总结他们研究方案和列出相关出版物的网页。如果你对这些网页或其他网页上的信息感兴趣，你要努力找到他们列出的参考文献和出版物。通常，当网站的作者能指出这些信息的研究来源时，你应对网站提供的信息十分有信心。你应对《心理学与生活》一书中的结论有信心，因为本书为所有观点都提供研究引文。要查明网页的信息是不是处于上述同样的标准！



一次专家的新闻访谈由于除去了上下文可能使合理的结论受到歪曲，或曲解了研究结论的简要描述。你怎样才能成为明智的媒体报道的消费者呢？

信仰。

- ◆ 总是对已提出的明显解释寻求其他的可能解释，特别是那些有利于提案人的解释。
- ◆ 认识到个人偏见能歪曲对现实的理解。
- ◆ 怀疑对复杂问题给出的简单答案，怀疑对复杂现象和难题给出的单一理由和对策。
- ◆ 质疑任何关于治疗、参与或产品效果的声明，办法是找到比较结果的基础：比较什么？
- ◆ 成为思想开明而又好怀疑的人：认识到大多数结论都

具有尝试性和不确定性；寻找新的证据来减少你的不确定性，同时使自己能不断变革和修正自己的观点。

- ◆ 向权威挑战，那些权威通常用个人的观点代替证据，而且又不接受建设性的批评。

我们希望你以开放的心态和怀疑的态度来读《心理学与生活》这本书。我们不希望你把心理学学习仅看做是获得一些事实知识。相反，我们希望你能分享观察、发现以及测试所学观点的乐趣。

要点

重述

发现的背景

- 在研究的发现阶段，观察、信念、信息以及一般知识导致研究者用一种新的方式思考某一现象。研究者由此形成待检验的理论和假设。

验证的背景：客观性的保障

- 验证是在某种确定的程度上检验、证明或反驳观点的阶段。
- 研究者为了检验他们的观点，使

用科学的方法和一整套减少误差的收集和解释证据的程序。

- 研究者依据标准程序和使用操作定义，他们强烈反对观察者的偏见。
- 实验研究方法决定待检验的假设所确定的变量间是否存在因果关系。
- 研究者去排除其他可能的解释。
- 相关研究方法确定了两个变量是否相关以及有多大程度的相关。相关关系并不意味着因果关系。

心理测量

- 研究者努力提出可信和有效的测量方法。
- 心理测量包括自我报告法和行为测量法。
- 人类和动物研究中的道德问题。
- 尊重人和动物被试的基本权利是所有研究者的义务。已经制定了各种法律措施确保被试受到道德的、仁慈的对待。

■ 成为明智的研究用户

- 成为明智的研究消费者包括了解如何进行批判性的思考以及知道如何评估研究者提出的观点。

■ 关键术语

A-B-A 设计

行为测量

被试间设计

个案研究

混淆变量

发现的背景

验证的背景

控制程序

相关系数

相关法

事后解说

因变量

决定论

双盲控制

期望效应

实验法

假设

自变量

观察者偏见

操作性定义

安慰剂控制

安慰剂效应

总体

信度

代表性的样本

样本

科学的方法

自我报告法

标准化

理论

效度

变量

被试内设计

统计附录

理解统计：

分析数据和形成结论

■ 分析数据
描述统计
推论统计

■ 成为一个明智的统计学
用户
关键术语



在第2章中，我们注意到心理学家运用统计学来解释他们收集的数据，同时也运用统计为他们得出的结论提供量化的基础。因此，了解统计学的一些知识有助于认识心理学知识的获得进程。从个人角度来讲，对统计学有基本的了解有利于你通过分析数据能够在摇摆不定的观点和行为中做更好的决定。

大多数学生把统计学看做是乏味的、无趣的课程。然而，统计学在你的生活中有许多至关重要的应用。为了证明这个观点，我们将追踪一项研究，从追踪它怎样来源于真实世界到为得出一般的结论而使用统计学。这项研究最初要求对一个刊登在报纸头版的故事做出反应，这是一个关于畏缩的人突然成为杀人犯的故事。故事如下：

一组研究者怀疑害羞以及其他的个性特征与暴力行为间可能存在一定的联系（Lee et al., 1997）。因此，这些研究者开始着手收集可能揭示这种联系的数据。研究者推论，表面上非暴力却突然犯谋杀罪的人很可能是非常害羞的、不具有攻击性的人，他们能很好地控制自己的情感和冲动。在大部分生活中，他们都忍气吞声。无论他们感觉多么气愤，他们都很少表达出自己的愤怒。从外表看，他们表现出与自己不相干，但是他们的内心可能正在为了控制强烈的愤怒而斗争。他们给人的印象是恬静的、被动的、有责任心的儿童或成人。由于他们是害羞的，他们可能不让其他人接近，所以没有人知道他们的真正情感。然后，突然爆发出来。最轻微的刺激——很小的侮辱，轻微的拒绝，一点点社会压力——点燃了导火索，他们释放出长久以来累积的那些受压抑的暴力。由于他们没学会通过讨论和言语谈判来解决人际间的矛盾，所以这些突然的谋杀者们把他们的愤怒付诸物质行动上。

研究者基于这些推理做出假设：与习惯性的杀人犯相比，害羞更可能是突然谋杀者特有的特性，这些人杀人但先前没有暴力历史或反社会行为，而习惯性的杀人犯先前有暴力犯罪行为的记录。另外，突然杀人犯比习惯暴力分子更能控制自己的冲动。最后，与习惯犯罪的罪犯相比，

“亲戚、同事和熟人形容弗雷德·考恩是一个正派的、恬静的人，一个喜欢孩子的绅士，一个真正的和蔼可亲的人。考恩上小学时的地方学校校长介绍考恩从前在礼貌、合作和信仰方面得了A。根据他的同事的介绍，考恩“从来不谈论任何人，他是你能随意摆布的人”。然而，考恩令每个认识他的人都很惊讶，在情人节，他手提半自动步枪来到工厂，开枪射击杀死4名同事、一名警员，最后自杀。

这个故事有个通常的情节：一个害羞的、恬静的人突然变成杀人犯，令每个认识他的人都很震惊。弗雷德·考恩与其他突然从绅士转变成暴力残忍的人有什么共同之处呢？这些人的个性品质与我们有什么区别呢？

突然谋杀者的被动性和依赖性更被表现在女性化和双性化（androgynous）的特性上，标准的性别角色调查已测量了这一点。

为了检验这些关于突然谋杀者的观点，研究者已获得批准对加州监狱的一组杀人犯施测心理问卷。19名狱犯（都是男性）同意参加这项研究。犯谋杀之前，一些人已犯过罪，而样本中的其他人先前没犯过罪。研究者从这两类被试身上收集三类数据：羞怯分数、性别角色认同分数和冲动控制分数。

使用斯坦福羞怯调查（Stanford Shyness Survey）收集羞怯分。问卷中最重要的项目是询问被试是否害羞；答案为是或否。量表中其他项目询问害羞的程度和类型，以及有关害羞起因的各种维度。

第二个问卷是贝姆性别角色问卷（Bem Sex-Role Inventory, BSRI），给被试呈现一系列形容词，如好斗的、深情的，要求被试评估每个形容词适合描述自己的程度（Bem, 1974, 1981）。一些形容词典型地与女性化有关，这些形容词的总分为被试的女性化分数。其他形容词评估男子气概，这些形容词的总分为被试的男性化分数。女性化分数减去男性化分数做为最终性别角色分数，这个分数反映了被试女性化与男性化的差异。男性化与女性化

分数的结合表明了被试的两性化分数。

第三个问卷是明尼苏达多项人格测验 (Minnesota Multiphasic Personality Inventory, MMPI)，这个问卷用于测量个性的不同方面 (见 14 章)。这个研究仅使用“自我过度控制” (ego-overcontrol) 量表，这个量表测量了个体控制冲动的程度。被试这个量表的得分越高，被试越表现出过分的自我控制。

研究者预期，与先前有犯罪记录的杀人犯相比，突然杀人者 (1) 经常在羞怯调查中描述自己是害羞的；(2) 在性别角色量表中选择更多的女性化特性；(3) 自我过度控制的分数更高。他们发现了什么呢？

在你得出结论之前，你要理解一些分析数据的基本程序。研究者收集的真实数据将作为原始材料，用于教你一些不同类型的统计分析和一些可能得到的结论。

分析数据

对于大多数心理学研究者来说，分析数据是令人兴奋的一步——统计分析使研究者得以发现他们的预期是否正确。在这一部分，我们将一步步分析来自突然杀人者研究的数据。如果你已做了预见，那么你将看到一些数字和方程。记住数学是工具；数学符号是一种为了阐明观点和概念操作的简略表达方法。

原始数据——实际分数或其他测量——来自于突然杀人者研究的 19 名狱犯的数据列在表 S. 1。由表可看出，突然杀人组有 10 名狱犯，习惯性罪犯组有 9 名狱犯。乍看这些数据，研究者能感受到你所体会到的困惑。这些分数意味着什么？在各种人格测验中，这两组杀人犯有什么差异？仅仅检验这种无组织的数据排列是很难知道的。

心理学家依据两类统计方法：描述统计和推论统计来解释数据并得出有意义的结论。描述统计 (descriptive statistics) 在客观的、同一的方法基础上使用数学程序描述数字数据的不同方面。如果你曾计算过你的年级平均成绩，那么你就使用过描述统计。推论统计 (inferential statistics) 利用概率论做出可靠的推论：什么样的结果可能仅仅是由于随机变异而产生的。

描述统计

描述统计是数据的概要描述模式。它用于描述来自于

表 S. 1 研究突然杀人犯 的原始数据

狱犯	羞怯	BSRI	MMPI
		女性化—男性化	自我的过度控制
组 1: 突然谋杀犯			
1	是	+5	17
2	否	-1	17
3	是	+4	13
4	是	+61	17
5	是	+19	13
6	是	+41	19
7	否	-29	14
8	是	+23	9
9	是	-13	11
10	是	+5	14
组 2: 惯犯谋杀者			
11	否	-12	15
12	否	-14	11
13	是	-33	14
14	否	-8	10
15	否	-7	16
16	否	+3	11
17	否	-17	6
18	否	+6	9
19	否	-10	12

一个实验被试或几组被试的数据。它也能描述变量间的相关。因而，研究者不必努力记住每个被试的所有得分，而是要得到每组被试的典型分数指标。研究者也要测量典型分数是如何变化的——这些是发散的还是成族聚集在一起的。让我们看一看研究者如何推导这些测量。

频次分布

你将如何总结表 S. 1 的数据呢？为了清楚描述各种分数的分布情况，我们得出频次分布 (frequency distribution) 状态——总结每类分数出现的频次。羞怯分数很容易总结。19 个分数中，9 个是 10 个否；在组 1 中几乎所有的反应都为是，在组 2 中几乎所有的反应都为否。然而，自我过度控制和性别角色分数并不容易区分成是和否两类。为了解这些数字反应的频率分布，我们可以比较两组被试的信息，我们主要讨论性别角色分数。

看表 S. 1 中的性别角色数据。最高分是 +61 (最女性化)，最低分是 -33 (最男性化)。19 个得分中，9 个是

正的，10个是负的——这意味着杀人犯中有9人描述自己是相对女性化的，10人描述自己是相对男性化的。但这些分数在两组被试中是如何分布的呢？对一组数据进行频率分布统计的第一步是把分数从高到低进行等级排序。对性别角色分数的等级排序如表 S. 2。第二步把这些排列后的分数进行分类，组成一些数目更小的类别称为组距。本研究分了 10 类，每一类可能包括 10 个分数。第三步是建构频次分布表，由高到低列出组距并记录频次——落入每组的个数。通过频次分布我们能看出性别角色分数大部分在 -20 到 +9 间（见表 S. 3）。大部分狱犯的得分都不太偏离 0，即他们的得分既不十分正也不十分负。

现在已把数据排列成几类。下一步研究者将使用图解表示频次分布。

表 S. 2 性别角色差异分数的等级排列

最高	+61	-1
	+41	-7
	+23	-8
	+19	-10
	+6	-12
	+5	-13
	+5	-14
	+4	-17
	+3	-29
		-33
		最低

注：+ 分表示更女性化；- 分表示更男性化。

表 S. 3 性别角色差异分数的频次分布

类别	频次
+60——+69	1
+50——+59	0
+40——+49	1
+30——+39	0
+20——+29	1
+10——+19	1
0——+9	5
-10——-1	4
-20——-11	4
-30——-21	1
-40——-31	1

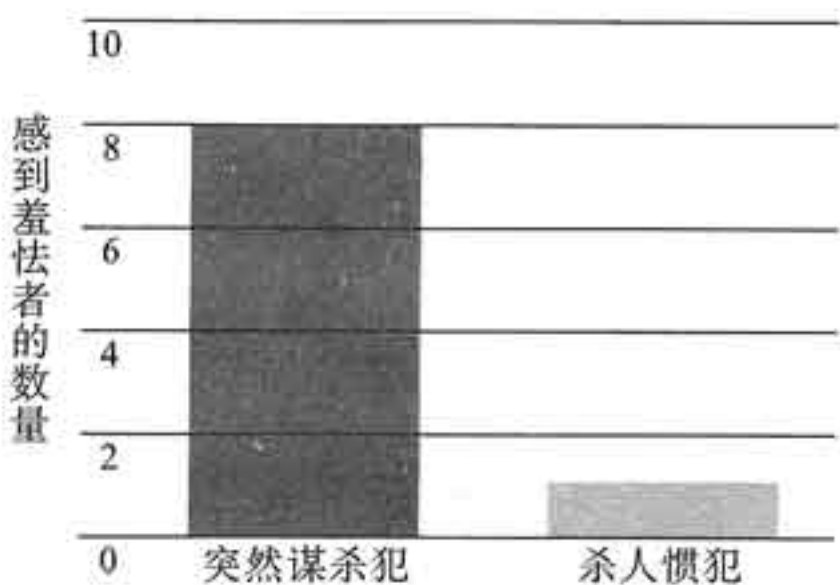


图 S. 1 两组谋杀犯的羞怯感得分（直条图）

图

当用图来表示数据分布时，常常可以让人更容易理解它们。最简单的一类图是直条图。这种图使我们能够看到数据中存在的模式。我们可以用一幅直条图表示出把自己描述为容易感到害羞的偶然杀人犯比杀人惯犯多出多少（见图 S. 1）。

对于更复杂的数据，如性别角色得分等，我们可以使用直方图。这种图类似于直条图，只是类别换成了间距——数据类别，而不是直条图中使用的名称类别。直方图以视觉方式提供了各间距中得分的数量。从直方图中的性别角色得分可以很容易地看出，这两组杀人犯的得分分布是有差别的（见图 S. 2）。

根据图 S. 1 和图 S. 2 可以看出，数据的基本分布情况服从研究者的两个假设。与惯犯相比，偶然的谋杀犯更可能会把自己描述成容易感到害羞，也更可能会用带有女性色彩的特质来描述自己。

集中趋势的度量

到现在为止，我们对这些数据的分布情况已经有了一个大致的印象。表和图增进了我们对研究结果的理解，但我们希望能了解更多——例如，最能代表这一组数据的数值。当我们比较两组或更多组数据时，这样的一个数值是非常有用的；比较两组数据中有代表性的数值要比比较整个数据分布容易得多。只用一个有代表性的分数来作为通过对组被试的测量所获得的多数典型分数的指标被称为**集中趋势的度量**(measure of central tendency)(它应该位于分布的中央，其他分数则分布在其周围。)心理学家主要使用三种不同的集中趋势测量方法：众数、中数和平均数。

众数(mode)是一个比其他数出现次数都要多的数值。对于羞怯感的测量值来说，突然杀人犯的众数反应是 yes——10 个人中有 8 个报告说自己是容易感到害羞的。而在惯犯当中，众数反应是 no。在突然的谋杀犯中，性别角色得分的众数为 +5。你能够计算出他们的自我过度

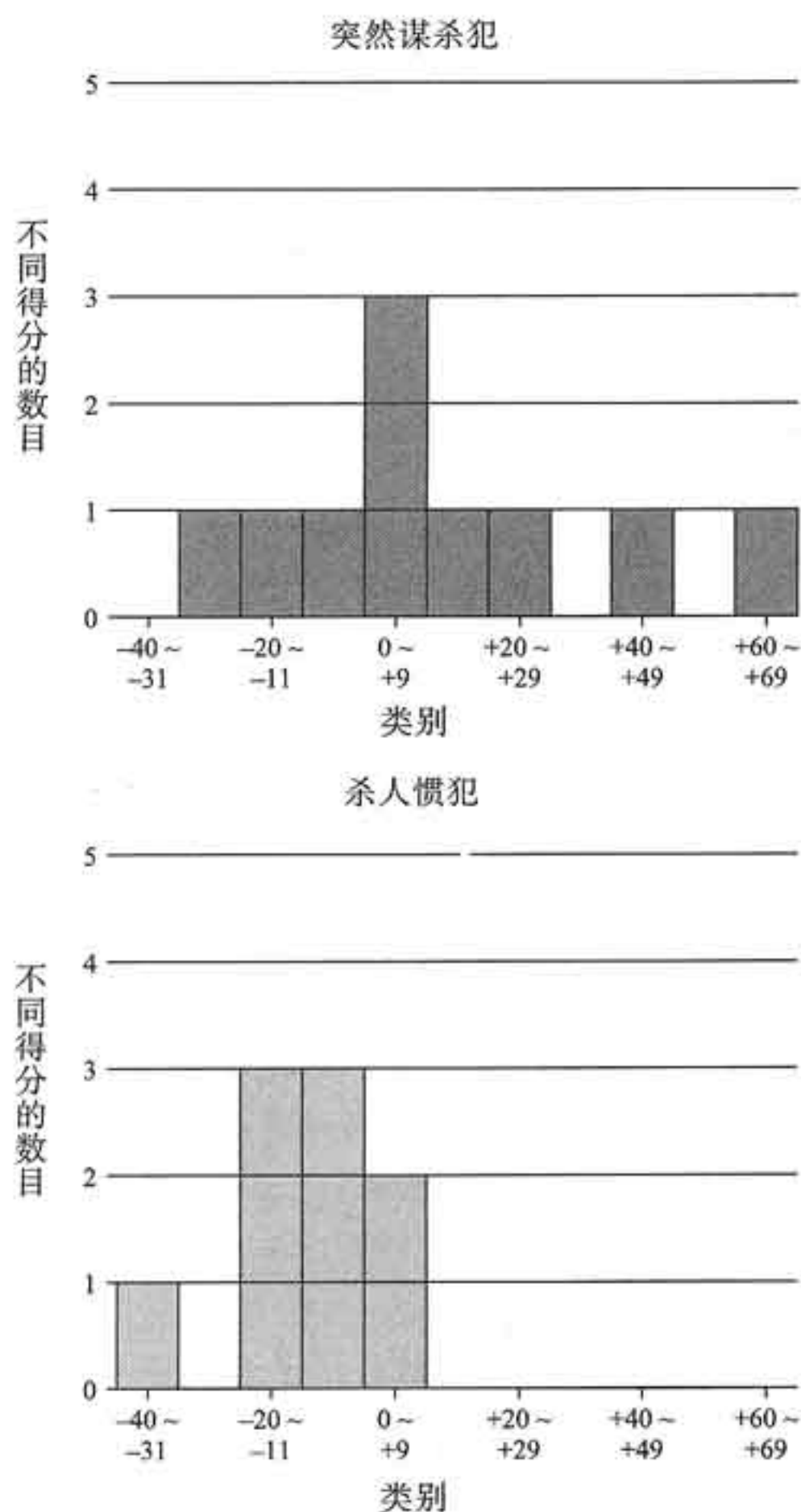


图 S.2 性别角色得分(直方图)

控制分数的众数吗？众数是最容易得出的集中趋势的指标，但常常又是用处最小的。如果你能够注意到过控得分中只有一个分数高于众数 17 但却有 6 个分数低于它这一情况的话，你就可能体会出众数用处很小的一个原因了。尽管 17 是频次最高的一个得分，但却不符合我们关于“代表性”或“集中趋势”的概念。

中数 (median) 更明显的也是一个代表集中趋势的度量；它将一组数据中高分的一半与低分的另一半区分开来。高出中数分数的数量与低于它分数的数量相等。当分数的个数为奇数时，中数是位于数据分布中间的那个分数；当分数的个数为偶数时，研究者常常以最中间的两个分数的平均值作为中数。例如，如果将杀人惯犯的性别角色得分按照高低顺序排列在单独的一张纸上，可以看出中数是 -10，分别有四个分数高于和低于这一数值。在突然

的杀人犯中，中数是 +5——第五和第六个分数的平均值，这两个分数恰巧都是 +5。中数不受极值的影响。例如，即使突然杀人犯中最高的性别角色得分是 +129 而不是这里的 +61，中数将仍然是 +5。这个分数仍然会把数据中高分的一半和低分的一半区分开来。中数始终处在数据分布的之间位置。

平均数 (mean) 是多数人听到平均这个词时常常会想到的。它同时还是最常用到的描述一组数据的统计量。要计算平均数的话，我们需要把所有数据加在一起，然后再除以这些数据的个数。这一操作可以用下面这个公式来表示：

$$M = (\sum X) / N$$

在该公式中， M 代表平均数， X 是单个的分数， $\sum$ (希腊字母 *sigma*) 表示把它后面的内容加在一起， N 则是所有分数的个数。由于所有性别角色分数的总和 ($\sum X$) 是 115，而分数的个数 (N) 是 10，所以突然杀人犯的性别角色得分的平均数 (M) 可以这样计算出来：

$$M = 115 / 10 = 11.5$$

可以试着自己计算一下这些犯人的过控得分的平均值。结果将会得到 14.4。

与中数不同，平均数会受到数据分布中特定分数的影响。改变某个极值的数值的确会改变平均值。例如，如果 4 号罪犯的性别角色得分是 +101 而不是这里的 +61，那么整组罪犯的分数平均值就会从 11.5 增加到 15.5。

离散性

除了了解哪一个分数最能代表整个数据分布外，了解这种集中趋势度量的代表性究竟如何也是很有用处的。其他分数大部分距离它很近还是非常分散？离散性的度量 (measures of variability) 是描述围绕在某些集中趋势度量周围的分数分布情况的统计量。

你能看出为什么关于离散性的度量非常重要吗？举个例子有助于说明这一点。假定你是一名小学教师。现在是一学年的开始，你将教授 30 名小学生阅读。了解到本班的儿童一般能够阅读一年级水平的课本，这将有助于你安排自己的课程。不过，如果你还了解这 30 名儿童的阅读能力的相同或不同程度的话，你就可以安排得更好。他们是否处在同一个水平上呢（低离散性）？如果是这样的话，那么你就可以安排一门很标准的二年级的课程。但如果有一些儿童能够阅读更深的材料而其他人却几乎不能看懂它们（高离散性）的话，该怎么办呢？平均水平现在已

经不能够代表整个班级的情况，因此你需要安排多种课程以满足这些儿童的不同需要。

关于离散性的最简单的一类度量是**全距**(range)，即频率分布中最高值与最低值之间的差值。对于突然杀人犯的性别角色得分来说，全距是 90：(+61) - (-29)。过控得分的全距则是 10：(+19) - (+9)。在计算全距时，我们只需要知道两个数值：最高值和最低值。

全距易于计算，但心理学家常常更喜欢用那些更敏感的、将所有数据都考虑进来而不是只考虑极端数值的度量。一个普遍使用的度量是**标准差**(standard deviation, SD)，它代表着所有分数与其平均数之间的平均差值。要计算标准差，我们需要知道数据的平均值和单个的具体分数。一般的步骤包括先用各个分数减去均值，然后再确定出这些离均差的平均值。公式如下：

$$SD = \sqrt{\frac{\sum(X-M)^2}{N}}$$

你应该能够根据计算平均数的那个公式认出这里的大多数符号。(X-M) 这个表达式的意思是“分数减去平均值”，通常称之为离均差。先从各个分数中减去平均值，然后将得到的结果进行平方（以消除负值）。把这些

表 S.4 计算突然杀人犯自我一过控分数的标准差

分数 (X)	离均差 (得分减去平均数) (X-M)	离均差的平方 (得分减去平均数) ² (X-M) ²
17	2.6	6.76
17	2.6	6.76
13	-1.4	1.96
17	2.6	6.76
13	-1.4	1.96
19	4.6	21.16
14	-.4	.16
9	-5.4	29.16
11	-3.4	11.56
14	-.4	.16
标准差 = SD = $\sqrt{\frac{\sum(X-M)^2}{N}}$		86.40 = $\sum(X-M)^2$
		$\sqrt{\frac{86.40}{10}} = \sqrt{8.64} = 2.94$
		SD = 2.94

数值相加(Σ)，然后再除以观测值的数目(N)，就得到了方差的均值。这个符号要求我们对包括在符号内的数值取平方根以抵消前面的平方操作。表 S.4 中计算出了突然杀人犯过控得分的标准差。回忆一下，这些分数的平均值是 14.4。由此，它就是那个必须用各个分数来进行相减以得出相应的离均差的数值。

标准差可以告诉我们一组分数的离散程度。标准差越大，则数据分布越分散。突然杀人犯中性别角色得分的标准差是 24.6，惯犯的标准差却只有 10.7。这表明，惯犯组的数据离散程度要低一些。与突然杀人犯相比，他们的分数距离平均值更紧密。当标准差很小时，平均数是整个数据分布的一个很好的代表值。而当标准差很大时，它对整组数据的代表性将减小。

相 关

解释心理学研究数据的另一个有用的工具是**相关系数**(correlation coefficient)，它是关于两个变量（如身高与体重或者性别角色得分与自我一过控得分）之间相关程度和性质的一个度量。相关系数可以告诉我们在某种度量上的分数与另一种度量上的分数之间的联系程度。如果在某个变量上获得高分的人倾向于在另一个变量上也获得高分，那么相关系数将为正值（大于 0）。如果在一个变量上获得高分的多数人在另一个变量上却倾向于得到低分，相关系数将会为负值（小于 0）。如果两个分数间不存在一致的关系，则相关系数将接近于 0（也请参见第 2 章）。

相关系数的取值范围从 +1（完全正相关）到 0 再到 -1（完全负相关）。在两个方向上离 0 越远，两个变量之间或正或负的联系就越紧密。相关系数越高，则根据一个变量的信息，可以更好的预测另一个变量。

在突然杀人犯中，性别角色得分与过控分数之间的相关（以 r 表示）为 +0.35。因此，这两者之间是正相关的——总的来说，那些认为自己具有更高的女性气质的人同时还倾向于具有高过控特征。不过，与可能出现的最高相关 +1.00 相比，这一相关只有中等水平，因此我们知道在这两者之间的关系中有很多例外情况。如果我们同时还测量了罪犯的自尊心，并且在过控分数与自尊心得分之间发现有 -0.68 的相关，这将意味着这两者之间存在着负相关。如果确实是这样的话，我们就可以说，那些具有获得高过控得分的个体往往自尊心较低。这将是一个比性别角色得分与过控得分之间相关系数更高的相关，因为 -0.68 比 +0.35 距离 0 这一代表着没有相关的点更远一些。

推论统计

我们已经使用了很多描述统计量来描述突然杀人犯研究所得到的数据，现在我们对这些结果的情况有了一定的了解。不过，仍然有一些基本问题没有得到回答。回忆一下，研究者曾经设想突然杀人犯可能比惯犯更容易感到害羞、更具有过控特征以及女性气质更重。在用描述统计比较了这两组罪犯的平均反应与离散程度后，看起来他们之间确实存在着一些差异。但我们怎么样才能够知道这种差异已经大到足够的程度而具有一定的意义了呢？如果我们以其他一些突然杀人犯和惯犯重复这项研究的话，能够预期会得到同样模式的结果吗？或者，已经得到的那些结果只是随机现象的产物？如果我们能够用某种方法对全部突然杀人犯与惯犯进行测量，所得到的平均值和标准差会和我们用小样本研究所得到的结果相同吗？

推论统计被用来回答上面这些问题。它们能够告诉我们可以根据样本研究做出什么样的推论，以及根据我们的数据可以合理地得出什么结论。推论统计利用概率论来确定一组数据完全由随机变化所得出的可能性。

正态曲线

要想理解推论统计是如何进行的，我们首先必须来看一下一种称之为正态曲线的数据分布的特别之处。当从大量个体身上收集关于某个变量（如身高、智商或过控性等）的数据时，数据的个数常常符合一条大致类似于如图 S.3 所示的曲线。注意，这条曲线是左右对称的（左半部分是右半部分的一个镜像），呈钟型——中间高，多数数据处于这一位置，离均值越远，曲线高度越低。这类曲线称之为正态曲线(normal curve)，或者叫做正态分布。（偏态分布是一条数据集中在某一端而不是围绕在中央位置的曲线。）

在正态曲线中，中数、众数以及平均数都是同一个数。可以预测，某个分数所占的具体百分比将落在曲线的不同区间里。图 S.3 给出了斯坦福—比奈智力测验所得到的智商分数。这些分数的均值是 100，标准差为 16。如果以沿坐标底线距离平均值的距离来表示标准差的话，我们可以发现，在平均值 100 上下 1 个标准差的范围内集中了 68% 多一点的分数。剩下的分数中，约有 27% 位于平均值下第一和第二标准差（IQ 分数 68 和 84 之间）以及平

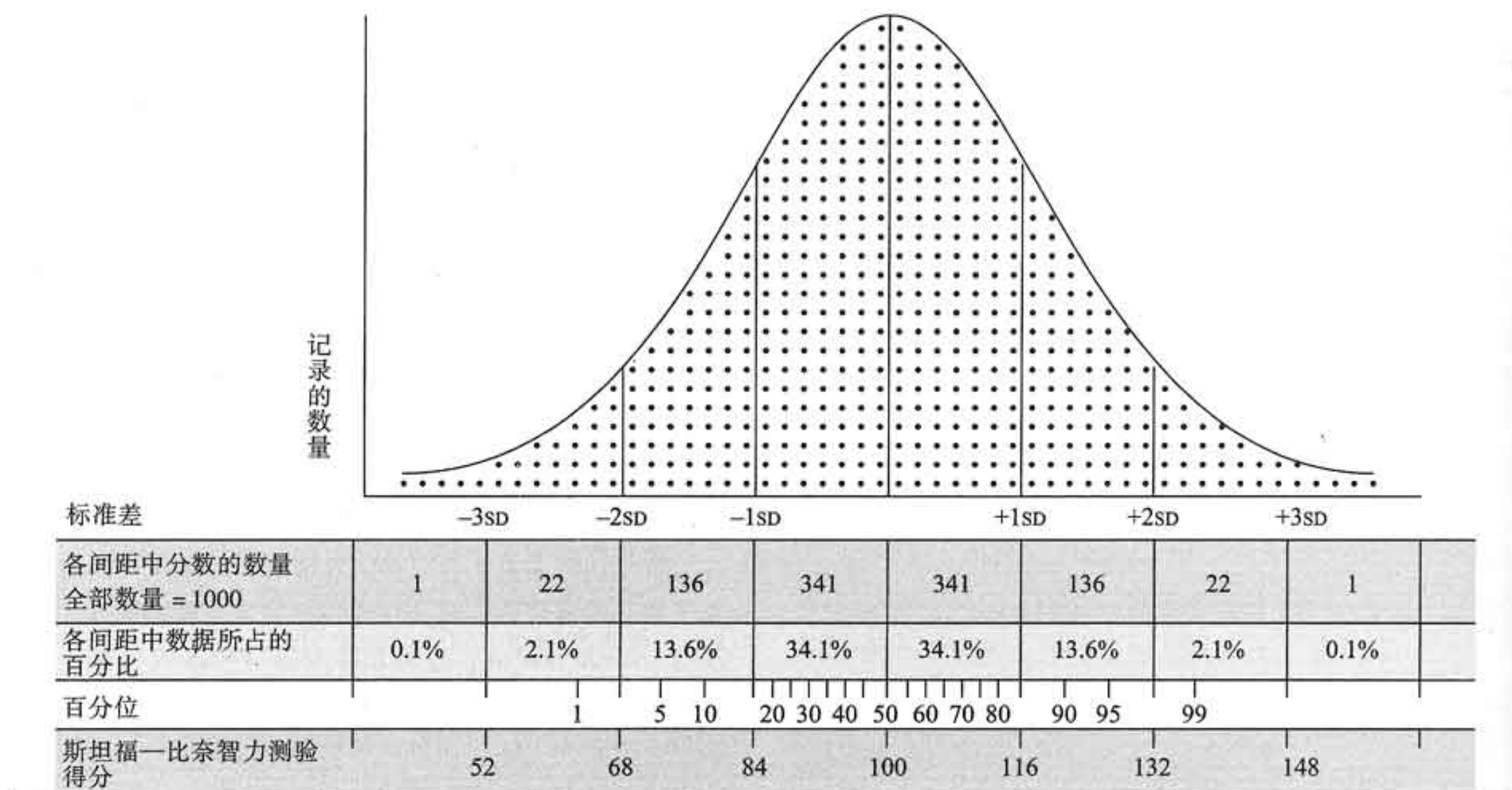


图 S.3 正态曲线

均值上第一和第二标准差 (IQ 分数 116 和 132 之间) 之间。另有不到百分之五的分数落入高于和低于平均值的第三个标准差范围内。很少有数据落在这个范围以外——只有约 0.25%。

推论统计指明了所得到的特定分数样本与你所要测量的内容之间存在着真正关联的概率,或者说它们是否只是由随机因素所产生的概率。例如,一个人的智商更有可能是 105 而不是 140,但 140 的智商出现的可能性又要高于智商 35 出现的可能性。

正态曲线也可以通过收集一系列完全由随机因素造成差异的度量来获得。如果你把同一枚硬币连续抛 10 次并分别记录下正反面朝上的次数,就很有可能各得到 5 次——多数时候是这样。如果像这样连续抛 100 组的话,有可能会得到一些全部为正面朝上或没有正面朝上的情况,也会得到更多的次数介于这两个极端数值之间的情况,其中最多的是正面或反面为五次左右的情况。如果把这 1 000 次抛硬币的情况做成一幅图的话,将会得到一条非常吻合正态的曲线,正如图中所示的那样。

统计显著性

当研究者在两组样本的均值之间发现有差异后,他肯定想知道这是一个真正的差异还是只是因随机因素而出现的差异。由于随机差异服从正态分布,研究者可以用正态曲线来回答这个问题。

举个简单的例子可以说明这一点。假定你的心理学教授希望知道监考人的性别对于男女学生的考试成绩是否会有影响。为此,教授随机把一半学生分配给一位男监考老师,另一半分配给一位女监考老师。他然后比较了两组学生的平均成绩。这两个平均数很可能会非常相似;如果有一些微小的差异,也很可能是随机出现的。这是为什么呢?如果只有随机因素在起作用且两组学生来自同一个总体(无差异)的话,那么男监考老师和女监考老师这两组样本的分数均值在大多数时候都应该是相当接近的。根据正态分布中不同区间里分数所占的百分比,我们知道,男监考老师组中只有不到三分之一的学生成绩会比女监考老师组中学生的平均成绩高出或低出一个标准差。男监考老师组中学生的平均成绩比女监考老师组中学生的平均成绩高出或低出三个标准差的可能性是非常小的。如果某位教授确实得到了这么大的一个差异,那么他可以相当有把握地认为这是一种真实的差异,在某种程度上与监考老师的性别有关。接下来的问题将是这个变量是如何影响测验成绩的。

如果男女学生被随机分配给两类监考人,就有可能来

分析监考人之间出现的差异在两组学生中是否一致或者只是局限于某一性别的学生。假定有数据表明,男监考老师给女学生的评分高于女监考老师的评分,而两者对男学生的评分却相同。教授就可以利用统计推论程序来估计某个观测到的差异因随机因素而出现的概率。这种计算的依据是差异量的大小和数据的分布情况。

根据已经达成的一致认识,当由随机因素导致的概率不足 5% (以 $p < 0.05$ 来表示) 时,心理学家将接受这个差异为“真”。显著差异 (significant difference) 是指符合这一标准的差异。不过,在某些情况下可能还会使用更严格的概率标准如 $p < 0.01$ (100 中不足 1 个) 和 $p < 0.001$ (1 000 中不足 1 个)。

当出现了统计意义上显著的差异时,研究者可以对所考察的行为做出一个结论。有很多不同的检验方法可以用来估计数据的统计显著性。选择何种检验方法将取决于研究设计、数据类型以及样本的大小。我们在这里只介绍一种最常使用的检验,即 t 检验。当研究者希望知道两组数据的均值差异是否达到统计上的显著性时,就可以使用这种检验。

我们可以用 t 检验来考察突然杀人犯的性别角色得分均值与惯犯的得分均值之间是否有显著差异。它使用一种数学程序来验证你可能已经根据图 S.2 得出的一个结论:两组罪犯性别角色得分的分布差异已经达到了为“真”的程度。如果我们做一项恰当的计算——它将两个均值之间的差异作为这些均值之间离散性的函数来进行处理——我们会发现,如果不存在真正差异的话,获得这样一个 t 值的概率很小,不足 5% ($p < 0.05$)。因此,这种差异在统计上是显著的,我们可以很肯定地认为这两组罪犯之间存在着一个真正的差异。与惯犯相比,突然杀人犯确实把自己评价为女性气质更重一些。另一方面,两组罪犯过控分数的差异并没有达到统计的显著性 ($p < 0.10$),因此我们在讨论这个差异时必须非常谨慎。在研究者所预测的方向上出现了一个趋势——这是一个在 100 次中会出现 10 次的差异。但是,这一差异并不在标准的 5% 的范围内。(在用另外一种统计检验对得分频率进行分析时,害羞感的差异达到了显著水平。)因此,通过使用推论统计,我们能够回答在开始进行研究时提出的一些基本问题,能够更深入地了解那些突然从性情温和、容易害羞的人变成杀人犯的个体心理。但是,任何结论都只是一种关于所研究的事件之间可能存在的相互关系的论断;它永远不会是确定性。科学研究中的真理都是暂时的,总会有后来更好的研究资料而被修正,由更好的假设而发展。

成为一个明智的统计学用户

既然我们已经学习了什么是统计、如何使用统计以及它们的意义，就应该简要地讨论一下它们可能被错误使用的情况。很多人往往会接受那些没有得到证实而只是由关于某种统计量的权威个人的看法所支持的“事实”。其他人则会简单地相信或者否定统计处理的结果，而根本不知道该如何考察那些用来支持某种产品、政治家或提案的数据。在第2章末尾，我们曾经提出了如何成为一位明智的研究使用者的建议。根据本章对统计学的简短介绍，我们可以将这些建议拓展到进行统计判断的情景中。

在进行统计分析时，有很多方面会给人造成一些有误导的印象。研究工作的各个阶段中所做的决策——从如何选择被试到如何设计研究、选择何种统计方法以及如何进行统计等各方面——对于从这些数据中得出的结论都有着复杂的影响。

在报告研究结果时，使用哪些被试往往会造成很大的差异，而这种差异很容易被忽略。例如，在调查人们关于堕胎权的观点时，在南部一个原教旨主义小社区里做这项研究会得到和在纽约市的大学里做这项研究很不同的结果。同样，职业生活组在调查成员的意见时所得出的结论很可能会和职业选择组在做同一调查时所得到的结论不同。

即使通过一定的方法对被试进行了随机选择而没有出现偏倚，如果不能满足统计学的基本假设，统计分析也会

得出一些误导结果。例如，假定有20个人参加智力测验；其中有19个人的得分在90~110之间，另有1人的得分是220。这组分数的均值将会因为这个偏离多数而极高的分数而被大大地抬高。在这类数据中，中数或众数能更准确地反映这组人的平均智力，而平均数却会使得结果看上去好像是这组人中的一般人都具有高智商。但另一方面，如果这组人的数量是2000而不是20的话，一个极端数值并不会造成很大的影响，此时平均数将是对该组人智力的一个合理概括。

避免出现此类错误的一个好方法是检查样本的大小——大样本比小样本出现误导结果的可能性要小。另一个方法是同时检查中数、众数以及平均数——当这三者近似而不是有很大差别时，可以更有把握地对结果进行解释。我们应该始终仔细地检查所使用的方法和报告出的研究结果。要注意研究者是否报告有样本大小、离散性的度量以及显著性水平。试着去检查他们所使用的方法是否准确且一致地测量了他们自己声称正在研究的内容。

统计是心理学研究的支柱。它们被用来理解观测到的结果，并被用来确定这些发现是否正确。运用我们介绍的这些方法，心理学家能够整理出数据的频次分布，并计算出这些分数的集中趋势和离散性。他们还可以用相关系数确定出几组分数间相互联系的强度和方向。最后，心理学研究者能够确定出观测值对总体的代表程度以及它们是否与总体有显著差异。统计学也有可能被错误地使用，误导那些不了解的人。但通过正确且合乎道德地应用统计学，它们将使得研究者能够扩展关于心理学的知识。

关键术语

相关系数
推论统计
众数

描述统计
平均数
正态曲线
频次分布
集中趋势的度量

全距
离散性的度量
显著差异
中数
标准差(SD)