样本量论证：多大的样本才合理？

Daniël Lakens

Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, Netherlands

原文信息：

译者/校对者列表：张译文（复旦大学心理学系）； 张火垠（深圳大学心理学院）；郑元瑞（昆明城市学院教育学院）；吴博文；李阳萍（西安交通大学外国语学院）；付芮嘉（陕西师范大学心理学院）；冯睿（华东师范大学认知神经科学研究所）；敖敦托雅（奥斯陆大学心理学系）；陈晓云（英国兰卡斯特大学心理学系）；王欣雨（南京师范大学心理学院）；胡传鹏（南京师范大学心理学院）

本文 word 版本链接：github.com/OpenSci-CN/OpenTransfer4Lakens/tree/main/Sample%20Size%20Justification

如何给我们反馈：

（1）在 GitHub 上提新的 issues；

（2）给以下邮件进行邮件反馈：王欣雨，xinyuwang28@outlook.com

本译文获得 Daniël Lakens 博士的允许和支持，在开放科学中文社区（Chinese Open Science Network，COSN）的开放翻译(OpenTransfer)小组的组织下完成，特此说明。更多关于 COSN 的信息，见 Jin, H., Wang, Q., Yang, Y.-F., Zhang, H., Gao, M., Jin, S., et al (2023, AMPPS, doi:10.1177/25152459221144986)

【摘 要】 我们进行实证研究时，一个重要的步骤是确定所需样本量的大小。而进行这一步骤的关键目的在于回答如下问题：**对于研究者要进行的(统计)推断目标来说，即将收集的数据在何种程度可以提供有价值的信息。**本文讨论了六种在定量实证性研究中用来论证样本量合理性的方法：1）对(近乎)总体进行数据收集，2）因有限的资源选择样本量，3）进行先验检验力分析来确定样本量，4）为达到研究所需的精确度而进行的计划，5）使用直觉的方法，或 6）直接承认没有进行任何的合理性论证。在论证样本量的合理性时，需要考虑的一个重要问题是，**哪些效应量是研究者们所感兴趣的，以及所收集的数据在多大程度上为我们对这些效应量的推断提供信息**。依据其所选的论证样本量合理性的方法，研究者们可以考虑以下几点：1）感兴趣的最小效应量是什么；2）哪个最小效应量将会在统计上显著，3）研究者期望哪些效应量的出现（以及这些期望的依据），4）根据效应量的置信区间，哪些效应量会被拒绝，5）根据检验力灵敏度分析，当前研究能检测的效应量在什么范围，以及 6）在某个具体的研究领域，预期会出现什么样的效应量。研究者可以使用本文提出的指南来改进其对样本量合理性的论证过程，例如，通过本文附带的在线应用程序，或许会因此而将某个研究所提供的信息与其(统计)推断目标一致。

一般来说研究者进行实证研究，收集有助于回答研究问题的数据。所收集的数据越多，为研究问题所提供的信息量也就越大。对样本量合理性的论证应该以**（统计）推断目标**为基础，推断目标可能是估计效应量的大小，或者是检验一个假设。虽然在论文提交指南、基金申请、伦理审查等中，都会要求进行样本量合理性的论证，但是，观测数据的样本量大小通常只是被***陈述***出来，而非经过***论证***得出。这让我们难以评估研究的信息量。为了避免问题的发现为时太晚（例如，得到不显著的假设检验结果)，研究人员应该在**数据收集前**仔细论证其样本量的合理性。

论证样本量合理性的六种方法

研究者通常很难论证其样本量的合理性（样本指的是被试量、观察值的数量及两者的组合）。本综述将讨论六种可用来回答定量研究中这个问题的方法。本文不求面面俱到，但涵盖了单项研究[1]所能用到的最常见且适用的方法。第一种论证样本量合理的方法，整个总体或者几乎整个总体的数据都被收集到了。第二种论证样本量合理的方法**以资源有限为核心**，资源有限是非常常见的，但很少明确地对其进行评估。第三种和第四种论证立足于研究者所期望的**统计检验力或精确度**。第五种论证则依赖于直觉，最后一种则是不做任何论证，直接选择一个样本量。以上每一种方法提供的合理性可强可弱，取决于研究者想要从计划采集的数据中推断出什么结论。

表 1，论证单个研究样本量合理性的方法概览

论证的类型	适用情况
测量整个总体	研究者能够具体化总体的情况，总体本身是有限的，（几乎）每个个体都可被测量
资源受限	资源受限是限制研究者所能收集到的样本量的首要原因
精确度	研究问题关注某个参数的大小，研究者收集足够的数据使估计值达到想要的精确度
先验检验力分析	研究目的是检验能否以理想的统计检验力在统计学上拒绝某些效应量
直觉	研究者根据文献描述或口头交流中获得的直觉、一般规则或常规做法来决定样本量
不做论证	研究者没有为选择特定的样本量说明任何理由，或没有明确指定的推断目标，且没有

[1] 关于元分析的检验力分析话题不在本文论述范围内，但可参考 Hedges 和 Pigott（2001）以及 Valentine, Pigott 和 Rothstein（2010）。

诚实地表达这一点

以上论证样本量合理性的方法——即使是“不做论证”的方法——都能让其他人了解到决定样本量大小的原因。显然，“直觉”和“不做论证”都不太可能给同行留下好印象。需要指出的是，这些论证的价值取决于我们从中获得了多少能够回答如下问题的信息：**最终的样本量在多大程度上允许研究者进行他们预期的研究推断？**也就是说，选择哪种具体的方法本身并不是关键。

上述论证方法在多大程度上能帮助研究者评估“数据”是否有信息，还取决于研究者在确定样本量时所提出的研究问题本身的细节及他们所选择的参数。一个非常差的先验统计检验力分析立刻会让研究变得只有非常低的信息量。当然，上述这六个论证方法并非互斥，设计一项研究时可以考虑多种方法并用。

六种方法，评估哪些效应量值得关注

我们收集的数据包含多大的信息量，这取决于研究者或同行（在某些情况下）所设定的推断目标。本文所考虑的各种不同的推断目标都有一个共同点，那就是研究者需要辨别出哪些**效应量**是有意义的。这也就意味着研究者需要对这些效应量进行评估。评估方式依赖于（某些方法的）统计特性以及某领域知识之间的相互结合。表 2 提供了六种可能的评估方式。**但本文并不是要做一个全面介绍，而是提供一些易上手、常见且实用的方式。**需要说明的是，这些评估方式并不是都与样本量的论证相关。由于这些评估都依赖相同的信息（如效应量、观察次数、标准差等），因此这六种评估方式可被视为一套互补的方法，来用于评估哪些效应量是值得引起关注的。本文附带的在线应用程序为广大研究者提供了一个可交互的表格，希望能够指导研究者完成样本量论证时需要注意的事项。

表 2，评估哪些效应量值得关注的方法概览

评估方式	研究者所面临的问题
感兴趣的最小效应量	从理论或实践层面上来说，有意义的最小效应量是多少？
统计上可得的最小效应量	给定测量方式和样本量，达到统计显著的临界效应量是多少？
预期效应量	根据理论预测或前人的研究，预期的效应量是多少？
置信区间宽度	在围绕着效应量设立置信区间时，有哪些效应量应当被排除在外？

灵敏度功效分析	在一系列潜在的效应量中，哪种效应在进行假设检验时最敏感？
某研究领域的效应量分布	在某个特定的研究领域，效应量的一般范围是多少?哪些效应是本来就不太可能被观察到的?

首先，研究者应当考虑他们所感兴趣的最小效应量是多少。第二，尽管这只与假设检验相关，研究者仍应当考虑哪些效应量会在统计上显著（在既定的显著性水平 α 和样本量之下）。第三，重要的是考虑预期效应量的范围，这需要思量预期效应的来源以及其中可能存在的偏差。第四，围绕总体可能的效应量来设置置信区间的宽度是有益的，因为我们可能需要用置信区间来拒绝其他可能的效应。第五，在做灵敏度功效分析时，应当广泛地估量潜在效应量的统计检验力。第六，应当考虑已发表的相关研究中效应量分布。

信息价值

几乎所有研究者都会面临资源的限制，因此他们需要在成本（收集额外数据的花费）与效益（额外数据所带来的的价值）之间进行权衡。通常，这被称为**“信息价值”**（Eckermann, Karnon, & Willan, 2010）。然而，衡量信息价值是极其困难的（Detsky，1990）。研究者不仅需要明确收集数据所需的成本，还需要去权衡这些成本所带来的收益。从信息价值的角度来看，并不是每个可收集的数据都具有同等的价值（J. Halpern 等，2001；Wilson，2015）。每当额外的数据并没有为推断目标提供更多的价值时，那么所花费的成本就会超过所需收益。

在大多数情况下，额外的信息（数据）并不止是发挥单一的作用，尤其是涉及多个推论目标时。研究者可能希望将某一效应与前人研究中所发现的大效应量或者基于理论预测得出的中等效应量以及具有实践意义的最小效应分别进行比较。在这种情况下，由于样本信息的价值差异，因此将导致**每个推理目标的最佳样本量不同**。在研究中，收集关于某个**大效应量**的数据信息是非常有价值的，随着数据的增加，额外数据的边际效益逐渐减少，甚至可能变为负值。然而，针对**中等效应量**的信息价值来说，继续收集数据到一定程度，额外数据的信息价值再次增加。对于**较小效应量**的信息价值来说，随着数据继续增加，会再次出现边际效益递减的情况，直到研究收集的数据越来越能够提供有关最小效应大小是否存在的信息价值（如果效应不存在，收集更多的数据也是无意义的）。

由于难以量化信息的价值，研究者在研究中通常使用不够规范的方式来证明他们样本量的合理性。尽管成本-效益分析在报告样本量合理性时总是模棱两可的，但信息价值的观点几乎隐含在所有论证样本量的理论框架中。故在随后对样本量论证的讨论中，我们将反复强调在（统计）推断目标

下信息价值的重要性。

测量（几乎）总体数据

在某些情况下，有可能会从总体收集全部（几乎）数据。例如，研究者可能会使用人口普查数据库，来收集某公司所有职工的数据，或者研究一小部分顶尖运动员。只要有可能测量总体，样本量论证的缘由就变得直截了当，因为研究者获得了所有可用的数据。

当测量总体时，就不需要进行假设检验了。毕竟，此时不存在可推论的群体[2]。当收集了总体数据后，那么总体的效应量就已知了，且不需要计算置信区间。如果总体规模已知，但并未测量全部的数据，那么随着被试量逐渐趋近于目标总体，置信区间宽度也将逐渐缩小到零。这被称为**估计方差的有限总体校正系数**(Kish, 1965)。样本均值的方差为$\frac{\sigma^2}{n}$，而在有限总体中，它要乘以标准误的有限总体校正系数:

$$FPC = \sqrt{\frac{(N-n)}{(N-1)}}$$

其中 N 为总体大小，n 为样本量。当 N 远大于 n 时，校正系数将趋近于 1(因此，当总体非常大时，这种校正通常被忽略，尽管总体是有限的)，并且不会对方差产生显著影响。当测量总体时，校正系数为 0，这样方差也为 0。例如，当总体为 100 名顶尖运动员时，采集的样本量为 35 名运动员，有限总体校正系数为$\sqrt{(100-35)/(100-1)} = 0.81$。superb R 包可以计算校正后的总体置信区间(Cousineau & Chiasson, 2019)。

有限的资源

在研究中，常因资源限制而使得数据无法合理收集(Lenth, 2001)。因此实际上，样本量总是受到资源的限制，纵使它并不是决定样本量的主要缘由，但也总是次要缘由。

尽管资源限制无处不在，但样本量的问题在实验设计环节却没有得到足够的重视(一个例外的例子，可见 Bulus 和 Dong(2021))。这导致人们觉得承认资源限制是不合时宜的，但事实并非如此，资源限制是一个很常见的问题，所以一个负责任的研究者在设计研究时，应当仔细评估资源限制所带来的影响。同样的，资源限制的评估也是**数据收集的成本**与**数据信息价值**之间的权衡。即使研究者没有明确量化这种权衡，但这也会在他们的实际操作中体现出来。例如，研究者很少把所有的资源都投在某一项研究上。鉴于资源有限，研究者们面临的是如何在多个研究项目上优化资源的分配。

时间和经费是所有研究者都要面临的限制。一个博士生有一定的时间来完成一篇博士论文，但

[2] 从某种程度上来说，即使测量了总体，我们仍然在做推论，因为我们测得的是众多可能性中的一个潜在的总体，见 Spiegelhalter（2019）。

通常在这段时间内也需完成多条研究线。除了时间限制，研究者的经费也很有限，这往往直接影响到数据收集的数量。在某些研究领域也存在的第三种限制，比如在研究患有罕见疾病的患者时，可能本身能够获取的数据量就非常少。总而言之，将**有限资源的优化**置于规划样本量的首要位置，并从研究者的可用资源出发。将其在既定的时间、成本内转化为研究者预期的样本量(N)。但问题就在于如何评估这 N 个测量值是否值得。如何判断一项研究是否提供了充足的信息，以及判断数据收集从何时起将没有意义?

当衡量资源受限是否使数据收集信息不足时，研究者需要明确他们在收集数据时的**推断目标是什么**(Parker & Berman, 2003)。但有数据总比没有数据强，所以从某种意义上来说，所有收集到的数据都是有价值的。然而，收集数据所花费的成本可能超过了数据所带来的价值。

无论有或没有数据，在确实需要做出决策时，对数据收集是否具有价值进行评估是最直接的方法。在这种情况下，任何额外的数据都将减少决策过程的错误率，哪怕只是一点点。例如，在没有数据的情况下，让我们猜测两种条件中哪个条件的真实平均分数更高，显然我们的猜测不会比猜抛硬币来的更加准确。但有了一些数据后，我们可以选择具有更高平均值的条件，并以此做出更准确的决策。尽管拥有少量的数据，但我们仍可能犯错误，但错误率比没有任何数据要小。在这些情况下，**只要错误率的降低优于数据收集的成本**，那么信息价值可能便是正向的。

小样本数据体现价值的另一种方式是，它可能将会被用来进行**元分析**(Maxwell & Kelley, 2011)。这种方式需小样本满足以下要求：1）研究者需公开数据，使得这些数据可用于日后的元分析；2）这些数据在未来有相当大的概率可被用于一个高质量的元分析(s．d．Halpern, Karlawish &Berlin, 2002)。然而，关于未来是否会有这样的元分析是不确定的，所以需要将这种不确定性与数据收集的成本进行权衡。

提高未来元分析可能性的一种方法是，由研究者自己来进行，他们可将进行的几项研究融合成一个小规模的元分析中(Cumming, 2014)。例如，研究者可能计划在接下来 12 年的授课中重复一项研究，并期望在 12 年后，对这 12 项研究进行元分析来求得有效推论(参考 ter Schure 和 Grünwald(2019))。此外，如果一个研究者无法自己收集所需数据，他们也可以尝试建立一个合作网，让同一领域内的其他研究者使用相同方法收集类似的数据。如果随着时间的推移，仍然不可能出现足够的数据来证实推断目标，那么对数据的收集就毫无意义。

即使研究者认为收集数据是有意义的，因为将来会进行元分析，同样他们也很可能会对当前数据进行统计分析。为了确保他们对分析结果的预期是准确的，首要考虑哪些**效应量**是有趣的，并进行灵敏度功效分析，以此估计感兴趣的效应出现**Ⅱ类错误**的概率。我们可从六个方面来评估效应量的意义，稍后会在本文的第二部分进行讨论。我们需要斟酌能达到统计显著的**最小效应量**，围绕效

应量的置信区间可能的宽度，以及在特定领域中可预期的效应。并在灵敏度功效分析中评估以上效应的**统计功效**。如果已经确定好了研究问题，那么可以考虑使用**折中检验力分析**来确定合适的错误率。

对资源受限的样本量估计进行阐述时，建议先从表 3 中提到的五个因素入手。明确地解决这些问题有助于评估数据是否值得收集。为了清晰地解决所有相关问题，可以在 https://shiny.ieis.tue.nl/sample_size_justification/找到一个交互式表格来进行。

表 3，基于资源限制进行样本量规划时的建议概览

需要解决什么？	怎么来解决？
日后是否会有相关的元分析？	考虑到未来可能会有相当多的类似研究出现，这就使元分析变得更有可能实现。
是否会考虑在现有数据（不论其是否可用）的情况下做出结论?	如果做出了决策，那么收集的任何数据都将降低错误率。需考虑使用折中检验力分析来确定 I 类和 II 类错误的错误率。为了降低错误率所付出的代价值得吗?
临界效应量是多少？	报告和解释临界效应量的大小，重点关注预期的效应量是否能产生显著的结果。如果不能，则表明对数据的解释将不能拘泥于 p 值。
置信区间的宽度是多少？	报告并解释置信区间的宽度。有这样不确定性的估计会有怎样的作用?如果零假设为真，那么拒绝置信区间外的效应是否值得(忽略低统计检验力的实验设计可能导致无法拒绝这些效应的情况)?
哪些效应量有良好的统计检验力？	报告灵敏度功效分析，并报告期望检验力水平范围内(例如，80%、90%和 95%)可以检测到的效应量大小，或绘制灵敏度分析图。

先验检验力分析

若一项研究以是否存在统计学意义上的显著为目标时，研究者往往希望确保他们的样本量足够大，以免对他们所关心的效应量做出错误结论。在上述论证样本量合理性的方式中，信息价值在于收集数据，从长远来说，可以收集数据直到得出错误结论的概率小于一个期望值。此时，如果研究者进行假设检验，有四种可能的结果:

1. 假阳性（I 类错误），由 α 水平决定。即使零假设为真，也会得到显著结果。

2. 假阴性（II 类错误），由 β，或 1-power 决定。即使备择假设为真，也会得到不显著的结果。

3. 真阴性，由 1-α 决定。当零假设为真时，得到不显著的结果。

4. 真阳性，由 1-β 决定。当备择假设为真时，得到显著的结果。

在既定的效应量、α 水平和统计检验力下，可以使用先验检验力分析来计算某一效应[3]所需要的样本量（在期望错误率之下）。图 1 给出了在双侧 α 水平为 0.05 的独立 t 检验中，统计检验力如何随着样本量（每组）的增加而增加。如果我们对 d＝0.5 的效应感兴趣，则每个条件下 90 的样本量将为我们提供超过 90% 的统计检验力。可以用统计检验力来确定被试的数量或项目的数量(Westfall, Kenny, & Judd, 2014)，同样也可以通过单个案例研究进行计算(Ferron & Onghena, 1996; McIntosh & Rittmo, 2020)。

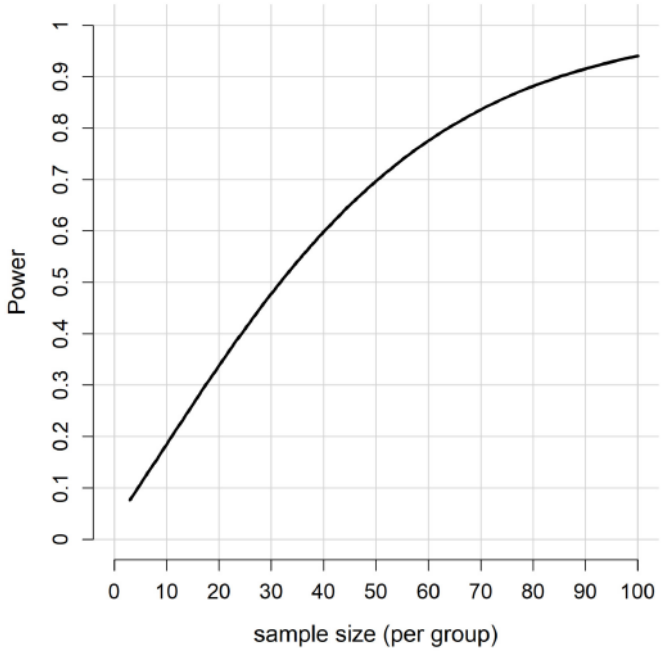

图 1，效应 d=0.5, α =0.05 的独立 t 检验的统计检验力曲线与样本量的关系

尽管普遍将 I 类错误设为 5% 、统计检验力设为 80%，但它们的设定应该是有理有据的(Lakens et al., 2018)。如“折中检验力分析”部分所述，默认 80%的统计检验力也缺乏依据。一般来说，错误率越低（从而统计检验力越高），研究的信息量越大，但需要的资源也就越多。在想要达到 90% 或 95% 检验力的研究中，研究者应考虑样本量的成本与降低错误率的收益之间的关系。除此之外，研究者还应考虑是否计划发表一篇由重复和拓展以往研究而构成的文章，在这种情况下，观察到多个 I 类错误的概率将非常低，但即使存在真实效应，得到混淆结果的概率也会增加（Lakens & Etz, 2017），这也是研究 II 类错误低的一个原因，鉴于此，或许可以稍微提高每个单项研究的 α 水平。

[3] 统计检验力分析可以根据标准化的效应量或前人的研究得到的效应量来进行。了解效应的标准差是很重要的（见 "了解你的测量方法 "一节），但我认为在论证样本量的情境下，探讨标准化效应更为方便。

图 2 呈现了两个分布。左边的分布（灰色虚线）以 0 为中心。这是**零假设**的模型。如果此时零假设为真，那么只有在极端效应量的情况下（在正向或负向的双侧检验中）会得到统计意义上显著的结果，但所有显著结果都犯了 I 类错误（曲线下深灰色的区域）。如果效应不存在，则零假设显著检验的统计检验力是不确定的。也就是说，在既定的 α 水平下，如果零假设为真，任何所得显著结果都是 I 类错误或假阳性。

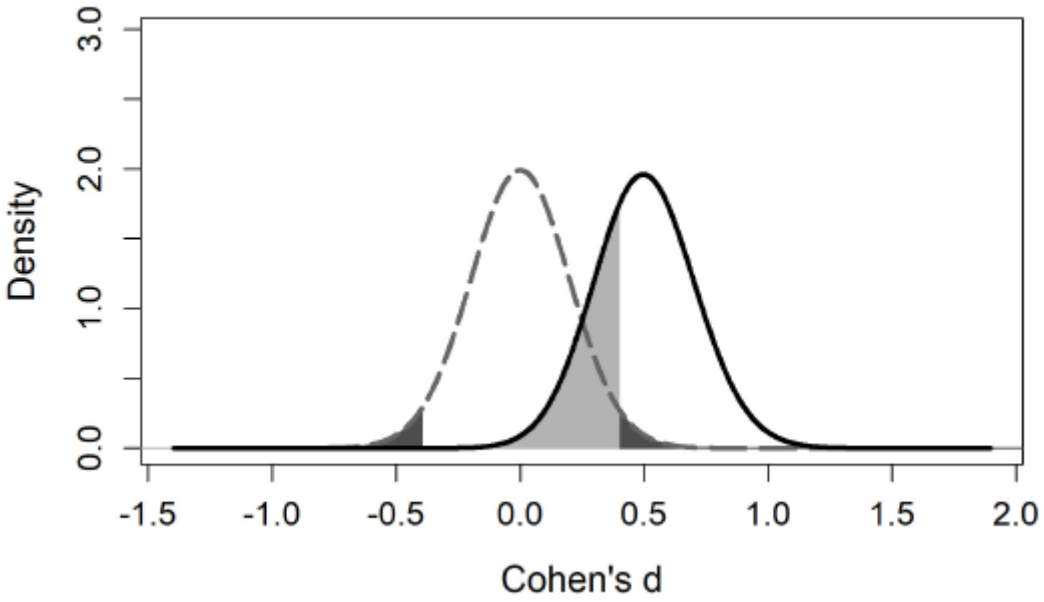

图 2，零假设（d=0,灰色虚线）和备择假设(d=0.5,黑色实线)，且 α =0.05，n=80 每组。

右边的分布（黑色实线）以 d＝0.5 的效应为中心。这是**备择假设**的模型，说明如果备择假设为真时，预期效应量为 d = 0.5。但是即使存在真正的效应，研究也不一定能得到具有统计学意义上的显著结果。由于**变异的随机性**，所得效应量接近于 0 而不具有统计学意义时，就会发生这种情况。这样的结果是**假阴性**（右侧曲线下方的浅灰色区域）。为了提高检验力，我们可以增加样本量。随着样本量的增加，分布变得更窄，从而降低了发生 II 类错误[4]的概率。

需要强调的是，先验检验力分析的目的并不是为真实效应提供足够的统计检验力。真实的效应量是未知的。先验检验力分析的目的是，**在研究者想要探究特定的效应量（小、中、大）时，能够获得足够的统计检验力**。就像 I 类错误是在零假设为真的条件下发生 I 类错误的最大概率一样，先验检验力分析是假设在特定效应量下进行计算的。且这个假设正确与否不得而知。研究者所能做的就是确保他们的假设是合理的。**研究（对 II 类错误进行控制）的统计推断是以特定效应量的假设为基础的。**他们允许推断，假设真实效应量至少与先验检验力分析中使用的一样大，则研究中的最大

[4] 上述图片可以在一个在线应用程序中进行复刻和调整：http://shiny.ieis.tue.nl/d_p_power/。

II 类错误率不高于先前的期望值。

如果在研究时，我们对“有”效应和“无”效应都进行了先验检验力分析，也许就能更好的进行说明。在设计研究时，必须考虑无效应的可能性（例如，平均值的差异为零）。**先验检验力分析既可以用于零假设显著性检验，也可以用于零效应的检验**，例如等价检验，可以通过拒绝“有”效应来为零假设提供统计学上的支持（Lakens, 2017; Meyners, 2012; Rogers, Howard, & Vessey, 1993）。当对同一样本进行多个预实验时，每次分析都需要独立的对样本量进行论证。如果可能的话，要确保收集的样本量（每次）为所有实验提供信息，也就是说，收集的样本量是基于全部先验检验力分析所得到的最大样本量。

例如，如果一项研究是以 90% 的统计检验力来接受或拒绝 d＝0.4 的效应，并且将双侧独立 t 检验的 α 水平定为 0.05，则研究者需要在每个条件下收集 133 个被试进行假设检验，以及每个条件 136 个被试进行等价检验。因此，研究者应争取收集 272 名被试进行两种检验来更好的做出结论。但这并不能保证某研究具有足够的统计检验力来获取真实效应（永远无法得知），但它保证了研究者具有足够的统计检验力来接受或拒绝关于某效应的假设。因此，只要研究者能够证明他们感兴趣的效应量是合理的，先验检验力分析就是有用的。

如果研究者在检验多个假设时矫正了 α 水平，则先验检验力分析应基于矫正后的 α 水平。例如，如果进行了四次检验，I 类错误率为 5%，进行 Bonferroni 校正后，则先验检验力分析应基于 0.0125 的校正 α 水平。

先验检验力分析可以通过分析法或计算模拟来进行。分析法速度更快，但灵活性较差。当研究者试图对更复杂的或不常见的检验方法进行统计检验分析时，他们将面临的一个共同挑战是，当前软件不能提供现成的解决方案。在这种情况下，**计算模拟**可以为任何检验方法提供一个灵活的解决方案（Morris, White, & Crowther, 2019）。以下代码是在 R 语言中进行计算模拟的示例，它对样本量为 20 的单样本 t 检验进行了 10000 次模拟迭代，此时假设真实的效应量为 d＝0.5。所有的模拟都由给定规则的随机数据组成（例如，均值为 0.5、标准差为 1 的正态分布），然后对数据进行检验。通过计算显著结果的百分比，就能算出任何检验方法的统计检验力。

```
p <- numeric(10000)   # to store p-values
for (i in 1:10000) {  #simulate 10k tests
  x <- rnorm(n = 20, mean = 0.5, sd = 1)
  p[i] <- t.test(x)$p.value # store p-value
}
sum(p < 0.05) / 10000 # Compute power
```

有各种各样的工具都能用来统计检验力分析。无论研究者决定使用哪种工具，都需要时间来学习如何正确使用该软件，再进行有意义的先验检验力分析。针对心理学领域进行检验力分析的教学资源非常丰富（Aberson, 2019; Cohen, 1988; Julious, 2004; Murphy, Myors, & Wolach, 2014），例如，有概括性的介绍（Baguley, 2004; Brysbaert, 2019; Faul, Erdfelder , Lang, & Buchner, 2007; Maxwell, Kelley, & Rausch, 2008; Perugini, Gallucci, & Costantini, 2018），以及现在越来越多的关于特定检验的实用教程（Brysbaert & Stevens, 2018; DeBruine & Barr, 2019; P. Green & MacLeod, 2016; Kruschke, 2013; Lakens & Caldwell, 2021; Schoemann, Boulton, & Short, 2017; Westfall, Kenny, & Judd, 2014）。学习统计检验力相关的基础知识非常重要，对学会如何执行基于计算模拟的检验力分析也大有裨益。我们也建议寻求专家的帮助，特别是当研究者对某特定检验的统计检验力分析缺乏经验时。

报告先验检验力分析时，请确保**统计检验力分析是完全可重复的**。如果在 R 中进行统计检验力分析，则可以共享分析脚本和相关包版本的信息。在许多软件中，可以直接将统计检验力分析导出为 PDF 文件。例如，可以在 G*Power 的“protocol of power analysis”选项下导出分析。如果软件没有提供导出的方法，请在补充文件中提供统计检验力分析的截图。

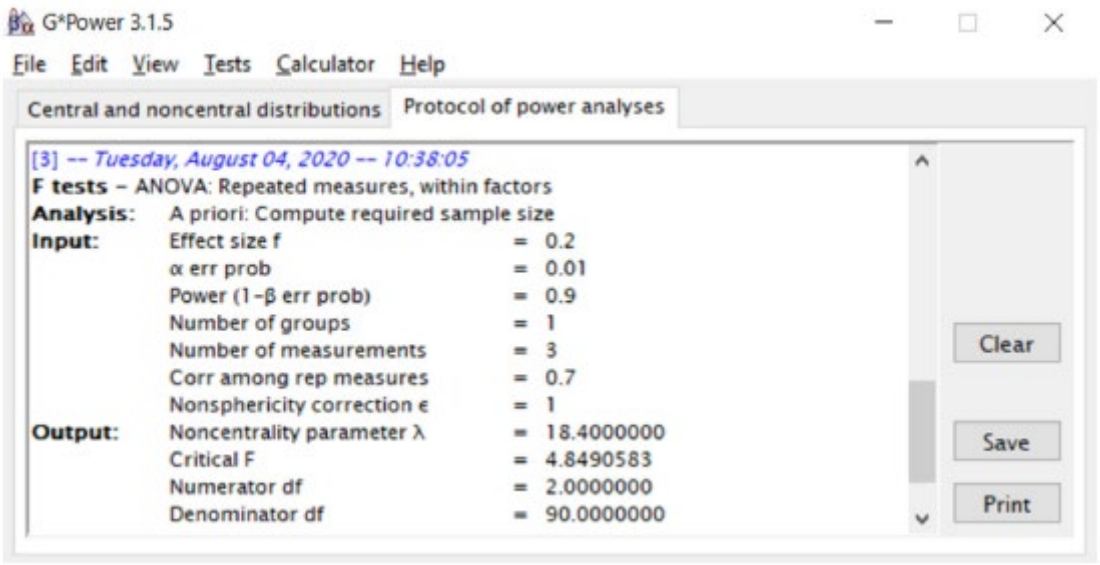

图 3，关于统计检验力分析的所有细节都可以在 G*Power 中导出。

有关可重复性的报告中需要附上**参数选择的理由**。如果检验力分析中使用的效应量是基于前人的研究，可依据表 5（如果效应是基于元分析）或表 6（如果效应是基于单项研究）中的观点进行探讨。如果效应量的估计基于现有文献，还请提供完整的引文，最好是直接引用进行样本量论证的文章。如果效应量是基于感兴趣的最小效应量，则不仅需要说明该值，还需证明该值是合理的（例如，基于理论预测或实践意义，请参见 Lakens, Scheel, and Isager (2018)）。有关先验检验力分析时应报告的所有方面，请参考表 4。

表 4，报告先验检验力分析时的建议

需要参考的事项？	需要怎么做？
列出计划进行的主要检验	对检验进行说明，且指明需要控制 I 类和 II 类错误
明确 α 水平	列出每个检验的 I 类错误并说明理由。确保在必要时对多重比较进行校正。
期望的统计检验力是多少？	列出每个检验的期望统计检验力（或者 II 类错误）并说明理由
对于每一个检验力分析，要说明效应量类型，效应量大小以及针对该效应量进行检验力的理由	报告效应量类型（如 Cohen's d，Cohen's f），效应量大小（如 0.3），选取该效应量的理由，以及是基于感兴趣的最小效应量、元分析估计的效应量还是以往单项研究估计的效应量，或其他来源。
考虑零假设为真的可能性	进行统计检验力分析，以考察是否存在有意义的效应（例如，等价检验的统计检验力）
确保统计检验力分析是可重复的	包括用于统计检验力分析的代码，或导出一份统计检验力分析的详细报告。

表 5，检验力分析中使用元分析来估计效应量时的建议

需要参考的事项？	需要怎么做？
与元分析中的研究是否相类似？	元分析中的研究在实验设计，测量方法以及被试选取方面与你所计划的研究非常相似吗？评估效应量估计值在您研究中的可推广性。
与元分析中的研究是否同质？	与元分析的研究是否存在异质性？如果是，请采用与你研究最相关的同质研究进行效应量估计
效率量的估计是否无偏？	原始研究是否报告了偏差检验测试(bias detection tests)偏差存在与否？如果存在，那么明智的做法可能是使用偏差校正手段来保守的估计效应量，并同时承认矫正后的估计效应量不等价于元分析的效应量估计

表 6，检验力分析中使用单项研究来估计效应量时的建议

需要参考的事项？	需要怎么做？
研究是否足够相似？	考虑研究之间在总体、实验设计、实验操纵、测量方法或其它方面的差异，从而导致期望效应有所不同。
是否存在偏差的风险？	如果效应量估计值较小，评估您不会采用（或不会发表）它的可能性。在统计检验力分析时，比较校正偏差前后的估计效应量的差异。
不确定性有多大？	被试量较少的研究具有更大的不确定性。考虑使用更加保守的效应量估计的可能性，以降低真实效应量的统计检验力不足的可能。

规划精确度

一些研究者建议根据所需的估计精确度水平来论证样本量的合理性（Cumming & Calin-Jageman, 2016; Kruschke, 2018; Maxwell, Kelley, & Rausch, 2008）。基于精确度来论证样本量合理性时，其目的是得到围绕**参数估计置信区间的宽度**。参数估计的置信区间宽度取决于标准差和数据量。研究者根据精确度来论证样本量合理性时，唯一需要论证清楚的是与推理目标相关的置信区间的理想宽度，以及他们对总体标准偏差的假设。

如果研究者已经确定了所需的精确度，并且对真实的标准偏差有个很好的估计，则可以直接依据精确度水平计算所需的样本量。例如，在测量一组被试的 IQ 时，从长远来看，研究者可能希望用围绕均值 2 个标准差之内的 95%的数据来估计智商分数。达到这种精确度水平（假设数据呈正态分布）所需的样本量可以通过下式计算：

$$N = (\frac{z * sd}{error})^2$$

其中 N 是观察次数（样本量），z 是与所需置信区间相关的临界值，sd 是总体 IQ 的标准差，error 是在期望错误率下，均值应落入的置信区间的宽度。在本例中，(1.96 × 15 / 2)^2 = 216.1 个观测值。如果研究者希望有 95% 的数据落在真实总体均值 2 个误差范围内，则应收集 217 个观测值（样本）。如果计算出非零均值差异（non-zero mean difference）的所需精度，则精度是非中心的 t 分布。对于这些计算，需要对期望效应量进行估计，但对所需样本量的影响相对较小（Maxwell, Kelley, & Rausch, 2008）。也可以将效应量的不确定性纳入样本量的计算中，这可称为“万全之策”(assurance)

（Kelley & Rausch, 2006）。 R 中 MBESS 包提供了各种函数来估计检验的样本量（Kelley，2007）。

但模棱两可的是**所需的精确度水平与推理目标之间关系**。没有任何文献可以帮助研究者选择所需的置信区间宽度。Morey (2020)说，大多数规划精确度的实例中都涉及到将所得效应与其他效应区分开的推断目标（关于贝叶斯观点，请参阅 Kruschke (2018)）。例如，研究者可能期望 $r = 0.4$ 的效应量，并会以不同的方式处理其相关性差异超出 0.2 的情况（因为 $0.2<r<0.6$），因为 $r = 0.6$ 或更大的效应也将会认为太大了，不可能是由假设的实验操纵所带来的（Hilgard，2021），而小于 $r = 0.2$ 的效应又被认为太小，从而无法支持理论预测。如果目标确实是为了获得一个足够精确的效应量估计，以便高概率的区分上述两种效应，那么推理目标实际上就是假设检验， 这就需要设计一个具有足够统计检验力的研究来拒绝效应（例如，检验相关性在 0.2 到 0.6 之间的预测范围）。

如果研究者不想检验假设，例如，可能他们更喜欢估计的方法而不是检验的方法，那么在没有明确的指导方针来帮助研究者确立精确度水平时，一种可行的方案可能是参照一个**公认的精确度标准**。该标准可能基于某一明确惯例（certain resolution），低于该惯例时，某研究领域中的测量将不再有差异性的结果。正如研究者通常使用 0.05 的 α 水平一样，他们可以对研究进行规划，在遵循惯例之下，得到期望效应的置信区间的理想宽度。那么未来的工作就是需要帮助研究者合理规划精确度置信区间的理想宽度。

以经验法则来确定样本量（Heuristics）

当研究者采用经验法则来确定样本量时，他们可能是因为自己无法评估样本量是否合理，但他们倾向于相信权威机构推荐的样本量。当我在 2005 年开始攻读博士时，大家通常会在每个条件下收集 15 名被试。当被问及为什么这么做时，没有人可以给出确切的答案，但人们相信这一做法在某篇文献中进行了合理的解释。但现在我意识到我们使用这样约定俗成的标准其实是没有任何的理论基础的。正如伯克利（1735）所说的：“人们从他人那里学习科研方法（准则）：每个学习者或多或少都会对权威有所敬畏，尤其是年轻的学习者，很少有人愿意在这些原则性的问题上纠结很久，而是倾向于遵循原则：从某种程度上来说，早期被承认且又经过重复的东西会变得具有说服力：这种“说服力”最终变为了“证据”。

关于样本量的选择，一些文章为研究者提供了一些简单的经验法则。这类文章显然满足了人们的需求，并且被大量引用，即使这些文章中提供的建议存在纰漏。例如，Wilson VanVoorhis 和 Morgan（2007 年）将 S.B.Green（1991 年）提出收集约 50 个观测值的经验法则进行了修改，改为建议在回归分析中至少使用 50+8 个观测值。实际上，Green 在他的文章中总结道:“总的来说，没有具体理论支持最小样本量或具有预测性的被试的 subjects-of-predicotrs 最小比率。”他在文中确实探讨了 $N = 50$

+ 8 这一经验法则如何为社会科学中的“典型”研究提供具体的最小观测量（样本量），因为这些研究具有“中等”效应量，正如 Green 引用 Cohen（1998）所提及的那样。但实际上 Cohen 并没有声称典型的社科研究具有‘中等’效应大小，而是说(1988，第 13 页):“许多在人格、社会和临床心理学研究中寻求的效应很可能是这里定义的小效应。”现在我们可看到一连串错误的引用最终是如何创造了一个具有误导性的经验法则。

经验法则似乎主要源于**错误的引用**和/或**过分简化的建议**。Simonsohn、Nelson 和 Simmons（2011）建议“作者必须对于每个条件收集至少 20 个观察值。”后来，同一作者在一次会议上提出了 n>50 的建议，除非你研究大的效应（Simmons、Nelson 和 Simon-sohn，2013）。遗憾的是，这一建议现在经常被错误地引用来解释在不考虑预期效应量的情况下，每个条件下不超过 50 个被试的理由。如果作者根据另一篇论文中的一般性建议来论证某个样本量（例如，n=50）的合理性，则可能是他们错误引用了该文章，或者他们引用的文章本身存在缺陷。

另一个常见的经验法则的方式是收集和前人研究相同的样本量。但有以下情况时，不建议采用这种方法，即在某领域普遍存在发表偏差时，或者主要是探索性的单项研究有新颖发现时。这种方法只有一种适用情况，那就是前人研究选取样本量的缘由也适用于当前的研究，这样的使用方法才是有效的。与其说你打算收集与前人研究相同的被试量，不如重复论证样本量的合理性，然后根据新发现进行更新（与前人研究的效应量讨论类似，见表 6）。

同行评审和编辑应该仔细审查采用经验法则来论证样本量的文章，因为这些经验法则可以让研究看起来（对于推断目标）具有很高的信息价值，即使该研究的结果并无更多的参考价值。每当遇到使用经验法则方式来规划样本量时，需要问问自己：“为什么使用这种方法？”知道其背后的逻辑是什么，以确定它是否适用于某特定情况是非常重要的。在大多数情况下，经验法则背后的逻辑性十分薄弱，且没有广泛的适用性。但一些领域可能会发展某些有效的经验法则用于样本量的规划。例如，某领域内的研究者可能达成共识，小于 d=0.3 的效应太小，不值得关注，而另一领域中所有研究都采用序列设计（见下文），其效应 d=0.3 时具有 90%的统计检验力。同样，可能存在某一领域，无论真实效应如何，都以预期精确度来收集数据。在这些情况下，使用有效的经验法则方法来收集数据将成为一种共识而存在。例如，Simonsohn（2015）建议在设计可重复的研究时，其样本量是原始研究的 2.5 倍，因为这为等价检验提供了 80%的检验力，此时的等价检验是假设真实效应量为 0，其界限设定为原始研究 33%的检验力进行检验。由于原作通常不会表明哪个效应量会推翻他们的假设，因此这种类似“小型望远镜”的经验法则，是可重复研究的一个很好的出发点，其推断目标是推翻早期前人研究中的某些大效应。研究者有义务补充足够理论知识，来区分某些经验法则是否有效且可靠，并出于特定研究的推断目标，来评估经验法则方法是否能得出有价值的结果。

不进行论证

这听起来像是一个无厘头的观点，但它也有存在的价值，它可用于研究者明确表明样本量的选取无特定缘由。或许有足够的资源收集更多的数据，但没有这么做。同样，研究者也能进行统计检验力分析，或规划精确度，但他们也没有这么做。**那么在这些情况下，与其假装进行了某种方式的样本量规划，不如坦白的说并没有。**这并不一定是坏事。他们仍然可以讨论感兴趣的最小效应量，统计学上可得到的最小效应量，以及效应量置信区间的宽度，并依据当前样本量分析并绘制出灵敏度功效分析。如果研究者在收集数据时确实没有明确的推断目标，那么根据同行收集数据时的合理推断目标来进行评估，这样的做法或许是可行的。

不要试图编造故事，让人觉得某研究的意义重大，其实不然。相反，**需要透彻的评估该研究感兴趣的效应量所具有的信息价值，并做到言行合一（结论与数据相符）**。不论证样本量的合理性可能没有问题，但这可能意味着该研究对大多数感兴趣的效应来说没很大的价值，这使得解释非显著的效应或者较大不确定性的估计值尤为困难。

你的推论目标是什么?

数据收集的推断目标通常与效应量大小有关。因此，为了设计一项信息量丰富的研究(informative study)，研究者需要确定**感兴趣的效应量**有哪些。首先，可以用三个效应量来帮助确定样本量。第一个是**研究者感兴趣的最小效应量**，第二个是**能够达到统计学显著的最小效应量**(仅在需要进行显著性检验的研究中)，第三个是**预期的效应量**。除了考虑这三个效应量之外，对效应量的范围估计也有助于确定样本量。通过计算感兴趣效应量的置信区间得到该范围(例如，效应量为零)，并得出可以拒绝的效应。类似地，绘制灵敏度曲线，估计研究检验力适中的效应量范围，以及检验力较低时的效应量范围，这些方法都有助于确定样本量。最后，在一些情况下，考虑在某个特定研究领域中可能出现的效应量范围同样有助于确定样本量。

什么是感兴趣的最小效应量？

强有力的样本量合理性依据是基于一个明确的感兴趣的最小效应量。感兴趣的最小效应量可以是基于理论预测的，也可以是基于实践的。一些方法学综述描述了在随机对照实验中如何确定感兴趣的最小效应量，详见 Cook et al. (2014) and Keefe et al. (2013)。也有一些综述采用了不同的方法来确定感兴趣的最小效应量，比如 King (2011) and Copay, Subach, Glassman, Polly, and Schuler (2007)。

更多的针对心理学研究的讨论，请见 Lakens et al. (2018).

当理论不太完备或研究问题远离实际应用时，确定感兴趣的最小效应量就很有挑战性，但此时仍然需要思考哪些效应小到可以忽略不计。接下来，第一步是与你的同行讨论在特定的研究方向下，哪些效应量是有意义的。对于效应量是否足够大，研究者们可能会有不同的看法 (Murphy et al., 2014)。因为每个学者认为值得研究的问题不同，每个学者对效应量是否足够大的看法也不同，不同领域的利益相关者认为有意义的效应量也有所不同 (Kelley & Preacher, 2012)。

尽管具有挑战性，但确定感兴趣的最小效应量非常有益。效应量的分布通常是不确定的（事实上，估计最小效应量通常是研究的目标之一），因此，当按照预期效应量进行研究时，统计检验力是否足够高以检验总体中真实效应，具有相当大的不确定性。然而，经过斟酌将感兴趣的最小效应量确定后，那么就有可能设计出一个具有足够统计检验力的研究（根据推断目标，依据既定的错误率来接受（detect）或拒绝感兴趣的最小效应量）。感兴趣的最小效应量可能是主观的 （如，一位研究者认为效应量小于 $d = 0.3$ 是有意义的，而另一个研究者可能对效应量大于 $d = 0.1$ 感兴趣），同时，确定感兴趣的最小效应所需要的参数也是不确定的（例如，进行成本-效益分析时）。但是，当确定了感兴趣的最小效应量之后，研究可以用II类错误来接受（detect）或拒绝该效应量。因此，当研究者能够确定感兴趣的最小效应量时，通常倾向于基于感兴趣的最小效应量来做先验检验力分析 (Aberson, 2019; Albers & Lakens, 2018; Brown, 1983; Cascio & Zedeck, 1983; Dienes, 2014; Lenth, 2001)。

最小统计检验效应（统计显著的最小效应量）

统计显著的最小效应量或临界效应量，提供了关于最小效应量的信息，如果可以得到这个效应，那么在给定的 α 水平和样本量下，这个效应将在统计学上是显著的（Cook et al., 2014）。对于任何临界 t 值（例如，对于 $t = 1.96$，$\alpha = 0.05$，大样本），我们可以计算临界均值差(mean difference)（Phillips et al., 2001），或临界标准效应量。对于双侧的独立样本 t 检验，临界均值差为：

$$M_{crit} = t_{crit}\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}$$

临界标准效应量为：

$$d_{crit} = t_{crit}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

在图 4 中，展示了当真实效应量为 $d = 0$ 或 $d = 0.5$ 时，每组 15 名被试的 Cohen's d 的分布。此图

与图 2 类似，但是增加了对临界 *d* 值的标注。**我们发现，在每组被试如此之少的情况下，只有效应大于 *d* = 0.75 时才能在统计学上显著。**这种效应量是否有意义，是否可以被实际预测，需要经过仔细地考量和论证。

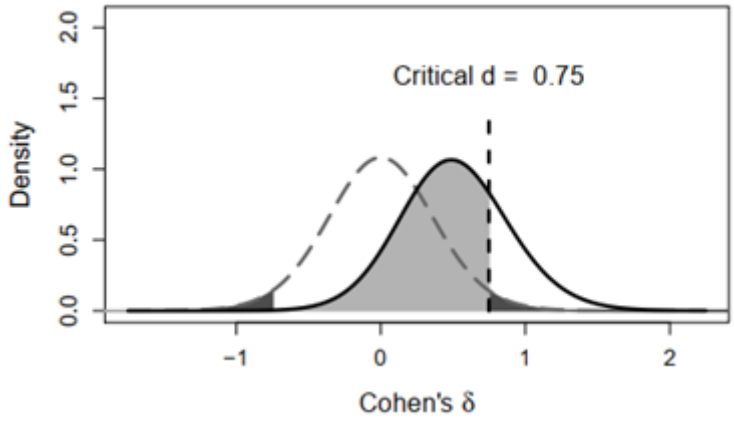

图 4，独立样本 t 检验的临界效应量（每组 n = 15, α = 0.05）

G*power 在统计检验力分析时提供了的临界的统计量(如临界 t 值)。例如，图 5 显示，在 $\alpha = 0.05$, N = 30 的双侧相关性检验中，只有效应大于 r = 0.361 或小于 r = -0.361 才能在统计学上显著。**这表明，当样本量相对较小时，需要有相当的大效应才能达到统计学显著。**

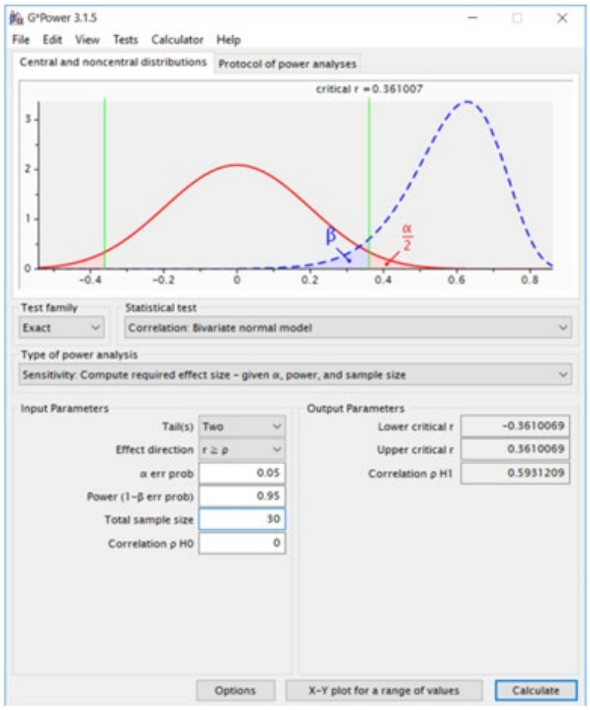

图 5，G*power 中相关性检验的临界值（n =30 和 α = 0.05）

重要的是要意识到，由于随机变异，每个研究都有概率产生大于临界效应量的效应，而真实的效应量很小 (当真正的效应量为 0 时，每个统计显著的效应都是一个 I 类错误)。计算**统计显著的最小效应量**（a minimal statistically detectable effect）对于没有进行先验检验力分析的研究是有用的，这

既适用于未报告样本量合理性的已发表研究 (Lakens 等人，2018 年)，也适用于通过经验法则进行样本量论证的研究者。

扪心自问，你的研究设计的临界效应量是否在实际可预测的效应范围内？如果不是，那么当已发表的文章效应显著时，要么是效应量出乎意料地大于预期，或者更可能的是高估了效应量。后者，鉴于**发表偏倚**，已发表的文章可能会导致效应量估计走偏。如果仍有可能继续增加样本量，例如忽略以往经验并进行先验检验力分析，那么就继续。如果不继续增加样本量了，例如资源有限等，那么应该要明确最小统计显著效应，且此时不应该拘泥于 p 值，而是关注效应量以及其置信区间（如表 3 所述）。

如果进行了先验检验力分析，计算最小统计显著效应同样也有用。例如，如果您认为真实效应量的最佳情况是 $d = 0.57$，并在先验检验力分析中采用了这个效应量，那么当您在一个包含两个独立组的研究设计中收集 50 个数据时，小于 $d = 0.4$ 的效应量将不会在统计学上显著。如果备择假设的最坏情况是真实效应量为 $d = 0.35$，那么当估计的效应量接近最坏情况时，您的研究设计将无法得到一个显著的结果。考虑到最小统计显著效应，你应该思考**假设检验能否得到一个有意义的结果**，以及**你目前论证样本量合理性的方式**（例如，使用经验法则，或者由资源支配来决定样本量大小）**能否使研究更有意义。**

预期的效应量是什么？

尽管效应量的真实分布往往是未知的，但在某些情况下，研究者将对其研究的效应量有一个合理的预期，并在先验检验力分析中使用这个预期效应量。即便从很大程度上来说，预期效应量是一种猜测，但斟酌哪些效应可以被预测是有用的。研究者可以根据他们预期的效应量来论证样本量的合理性，尽管从感兴趣的最小效应量的角度来看，这不会有很大的参考价值。因为在这种情况下，对推断目标是有价值的(检验预期效应是否存在)，但对于次要目标 (检验感兴趣的最小效应量是否存在) 来说则没那么多的价值。

对于效应量分布的预测通常有三个来源:元分析、前人研究或理论模型。研究者倾向于在先验检验力分析时设置较高的预期效应量，因为较高的效应量需要的样本量较少，但对效应量的预期太过乐观将会增加结果假阴性的概率。在论证样本量（基于先验检验力分析）的合理性时，重要的是批判性地估量检验力分析中所使用的预期效应量。

使用来自元分析中的评估

通过元分析来估计效应量是最完美的，可以为研究者提供最准确的信息，来表明哪些效应是可

预期的。由于学术领域普遍存在发表偏倚，来自元分析的效应量估计不一定是准确的。它们可能是有偏的，甚至有时是很大的偏差。此外，**元分析通常具有相当大的异质性**，这意味着元分析估计出的效应量与组成元分析的子集是不同的。因此，想要在检验力分析中使用元分析估计出的效应量，需要非常谨慎。

如果研究者想要在先验检验力分析中使用元分析估计出的效应量。他们需要考虑三个因素（见表 5）。首先，**元分析中的研究和他们想要做的研究应该是比较相似的**，这样才能够期望得到一个合理且相似效应量。本质上，这需要评估效应量的估计值在新研究中的可推广性。重要的是需要考量元分析中的研究和所计划的研究之间的差异，这涉及到实验操纵、测量、总体及其他相关变量。

其次，**研究者应该检查与元分析所报告的效应量是否是同质的**。如果不是，且异质性相当大，这意味着所估计出的效应量与真实值可能有所出入。使用元分析进行评估时，应采用最接近研究计划的研究子集。请注意，异质性仍有可能存在(当未测量的变量调节了样本的效应量时，即使是完全重复性的研究也可能表现出异质性(Kenny & Judd, 2019; Olsson-Collentine et al., 2020))，所以选择相似研究的主要目的是通过现有的数据来增加预估准确的概率，但不能保证它是绝对正确的。

最后，元分析的效应量估计应该尽可能没有偏差。核对元分析中报告的偏差检测测(bias detection test)是否是达到最高标准，或自己进行多次偏差检验测试（Carter, Schönbrodt, Gervais, & Hilgard, 2019），并采纳偏差校正后的效应量估计值（这些估计可能仍然存在偏差，并且不一定能反映真实的效应量分布）。

根据前人研究来估计样本量 ·

如果没有元分析研究作为参考，研究者通常会用**某个前人研究的效应量**来做先验检验力分析。首要需要考虑的问题是两个研究之间是否足够相似。与使用元分析估计效应量类似，研究者需要考虑不同研究在总体、实验设计、实验操纵、测量以及其他可能影响效应量的因素上是否存在差异。例如，个体的反应时差异会随着年龄的增加而增加。因此，相比于年轻被试样本的研究，年长被试样本的研究标准效应量更小。 其次，如果前人研究采用了一个强操纵，而你计划使用一个相对弱的实验操纵，那么将效应量预估地相对小一点会更适合。最后，**效应量不能在不同实验设计的研究之间推广**。 例如，组间对比实验的效应量通常与后续实验中交互作用的效应量是不一样的，后者往往会在原实验设计的基础上添加新的变量，从而导致效应量的差异 (Lakens & Caldwell, 2021)。

即便实验设计再相似，统计学家们也反对用预实验的效应量做检验力分析。Leon, Davis 和 Kraemer (2011) 认为：

与之相反，由于小样本固有的不精确性，预实验不能为后续的正式研究提供有意义的效应量估

计值。

利用已发表文章的效应量做检验力分析需要谨慎考虑，主要有以下两个原因：由于随机变异，前人研究中的效应量可能与真实的效应量分布不同；由于发表偏倚往往会夸大研究的效应量。图 6 展示了某研究η_p^2的分布，该研究一共 3 个条件，每个条件下 25 名被试，在零假设为真的情况下，存在“中等”的“真实”效应 (η_p^2= 0.0588) (Richardson, 2011)。图 4 标注了临界效应量，在小样本量下，观测到的效应量小于η_p^2 = 0.08 时在统计学上不显著。如果零假设为真，效应值大于η_p^2 = 0.08 会犯 I 类错误 (深灰色区域)。如果备择假设为真，效应值小于η_p^2 = 0.08 会犯 II 类错误 (浅灰色区域)。显然，显著结果的效应量都比真实效应量 (η_p^2 = 0.0588) 要大 。因此，依据显著结果的效应量进行统计检验力分析(通常只有显著结果的研究能够被发表) 往往会高估真实效应量，从而产生偏差。

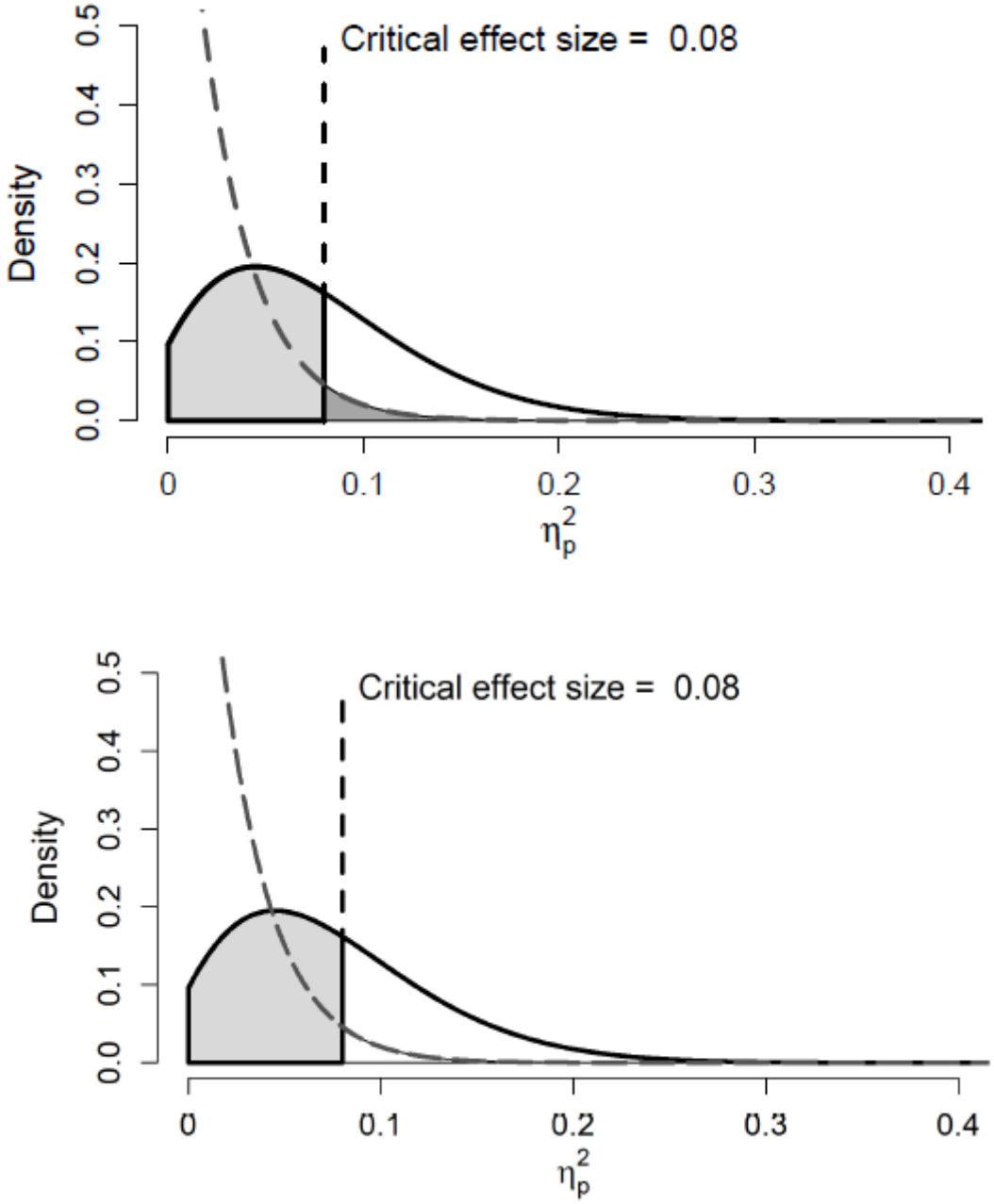

图 6，零假设为真（虚线灰色曲线）以及真实效应量中等大小（η_p^2 = 0.0588，实线黑色曲线）情况下的η_p^2分布（3 组，各 25 名被试）。

尽管测量了所有的效应量 (e.g., 从预实验中所得)，由于随机差异，所得效应量有时还是会很小。如图 6 所示，即使真实效应量只有 0.0588，小样本的预实验所得效应量很可能有η_p^2= 0.01。将估计出的效应量η_p^2 = 0.01 用在先验检验力分析中，如果后续研究想要达到 80%的统计检验力，将会得出需要的总样本量为 957。如果研究者仅根据预实验的效应量来计算（检验力分析）后续实验所需被试量，这样会高估效应量，从而导致后续实验的检验力系统性地低于我们的假设(Albers & Lakens, 2018)。

归根结底，通过小样本研究来估计效应量（用于先验检验力分析）存在本质上的问题，由于发

表偏倚或后续的偏差，也就是说研究者最终用于检验力分析的效应量并不是基于完整的 F 分布，而是基于截断的 F(truncated)分布 (Taylor &Muller, 1996)。例如，假设图 6 所展示的是一个极端的发表偏倚情况。研究者只能得到图中$\eta_p^2 > 0.08$ 的部分分布，这些结果均在统计学上达到显著。这种情况下，也能够求得基于特定假设且经过偏差矫正后的效应量（估计的）。假设，单因素 3 水平的研究的方差分析结果是 $F(2, 42) = 0.017$, $\eta_p^2 = 0.176$。如果我们在先验检验力分析中使用这个效应量，那么为了得到 80%的统计检验力，每个条件需要的样本量为 17 名被试。

然而，偏差已然存在，我们可以用 BUCSS R 包 (S. F. Anderson, Kelley, & Maxwell, 2017) 进行检验力分析来尝试校正偏差。校正偏差后的检验力分析结果显示（基于特定发表偏倚模型的截断 F 分布，即只有显著结果能被发表），每个条件下需要的样本量为 73 名被试。当用于计算检验力的非中心参数的偏差校正估计值为零时，此方法则不再适合。相对的，只要研究者认为存在偏差，就可以做一个偏差校正模型。一个相对简单且更保守的方式是，用效应量估计出 60%双侧置信区间的下限来进行检验力分析。此法被 Perugini, Gallucci 和 Costantini (2014)称为**“保底检验力”**(safeguard power)。上述提到的两种方式都是相对保守的检验力分析，但并不能一定是更准确的估计。因为基于一个可能存在偏差的且/或样本量较小的研究效应量，几乎不可能完成一个准确的检验力分析 (Teare et al., 2014)。在无法得到感兴趣的最小效应量时，对效应量的估计存在极大的不确定性。这种情况下用**序列设计(sequential design)**来进行实验相对更高效。

总而言之，**如果满足以下三个条件（表 6），则可以利用前人研究的效应量做检验力分析**。第一，研究设计与前人研究足够相似。第二，偏差存在的风险较低。（例如，效应量的估计是源于预注册的报告，或者是没有影响发表可能性的实验结果报告，即无关发表偏倚）。第三，基于 95%置信区间的所得效应量，样本量足够大，可以得到相对准确的效应量估计。效应量估计总是伴随着不确定性，因此进行先验检验力分析时，将 95%置信区间的上下限都考虑进来，可能会为这些不确定性提供一些有效的信息。

根据理论模型来估计样本

你可以根据一个足够详细、具体的理论模型来搭建一个计算模型，并根据这个计算模型来估计效应量，前提是你十分了解模型中与数据收集相关的核心参数有哪些。例如，如果研究者对每个刺激特征之间的异同的权重十分了解，那么可以通过 Tversky (1977) 的**对比模型**计算**每对刺激的相似性判断预测** (predicted similarity judgement)，以及估计**不同条件之间差异的预期效应量**。尽管可以做点估计的计算模型相对稀少，但合适的模型常常可以为研究者的预期效应量提供强有力的证据。

计算效应量的置信区间宽度

如果研究者能够估计数据的标准差，那么就有可能预先估计出效应量的 95%置信区间(Kelley, 2007)。置信区间表示的是一个估计值的范围，这个范围足够宽，真正的总体参数将会落在置信区间（100-α）%的范围内。在任何单项研究中，真正的总体效应要么落在置信区间内，要么不在置信区间内，但总地来说，**人们可以认为置信区间包括了真实的总体效应(要记得存在犯错的可能)**。Cumming(2013)将得到的效应量与 95%置信区间上限(或 95%置信区间下限)之间的差距称为**误差幅度（margin of error）**。

如果我们根据 t 值和样本量计算效应量为 d=0 的 95%置信区间(Smithson, 2003)，就会发现，当独立样本 t 检验的每个条件下各有 15 个观测值时，95%置信区间的范围从 d=-0.72 到 d=0.72[5]。误差幅度是 95%置信区间宽度的一半，即 0.72。使用无先验信息的贝叶斯估计将计算出一个具有相同(或非常相似)上限和下限的置信区间(Albers et al., 2018; Kruschke, 2011)，并且在收集完数据后，可能得出一个包含总体效应的范围，但范围太大并不能提供信息。不论在分析数据时基于哪种统计学派，在每组只有 15 个观测值的情况下所得的置信区间范围，并不能使我们获得更多信息。

对置信区间宽度的有效解释之一是，当真实效应量为 0 时，效应量多大时你可以拒绝该效应。换句话说，**如果效应不存在，根据收集到的数据情况，你能够拒绝哪些效应量？哪些效应量不会被拒绝？**以下这些研究有 d＝0.7 的效应量，比如“人们在被激怒时变得具有攻击性”，“相比于其他群体，人们更喜欢自己的群体”，以及“恋爱对象在外表吸引力上彼此相似”(Richard, Bond, & Stokes-Zoota, 2003)。根据置信区间的宽度，只能拒绝过大的效应，如果效应真实存在，那么应该已经被发现了。如果你研究的大多数效应比 d=0.7 小得多，那么通过 n=15 的研究，很可能发现不了任何东西。在大多数研究领域，过大的效应通常被认为是不合理的(尽管合理的效应量大小在不同领域之间是不同的，如下所述)。然而，例如在大样本中，如果零假设是真的，研究者可以拒绝大于 d＝0.2 的效应。而根据置信区间宽度的分析，许多研究领域的同行可能会认为这项研究是有价值的。

我们发现，误差幅度几乎与统计上可检测到的最小效应(d=0.75)相同，但并不完全相同。这个小的变异来源于根据 t 分布计算的 95%置信区间。如果真正的效应量不为零，则根据非中心 t 分布计算置信区间，那么得到的 95%置信区间是不对称的。图 7 显示了三个 t 分布，一个在 0 处对称分布，另外两个分别在 2 和 3 的非中心参数(均值之间的标准化差异)的不对称分布。这种不对称性在非常小的样本中最为明显(图中的分布自由度为 5)，但在计算置信区间和统计检验力时，在较大的样本中

[5] 效应量的置信区间可以使用在线应用程序来计算：https://www.aggieerin.com/shiny-server/

也很明显。例如，当真实效应量为 d=0.5 时，每组 15 个观测值的效应量为d_s=0.50，95% 置信区间为 [-0.23, 1.22]。如果我们计算临界效应量的 95%置信区间，将得到d_s=0.75，95%置信区间为[0.00, 1.48]。95%置信区间的范围从 0.00 到 1.48，这与置信区间和 p 值之间的关系一致，也就是说如果检验有统计学意义（结果显著），那么 95%的置信区间不包括 0。正如前面提到的，这里推荐的不同方法，通常是基于相同的信息来评估研究价值。

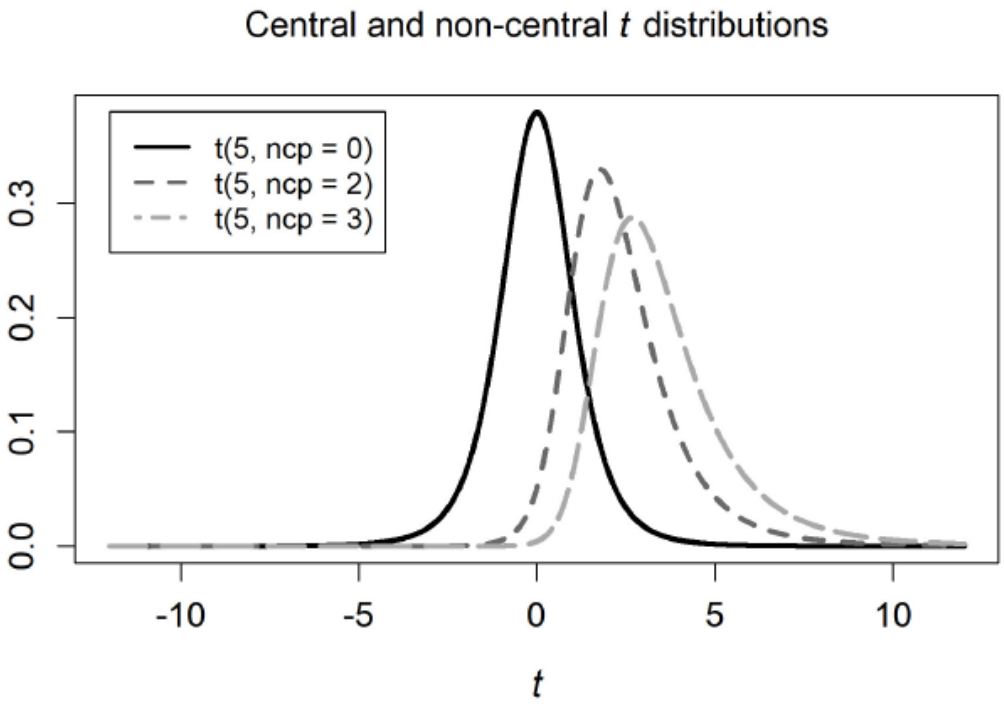

图 7，中心的（黑色）和 2 个非中心的（深灰和亮灰）t 分布

绘制灵敏度功效分析图

灵敏度功效分析确定了样本量、期望检验力和 α 水平，并回答了在期望检验力下研究能获得多少效应量的问题。因此，当样本量已知时，就可以进行灵敏度分析。有时，我们使用的是已经回答过不同研究问题的现成数据，或者从现有数据库中抽取的一部分，此时样本量已知，你可以为新的统计分析进行灵敏度功效分析。而其他时候，你可能在最初收集数据时没有仔细考虑样本量的问题，但希望在分析结果时反映出该研究对感兴趣的效应量(范围)的统计检验力。最后，虽然有可能在未来会收集到足够的样本量，但是由于资源限制，你知道你能够收集的最大样本量是有限的。你希望对于你认为合理且有趣的效应（例如感兴趣的最小效应大小或预期的效应大小）是否具有足够的统计检验力进行反思。

假设某研究者正进行一项研究，总共将收集 30 个样本，每个条件下各 15 名被试。图 8 显示了

如何在 G* power 中进行灵敏度功效分析，在该研究中，我们决定使用 5%的 α 水平，并希望获得 90%的检验力。灵敏度功效分析结果显示，研究设计有 90%的检验力来检测 d=1.23 以上的效应。有研究者认为 90%的期望检验力是相当高的，并且认为如果统计检验力更低的话，仍然可以开展一项有趣的研究。**因此，在小效应量范围内绘制灵敏度曲线是很有用的。**

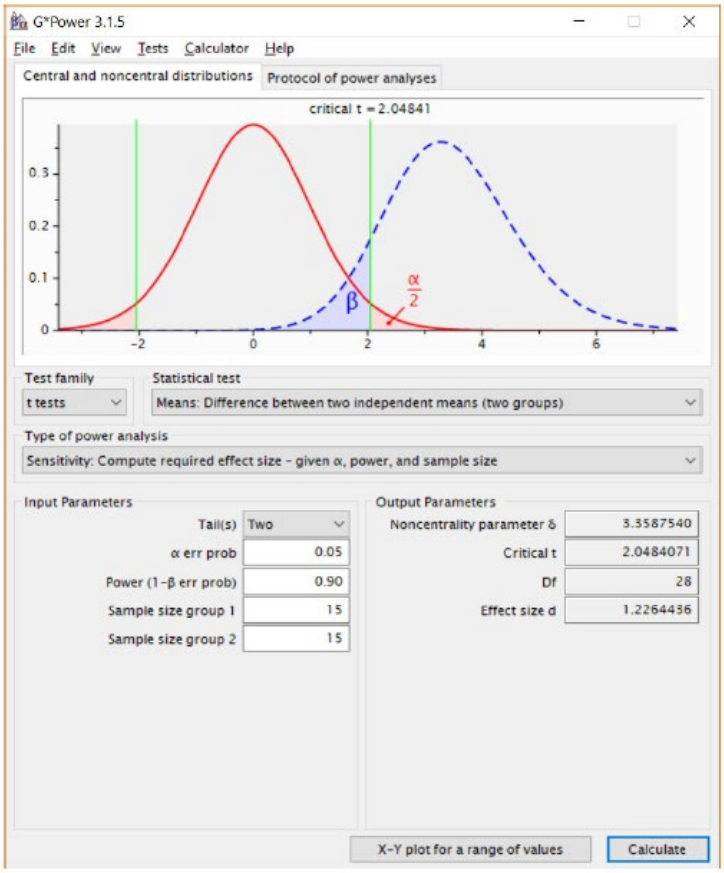

图 8，G* power 软件中的灵敏度功效分析

在灵敏度检验力分析中，有重要的两个维度：**效应量**以及**检验力**（即特定的效应量显著时的检验力）。**这两个维度可以共同绘制灵敏度曲线。**例如，在 G*Power 中，可以通过点击“X-Y plot for a range of values”按钮来绘制灵敏度曲线，如图 9 所示。研究者可以检测合理的先验效应量范围的检验力，或者他们可以检测哪些效应量将提供合理的检验力水平。在基于模拟的检验力分析方法中，可以通过对一系列可能的效应量进行检验力分析来创建灵敏度曲线。即使 50%的检验力被认为是可以接受的(此时如若得到不显著的结果之后，是否接受零假设，是一个相对复杂的决策过程)，图 9 显示了一个检验力非常低的研究，这对于大多数领域来说是属于较大合理范围内的效应量。因此，灵敏度功效分析为评估研究价值提供了额外的方法，并可以提示研究者，对于在某些特定的实验设计下，一些实际预期范围内的效应不太可能产生显著的结果。

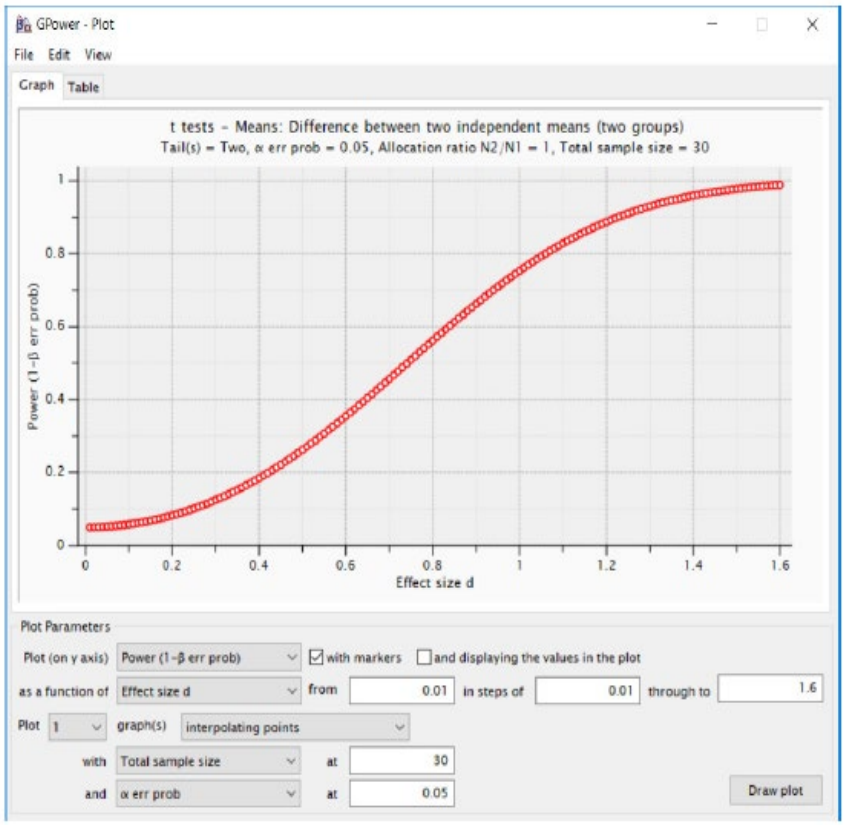

图 9，当每组 n=15，α =0. 05 时，效应量与期望检验力的关系图

如果每组的被试量更大，评估结果可能会更好。虽然我们不知道效应量究竟多大，但如果每组收集 150 个被试，灵敏度分析结果显示感兴趣效应的检验力是足够的，此外，对于相当小的效应，仍有大约为 50%的检验力。为了使灵敏度分析有意义，灵敏度曲线应该与感兴趣的最小效应量或预期的效应量范围进行比较。灵敏度功效分析没有明确的界限可供参考(Bacchetti, 2010)。但我们可以在人们观测到的或关心的不同效应量及其相关的统计检验力之间进行总体权衡。

效应量在研究领域的分布

根据我个人的经验，在独立样本 t 检验的先验检验力分析中，最常输入效应量估计是 Cohen’s 基准的“中等”效应量，这是默认的一个效应量。当你打开 G*Power 时，“中等”效应是先验检验力分析的默认选项。其实，Cohen’s 基准的小、中、大效应不应该用于先验检验力分析 (Cook et al., 2014; Correll, Mellinger, McClelland, & Judd, 2020)，而且 Cohen 本人后悔提出了这些基准(Funder & Ozer, 2019)。研究主题的多样性意味着，任何用于计算统计检验力的“默认的”或“经验法则式方法的”不仅不符合你的实际情况，还可能导致样本量与你试图用数据来回答的研究问题之间有很大的差距。

研究者想知道，如果没有其他依据来确定先验检验力分析的效应量，那么如何选默认值更好呢？Brysbaert(2019)建议心理学领域可以将 d=0.4 作为默认值，这是在可重复的项目和若干元分析中观测

到的平均水平。我们无法知道这个平均效应量是否可行，但很明显，各个领域和研究问题之间存在着巨大的异质性。平均效应量通常都与研究中预期的效应量有很大偏差。一些研究者建议根据效应量在特定领域的分布来更新 Cohen's 基准(Bosco, Aguinis, Singh, Field, & Pierce, 2015; Funder & Ozer, 2019; Hill, Bloom, Black, & Lipsey, 2008; Kraft, 2020; Lovakov & Agadullina, 2017)。当我们基于已发表的文章来估计效应量时，需要考量效应量由于发表偏倚而被夸大的可能性。由于某一特定研究领域内的效应量存在较大的差异，所以基于某一领域内效应量的经验分布，选择一个大、中、小的效应量基准来进行检验力分析的用处不大。

在解释效应量的置信区间时，应该了解一些文献中效应量的分布。如果在一个特定的研究领域里，几乎没有效应大于你在等价检验中可以拒绝的效应量(例：如果观测到的效应量为 0，设计将只拒绝大于如 d=0.7 的效应)，那么，此时收集到的数据不太可能获得有用的信息。

我们很难找到依据来证明的是：从效应量的经验分布中推导出的特定效应量，可以作为先验检验力分析中使用的效应量。有人可能会说，使用文献中效应量分布的效应量基准比胡乱猜测要好，但这并不是论证样本量合理性的强证据。研究者们必须承认，**在预期效应量不明确的情况下，不能进行先验检验力分析**(Scheel, Tiokhin, Isager, & Lakens, 2020)。而其他论证样本合理性的理由，比如资源限制，或者结合序列研究设计（a sequential study design），可能可以更好地来满足研究的推断目标。

设计定性研究（informative study）时的注意事项

到目前为止，我们一直把重点放在定量研究的样本量论证。其中有许多相关的主题对设计定性研究也有一定的帮助。首先，除了先验检验力分析以及灵敏度功效分析之外，探讨折中检验力分析（compromise power analysis）（有帮助的）和事后检验力分析(没有帮助的，例如，Zumbo and Hubley (1998), Lenth (2007))同样十分重要。如果能够通过先验检验力分析来确定样本量，那么序列设计的数据收集会非常有效，因为我们随时对收集到的数据进行分析来决定是否继续实验。此外，在不增加样本量的前提下提高研究统计检验力的方法非常有价值。另外，研究者应当充分了解自己实验的因变量，尤其是因变量的标准差。最后，**样本量的规划在定性研究中同样重要**，尽管在定性研究的领域内有关样本量规划的研究较少，但我们给出了一些建议，可供研究者们设计定性研究时参考。接下来我们将依次讨论上述每一点。

折中检验力分析(compromise power analysis)

在折中检验力分析中，样本量和效应量是固定的，且检验的错误率是依据 I 类错误和 II 类错误

之间的期望比(相对重要程度之比，desired ratio）所计算出来的。**当需要收集大量的数据或者只能收集少量数据时，折中检验力分析都是有用的。**

前者，因为研究者们十分幸运地能够收集到足够多的数据，所以对于研究者们所有感兴趣的效应量而言，研究的统计检验力都很高。比如，研究者想在某一公司测试一种能够降低压力水平的干预措施，因此招募 2000 名雇员在公司的年度评估中回答了一系列问题。小于 d=0.2 的效应不足以引起个体的主观注意(Jaeschke, Singer, & Guyatt, 1989)。当 α 水平为 0.05 时，统计检验力为 0.99，即犯 II 类错误的概率为 0.01。这意味着当感兴趣的最小效应量为 d = 0.2 时，研究者犯 I 类错误的可能性是犯 II 类错误可能性的 8.30 倍。

尽管研究者最初提出控制 I 类错误和 II 类错误，是为了论证其错误率合理性(Neyman & Pearson, 1933)，**但一个常见的错误想法是：将 I 类错误设定为 0.05，II 类错误为 0.2，意味着 II 类错误发生的概率是 I 类错误的 4 倍**。通常，默认使用 80%的检验力（或者 20%的 II 类错误）是基于 Cohen（1988）个人的偏好，他对此解释道：

这是一个惯例，当研究者没有其他依据来设置所需的检验力值时，默认采用 0.80。这意味着 β（II 类错误）被设定为 0.20。采取这个值有以下几个原因(Cohen, 1965，第 98-99 页)。首先，主要考虑到 α=0.05 这个隐含条件。其次，选取 0.20 的 β 值是考虑到这两种误差的相对严重性之比为 0.20/0.05。即第一类错误的严重程度是第二类错误的四倍。之所以提出 0.80 的检验力值是因为当研究者找不到依据来确定检验力时可以参考，但当研究者在其具体的研究调查中找到实质性的依据来选择一个特定值时，可以不采用这个值（This .80 desired power convention is offered with the hope that it will be ignored whenever an investigator can find a basis in his substantive concerns in his specific research investigation to choose a value ad hoc.）。

我们可以看到，约定是基于其他约定之上的：即，80%检验力的标准是基于 α 水平为 0.05 的标准之上的。因此，我们从 Cohen 那学到的不应是以 80%的检验力为目标，而是应当基于每个错误的相对严重性，来论证错误率的合理性。这就是折中检验力分析的用处。如果你和 Cohen 有一样的信念（即 I 类错误的严重程度是 II 类错误的四倍），那么回到之前所提出的 2000 名雇员的研究，对所有感兴趣的效应量而言，当 II 类错误率较低时，调整 I 类错误率是有意义的(Cascio & Zedeck, 1983)。实际上，Erdfelder，Faul 和 Buchner (1996)开发 G*Power 软件，在一定程度上为研究者们提供了一个做折中检验力分析的工具。

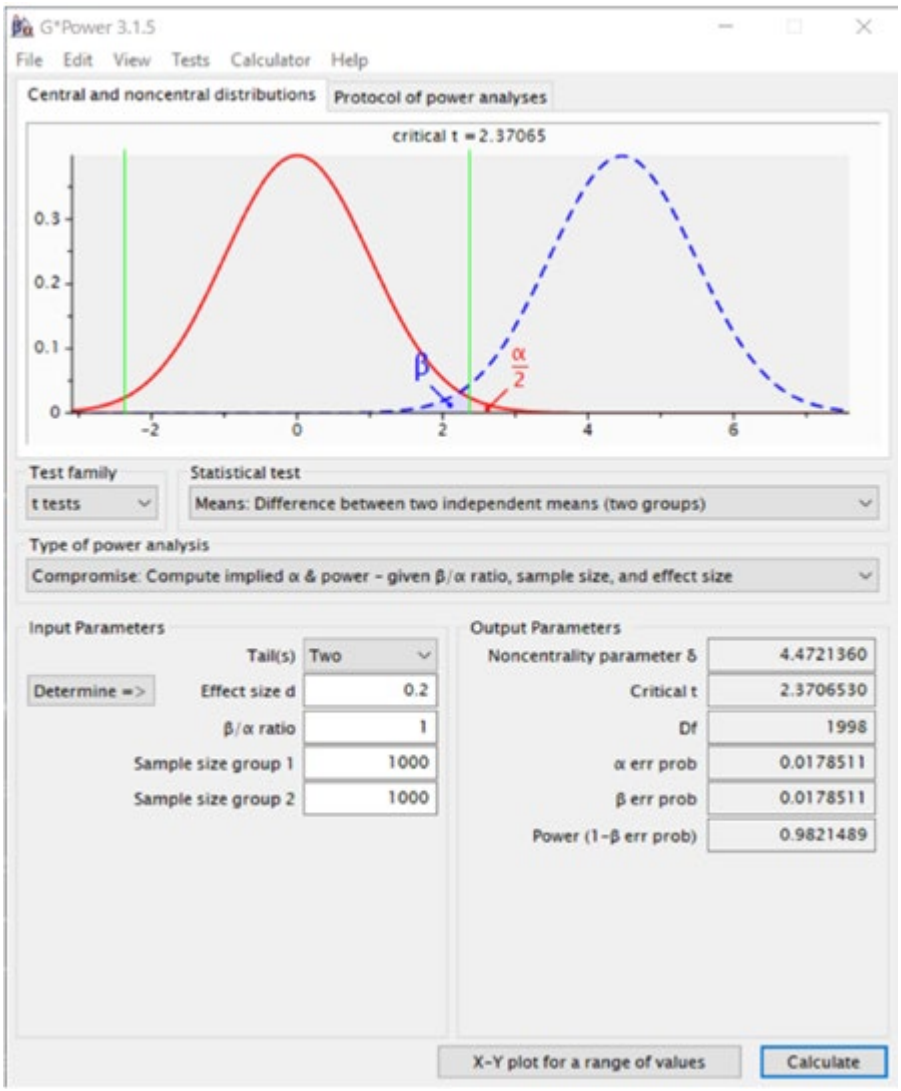

图 10 G* power 中的折中检验力分析

图 10 展示了当 I 类错误的代价等同于 II 类错误时（β/α ratio 为 1 时），在 G*Power 中做折中检验力分析的情况，当每个条件有 1000 个被试时，I 类错误和 II 类错误均为 0.0179。正如 Faul，Erdfelder，Lang 以及 Buchner (2007) 所描述的：

显然，折中检验力分析很容易导致非常规的显著性水平，也就是大于 α =0.05（在小样本或小效应量的情况下）或小于 α =0.001（在大样本或者大效应量的情况下）。然而我们相信，平衡 I 类错误和 II 类错误风险的益处能够弥补违反显著性水平约定（α =0.05）的代价。

这将指引我们了解折中检验力分析发挥其他作用的情况，即当我们知道研究的统计检验力很低时。虽然在错误率很高时做决定是非常不可取的，但当研究者发现自己必须基于较少的信息作出决策时，Winter（1962）认为：

经常使用 0.05 或 0.01 作为显著性水平是一种约定俗成的惯例，但基本上没有科学性和逻辑性可言。在上述显著性水平下，如果研究的检验力很低，并且当 I 类错误和 II 类错误的重要性大致相同时，0.30 和 0.20 的显著性水平可能比 0.05 和 0.01 的显著性水平更为合适。

例如，我们计划做一个双侧 t 检验，每个独立组最多能够收集 50 个样本，且预期总体效应量为 0.5，如果将 α 水平设置为 0.05，检验力将会达到 70%。同样，我们可以平衡两类错误（使其重要性相等），并将 α 水平设置为 0.149，最终得到效应量为 d = 0.5 且统计检验力为 0.851（即给定一个 0.149 的 II 类错误率）。**在折中检验力分析下 α 和 β 的选择，可以扩展到虚无假设和备择假设先验概率的考量当中**(Maier & Lakens, 2022; Miller & Ulrich,2019; Murphy, Myors, & Wolach, 2014)。

折中检验力分析需要研究者确定**样本量**的大小。这个样本量的大小就需要进行合理性论证，因此折中检验力分析通常与资源受限的样本量一起进行权衡。如果你的资源有限，又迫切的需要做出决策，那么折中检验力分析非常关键。在这种情况下，研究者应该认真考虑一个可接受的 I 类和 II 类错误率。然而，当一个研究的样本量很大，但研究者仍不能自由设置样本量时，折中检验力分析仍然是有意义的。例如，收集的是一项较大的国际研究中的一部分数据，且样本量是基于其他研究问题得来。在 II 类错误率非常低（且检验力很高）的研究设计中，一些统计学者认为还应当降低 α 水平来防止林德利悖论（Lindley's paradox），在林德利悖论中，效应显著($p < \alpha$)是对零假设的某种证明(Good, 1992; Jeffreys, 1939)。降低统计检验力分析中的 α 水平可以防止该悖论的产生，这就为大样本量的折中检验力分析提供了另一种依据(Maier & Lakens, 2022)。最后，折中检验力分析需要对**效应量**进行合理论证，要么基于感兴趣的最小效应量，要么基于预期的效应量。表 7 列出了应该与折中检验力分析一起讨论的三个方面。

表 7 依据折中检验力分析来权衡错误率合理性时的一些建议

需要参考的事项	该怎么做？
样本量的合理论证	说明为什么要收集特定的样本量（例如，基于资源的限制或决定样本量的其他因素）。
效应量的合理论证	效应量的大小是基于感兴趣的最小效应量，还是所预期的效应量。
I 类错误和 II 类错误之间的期望比（相对重要程度之比，desired ratio）	通过仔细评估每类错误的后果来权衡二者的相对代价

如果期刊编辑要求事后统计检验力，该怎么办？

事后检验力、回溯性检验力或观察性检验力被用于描述效应量（假设从收集到的数据估算出的效应量是真实的效应量）的统计检验力(Lenth, 2007; Zumbo & Hubley, 1998)。所以，在收集数据之前，无法计算事后检验力，它并不像先验检验力分析那样，可以基于感兴趣的效应量来进行估计，而且

它也不像灵敏度功效分析那样，可以对一系列感兴趣的效应量进行估计。因为事后检验力是基于已收集数据的效应量，除了报告的 *p* 值之外，没有增添其他信息，但它以不同的方式呈现了相同的信息。**编辑和审稿人经常要求作者用事后检验力分析来解释不显著的结果。这不是一个合理的要求，无论何时提出，你都不应该遵从。**相反，你应该进行**灵敏度功效分析**，并讨论感兴趣的最小效应量的检验力，以及预期效应量的一个实际范围。

事后检验力与统计检验中的 *p* 值直接相关(Hoenig & Heisey, 2001)。对于 *p* 值恰好为 0.05 的 *z* 检验，事后检验力始终为 50%。产生这种关系的原因是，当所得 *p* 值等于 *α* 水平（例如 0.05）时，所得 z 分数正好等于检验显著的临界值（例如，在一个 *α* 水平为 5%的双侧检验中，*z* =1.96）。当备择假设以临界值为中心时，如果备择假设为真，我们预期所得数据的一半数据会低于临界值，而另一半数据会高于临界值。所以，在事后统计检验力分析中，如果假定分析的效应量为真，那么 *p* 值与 α 水平相同的检验，其统计检验力恰好为 50%。

对于其他统计检验来说，当备择假设的分布不对称时（例如 t 检验，备择假设遵循非中心化的 t 分布，如图 7 所示），一个 *p*=0.05 时检验力不为 50%，但是通过对 p 值和检验力绘图，我们发现这两个统计量总是直接相关的。如图 11 所示，如果 *p* 值显示不显著（即，大于 0.05 时），那么可得知在 t 检验中检验力将低于 50%。同样的，**Lenth(2007）在 F 检验中也说明了检验力（事后）是由当前的 *p* 值决定的**，尽管不显著时检验力低于 50%的说法不成立。

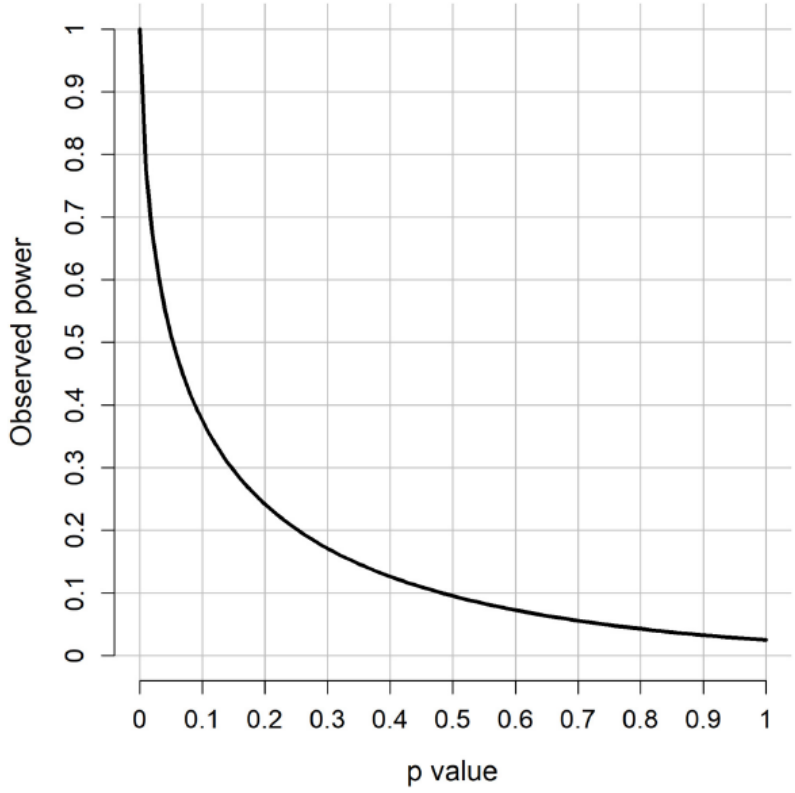

图 11　α =0. 05 且 n = 10 时，独立 *t* 检验的 *p* 值与检验力之间的关系

期刊编辑或审稿人可能会要求研究者报告事后检验力，以区分真阴性（确实没有效应）和假阴性（也就是 II 类错误，效应真实存在但未被发现）。但实际上事后检验力只是报告 p 值的另一种方式，报告事后检验力不足以解决编辑所提出的问题(Hoenig & Heisey, 2001; Lenth, 2007; Schulz & Grimes, 2005; Yuan &Maxwell, 2005)。为了得出有意义的效应确实不存在的结论，研究者应该进行**等价检验**，并设计一个高检验力的研究来验证感兴趣的最小效应不存在。或者，研究设计之初并没有确定感兴趣的最小效应量时，研究者可以报告**灵敏度功效分析**。

序列分析（Sequential Analyses）

在序列设计中，利用先验检验力来衡量样本量是否合理是非常有效的。序列设计可以在数据收集过程中多次分析来**控制错误率**（例如，在收集了 50、100 和 150 个观测数据之后），与一次性的设计（fixed design）相比，序列设计可以在一定程度上**减少预期的平均样本量**，因为一次性设计通常需要在收集大量数据后才进行数据分析（Proschan, Lan, &Wittes, 2006;Wassmer & Brannath, 2016）。序列设计有很长的历史(Dodge & Romig, 1929)，产生了一些的演变，比如说，序列概率比检验（(Wald, 1945）；结合独立性统计检验（independent statistical tests）（Westberg，1985），成组序列设计（group sequential designs）(Jennison & Turnbull，2000），序列贝叶斯因子(sequential Bayes factors）(Schönbrodt,Wagenmakers, Zehetleitner, & Perugini，2017）以及和安全性检验（safe testing）（Grünwald, deHeide, & Koolen，2019）。在这些方法中，如果边实验边分析数据，使用序列概率比检验最有效（Schnuerch & Erdfelder, 2020）。成组序列设计，即分批收集数据，在数据收集、误差控制和效应量估计的调整等方面更灵活（Wassmer & Brannath，2016）。如果变量之间存在某种依存关系（dependecies），安全性检验的灵活性最佳（ter Schure & Grünwald, 2019）。

当效应量非常不确定时，或者当真正的效应量可能大于研究的最小效应量时，序列设计将非常有用（Lakens，2014）。在这种情况下，如果效应量大于感兴趣的最小效应量，就可以提前结束数据收集，但如果需要的话，仍然可以继续收集到最大样本量。序列设计可以在假设检验过程中避免无效工作量，可以在确实存在效应（拒绝零假设）时停止收集数据，也可以在确实无效应（拒绝备择假设）时停止。**成组序列设计**是目前最广泛使用的序列分析方法，可以用 rpact（Wassmer & Pahlke, 2019）或 gsDesign（K. M. Anderson,2014）进行计划和分析[6]。

在不增加样本量的情况下提高检验力

提升研究价值最直接的方法是增加样本量。通常因为资源有限，所以在不增加样本量的情况下，探索不同的方法来提高检验力也很有价值。第一个选择是使用**相关的方向性检验**。研究者通常会做

[6] 在线应用程序可用于 rpact: https://rpact.shinyapps.io/public/ 和 gsDesign: https://gsdesign.shinyapps.io/prod/

出方向性的预测，比如“我们预测 X 大于 Y”。从逻辑上来说，由这个预测得出的统计检验是方向性（或单侧）的 t 检验。方向性检验会将I类错误率移动到分布尾部的一侧上，这会使得临界值变小，因此只需要较少的样本量就能获得相同的统计检验力。

虽然有一些关于方向性检验何时适用的讨论，都在用 Neyman-Pearson 对假设检验的观点来支持自己的想法(Cho & Abe, 2013)，这使得（预注册的）方向性检验成为一种最直接的方法，既能提高检验能力，但也会加大预测风险。然而，在某些情况下，你可能没办法提出一个方向性问题。特别是在具有应用价值的研究中，尽管结果与预期方向相反，但检验结果是否能够拒绝零效应也很重要。例如，当你正在评估最近引入的一项教育干预措施，并预测该干预措施将提高学生的表现，你可能想要探究一下学生表现更差的可能性，以便能够建议学校放弃这项新的干预措施。在这种情况下，也可能以“不平衡”的方式分配错误率，例如，相比于积极方向，将更严格的错误率分配给消极方向(Rice & Gaines, 1994)。

在不增加样本量的情况下，增加检验力的另一种方法是**提高检验的 α 水平**，如折中检验力分析部分所述，显然，这增加了犯 I 类错误的概率。我们应当认真权衡犯任何一类错误的风险，这通常需要考虑零假设为真的先验概率(Cascio & Zedeck, 1983; Miller & Ulrich, 2019; Mudge, Baker, Edge, & Houlahan, 2012; Murphy, Myors, & Wolach, 2014)。如果你“必须”做出决定，或者想要提出一种观点，而你能收集到的数据又确实有限，那么无论是基于折中检验力分析或是成本-效益分析（cost-benefit analysis），提高 α 水平都是合理的(Baguley, 2004; Field, Tyre, Jonzén, Rhodes & Possingham, 2004)。

另一种被广泛推荐的提高研究检验力的方法是**尽可能使用被试内设计**。几乎在所有情况下，当研究者对组间差异感兴趣时，被试内设计需要的被试比被试间设计少。可以从 Maxwell、Delaney 和 Kelley(2017)给出的等式来解释样本量减少的原因。假设总体正态分布，一个两组的被试内设计（NW）的被试量与一个两组的被试间设计（NB）所需的被试量相关：

$$NW = \frac{NB(1-\rho)}{2}$$

被试间设计所需的被试量是被试内设计的 2 倍，因为在一个具有两种条件的被试内设计中，每个被试提供两个数据点。与被试间设计相比，在多大程度上减少样本量还取决于因变量之间的相关性（例如，控制组与实验组数据之间的相关性），这一点体现在方程的（1-ρ）部分。如果相关性为 0，则被试内设计只需要被试间设计被试数量的一半（例如，被试内 64 名被试，被试间 128 名被试）。相关性越高，被试内设计的相对效益就越大，当相关性为负（高达-1）时，相对效益就会消失。特别是当被试内设计中的因变量是正相关时，基于可得的样本量，被试内设计将极大地提高检验力。**尽可能使用被试内设计，但要权衡更高的检验力所带来的好处与被试内设计中产生的顺序效应或遗留**

效应（carryover effect，即练习效应和疲劳效应）所带来的负面影响(Maxwell, Delaney & Kelley, 2017)[7]。对于多因素多水平的设计，可能很难给出完整的相关矩阵（即每对变量间的相关性所构成的矩阵）(Lakens & Caldwell, 2021)。在这些情况下，序列分析也许能够提供解决方案。

一般来说，变异越小，标准化效应量就越大（将原始效应除以较小的标准差），因此在样本量相同的情况下，检验力就越高。文献中提供了一些额外的建议(Allison, Allison, Faith, Paultre & PiSunyer, 1997; Bausell & Li, 2002; Hallahan & Rosenthal, 1996)，例如：

1. 参与实验之前，如果需要对被试进行筛选，建议使用更高效的方法进行筛选。
2. 将被试不均等地分配到不同条件（例如，控制组下的数据比实验组的数据更易收集）。
3. 采用较低误差的可靠测量方法(Williams, Zimmerman & Zumbo, 1995)。
4. 巧妙使用预注册的协变量(Meyvis & Van Osselaer, 2018)。

重要的是要考虑，**减少数据变异的这些方法是否会损耗过多的外部效度**。例如，在随机控制试次的意向性治疗分析（ntention-to-treat analysis）中，不遵守协议的被试将被保留在分析中，这样研究的效应能准确地代表在人群中实施干预后所得到的效应，而不是只代表了那些完全遵守协议的人的干预效应(Gupta, 2011）。在减少变异和外部效度两方面上，其他研究领域也存在类似的权衡。

了解你的测量方法

虽然讨论标准化的效应量大小很方便，但如果研究者能够用原始（非标准化）分数来解释效应，并了解其测量的标准偏差，相对来说是更好的(Baguley, 2009; Length, 2001)。为了使学术界能够对实验数据的标准偏差有一个实际预期，同领域内的研究者使用相同效度的测量方式将更有益。这将提供更加可靠的信息，使得期望精确度的设计更容易，也能够在先验检验力分析中使用一个非标准化的感兴趣的最小效应量。

除了对标准偏差的了解之外，了解因变量之间的相关性也很重要（例如，因为一个因变量 t 检验的 Cohen's d_z 依赖于均值之间的相关性）。在进行预测时，模型越复杂，就需要了解数据生成过程的更多方面。例如，在层级模型中，研究者需要了解变异的成分以进行检验力分析(DeBruine & Barr, 2019; Westfall, Kenny & Judd, 2014)。最后，研究所用测量方法的信度很重要(Parsons, Kruijt & Fox, 2019)，尤其是在参考一项已发表研究的效应量时，而你和它所使用的测量方法信度不同，或者同一测量方法用于不同的群体时，这时，不同群体之间的测量信度可能不同。随着开放数据的增加，通过以往研究数据来估计这些参数将会更容易。

[7] 你可以在这个在线应用程序中比较被试内和被试间设计：http://shiny.ieis.tue.nl/within_between

如果我们计算样本的标准偏差，这个值是对总体真实值的估计。在小样本中，我们的估计值与真实值可能有较大差距，然而由于大数定律，随着样本量的增加，我们对标准偏差的估计将更加精确。由于样本标准差是一个不确定的估计值，所以我们可以围绕估计值计算出**置信区间**(Smithson, 2003)，也可以设计**小样本的预实验**（pilot study），得出可靠的标准偏差估计值。方差 σ^2 的置信区间如下公式所示，标准偏差的置信区间则为这些值的平方根：

$$(N-1)s^2/\chi^2_{N-1;\alpha/2}, (N-1)s^2/\chi^2_{N-1;1-\alpha/2}$$

当参数存在不确定性时，研究者可以使用序列设计进行**内部的预实验**（internal pilot study）(Wittes & Brittain, 1990)。内部预实验的理念是，研究者为研究指定一个暂定的样本量，进行中期分析，使用内部预实验的数据来更新参数，如实验的方差，最后得出最终的样本量。只要对数据的中期考察是盲目的（例如，不考虑相关条件的信息），就可以根据新的方差估计结果对样本量进行调整，而不会对 I 类错误产生任何实际影响(Friede & Kieser, 2006; Proschan, 2005)。因此，如果研究者想设计一个定性研究，其中 I 类和 II 类错误已经得到控制，但他们缺乏关于标准偏差的信息，内部预实验可能是一个值得考虑的方法(Chang, 2016)。

约定俗成的元经验法则（Conventions as meta-heuristics）

即使研究者可能不会直接使用经验法则式的方法来确定研究中的样本量，但经验法则也会间接地在样本量规划中发挥作用。基于推断目标的样本量论证（如检验力分析、准确度或决策），都要求研究者确定 I 类和 II 类错误、精确度以及感兴趣的最小效应量的预期数值。尽管有时可以证实上述数值的合理性（例如，基于成本-效益分析），但这些数值的可靠程度可能需要更专门的研究来验证。这样更专门的研究可能很难实现，因为这些研究本身可能就不值得花钱（例如，用大多数同行认为保守的 α 水平进行研究，比基于成本-收益分析收集数据来确定所需 α 水平，所花费的更少）。所以在这些情况下，研究者倾向于使用惯例的数值。

当涉及到计算样本量所需的置信区间宽度、期望检验力或任何其他输入值时，透明且公开地报告如何使用经验法则或惯例（例如通过使用本文所附带的在线应用程序）**非常重要**。例如，通常在没有进行合理性论证的情况下，使用 5%的I类错误和 80%的检验力实际上是同行所能接受的最小信息价值的一个较低的阈值（而对样本量进行合理性论证时，同行也可以接受更高的错误率）。重要的是我们需要认识到，这些数值不是固定的。期刊在投稿指南中可以任意规定一个他们所希望的更高效准确的信息价值（例如，Nature Human Behavior 杂志要求投稿的研究设计要达到 95%的统计检验

力，我自己所在的部门要求研究者提交 ERB 提案，尽可能使研究设计达到 90%的统计检验力）。如果某研究者所报告的信息价值高于以往研究的最低值，应当给予一定鼓励。

在过去，一些领域已经改变了以往的惯例，比如现在在物理学中用 5σ 阈值来宣布一个发现，而不再使用 5%的 I 类错误。在其他领域，尚未有这种尝试获得成功的（例如，Johnson(2013)）。改进后的惯例应视具体情况而定，通过学会的学术会议来确定惯例可能更明智(Mullan & Jacoby, 1985)。学会会议在医学领域中很常见，并已被用于确定感兴趣的最小效应量（例如，Fried, Boers & Baker(1993)）。在许多研究领域，现行惯例都可以进行改进。例如，单项研究和元分析的默认 α 水平为 5%似乎很奇怪，可以想象，未来元分析的默认 α 水平将远低于 5%。在特定情况下，让大家更清楚什么样的输入值缺乏合理的缘由，这将会促使各个领域开始讨论该如何改进现行惯例。日后如果可能的话，在线应用程序将会链接到更好示例，并随之更新。

定性研究中的样本量规划

样本量规划对于定性研究来说也很重要。在定性研究中，样本量的规划应该基于如下的考虑：花费成本收集更多被试的数据但并不能产生更多信息了，之前的信息已经足以实现推理目标。这种观点的一种广泛应用被称为饱和，这意味着，新数据重复了早期的观测结果而没有添加新信息(Morse, 1995)。例如，假设我们问人们他们为什么养宠物，通过访谈，结果能得到几类原因，但在采访了 20 个人之后，没有新的原因分类出现，那么此时已经达到了饱和。定性研究还有其他的思想体系(philosophies)存在，并非所有思想体系存在饱和的问题。遗憾的是，还没有针对这些思想体系提出合适的样本量规划的方法(Marshall, Cardon, Poddar & Fonteno, 2013)。

采样时，通常不是选择具有代表性的样本，而是选择一个包含足够多样化的被试样本，以便有效地达到饱和。Fugard 和 Potts(2015)展示了在定性研究中，如何对样本量进行更高效的规划，1）群体中存在的编码数量（例如，人们养宠物的原因数量），2）从单个信息源中得到编码的概率（例如，你采访某个人，他可能所提到的每个养宠原因的概率），3）你想要得到的每个编码的次数（即每个原因出现的频次）。他们在 R 中提供了一个基于二项分布的公式来计算所需样本量，以便获取所需编码的期望概率。

Rijnsoever(2017)采用了一种更先进的方法，该方法也探讨了不同抽样策略的重要性。一般来说，相对于随机采样，有目的地从你所期望的样本中采样来获取信息将更加高效，但这需要你对预期的编码和每个编码的子群体都有很好的理解。有时我们也许能够确定，在采访某个信息源时至少会产生一个新的编码（例如，基于采访前的非正式沟通）。在定性研究中，一个好的样本量规划是基于：1）对总体（包括所有子集）的识别，2）对总体（子集）中编码数量的估计，3）在信息源中获得一

个编码的概率，4）所使用的抽样策略。

讨论

要设计一项内容丰富的研究，论证样本量的合理性是必不可少的步骤。根据数据收集的目标、可用的资源和统计分析的方法，有多种途径来证明研究样本量设置的合理性。**所有这些方法的首要原则是：研究者应将他们所收集的信息价值与他们的推理目标联系起来**。

在研究设计的样本量合理性论证过程中，有时会得出这样的结论：收集数据是不值得的，因为这项研究付出的成本并不能收获足够的信息价值。在某些情况下，不太可能有足够的数据来进行元分析(例如，因为大众对主题缺乏普遍的兴趣)，这些信息将不会被用于做出决定或声明，且统计检验不允许你以合理的错误率来检验假设，也不允许你以足够的精确度来估计效应量。**如果没有足够的理由去收集尽可能多的数据，那么无论如何进行这项研究都是在浪费时间和金钱**(Brown, 1983; Button et al., 2013; Halpern et al.,2002)。

越来越多的心理学家意识到，在过去的研究中，样本量往往太小，无法实现推断目标(Button et al., 2013; Fraley & Vazire,2014; Lindsay, 2015; Sedlmeier & Gigerenzer, 1989)。随着越来越多的期刊开始要求论证样本量，一些研究者也将意识到他们需要收集比过去更多的样本量。这意味着研究者需要在资助提案中要求更多用于被试费的资金，或者需要更多的合作(Moshontz et al., 2018)。如果你认为你的研究问题很重要，但以你现有的资源无法回答这个研究问题，进而可以考虑与同行进行合作研究，共同寻求这个问题的答案。

论证样本量的合理性不应该被视为研究者在申请资助、通过伦理审查或发表手稿之前所需要克服的障碍。如果只是简单陈述样本量，而没有仔细地论证，那么会很难评估研究者收集的信息价值是否超过数据收集的成本。**能够对样本量做出强有力的论证，意味着研究者知道他们想从研究中了解什么，并且能够设计出一项为科研问题提供丰富答案的研究。**

参考文献

略