|  |  |
| --- | --- |
| **9.4 Power Analysis for Interval Hypothesis Tests** | **9.4 区间假设检验的检验力分析** |
| When designing a study it is a sensible strategy to always plan for both the presence and the absence of an effect. Several scientific journals require a sample size justification for Registered Reports where the statistical power to reject the null hypothesis is high, but where the study is also capable of demonstrating the absence of an effect, for example by also performing a power analysis for an equivalence test. As we saw in the chapter on error control and likelihoods null results are to be expected, and if you only think about the possibility of observing a null effect when the data has been collected, it is often too late. | 在设计研究时，一种明智的策略是始终计划考虑效应的存在与否。在注册报告中，一些科学期刊要求提供样本量的合理性，其中拒绝零假设的统计检验力很高，但研究也能够证明不存在影响，例如通过对等效性检验进行检验力分析。正如我们在误差控制和似然性的章节中看到的那样，零结果是意料之中的，如果您只在收集数据时考虑观察到零效应的可能性，通常为时已晚。 |
| The statistical power for interval hypotheses depend on the alpha level, the sample size, the smallest effect of interest you decide to test against, and the true effect size. For an equivalence test, it is common to perform a power analysis assuming the true effect size is 0, but this might not always be realistic. The closer the expected effect size is to the smallest effect size of interest, the larger the sample size needed to reach a desired power. Don’t be tempted to assume a true effect size of 0, if you have good reason to expect a small but non-zero true effect size. The sample size that the power analysis indicates you need to collect might be smaller, but in reality you also have a higher probability of an inconclusive result. Earlier versions of TOSTER only enabled researchers to perform power analyses for equivalence tests assuming a true effect size of 0, but a new power function by Aaron Caldwell allows users to specify delta, the expected effect size. | 区间假设的统计检验力取决于alpha水平、样本量、您决定检验的最小感兴趣效应以及真实效应大小。对于等效性检验，通常假定真实效应大小为0来执行检验力分析，但这可能并不总是现实的。预期效应量越接近感兴趣的最小效应量，需要达到所需检验力的样本量就越大。如果您有充分的理由预期一个小但非零的真实效应量，请不要试图假定真实效应量为0。检验力分析表明您需要收集的样本量可能较小，但实际上您也更有可能得到不确定的结果。早期版本的 TOSTER 仅允许研究人员在假设真实效应大小为 0 的情况下对等效性检验执行检验力分析，但 Aaron Caldwell 的新检验力函数允许用户指定 delta，即预期的效应量。 |
| Assume a researchers desired to achieve 90% power for an equivalence test with an equivalence range from -0.5 to 0.5, with an alpha level of 0.05, and assuming a population effect size of 0. A power analysis for an equivalence test can be performed to examine the required sample size. | 假设研究人员希望等效性检验达到 90% 的检验力，等效性范围为 -0.5 到 0.5，alpha 水平为 0.05，并假设总体效应量为 0。可以进行等效性检验的检验力分析，从而确定所需的样本量。 |
| TOSTER::power\_t\_TOST(power = 0.9, delta = 0,  alpha = 0.05, type = "two.sample",  low\_eqbound = -0.5, high\_eqbound = 0.5) | NA |
| Two-sample TOST power calculation  power = 0.9  beta = 0.1  alpha = 0.05  n = 87.26261  delta = 0  sd = 1  bounds = -0.5, 0.5  NOTE: n is number in \*each\* group | NA |
| We see that the required sample size is 88 participants in each condition for the independent t-test. Let’s compare this power analysis to a situation where the researcher expects a true effect of d = 0.1, instead of a true effect of 0. To be able to reliably reject effects larger than 0.5, we will need a larger sample size, just as how we need a larger sample size for a null hypothesis test powered to detect d = 0.4 than a null hypothesis test powered to detect d = 0.5. | 我们看到，对于独立样本t检验，所需的样本量为每个条件88位参与者。现在，让我们将这个检验力分析与研究人员预期真实效应为d = 0.1的情况进行比较，而不是真实效应为0。为了能够可靠地拒绝大于0.5的效应，我们将需要更大的样本量，就像我们需要更大的样本量来检测检验力为d = 0.4的零假设，而不是检验力为d = 0.5的零假设一样。 |
| TOSTER::power\_t\_TOST(power = 0.9, delta = 0.1,  alpha = 0.05, type = "two.sample",  low\_eqbound = -0.5, high\_eqbound = 0.5) | NA |
| Two-sample TOST power calculation  power = 0.9  beta = 0.1  alpha = 0.05  n = 108.9187  delta = 0.1  sd = 1  bounds = -0.5, 0.5  NOTE: n is number in \*each\* group | NA |
| We see the sample size has now increased to 109 participants in each condition. As mentioned before, it is not necessary to perform a two-sided equivalence test. It is also possible to perform a one-sided equivalence test. An example of a situation where such a directional test is appropriate is a replication study. If a previous study observed an effect of d = 0.48, and you perform a replication study, you might decide to consider any effect smaller than d = 0.2 a failure to replicate - including any effect in the opposite direction, such as an effect of d = -0.3. Although most software for equivalence tests requires you to specify an upper and lower bound for an equivalence range, you can mimic a one-sided test by setting the equivalence bound in the direction you want to ignore to a low value so that the one-sided test against this value will always be statistically significant. This can also be used to perform a power analysis for a minimum effect test, where one bound is the minimum effect of interest, and the other bound is set to an extreme value on the other side of the expected effect size. | 我们看到，样本量现在增加到每个条件109位参与者。如前所述，并不需要执行双侧等效性检验。也可以执行单侧等效性检验。单侧等效性检验适用的一个例子是重复性研究。如果之前的研究观察到d = 0.48的效应，并且您执行了一项重复性研究，您可能决定将任何小于d = 0.2的效应视为重复失败——包括任何相反方向的效应，例如d = -0.3的效应。虽然大多数等效性检验软件需要您为等效性范围指定一个上限和下限，但您可以通过将您想要忽略的方向的等效性界限设置为一个低值来模拟单侧检验，使得对这个值的单侧检验始终具有统计学意义。这也可以用来执行最小效应检验的检验力分析，其中一个界限是感兴趣的最小效应，另一个界限则设置为预期效应量的另一侧的极端值。 |
| In the power analysis for an equivalence test example below, the lower bound is set to -5 (it should be set low enough such that lowering it even further has no noticeable effect). We see that the new power function in the TOSTER package takes the directional prediction into account, and just as with directional predictions in a nil null hypothesis test, a directional prediction in an equivalence test is more efficient, and only 70 observations are needed to achieve 90% power. | 在下面的等效性检验的检验力分析示例中，下限被设定为-5（应该将其设置得足够低，以便进一步降低对结果没有明显影响）。我们可以看到TOSTER软件包中的新检验力函数考虑了方向性预测，与在零假设检验中的方向预测一样，等效性检验中的方向预测更有效，只需要70个观测值即可达到90%的检验力。 |
| # New TOSTER power functions allows power for expected non-zero effect.  TOSTER::power\_t\_TOST(power = 0.9, delta = 0,  alpha = 0.05, type = "two.sample",  low\_eqbound = -5, high\_eqbound = 0.5) | NA |
| Two-sample TOST power calculation  power = 0.9  beta = 0.1  alpha = 0.05  n = 69.19784  delta = 0  sd = 1  bounds = -5.0, 0.5  NOTE: n is number in \*each\* group | NA |
| Statistical software offers options for power analyses for some statistical tests, but not for all tests. Just as with power analysis for a nil null hypothesis test, it can be necessary to use a simulation-based approach to power analysis. | 统计软件为某些统计检验提供了检验力分析选项，但并非所有检验都具备此功能。正如在零假设检验中进行检验力分析一样，有时需要使用基于模拟的方法进行检验力分析。 |
| **9.5 The Bayesian ROPE procedure** | **9.5 贝叶斯 ROPE 程序** |
| In Bayesian estimation, one way to argue for the absence of a meaningful effect is the region of practical equivalence (ROPE) procedure (Kruschke (2013)), which is “somewhat analogous to frequentist equivalence testing” (Kruschke & Liddell (2017)). In the ROPE procedure, an equivalence range is specified, just as in equivalence testing, but the Bayesian highest density interval based on a posterior distribution (as explained in the chapter on Bayesian statistics) is used instead of the confidence interval. | 在贝叶斯估计中，一种论证缺乏有意义效应的方法是使用实用等效区间（ROPE）程序（Kruschke（2013）），它“有点类似于频率学派的等效性检验”（Kruschke＆Liddell（2017））。在ROPE程序中，指定等效性范围，就像在等效性检验中一样，但是基于后验分布的贝叶斯最高密度区间（如在贝叶斯统计的章节中所解释的）被用来代替置信区间。 |
| If the prior used by Kruschke was perfectly uniform, and the ROPE procedure and an equivalence test used the same confidence interval (e.g., 90%), the two tests would yield identical results. There would only be philosophical differences in how the numbers are interpreted. The BEST package in R that can be used to perform the ROPE procedure by default uses a ‘broad’ prior, and therefore results of the ROPE procedure and an equivalence test are not exactly the same, but they are very close. One might even argue the two tests are ‘practically equivalent’. In the R code below, random normally distributed data for two conditions is generated (with means of 0 and a standard deviation of 1) and the ROPE procedure and a TOST equivalence test are performed. | 如果Kruschke使用的先验分布完全均匀，并且ROPE程序和等效性检验使用相同的置信区间（例如90％），那么两个检验将产生相同的结果。在如何解释数字方面只会存在哲学上的差异。在R中可以使用BEST软件包执行ROPE程序，该软件包默认使用“广泛”的先验分布，因此ROPE程序和等效性检验的结果并不完全相同，但它们非常接近。有人甚至可能会争辩说这两个检验“实际上是等价的”。在下面的R代码中，生成了两个条件下的随机正态分布数据（均值为0，标准差为1），并执行了ROPE程序和TOST等效性检验。 |
| Waiting for parallel processing to complete...done. | 等待并行处理完成...完成。 |
|  | NA |
|  | NA |
| The 90% HDI ranges from -0.06 to 0.39, with an estimated mean based on the prior and the data of 0.164. The HDI falls completely between the upper and the lower bound of the equivalence range, and therefore values more extreme than -0.5 or 0.5 are deemed implausible. The 95% CI ranges from -0.07 to 0.36 with an observed mean difference of 0.15. We see that the numbers are not identical, because in Bayesian estimation the observed values are combined with a prior, and the mean estimate is not purely based on the data. But the results are very similar, and will in most cases lead to similar inferences. The BEST R package also enables researchers to perform simulation based power analyses, which take a long time but, when using a broad prior, yield a result that is basically identical to the sample size from a power analysis for an equivalence test. The biggest benefit of ROPE over TOST is that it allows you to incorporate prior information. If you have reliable prior information, ROPE can use this information, which is especially useful if you don’t have a lot of data. If you use informed priors, check the robustness of the posterior against reasonable changes in the prior in sensitivity analyses. | 90% HDI范围为-0.06到0.39，基于先验和数据估计的平均值为0.164。HDI完全落在等效范围的上限和下限之间，因此超过-0.5或0.5的值被认为不可信的。95% CI范围为-0.07到0.36，观察到的平均差为0.15。我们看到这些数字不是完全相同的，因为在贝叶斯估计中，观察到的值与先验结合，平均估计值不仅仅基于数据。但结果非常相似，并且在大多数情况下会导致相似的推论。BEST R软件包还使研究人员能够执行基于模拟的检验力分析，这需要很长时间，但是使用广泛的先验时，结果与等效性检验的检验力分析的样本量基本相同。ROPE相对于TOST的最大优势在于它允许您纳入先验信息。如果您具有可靠的先验信息，ROPE可以使用此信息，这在您没有大量数据时尤其有用。如果您使用了知情先验，建议进行敏感性分析，检查后验在先验合理变化下的稳健性。 |
| **9.6 Which interval width should be used?** | **9.6 应该使用哪个区间宽度？** |
| Because the TOST procedure is based on two one-sided tests, a 90% confidence interval is used when the one-sided tests are performed at an alpha level of 5%. Because both the test against the upper bound and the test against the lower bound needs to be statistically significant to declare equivalence (which as explained in the chapter on error control is an intersection-union approach to multiple testing) it is not necessary to correct for the fact that two tests are performed. If the alpha level is adjusted for multiple comparisons, or if the alpha level is justified instead of relying on the default 5% level (or both), the corresponding confidence interval should be used, where CI = 100 - (2 \* \(\alpha\)). Thus, the width of the confidence interval is directly related to the choice for the alpha level, as we are making decisions to reject the smallest effect size of interest, or not, based on whether the confidence interval excluded the effect that is tested against. | 因为 TOST 程序基于两个单侧检验，所以当在 5% 的 alpha 水平下执行单侧检验时，将使用 90% 的置信区间。因为针对上限的检验和针对下限的检验都需要具有统计显着性才能声明等效性（正如在误差控制一章中所解释的那样，多重检验的交集-并集方法），所以不必为进行了两次检验而校正。如果针对多重比较调整了 alpha 水平，或者如果 alpha 水平是合理的而不是依赖于默认的 5% 水平（或两者），则应使用相应的置信区间，其中CI = 100 - (2 \* \(\alpha\))  。因此，置信区间的宽度与alpha水平的选择直接相关，因为我们基于置信区间是否排除所检验的效应来决定是否拒绝感兴趣的最小效应量。 |
| When using a Highest Density Interval from a Bayesian perspective, such as the ROPE procedure, the choice for a width of a confidence interval does not follow logically from a desired error rate, or any other principle. Kruschke (2014) writes: “How should we define ‘reasonably credible’? One way is by saying that any points within the 95% HDI are reasonably credible.” McElreath (2016) has recommended the use of 67%, 89%, and 97%, because “No reason. They are prime numbers, which makes them easy to remember.”. Both these suggestions lack a solid justification. As Gosset (or Student), observed (1904): | 当从贝叶斯角度使用最高密度区间时，比如ROPE程序，置信区间宽度的选择在逻辑上不符合所需的错误率或任何其他原则。Kruschke（2014）写道：“我们应该如何定义‘合理可信’？一种方法是说，任何在95% HDI内的点都是合理可信的。”McElreath （2016）推荐使用67%、89%和97%，因为“没有理由。它们是质数，因此很容易记住。” 这两种建议都缺乏坚实的依据。正如Gosset（或学生）观察到的（1904）： |
| Results are only valuable when the amount by which they probably differ from the truth is so small as to be insignificant for the purposes of the experiment. What the odds selected should be depends-  1. On the degree of accuracy which the nature of the experiment allows, and  2. On the importance of the issues at stake. | 结果仅在它们可能与真相相差的程度足够小以至于在实验目的上可以忽略不计时才有价值。选定的赔率应取决于以下两点：  1.实验允许的精度程度，以及  2.相关问题的重要性。 |
| There are only two principled solutions. First, if a highest density interval width is used to make claims, these claims will be made with certain error rates, and researchers should quantify the risk of erroneous claims by computing frequentist error rates. This would make the ROPE procedure a Bayesian/Frequentist compromise procedure, where the computation of a posterior distribution allows for Bayesian interpretations of which parameters values are believed to be most probable, while decisions based on whether or not the HDI falls within an equivalence range have a formally controlled error rate. Note that when using an informative prior, an HDI does not match a CI, and the error rate when using an HDI can only be derived through simulations. The second solution is to not make any claims, present the full posterior distribution, and let readers draw their own conclusions. | 有两种原则性的解决方案。首先，如果使用最高密度区间宽度来做出声明，这些声明将具有一定的错误率，研究人员应该通过计算频率主义的错误率来量化错误声明的风险。这将使ROPE程序成为贝叶斯/频率派的折衷程序，其中后验分布的计算允许贝叶斯解释哪些参数值被认为是最可能的，而基于 HDI 是否落在等价范围内的决策具有 一个正式控制的错误率。请注意，当使用信息先验时，HDI与CI不匹配，并且使用HDI时的错误率只能通过模拟来推导。第二种解决方案是不做任何声明，呈现完整的后验分布，并让读者自己得出结论。 |
| **9.7 Setting the Smallest Effect Size of Interest** | **9.7 设置感兴趣的最小效应大小** |
| To be able to falsify our predictions using an equivalence test is to specify which observed values would be too small to be predicted by our theory. We can never say that an effect is exactly zero, but we can examine whether observed effects are too small to be theoretically or practically interesting. This requires that we specify the smallest effect size of interest (SESOI). The same concept goes by many names, such as a minimal important difference, or clinically significant difference (King, 2011). Take a moment to think about what the smallest effect size is that you would still consider theoretically or practically meaningful for the next study you are designing. It might be difficult to determine what the smallest effect size is that you would consider interesting, and the question what the smallest effect size of interest is might be something you have never really thought about to begin with. However, determining your smallest effect size of interest has important practical benefits. First, if researchers in a field are able to specify which effects would be too small to matter, it becomes very straightforward to power a study for the effects that are meaningful. The second benefit of specifying the smallest effect size of interest is that it makes your study falsifiable. Having your predictions falsified by someone else might not feel that great for you personally, but it is quite useful for science as a whole (Popper, 2002). After all, if there is no way a prediction can be wrong, why would anyone be impressed if the prediction is right? | 能够使用等效性检验来验证我们的预测是否正确，就需要明确规定哪些观察值太小而无法用我们的理论预测。我们永远无法说效应完全为零，但我们可以检查观察到的效应是否太小而不具备理论或实际上的重要性。这需要我们指定感兴趣的最小效应量（SESOI）。同样的概念有许多名称，比如最小重要差异或临床显著差异（King，2011）。花些时间思考一下，对于您正在设计的下一项研究，最小效应量是多少才会认为是理论或实际上有意义的？确定您感兴趣的最小效应量可能很困难，而确定您感兴趣的最小效应量对于实践有重要的好处。首先，如果某个领域的研究人员能够确定哪些效应太小而不重要，那么就可以非常直接地为有意义的效应开展研究。其次，指定感兴趣的最小效应量的好处是可以使您的研究具有可证伪性。您的预测被别人证伪对您个人来说可能感觉不太好，但对整个科学来说却非常有用（Popper，2002）。毕竟，如果没有任何方法可以证明预测是错误的，那么如果预测是正确的，谁会感到惊讶呢？ |
| To start thinking about which effect sizes matter, ask yourself whether any effect in the predicted direction is actually support for the alternative hypothesis. For example, would an effect size of a Cohen’s d of 10 be support for your hypothesis? In psychology, it should be rare that a theory prediucts such a huge effect, and if you observed a d = 10, you would probably check for either a computation error, or a confound in the study. On the other end of the scale, would an effect of d = 0.001 be in line with the theoretically proposed mechanism? Such an effect is incredibly small, and is well below what an individual would notice, as it would fall below the just noticeable difference given perceptual and cognitive limitations. Therefore, a d = 0.001 would in most cases lead researchers to conclude “Well, this is really too small to be something that my theory has predicted, and such a small effect is practically equivalent to the absence of an effect.” However, when we make a directional prediction, we say that these types of effects are all part of our alternative hypothesis. Even though many researchers would agree such tiny effects are too small to matter, they still officially support for our alternative hypothesis if we have a directional prediction with a nil null hypothesis. Furthermore, researchers rarely have the resources to statistically reject the presence of effects this small, so the claim that such effects would still support a theoretical prediction makes the theory practically unfalsifiable: A researcher could simply respond to any replication study showing a non-significant small effect (e.g., d = 0.05) by saying: “That does not falsify my prediction. I suppose the effect is just a bit smaller than d = 0.05”, without ever having to admit the prediction is falsified. This is problematic, because if we do not have a process of replication and falsification, a scientific discipline risks a slide towards the unfalsifiable (Ferguson & Heene, 2012). So whenever possible, when you design an experiment or you have a theory and a theoretical prediction, carefully think about, and clearly state, what the smallest effect size of interest is. | 开始思考哪些效应量是重要的，可以问自己预测方向的任何效应是否实际上支持备择假设？例如，Cohen's d 为 10 的效应量是否支持您的假设？在心理学中，理论很少预测如此巨大的效应量，如果您观察到 d = 10，您可能会检查计算错误或研究中的混淆变量。另一方面，d = 0.001 的效应量是否符合理论提出的机制？这样的效应量非常小，远低于个人能注意到的水平，因为它会低于感知和认知限制的刚刚可察觉的差异。因此，在大多数情况下，d = 0.001 会导致研究人员得出结论：“嗯，这实在是太小了，根本不是我的理论所预测的，这么小的效果，几乎等同于没有效果。”然而，当我们做出方向性预测时，我们说这些类型的效应都是我们备择假设的一部分。尽管许多研究人员会同意这种微小的影响太小而不重要，但如果我们有一个零假设的定向预测，它们仍然是支持我们备择假设的证据。此外，研究人员很少有资源从统计上拒绝如此小的效应的存在，因此声称这种影响仍然支持理论预测使得该理论实际上不可证伪：研究人员可以简单地回应任何显示出非显著小效应的重复性研究（例如 d = 0.05）：“这并没有证伪我的预测，我想效应只是比 d = 0.05 稍微小一些”，而无需承认预测已被证伪。这是有问题的，因为如果我们没有重复性和证伪的过程，科学学科就有滑向不可证伪的风险（Ferguson & Heene, 2012）。因此，只要有可能，当您设计实验或有理论和理论预测时，请仔细考虑并清楚地说明，感兴趣的最小效应量是多少。 |
| **9.8 Specifying a SESOI based on theory** | **9.8根据理论来指定SESOI** |
| One example of a theoretically predicted smallest effect size of interest can be found in the study by Burriss et al. (2015), who examined whether women displayed increased redness in the face during the fertile phase of their ovulatory cycle. The hypothesis was that a slightly redder skin signals greater attractiveness and physical health, and that sending this signal to men yields an evolutionary advantage. This hypothesis presupposes that men can detect the increase in redness with the naked eye. Burriss et al. collected data from 22 women and showed that the redness of their facial skin indeed increased during their fertile period. However, this increase was not large enough for men to detect with the naked eye, so the hypothesis was falsified. Because the just-noticeable difference in redness of the skin can be measured, it was possible to establish a theoretically motivated SESOI. A theoretically motivated smallest effect size of interest can be derived from just-noticeable differences, which provide a lower bound on effect sizes that can influence individuals, or based on computational models, which can provide a lower bound on parameters in the model that will still be able to explain observed findings in the empirical literature. | 一个理论预测的感兴趣的最小效应量的例子可以在Burriss等人（2015）的研究中找到，他们研究了女性在排卵周期的育龄期间是否面部出现增加的红晕。他们的假设是，略微红润的皮肤可以传递更高的吸引力和身体健康性的信号，并且将这种信号发送给男性会产生进化优势。这个假设的前提是男性可以用肉眼检测出红晕的增加。Burriss等人从22名女性收集了数据，结果表明她们面部的红晕确实在育龄期间增加了。然而，这种增加对男性来说不足以用肉眼检测出来，因此假设被证伪。因为可以测量皮肤发红的细微差别，所以有可能建立一个理论上有动机的 SESOI。理论上推动的感兴趣的最小效应量可以从刚好可察觉差异中推导出来，它提供了能够影响个体的效应量的下限，或者基于计算模型，它可以提供模型中参数的下限，该参数仍然会能够解释实证文献中观察到的发现。 |
| **9.9 Anchor based methods to set a SESOI** | **9.9锚定法设置 SESOI** |
| Building on the idea of a just-noticeable difference, psychologists are often interested in effects that are large enough to be noticed by single individuals. One procedure to estimate what constitutes a meaningful change on an individual level is the anchor-based method (Jaeschke et al., 1989; King, 2011; Norman et al., 2004). Measurements are collected at two time points (e.g., a quality of life measure before and after treatment). At the second time point, an independent measure (the anchor) is used to determine if individuals show no change compared to time point 1, or if they have improved, or worsened. Often, the patient is directly asked to answer the anchor question, and indicate if they subjectively feel the same, better, or worse at time point 2 compared to time point 1. Button et al. (2015) used an anchor-based method to estimate that a minimal clinically important difference on the Beck Depression Inventory corresponded to a 17.5% reduction in scores from baseline. | 基于刚刚可察觉差异的想法，心理学家通常对大到足以被单个个体注意到的影响感兴趣。锚定法是估计个体层面上何为有意义变化的程序（Jaeschke等人，1989；King，2011；Norman等人，2004）。该方法需要在两个时间点收集测量数据（例如，治疗前后的生活质量测量）。在第二个时间点，使用独立测量（锚点）来确定与时间点 1 相比个人是否有变化，或者他们是否有所改善或恶化。通常，患者会被直接问及锚定问题，并指出与时间点 1 相比，他们在时间点 2 的主观感觉是否相同、更好或更差。Button等人（2015）使用锚定法估计贝克抑郁量表的最小临床显著差异对应于基线分数降低 17.5%。 |
| Anvari and Lakens (2021) applied the anchor-based method to examine a smallest effect of interest as measured by the widely used Positive and Negative Affect Scale (PANAS). Participants completed the 20 item PANAS at two time points several days apart (using a Likert scale going from 1 = “very slightly or not at all”, to 5 = “extremely”). At the second time point they were also asked to indicate if their affect had changed a little, a lot, or not at all. When people indicated their affect had changed “a little”, the average change in Likert units was 0.26 scale points for positive affect and 0.28 scale points for negative affect. Thus, an intervention to improve people’s affective state that should lead to what individuals subjectively consider at least a little improvement might set the SESOI at 0.3 units on the PANAS. | Anvari和Lakens（2021）应用了锚定法来研究广泛使用的积极和消极情绪量表（PANAS）测量的感兴趣的最小效应量。参与者在相隔几天的两个时间点完成了 20 个项目的 PANAS调查（使用李克特量表，从1 =“非常轻微或根本没有”到5 =“极度”）。在第二个时间点，他们还被问及他们的情感是否发生了一点、很多或根本没有变化。当人们表示他们的情绪“有一点”变化时，积极情绪的平均变化是0.26分，消极情绪的平均变化是0.28分。因此，用于改善人们情绪状态的干预措施，应该导致个体主观上认为至少有一点改善，可以将SESOI设置为PANAS量表上的0.3个单位。 |
| **9.10 Specifying a SESOI based on a cost-benefit analysis** | **9.10 根据成本效益分析确定SESOI** |
| Another principled approach to justify a smallest effect size of interest is to perform a cost-benefit analysis. Research shows that cognitive training may improve mental abilities in older adults which might benefit older drivers (Ball et al., 2002). Based on these findings, Viamonte, Ball, and Kilgore (2006) performed a cost-benefit analysis and concluded that based on the cost of the intervention ($247.50), the probability of an accident for drivers older than 75 (p = 0.0710), and the cost of an accident ($22,000), performing the intervention on all drivers aged 75 or older was more efficient than not intervening or only intervening after a screening test. Furthermore, sensitivity analyses revealed that intervening for all drivers would remain beneficial as long as the reduction in collision risk is 25%. Therefore, a 25% reduction in the probability of elderly above 75 getting into a car accident could be set as the smallest effect size of interest. | 证明感兴趣的最小效应量合理的另一种原则性方法是执行成本效益分析。研究表明，认知训练可能改善老年人的心智能力，从而可能使老年驾驶员受益（Ball等人，2002）。基于这些发现，Viamonte、Ball和Kilgore（2006）进行了成本效益分析并得出结论：根据干预措施的成本（247.50美元），75岁以上的老年驾驶员发生事故的概率（p = 0.0710）和一次事故的成本（22,000美元）相比，对所有75岁及以上的驾驶员进行干预比不干预或仅在筛查检验后干预更为有效。此外，敏感性分析表明，只要碰撞风险降低了25%，对所有驾驶员进行干预仍将是有益的。因此，可以将 75 岁以上老年人发生车祸的概率降低 25% 设置为感兴趣的最小效应量。 |
| For another example, economists have examined the value of a statistical life, based on willingness to pay to reduce the risk of death, at $1.5 - $2.5 million (in the year 2000, in western countries, see Mrozek & Taylor (2002)). Building on this work, Abelson (2003) calculated the willingness to pay to prevent acute health issues such as eye irritation at about $40-$50 per day. A researcher may be examining a psychological intervention that reduces the amount of times people touch their face close to their eyes, thereby reducing eye irritations caused by bacteria. If the intervention costs $20 per year to administer, it therefore should reduce the average number of days with eye irritation in the population by at least 0.5 days for the intervention to be worth the cost. A cost-benefit analysis can also be based on the resources required to empirically study a very small effect when weighed against the value this knowledge would have for the scientific community. | 另一个例子，经济学家根据人们为降低死亡风险愿意支付的费用，计算出一份统计生命价值在150万到250万美元之间（2000 年，在西方国家，参见 Mrozek & Taylor (2002)）在这项工作的基础上，Abelson（2003 年）计算出为预防眼睛刺激等急性健康问题而支付的意愿约为每天 40-50 美元。 研究人员可能正在研究一种心理干预措施，可以减少人们将脸靠近眼睛的次数，从而减少细菌引起的眼睛刺激。 如果干预每年花费 20 美元，那么应该将人群中眼部刺激的平均天数减少至少 0.5 天，干预才值得花费。成本效益分析也可以基于研究非常小的效应所需的资源与这种知识对科学界的价值之间的权衡。 |
| **9.11 Specifying the SESOI using the small telescopes approach** | **9.11使用小型望远镜法确定 SESOI** |
| Ideally, researchers who publish empirical claims would always specify which observations would falsify their claim. Regrettably, this is not yet common practice. This is particularly problematic when a researcher performs a close replication of earlier work. Because it is never possible to prove an effect is exactly zero, and the original authors seldom specify which range of effect sizes would falsify their hypotheses, it has proven to be very difficult to interpret the outcome of a replication study (Anderson & Maxwell, 2016). When does the new data contradict the original finding? | 理想情况下，发表经验声明的研究人员总是会指定哪些观察结果会证伪他们的声明。 遗憾的是，这还不是普遍做法。当研究人员对早期工作进行近似重复性研究时，这尤其成问题。因为永远无法证明一个效应确切等于零，而原作者很少指定哪种效应量的范围将推翻他们的假设，所以很难解释重复性研究的结果(Anderson和Maxwell，2016)。新数据何时与原始发现相矛盾？ |
| Consider a study in which you want to test the idea of the wisdom of crowds. You ask 20 people to estimate the number of coins in a jar, expecting the average to be very close to the true value. The research question is whether the people can on average correctly guess the number of coins, which is 500. The observed mean guess by 20 people is 550, with a standard deviation of 100. The observed difference from the true value is statistically significant, t(19)=2.37, p = 0.0375, with a Cohen’s d of 0.5. Can it really be that the group average is so far off? Is there no Wisdom of Crowds? Was there something special about the coins you used that make it especially difficult to guess their number? Or was it just a fluke? You set out to perform a close replication of this study. | 考虑一项研究，您想在其中检验群体智慧的想法。您让 20 个人估计一个罐子里的硬币数量，期望平均值非常接近真实值。 研究问题是人们是否能平均正确猜出硬币数量，即 500。观察到的 20 人平均猜测为 550，标准差为 100。观察到的与真实值的差异具有统计显着性，t (19)=2.37，p = 0.0375，Cohen's d 为 0.5。 小组平均水平真的相差如此之远吗？ 没有群体的智慧吗？ 您使用的硬币有什么特别之处使您很难猜出它们的数量吗？ 还是只是侥幸？您用近似重复性方法来重新进行这项研究。 |
| You want your study to be informative, regardless of whether there is an effect or not. This means you need to design a replication study that will allow you to draw an informative conclusion, regardless of whether the alternative hypothesis is true (the crowd will not estimate the true number of coins accurately) or whether the null hypothesis is true (the crowd will guess 500 coins, and the original study was a fluke). But since the original researcher did not specify a smallest effect size of interest, when would a replication study allow you to conclude the original study is contradicted by the new data? Observing a mean of exactly 500 would perhaps be considered by some to be quite convincing, but due to random variation you will (almost) never find a mean score of exactly 500. A non-significant result can’t be interpreted as the absence of an effect, because your study might have too small a sample size to detect meaningful effects, and the result might be a Type 2 error. So how can we move forward and define an effect size that is meaningful? How can you design a study that has the ability to falsify a previous finding? | 您希望您的研究不论是否存在效应都能提供有意义的信息。这意味着您需要设计一项重复性研究，都能够得出有意义的结论，无论备择假设是否正确（人群无法准确猜测硬币的数量）或原假设是否正确（人群会猜测500枚硬币，原始的研究是巧合）。但是，由于原始研究人员没有指定感兴趣的最小效应量，复制研究什么时候可以让您得出原始研究与新数据相矛盾的结论？观察到恰好 500 的平均值可能会被某些人认为是非常有说服力的，但由于随机变化，您将（几乎）永远不会找到恰好 500 的平均值。一个不显著的结果不能解释为没有效应，因为您的研究可能样本量太小，无法检测到有意义的效应，结果可能是第二类错误。那么我们应该如何前进并定义一个有意义的效应大小呢？您如何设计一项研究，具有证伪先前研究的能力？ |
| Uri Simonsohn (2015) defines a small effect as “one that would give 33% power to the original study”. In other words, the effect size that would give the original study odds of 2:1 against observing a statistically significant result if there was an effect. The idea is that if the original study had 33% power, the probability of observing a significant effect, if there was a true effect, is too low to reliably distinguish signal from noise (or situations where there is a true effect from situations where there is no true effect). Simonsohn (2015, p. 561) calls this the small telescopes approach, and writes: “Imagine an astronomer claiming to have found a new planet with a telescope. Another astronomer tries to replicate the discovery using a larger telescope and finds nothing. Although this does not prove that the planet does not exist, it does nevertheless contradict the original findings, because planets that are observable with the smaller telescope should also be observable with the larger one.” | Uri Simonsohn（2015）将小效应定义为“能够给原始研究提供33％的检验力”。换句话说，如果存在效应，原始研究获得2:1的胜算来观察到统计显著性结果的效应大小。这个想法是，如果原始研究有33％的检验力，那么如果存在真正的效应，观察到显著效应的概率太低，不能可靠地区分信号和噪音（或存在真正效应的情况和不存在真正效应的情况）。 Simonsohn（2015，第561页）称之为小型望远镜法，并写道：“想象一位天文学家使用望远镜声称发现了一个新行星。另一位天文学家试图使用更大的望远镜复制发现，但没有发现任何东西。尽管这并不能证明行星不存在，但它确实与原始发现相矛盾，因为使用较小望远镜可以观测到的行星也应该可以用更大望远镜观测到。” |
| Although this approach to setting a smallest effect size of interest (SESOI) is arbitrary (why not 30% power, or 35%?) it suffices for practical purposes (and you are free to choose a power level you think is too low). The nice thing about this definition of a SESOI is that if you know the sample size of the original study, you can always calculate the effect size that study had 33% power to detect. You can thus always use this approach to set a smallest effect size of interest. If you fail to find support for an effect size the original study has 33% power to detect, it does not mean there is no true effect, and not even that the effect is too small to be of any theoretical or practical interest. But using the small telescopes approach is a good first step, since it will get the conversation started about which effects are meaningful and allows researchers who want to replicate a study to specify when they would consider the original claim falsified. | 虽然这种方法设定感兴趣的最小效应量是随意的（为什么不是30%或35%？），但它足以满足实际目的（您可以自由选择您认为过低的检验力水平）。SESOI 的这个定义的好处是，如果您知道原始研究的样本量，您总是可以计算出该研究具有 33% 检测能力的效应量。 因此，您始终可以使用这种方法来设置感兴趣的最小效应量。如果您未能找到对原始研究具有 33% 检测检验力的效应量的支持，这并不意味着没有真正的效应，甚至也不意味着效应太小以至于没有任何理论或实践意义。但是使用小型望远镜法是很好的开端，因为它将开始讨论哪些影响是有意义的，并允许想要进行重复研究的研究人员指定他们何时会认为原始声明是伪造的。 |
| With the small telescopes approach, the SESOI is based only on the sample size in the original study. A smallest effect size of interest is set only for effects in the same direction. All effects smaller than this effect (including large effects in the opposite direction) are interpreted as a failure to replicate the original results. We see that the small telescopes approach is a one-sided equivalence test, where only the upper bound is specified, and the smallest effect size of interest is determined based on the sample size of the original study. The test examines if we can reject effects as large or larger than the effect the original study has 33% power to detect. It is a simple one-sided test, not against 0, but against a SESOI. | 使用小型望远镜法，SESOI 仅基于原始研究中的样本量。仅针对相同方向的效果设置感兴趣的最小效应量。 所有小于此效应的（包括相反方向的大效应）都被解释为无法重复原始结果。我们可以看到，小型望远镜法是一个单侧等价性检验，只指定了上限，感兴趣的最小效应量是基于原始研究的样本大小确定的。该检验检查我们是否可以拒绝与原始研究有33％检验力检测到的效应一样大或更大的效应。它是一个简单的单侧检验，不是针对0，而是针对SESOI。 |
| For example, consider our study above in which 20 guessers tried to estimate the number of coins. The results were analyzed with a two-sided one-sample t-test, using an alpha level of 0.05. To determine the effect size that this study had 33% power for, we can perform a sensitivity analysis. In a sensitivity analysis we compute the required effect size given the alpha, sample size, and desired statistical power. Note that Simonsohn uses a two-sided test in his power analyses, which we will follow here – if the original study reported a pre-registered directional prediction, the power analysis should be based on a one-sided test. In this case, the alpha level is 0.05, the total sample size is 20, and the desired power is 33%. We compute the effect size that gives us 33% power and see that it is a Cohen’s d of 0.358. This means we can set our smallest effect size of interest for the replication study to d = 0.358. If we can reject effects as large or larger than d = 0.358, we can conclude that the effect is smaller than anything the original study had 33% power for. The screenshot below illustrates the correct settings in G\*Power, and the code in R is: | 例如，考虑我们上面的研究，其中 20 位猜测者试图估计硬币的数量。 使用 0.05 的 alpha 水平，使用双侧单样本 t 检验分析结果。为了确定本研究具有 33% 检验力的效应量，我们可以进行敏感性分析。 在敏感性分析中，我们根据 alpha、样本量和所需的统计检验力计算所需的效应量。 请注意，Simonsohn 在他的检验力分析中使用了双侧检验，我们将在此处遵循——如果原始研究报告了预先登记的方向预测，则检验力分析应基于单侧检验。 在本例中，alpha 水平为 0.05，总样本量为 20，所需检验力为 33%。 我们计算给我们 33% 检验力的效应大小，发现它是 Cohen 的 d 值 0.358。 这意味着我们可以将重复研究感兴趣的最小效应量设置为 d = 0.358。 如果我们可以拒绝大于或大于 d = 0.358 的效应，我们可以得出结论，该效应小于原始研究具有 33% 检验力的任何效应。 下面的屏幕截图说明了 G\*Power 中的正确设置，R 中的代码是： |
| library("pwr")  pwr::pwr.t.test(  n = 20,  sig.level = 0.05,  power = 0.33,  type = "one.sample",  alternative = "two.sided"  ) | NA |
| One-sample t test power calculation  n = 20  d = 0.3577466  sig.level = 0.05  power = 0.33  alternative = two.sided | NA |
|  | NA |
| Figure 9.6: Screenshot illustrating a sensitivity power analysis in G\*Power to compute the effect size an original study had 33% power to detect. | 图9.6：G\*Power中演示敏感性检验力分析的截图，用于计算原始研究能够检测到33%检验力的效应大小。 |
| Determining the SESOI based on the effect size the original study had 33% power to detect has an additional convenient property. Imagine the true effect size is actually 0, and you perform a statistical test to see if the data is statistically smaller than the SESOI based on the small telescopes approach (which is called an inferiority test). If you increase the sample size by 2.5 times, you will have approximately 80% power for this one-sided equivalence test, assuming the true effect size is exactly 0 (e.g., d = 0). People who do a replication study can follow the small telescope recommendations, and very easily determine both the smallest effect size of interest, and the sample size needed to design an informative replication study, assuming the true effect size is 0 (but see the section above for a-priori power analyses where you want to test for equivalence, but do not expect a true effect size of 0). | 基于原始研究的效应大小能够检测到33%的检验力，确定SESOI具有额外的方便性质。想象一下，真实的效应大小实际上为0，并且您执行统计检验以查看数据是否在统计上小于基于小型望远镜法的 SESOI（这称为劣势检验）。 如果将样本量增加 2.5 倍，假设真实效应量恰好为 0（例如，d = 0），则此单侧等价检验的检验力约为 80%。进行重复研究的人可以遵循小型望远镜的建议，并且可以很容易地确定感兴趣的最小效应量和设计信息复制研究所需的样本量，假设真实效应量为 0（但请参阅上一节对于先验检验力分析，您想要检验等价性，但不要期望真实效应大小为 0）。 |
| The figure below, from Simonsohn (2015) illustrates the small telescopes approach using a real-life example. The original study by Zhong and Liljenquist (2006) had a tiny sample size of 30 participants in each condition and observed an effect size of d = 0.53, which was barely statistically different from zero. Given a sample size of 30 per condition, the study had 33% power to detect effects larger than d = 0.401. This “small effect” is indicated by the green dashed line. In R, the smallest effect size of interest is calculated using: | 下图来自 Simonsohn（2015 ），使用现实生活中的例子说明了小型望远镜法。 Zhong 和 Liljenquist（2006 ）的最初研究在每种情况下的样本量很小，只有 30 名参与者，观察到的效应量为 d = 0.53，这与零几乎没有统计学差异。 假设每个条件的样本量为 30，则该研究有 33% 的检验力来检测大于 d = 0.401 的影响。 这种“小效果”由绿色虚线表示。 在 R 中，感兴趣的最小效应量是使用以下方法计算的： |
| pwr::pwr.t.test(  n = 30,  sig.level = 0.05,  power = 1/3,  type = "two.sample",  alternative = "two.sided"  ) | NA |
| Two-sample t test power calculation  n = 30  d = 0.401303  sig.level = 0.05  power = 0.3333333  alternative = two.sided  NOTE: n is number in \*each\* group | NA |
| Note that 33% power is a rounded value, and the calculation uses 1/3 (or 0.3333333…). | 请注意，33%的统计检验力是一个取整的值，计算时使用了1/3（或0.3333333 ...）。 |
|  | NA |
| Figure 9.7: Example used in Simonsohn (2015) of an original study and two replication studies. | 图 9.7：Simonsohn (2015) 在一项原始研究和两项重复研究中使用的示例。 |
| We can see that the first replication by Gámez and colleagues also had a relatively small sample size (N = 47, compared to N = 60 in the original study), and was not designed to yield informative results when interpreted with a small telescopes approach. The confidence interval is very wide and includes the null effect (d = 0) and the smallest effect size of interest (d = 0.401). Thus, this study is inconclusive. We can’t reject the null, but we can also not reject effect sizes of 0.401 or larger that are still considered to be in line with the original result. The second replication has a much larger sample size, and tells us that we can’t reject the null, but we can reject the smallest effect size of interest, suggesting that the effect is smaller than what is considered an interesting effect based on the small telescopes approach. | 我们可以看到，Gámez及其同事进行的第一次重复研究也具有相对较小的样本量（N = 47，相对于原始研究中的N = 60），并且不是为了通过小型望远镜法产生有意义的结果而设计的。置信区间非常宽，包括零效应（d = 0）和感兴趣的最小效应量（d = 0.401）。因此，这项研究是无法确定的。我们不能否认零值，但我们也不能否认大于0.401的效应量，因为这仍然符合原始结果。第二次重复研究具有更大的样本量，并告诉我们不能否认零值，但我们可以拒绝感兴趣的最小效应量，这表明该效应小于根据小型望远镜法认为有趣的效应。 |
| Although the small telescope recommendations are easy to use, one should take care not to turn any statistical procedure into a heuristic. In our example above with the 20 referees, a Cohen’s d of 0.358 would be used as a smallest effect size of interest, and a sample size of 50 would be collected (2.5 times the original 20), but if someone would make the effort to perform a replication study, it would be relatively easy to collect a larger sample size. Alternatively, had the original study been extremely large, it would have had high power for effects that might not be practically significant, and we would not want to collect 2.5 times as many observations in a replication study. Indeed, as Simonsohn writes: “whether we need 2.5 times the original sample size or not depends on the question we wish to answer. If we are interested in testing whether the effect size is smaller than d33%, then, yes, we need about 2.5 times the original sample size no matter how big that original sample was. When samples are very large, however, that may not be the question of interest.” Always think about the question you want to ask, and design the study so that it provides an informative answer for a question of interest. Do not automatically follow a 2.5 times n heuristic, and always reflect on whether the use of a suggested procedure is appropriate in your situation. ::: | 虽然小望远镜建议易于使用，但应注意不要将任何统计程序变成启发式程序。 在我们上面关于20名裁判的例子中，Cohen's d为0.358将用作感兴趣的最小效应量，并且将收集50个样本量（原始20个的2.5倍），但如果有人付出努力进行重复性研究，则收集更大的样本量将相对容易。或者，如果原始研究非常大，则对于可能不太实际的效应具有很高的检验力，我们将不希望在重复研究中收集2.5倍于原始观测的数量。事实上，正如Simonsohn所写：“我们是否需要原始样本量的2.5倍取决于我们希望回答的问题。如果我们想检验效应量是否小于d33％，那么，无论原始样本大小如何，我们都需要大约2.5倍的原始样本量。但是，当样本非常大时，这可能不是我们感兴趣的问题。”始终考虑您想要问的问题，并设计研究，以便为感兴趣的问题提供信息丰富的答案。不要自动遵循2.5倍n的启发式方法，并且始终反思建议程序在您的情况下是否适当。 ::: |
| ::: {.webex-check .webex-box} ## Setting the Smallest Effect Size of Interest to the Minimal Statistically Detectable Effect | NA |
| Given a sample size and alpha level, every test has a minimal statistically detectable effect. For example, given a test with 86 participants in each group, and an alpha level of 5%, only t-tests which yield a t ≥ 1.974 will be statistically significant. In other words, t = 1.974 is the critical t-value. Given a sample size and alpha level, the critical t-value can be transformed into a critical d-value. As visualized in Figure 9.8, with n = 50 in each group and an alpha level of 5% the critical d-value is 0.4. This means that only effects larger than 0.4 will yield a p < α. The critical d-value is influenced by the sample size per group, and the alpha level, but does not depend on the the true effect size. | 给定样本量和 alpha 水平，每个检验都具有最小的统计可检验效应。例如，给定每组有 86 名参与者的检验，且 alpha 水平为 5%，只有 t ≥ 1.974 的 t 检验才具有统计显著性。 换句话说，t = 1.974 是临界 t 值。 给定样本大小和 alpha 水平，可以将临界 t 值转换为临界 d 值。如图 9.8 所示，每组 n = 50，alpha水平为 5%，临界 d 值为 0.4。这意味着只有大于 0.4 的效应才会产生 p < α。 临界 d 值受每组样本量和 alpha 水平的影响，但不取决于真实效应量。 |
|  | NA |
| Figure 9.8: Null and alternative distribution with Type 1 and Type 2 error indicating the smallest effect size that will be statistically significant with n = 50 per condition. | 图 9.8：具有一类和二类错误的零分布和备择分布表明最小的效应量将具有统计显著性，每个条件 n = 50。 |
| It is possible to observe a statistically significant test result if the true effect size is smaller than the critical effect size. Due to random variation, it is possible to observe a larger value in a sample than is the true value in the population. This is the reason the statistical power of a test is never 0 in a null hypothesis significance test. As illustrated in Figure 9.9, even if the true effect size is smaller than the critical value (i.e., if the true effect size is 0.2) we see from the distribution that we can expect some observed effect sizes to be larger than 0.4 when the true population effect size is d = 0.2 – if we compute the statistical power for this test, it turns out we can expect 16.77% of the observed effect sizes will be larger than 0.4, in the long run. That is not a lot, but it is something. This is also the reason why publication bias combined with underpowered research is problematic: It leads to a large overestimation of the true effect size when only observed effect sizes from statistically significant findings in underpowered studies end up in the scientific literature. | 如果真实效应量小于临界效应量，则可以观察到统计显著的检验结果。由于随机变化，有可能在样本中观察到比总体中的真实值更大的值。这就是为什么在零假设显著性检验中，检验的统计检验力永远不为零的原因。正如图9.9所示，即使真实效应量小于临界值（例如，真实效应量为0.2），我们从分布中可以看出，当真实效应量为 总体效应量为 d = 0.2——如果我们计算此检验的统计检验力，结果表明从长远来看，我们可以预期观察到的效应量中有 16.77% 会大于 0.4。 这不是很多，但它是一些东西。 这也是为什么发表偏倚与检验力不足的研究相结合会产生问题：当只有在检验力不足的研究中从具有统计学意义的发现中观察到的效应量最终出现在科学文献中时，它会导致对真实效应量的大幅高估。 |
|  | NA |
| Figure 9.9: Null and alternative distribution with Type 1 and Type 2 error indicating the smallest effect size that will be statistically significant with n = 50 per condition. | 图 9.9：具有一类和二类错误的零分布和备择分布表明最小效应量将具有统计显著性，每个条件 n = 50。 |
| We can use the minimal statistically detectable effect to set the SESOI for replication studies. If you attempt to replicate a study, one justifiable option when choosing the smallest effect size of interest (SESOI) is to use the smallest observed effect size that could have been statistically significant in the study you are replicating. In other words, you decide that effects that could not have yielded a p-value less than α in an original study will not be considered meaningful in the replication study. The assumption here is that the original authors were interested in observing a significant effect, and thus were not interested in observed effect sizes that could not have yielded a significant result. It might be likely that the original authors did not consider which effect sizes their study had good statistical power to detect, or that they were interested in smaller effects but gambled on observing an especially large effect in the sample purely as a result of random variation. Even then, when building on earlier research that does not specify a SESOI, a justifiable starting point might be to set the SESOI to the smallest effect size that, when observed in the original study, could have been statistically significant. Not all researchers might agree with this (e.g., the original authors might say they actually cared just as much about an effect of d =0.001). However, as we try to change the field from the current situation where no one specifies what would falsify their hypothesis, or what their smallest effect size of interest is, this approach is one way to get started. In practice, as explained in the section on post-hoc power, due to the relation between p = 0.05 and 50% power for the observed effect size, this justification for a SESOI will mean that the SESOI is set to the effect size the original study had 50% power to detect for an independent t test. This approach is in some ways similar to the small telescopes approach by Simonsohn (2015), except that it will lead to a somewhat larger SESOI. | 我们可以使用最小可检测效应来设置重复研究的SESOI。如果您试图重复一项研究，选择感兴趣的最小效应量（SESOI）的一个合理选项是使用在您重复的研究中可能具有统计显著性的最小观察到的效应量。换句话说，您决定在复制研究中不考虑那些在原始研究中无法产生小于α的P值的影响是没有意义的。这里的假设是原作者希望观察到显著效应，因此对观察到的无法产生显著结果的效果量不感兴趣。原始作者可能没有考虑他们的研究具有良好的统计能力来检测哪些效应大小，或者他们对较小的效应感兴趣，但赌注观察样本中纯粹由于随机变异而产生的特别大的效应。即使那样，当建立在未指定 SESOI 的早期研究的基础上时，合理的起点可能是将 SESOI 设置为最小效应量，当在原始研究中观察到时，该效应量可能具有统计显著性。并非所有研究人员都会同意这一点（例如，原始作者可能会说他们实际上也关心d = 0.001的效应）。然而，当我们试图改变目前没有人指定什么会证伪他们的假设，或者他们感兴趣的最小效应量是什么的情况时，这种方法是一种开始的方式。实际上，如事后检验力部分所述，由于观察到的效应量的 p = 0.05 和 50% 检验力之间的关系，这种对 SESOI 的证明将意味着 SESOI 被设置为原始效应量研究有 50% 的能力来检测独立的t检验。这种方法在某些方面类似于 Simonsohn (2015) 的小型望远镜法，只是它会导致更大的 SESOI。 |
| Setting a smallest effect size of interest for a replication study is a bit like a tennis match. Original authors serve and hit the ball across the net, saying ‘look, something is going on’. The approach to set the SESOI to the effect size that could have been significant in the original study is a return volley which allows you to say ‘there does not seem to be anything large enough that could have been significant in your own original study’ after performing a well-designed replication study with high statistical power to reject the SESOI. This is never the end of the match – the original authors can attempt to return the ball with a more specific statement about effects their theory predicts, and demonstrate such a smaller effect size is present. But the ball is back in their court, and if they want to continue to claim there is an effect, they will have to support their claim by new data. | 为重复研究设置感兴趣最小效应量有点像网球比赛。原始作者发球并把球打过网，说“看，有些东西正在发生”。将SESOI设置为原始研究中可能会显著的效应大小的方法是回球，这样在进行设计良好、统计检验力高的重复研究后，您可以说“在您的原始研究中似乎没有足够大的效应能够显著”，这并不是比赛的终点——原始作者可以尝试以更具体的方式击回球，陈述说明其理论预测的效应，并证明存在这样更小的效应量。但球回到他们这边了，如果他们想继续声称存在效应，他们将不得不通过新数据支持自己的主张。 |
| Beyond replication studies, the amount of data that is collected limits the inferences one can make. It is also possible to compute a minimal statistically detectable effect based on the sample sizes that are typically used in a research field. For example, imagine a line of research in which a hypothesis has almost always been tested by performing a one-sample t-test, and where the sample sizes that are collected are always smaller than 100 observations. A one-sample t-test on 100 observations, using an alpha of .05 (two sided), has 80% power to detect an effect of d = 0.28 (as can be calculated in a sensitivity power analysis). In a new study, concluding that one can reliably reject the presence of effects more extreme than d = 0.28 suggests that sample sizes of 100 might not be enough to detect effects in such research lines. Rejecting the presence of effects more extreme than d = 0.28 does not test a theoretical prediction, but it contributes to the literature by answering a resource question. It suggests that future studies in this research line will need to change the design of their studies by substantially increasing the sample size. Setting the smallest effect size of interest based on this approach does not answer any theoretical question (after all, the SESOI is not based on any theoretical prediction). But informing peers that given the sample size commonly collected in a field in a field, the effect is not large enough so that it can be reliably studied is a useful contribution to the literature. It does not mean that the effect is not interesting per se, and a field might decide that it is time to examine the research question collaboratively, by coordinating research lines, and collecting enough data to reliably study whether a smaller effect is present. | 除了重复研究之外，收集的数据量限制了人们能够做出的推论。根据研究领域通常使用的样本量，也可以计算出最小的统计可检测效应。例如，假设一个研究领域中的假设几乎总是通过执行单样本 t 检验来检验，并且收集的样本大小始终小于 100个观测值。对 100 个观察值的单样本 t 检验，使用 0.05 的 alpha（双侧），具有 80% 的检验力来检测 d = 0.28 的影响（可以在灵敏度检验力分析中计算）。在一项新研究中，得出结论认为可以可靠地拒绝存在比 d = 0.28 更极端的影响，这表明 100 的样本量可能不足以检测此类研究系列中的影响。拒绝比 d = 0.28 更极端影响的存在并不能检验理论预测，但它通过回答资源问题对文献做出贡献。这表明该研究领域的未来研究将需要通过大幅增加样本量来改变研究设计。基于这种方法设置感兴趣的最小效应量并不能回答任何理论问题（毕竟，SESOI 不基于任何理论预测）。但是，告知同行，在给定研究领域通常收集的样本量的情况下，效应不足以进行可靠地研究，这是对文献的有益贡献。这并不意味着该效应本身并不有趣，并且一个领域可能会决定是时候通过协调研究路线并收集足够的数据来可靠地研究是否存在较小的效应来协作检查研究问题。 |