**9 Equivalence Testing and Interval Hypotheses**

**9 等效性检验与区间假设**

Most scientific studies are designed to test the prediction that an effect or a difference exists. Does a new intervention work? Is there a relationship between two variables? These studies are commonly analyzed with a null hypothesis significance test. When a statistically significant *p*-value is observed, the null hypothesis can be rejected, and researchers can claim that the intervention works, or that there is a relationship between two variables, with a maximum error rate. But if the *p*-value is not statistically significant, researchers very often draw a logically incorrect conclusion: They conclude there is no effect based on *p* > 0.05.

大多数科学研究都是为了检验效应或差异存在的预测。新的干预措施有效吗？两个变量之间有关系吗？这些研究通常采用零假设显著性检验进行分析。当观察到具有统计学意义的*p*值时，研究人员能拒绝零假设，并且研究人员可以以最大的错误率声称干预有效，或者两个变量之间存在关系。但是，如果*p*值在统计意义上不显著，研究人员往往会得出一个逻辑上不正确的结论：他们基于*p* ＞ 0.05的结果得出结论，研究不存在任何效应。

Open a result section of an article you are writing, or the result section of an article you have recently read. Search for “*p* > 0.05”, and look carefully at what you or the scientists concluded (in the results section, but also check which claim they make in the discussion section). If you see the conclusion that there was ‘no effect’ or there was ‘no association between variables’, you have found an example where researchers forgot that *absence of evidence is not evidence of absence* (Altman & Bland, 1995). A non-significant result in itself only tells us that we cannot reject the null hypothesis. It is tempting to ask after *p* > 0.05 ‘so, is the true effect zero’? But the *p*-value from a null hypothesis significance test cannot answer that question. It might be useful to think of the answer to the question whether an effect is absent after observing *p* > 0.05 as (mu), used as a non-dualistic answer, neither yes nor no, or ‘unasking the question’. It is simply not possible to answer the question whether a meaningful effect is absent based on *p* > 0.05.

打开你正在写的一篇论文的结果部分，或者你最近读过的一篇论文的结果部分。搜索“*p*>0.05”，仔细查看你或科学家得出的结论（在结果部分，但也要检查他/她们在讨论部分的说法）。如果你看到“没有效应”或“变量之间没有关联”的结论，你就会发现一个例子，研究人员忘记了*证据缺乏不等于没有证据*(Altman & Bland, 1995)。一个不显著的结果本身只是告诉我们，我们不能拒绝零假设。在观察到*p*＞0.05之后，人们很容易问，“那么，真正的效应是零吗？”但来自零假设显著性检验的*p*值不能回答这个问题。在观察到*p*＞0.05后，将是否存在效应的问题的答案视为（mu），用作非二元答案，既不是“是”也不是“否”，或者“未提出问题”，这可能是有用的。基于*p*＞0.05，根本不可能回答一个有意义的效应是否不存在的问题。

There should be many situations where researchers are interested in examining whether a meaningful effect is absent. For example, it can be important to show two groups do not differ on factors that might be a confound in the experimental design (e.g., examining whether a manipulation intended to increase fatigue did not affect the mood of the participants, by showing that positive and negative affect did not differ between the groups). Researchers might want to know if two interventions work equally well, especially when the newer intervention costs less or requires less effort (e.g., is online therapy just as eﬀicient as in person therapy?). And other times we might be interested to demonstrate the absence of an effect because a theoretical model predicts there is no effect, or because we believe a previously published study was a false positive, and we expect to show the absence of an effect in a replication study (Dienes, 2014). And yet, when you ask researchers if they have ever designed a study where the goal was to show that there was no effect, for example by predicting that there would be no difference between two conditions, many people say they have never designed a study where their main prediction was that the effect size was 0. Researchers almost always predict there is a difference. One reason might be that many researchers would not even know how to statistically support a prediction of an effect size of 0, because they were not trained in the use of equivalence testing.

在许多情况下，研究人员都有兴趣检验一个有意义的效应是否不存在。例如，重要的是去证明两个组别在实验设计中可能混淆的因素上没有差异（例如，通过证明两组之间的积极和消极影响没有差异，检验旨在增加疲劳的操作是否不会影响被试的情绪）。研究人员可能想知道两种干预措施是否同样有效，尤其是当新的干预措施成本更低或需要更少的努力时（例如，线上治疗和面对面治疗一样有效吗？）。并且其他时候，我们可能有兴趣证明效应不存在，因为理论模型预测没有效应，或者因为我们认为之前发表的研究是假阳性，我们希望在重复研究中证明效应不存在(Dienes, 2014)。然而，当你问研究人员，他们是否设计过一项旨在证明没有效应的研究，例如预测两种条件之间没有差异时，许多人说，他们从未设计过一个主要预测是效应大小为0的研究。研究人员几乎总是预测会有差异。其中一个原因可能是许多研究人员甚至不知道如何在统计上支持一个效应大小为0的预测，因为他们没有接受过使用等效性检验的训练。

It is never possible to show an effect is *exactly* 0. Even if you collected data from every person in the world, the effect in any single study will randomly vary around the true effect size of 0 - you might end up with a mean difference that is very close to, but not exactly, zero, in any finite sample. Hodges & Lehmann (1954) were the first to discuss the statistical problem of testing whether two populations have the same mean. They suggest (p. 264) to: “test that their means do not differ by more than an amount specified to represent the smallest difference of practical interest”. Nunnally (1960) similarly proposed a ‘fixed-increment’ hypothesis where researchers compare an observed effect against a range of values that is deemed too small to be meaningful. Defining a range of values considered practically equivalent to the absence of an effect is known as an **equivalence range** (Bauer & Kieser, 1996) or a **region of practical equivalence** (Kruschke, 2013). The equivalence range should be specified in advance, and requires careful consideration of the smallest effect size of interest.

永远不可能证明一个效应大小正好是0。即使你从世界上每个人那里收集到数据，任何一项研究中的效应都会在真实效应量0左右随机变化——在任何有限的样本中，你最终可能会得到非常接近但不完全为0的平均数差异。Hodges和Lehmann（1954）是第一个讨论检验两个群体是否具有相同平均值的统计问题的人。他们建议（第264页）：“检验其平均值差异不超过规定的代表实际关注的最小差异”。Nunnally（1960）同样提出了一个“固定增量”假设，研究人员将观察到的效应与一个被认为太小而没有意义的值的范围进行比较。定义一个被认为实际上等同于没有效应的值的范围被称为一个**等效范围**(Bauer & Kieser, 1996)或**实际等效区域**(Kruschke, 2013)。等效范围应提前规定，并需要仔细考虑关注的最小效应量。

Although researchers have repeatedly attempted to introduce tests against an equivalence range in the social sciences (Cribbie et al., 2004; Hoenig & Heisey, 2001; Levine et al., 2008; Quertemont, 2011; J. L. Rogers et al., 1993), this statistical approach has only recently become popular. During the replication crisis, researchers searched for tools to interpret null results when performing replication studies. Researchers wanted to be able to publish informative null results when replicating findings in the literature that they suspected were false positives. One notable example were the studies on pre-cognition by Daryl Bem, which ostensibly showed that participants were able to predict the future (Bem, 2011). Equivalence tests were proposed as a statistical approach to answer the question whether an observed effect is small enough to conclude that a previous study could not be replicated (S. F. Anderson & Maxwell, 2016; Lakens, 2017; Simonsohn, 2015). Researchers specify a smallest effect size of interest (for example an effect of 0.5, so for a two-sided test any value outside a range from -0.5 to 0.5) and test whether effects more extreme than this range can be rejected. If so, they can reject the presence of effects that are deemed large enough to be meaningful.

尽管研究人员一再试图在社会科学中引入针对等效范围的检验(Cribbie et al., 2004; Hoenig & Heisey, 2001; Levine et al. , 2008; Quertemont, 2011; J.L.Rogers et al., 1993)，但这种统计方法直到最近才流行起来。在可重复性危机期间，研究人员在进行重复研究时寻找解释无效结果的工具。研究人员希望在重复他们怀疑是假阳性的文献中的发现时，能够发布信息丰富的无效结果。一个值得注意的例子是Daryl Bem对前认知的研究，该研究表面上表明被试能够预测未来(Bem, 2011)。等效性检验被提议作为一种统计方法，以回答观察到的效应是否小到足以得出先前研究无法重复的结论的问题(S.F.Anderson & Maxwell, 2016; Lakens, 2017; Simonsohn, 2015)。研究人员指定了关注的最小效应量（例如0.5的效应，因此对于双侧检验来说，是在-0.5到0.5范围之外的任何值），并检验是否可以拒绝比这个范围更极端的效应。如果是这样，他们可以拒绝那些被认为足够大而有意义的效应的存在。

One can distinguish a **nil null hypothesis**, where the null hypothesis is an effect of 0, from a **non-nil null hypothesis**, where the null hypothesis is any other effect than 0, for example effects more extreme than the smallest effect size of interest (Nickerson, 2000). As Nickerson writes:

人们可以将**0零假设**与**非0零假设**区分开来，其中零假设是效应为0，非0零假设是除0之外的任何其他效应，例如比关注的最小效应量更极端的效应（Nickerson，2000）。正如尼克森所写：

The distinction is an important one, especially relative to the controversy regarding the merits or shortcomings of NHST inasmuch as criticisms that may be valid when applied to nil hypothesis testing are not necessarily valid when directed at null hypothesis testing in the more general sense.

这种区别是一个重要的区别，尤其是相对于有关NHST优点或缺点的争议，因为当应用于0假设检验时可能有效的批评在更普遍意义上针对零假设检验的时候不一定有效。

Equivalence tests are a specific implementation of **interval hypothesis tests**, where instead of testing against a null hypothesis of no effect (that is, an effect size of 0; **nil null hypothesis**), an effect is tested against a null hypothesis that represents a range of non-zero effect sizes (**non-nil null hypothesis**). Indeed, one of the most widely suggested improvements that mitigates the most important limitations of null hypothesis significance testing is to replace the nil null hypothesis with the test of a range prediction (by specifying a non-nil null hypothesis) in an interval hypothesis test (Lakens, 2021). To illustrate the difference, Panel A in Figure 9.1 visualizes the results that are predicted in a two-sided null hypothesis test with a nil hypothesis, where the test examines whether an effect of 0 can be rejected. Panel B shows an interval hypothesis where an effect between 0.5 and 2.5 is predicted, where the non-nill null hypothesis consists of values smaller than 0.5 or larger than 2.5, and the interval hypothesis test examines whether values in these ranges can be rejected. Panel C illustrates an equivalence test, which is basically identical to an interval hypothesis test, but the predicted effects are located in a range around 0, and contain effects that are deemed too small to be meaningful.

等效性检验是**区间假设检验**的一种具体实施方式，在这种检验中不是针对无效应的零假设（即效应量为**0**；**0零假设**）进行检验，而是针对代表一系列非0效应量的零假设来检验效应（**非0零假设**）。事实上，零假设显著性检验最重要的局限性最广泛建议的缓解措施之一是用区间假设检验中的范围预测检验（通过指定非0零假设）取代0零假设(Lakens，2021)。为了说明这种差异，图9.1中的面板A可视化了在具有0假设的双侧零假设检验中预测的结果，在该检验中，检验是否可以拒绝效应为0的假设。面板B显示了区间假设，其中预测了一个在0.5和2.5之间的效应，其中非0零假设由小于0.5或大于2.5的值组成，并且区间假设检验检验这些范围内的值能否被拒绝。面板C展示了一个等效性检验，它基本上与区间假设检验相同，但预测的效应位于0左右的范围内，并且包含被认为太小而没有意义的效应。

When an equivalence test is reversed, a researcher designs a study to reject effects less extreme than a smallest effect size of interest (see Panel D in Figure 9.1), it is called a **minimum effect test** (Murphy & Myors, 1999). A researcher might not just be interested in rejecting an effect of 0 (as in a null hypothesis significance test) but in rejecting a range of effects that are too small to be meaningful. All else equal, a study designed to have high power for a minimum effect requires more observations than if the goal had been to reject an effect of zero. As the confidence interval needs to reject a value that is closer to the observed effect size (e.g., 0.1 instead of 0) it needs to be more narrow, which requires more observations.

当等效性检验被逆转时，研究人员设计了一项研究，以拒绝比关注的最小效应量更不极端的效应（见图9.1中的面板D），这被称为**最小效应检验**(Murphy & Myors，1999)。研究人员可能不仅对拒绝一个效应为0的假设（如零假设显著性检验）关注，而且对拒绝太小而没有意义的效应范围关注。在其他条件相同的情况下，一项旨在为最小效应提供高检验力的研究比目标是拒绝效应为0的假设时需要更多的观测。由于置信区间需要拒绝更接近观测到的效应量的值（例如，0.1而不是0），因此它需要更加收缩，这需要更多的观测。

One benefit of a minimum effect test compared to a null hypothesis test is that there is no distinction between statistical significance and practical significance. As the test value is chosen to represent the minimum effect of interest, whenever it is rejected, the effect is both statistically and practically significant (Murphy et al., 2014). Another benefit of minimum effect tests is that, especially in correlational studies in the social sciences, variables are often connected through causal structures that result in real but theoretically uninteresting nonzero correlations between variables, which has been labeled the ‘crud factor’ (Meehl, 1990a; Orben & Lakens, 2020). Because an effect of zero is unlikely to be true in large correlational datasets, rejecting a nil null hypothesis is not a severe test. Even if the hypothesis is incorrect, it is likely that an effect of 0 will be rejected due to ‘crud’. For this reason, some researchers have suggested to test against a minimum effect of *r* = 0.1, as correlations below this threshold are quite common due to theoretically irrelevant correlations between variables (Ferguson & Heene, 2021).

与零假设检验相比，最小效应检验的一个好处是在统计显著性和实际显著性之间没有区别。由于检验值被选择来表示关注的最小效应，无论何时被拒绝，这种影响在统计上和实际上都是显著的（Murphy et al.，2014）。最小效应检验的另一个好处是，特别是在社会科学的相关性研究中，变量往往通过因果结构联系在一起，导致变量之间真实但理论上不关注的非零相关性，这被称为“粗糙因素”(Meehl, 1990a; Orben & Lakens, 2020)。由于0效应在大型相关数据集中不太可能成立，因此拒绝0零假设并不是一个严格的检验。即使假设不正确，0效应的假设也可能因“粗糙”而被拒绝。出于这个原因，一些研究人员建议针对*r* = 0.1的最小效应进行检验，因为由于变量之间理论上不相关的相关性，低于该阈值的相关性非常常见(Ferguson & Heene, 2021)。

Figure 9.1 illustrates two-sided tests, but it is often more intuitive and logical to perform one-sided tests. In that case, a minimum effect test would, for example, aim to reject effects smaller than 0.1, and an equivalence test would aim to reject effects larger than for example 0.1. Instead of specifying an upper and lower bound of a range, it is suﬀicient to specify a single value for one-sided tests. A final variation of a one-sided non-nil null hypothesis test is known as a test for **non-inferiority**, which examines if an effect is larger than the lower bound of an equivalence range. Such a test is for example performed when a novel intervention should not be noticeably worse than an existing intervention, but it can be a tiny bit worse. For example, if a difference between a novel and existing intervention is not smaller than -0.1, and effects smaller than -0.1 can be rejected, one can conclude an effect is non-inferior (Mazzolari et al., 2022; Schumi & Wittes, 2011). We see that extending nil null hypothesis tests to non-nil null hypotheses allow researchers to ask questions that might be more interesting.

图9.1说明了双侧检验，但做单侧检验通常更直观、更合乎逻辑。在这种情况下，例如，最小效应检验的目标是拒绝小于0.1的效应，而等效性检验的目标是拒绝大于例如0.1的效应。与其指定范围的上限和下限，不如为单侧检验指定一个值。单侧非0零假设检验的最后一种变体被称为**非劣效性**检验，它检查效应是否大于等效范围的下限。例如，当一种新的干预措施不应该明显比现有的干预措施差，但可能会差一点点时，就会进行这样的测试。例如，如果新的干预措施和现有的干预措施之间的差异不小于-0.1，并且小于-0.1的效应可以被拒绝，则可以得出结论，效果是非劣效的(Mazzolari et al., 2022; Schumi & Wittes, 2011)。我们发现，将0零假设检验扩展到非0零假设可以让研究人员提出可能更有趣的问题。



**A: Two−sided NHST**

**A:** **Two−sided** **NHST**

H0

H0

H1

H1

H1

H1

−4

−4

−3

−3

−2

−2

−1

−1

0

0

1

1

2

2

3

3

4

4

observed difference

observed difference

**B: Interval Hypothesis Test**

**B：** **区间假设检验**

H0

H0

H1

H1型

H0

H0

−4

−4

−3

−3

−2

−2

−1

−1

0

0

1

1.

2

2.

3

3.

4

4.

observed difference

观测到的差异

**C: Equivalence Test**

**C：** **等效性测试**

H0

H0

H1

H1型

H0

H0

−4

−4

−3

−3

−2

−2

−1

−1

0

0

1

1.

2

2.

3

3.

4

4.

observed difference

观测到的差异

**D: Minimum Effect Test**

**D：** **最小影响测试**

H1

H1型

H0

H0

H1

H1型

−4

−4

−3

−3

−2

−2

−1

−1

0

0

1

1.

2

2.

3

3.

4

4.

observed difference

观测到的差异

Figure 9.1: Two-sided null hypothesis test (A), interval hypothesis test (B), equivalence test (C) and minimum effect test (D).

**9.1 Equivalence tests**

# **等效性检验**

Equivalence tests were first developed in pharmaceutical sciences (Hauck & Anderson, 1984; Westlake, 1972) and later formalized as the **two one-sided tests (TOST)** approach to equivalence testing (Schuirmann, 1987; Seaman & Serlin, 1998; Wellek, 2010). The TOST procedure entails performing two one-sided tests to examine whether the observed data is surprisingly larger than a lower equivalence boundary (∆𝐿), or surprisingly smaller than an upper equivalence boundary (∆*U*):

等效性检验最早是在药物科学中发展起来的（Hauck&Anderson，1984；Westlake，1972），后来正式成为等效性检验的**两个单侧检验(TOST)**方法(Schuirmann，1987；Seaman & Serlin，1998；Wellek，2010）。TOST程序需要进行两次单侧检验，以检验观察到的数据是否出乎意料地大于等效下限(∆𝐿), 或者出乎意料地小于等效上限(∆*U*)：

and

和

where *M* indicates the means of each sample, *n* is the sample size, and is the pooled standard deviation:

其中*M*表示每个样本的平均值，*n*是样本量，σ是合并的标准偏差：

If both one-sided tests are significant, we can reject the presence of effects large enough to be meaningful. The formulas are highly similar to the normal formula for the *t*-statistic. The difference between a NHST *t*-test and the TOST procedure is that the lower equivalence boundary and the upper equivalence boundary are subtracted from the mean difference between groups (in a normal *t*-test, we compare the mean difference against 0, and thus the delta drops out of the formula because it is 0).

如果这两个单侧检验都是显著的，我们可以拒绝足够大而有意义的效应的存在。这些公式与*t*统计量的正态公式高度相似。NHST *t*检验和TOST程序之间的区别在于，从组别之间的平均差中减去等效下限和等效上限（在正常的*t*检验*中*，我们将平均差与0进行比较，因此∆从公式中删除，因为它是0）。

To perform an equivalence test, you don’t need to learn any new statistical tests, as it is just the well-known *t*-test against a different value than 0. It is somewhat surprising that the use of *t-*tests to perform equivalence tests is not taught alongside their use in null hypothesis significance tests, as there is some indication that this could prevent common misunderstandings of *p*-values (Parkhurst, 2001). Let’s look at an example of an equivalence test using the TOST procedure.

要进行等效性检验，你不需要学习任何新的统计检验，因为它只是针对不同于0的值进行的众所周知的*t*检验。令人有些惊讶的是，使用*t*检验进行等效性检验并没有与在零假设显著性检验中使用*t*检验一起进行教学，因为有一些迹象表明，这可以防止对*p*值的常见误解(Parkhurst，2001)。让我们来看一个使用TOST程序进行等效性检验的例子。

In a study where researchers are manipulating fatigue by asking participants to carry heavy boxes around, the researchers want to ensure the manipulation does not inadvertently alter participants’ moods. The researchers assess positive and negative emotions in both conditions, and want to claim there are no differences in positive mood. Let’s assume that positive mood in the experimental fatigue condition (𝑚1 = 4.55, *sd*1 = 1.05, *n*1= 15) did not differ from the mood in the control condition (𝑚2 = 4.87, 𝑠𝑑2 = 1.11, 𝑛2 = 15). The researchers conclude: “Mood did not differ between conditions, *t* = -0.81, *p* = .42”. Of course, mood did differ between conditions, as 4.55 - 4.87 = -0.32. The claim is that there was no *meaningful* difference in mood, but to make such a claim in a correct manner, we first need to specify which difference in mood is large enough to be meaningful. For now, let’s assume the researcher consider any effect less extreme half a scale point too small to be meaningful. We now test if the observed mean difference of -0.32 is small enough such that we can reject the presence of effects that are large enough to matter.

在一项研究中，研究人员通过让被试随身携带沉重的盒子来操纵疲劳，研究人员希望确保这种操作不会无意中改变被试的情绪。研究人员评估了这两种情况下的积极情绪和消极情绪，并声称在积极情绪上没有差异。让我们假设实验性疲劳条件下的积极情绪(𝑚1 = 4.55, *sd*1 = 1.05, *n*1= 15)与控制条件下的情绪(𝑚2 = 4.87, 𝑠𝑑2 = 1.11, 𝑛2 = 15)没有差异。研究人员得出结论：“不同条件下的情绪没有差异，*t*=-0.81，*p*=.42”。当然，不同条件下的情绪确实不同，因为4.55-4.87=-0.32。这种说法是指在情绪上无*有意义的*差异，但要以正确的方式得出这样的说法，我们首先需要指定哪种情绪差异足够大，才能视为是有意义的。目前，让我们假设研究人员认为任何不那么极端的效应——半个标度点太小而没有意义。我们现在检验观察到的-0.32的平均差异是否足够小，以便我们可以拒绝大到要去重视的效应的存在。

The TOSTER package (originally created by myself but recently redesigned by [Aaron Caldwell](https://aaroncaldwell.us/)) can be used to plot two *t*-distributions and their critical regions indicating when we can reject the presence of effects smaller than -0.5 and larger than 0.5. It can take some time to get used to the idea that we are rejecting values more extreme than the equivalence bounds. Try to consistently ask in any hypothesis test: Which values can the test reject? In a nil null hypothesis test, we can reject an effect of 0, and in the equivalence test in the Figure below, we can reject values lower than -0.5 and higher than 0.5. In Figure [9.2](#_bookmark1) we see two *t*-distributions centered on the upper and lower bound of the specified equivalence range (-0.5 and 0.5).

TOSTER软件包（最初由我创建，但最近由[Aaron Caldwell­](https://aaroncaldwell.us/)重新设计）可用于绘制两个*t*分布及其临界区域的图表，指示我们何时可以拒绝小于-0.5和大于0.5的效应。我们可能需要一些时间来习惯这样一种想法，即我们拒绝的值比等效边界更极端。在任何假设检验中，试着始终提问：检验可以拒绝哪些值？在0零假设检验中，我们可以拒绝效应为0的假设，在下图的等效性检验中，可以拒绝低于-0.5和高于0.5的值。在图[9.2](#_bookmark1)中，我们看到两个*t*分布集中在指定等效范围的上限和下限（-0.5和0.5）。

SMD cannot be plotted if type = "tnull"

Below the two curves we see a line that represents the confidence interval ranging from -0.99 to 0.35, and a dot on the line that indicates the observed mean difference of -0.32. Let’s first look at the left curve. We see the green highlighted area in the tails that highlights which observed mean differences would be extreme enough to statistically reject an effect of -0.5. Our observed mean difference of -0.32 lies very close to -0.5, and if we look at the left distribution, the mean is not far enough away from -0.5 to fall in the green area that indicates when observed differences would be statistically significant. We can also perform the equivalence test using the TOSTER package, and look at the results.

在这两条曲线下面，我们看到一条线表示-0.99至0.35的置信区间，线上的一个点表示观察到的-0.32的平均差异。让我们先看看左边的曲线。我们在尾部看到绿色突出显示区域，突出显示观察到的平均差异将非常极端，足以在统计上拒绝-0.5的效应。我们观察到的-0.32的平均差异非常接近-0.5，如果我们看左边的分布，平均值离-0.5不远，不足以落在表明观察到的差异何时具有统计学意义的绿色区域。我们还可以使用TOSTER软件包进行等效性检验，并查看结果。

TOSTER:：tsum\_TOST（m1=4.55，

m2=4.87，

sd1=1.05，

sd2=1.11，

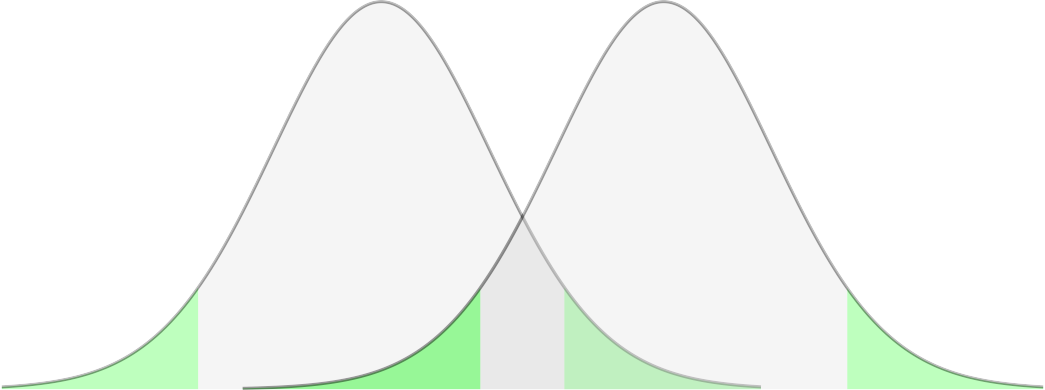
n1=15，

n2=15，

low\_eqbound=-0.5，

high\_eqbound=0.5）

**−0.5 0.5**



**Mean Difference**

**Mean** **Difference**

**−2 −1 0 1 2**

Note: green indicates rejection region for null equivalence and MET hypotheses

注：绿色表示零等效和MET假设的拒绝区域

Figure 9.2: The mean difference and its confidence interval plotted below the *t*-distributions used to perform the two-one-sided tests against -0.5 and 0.5.

图9.2：绘制在*t*分布下方的平均差及其置信区间，用于对-0.5和0.5进行两次单侧检验。

Welch Modified Two-Sample t-Test

The equivalence test was non-significant, t(27.91) = 0.456, p = 3.26e-01

The null hypothesis test was non-significant, t(27.91) = -0.811, p = 4.24e-01

NHST: don't reject null significance hypothesis that the effect is equal to zero

TOST: don't reject null equivalence hypothesis

TOST Results

t df p.value

t-test -0.8111 27.91 0.424

TOST Lower 0.4563 27.91 0.326

TOST Upper -2.0785 27.91 0.023

Effect Sizes

Estimate SE C.I. Conf. Level

Raw -0.3200 0.3945 [-0.9912, 0.3512] 0.9

Hedges's g(av) -0.2881 0.3930 [-0.8733, 0.3021] 0.9

Note: SMD confidence intervals are an approximation. See vignette("SMD\_calcs").

In the line ‘t-test’ the output shows the traditional nil null hypothesis significance test (which we already knew was not statistically significant: *t* = 0.46, *p* = 0.42. Just like the default *t*-test in R, the tsum\_TOST function will by default calculate Welch’s *t*-test (instead of Student’s *t*-test), which is a better default (Delacre et al., 2017), but you can request Student’s *t*-test by adding var.equal = TRUE as an argument to the function.

在“t检验”一行中，输出结果显示了传统的0零假设显著性检验（我们已经知道这在统计学上并不显著：*t=*0.46，*p*=0.42。就像R中的默认*t*检验一样，tsum\_TOST函数在默认情况下会计算Welch’s *t*检验（而不是Student’s *t*检验）*，*这是一个更好的默认值(Delacre et al., 2017)，但你可以通过添加var.equal = TRUE作为函数的参数来请求Student’s *t检*验。

We also see a test indicated by TOST Lower. This is the first one-sided test examining if we can reject effects lower than -0.5. From the test result, we see this is not the case: *t* = 0.46, *p* = 0.33. This is an ordinary *t*-test, just against an effect of -0.5. Because we cannot reject differences more extreme than -0.5, it is possible that a difference we consider meaningful (e.g., a difference of -0.60) is present. When we look at the one-sided test against the upper bound of the equivalence range (0.5) we see that we can statistically reject the presence of mood effects larger than 0.5, as in the line TOST Upper we see *t* = -2.08, *p* = 0.02. Our final conclusion is therefore that, even though we can reject effects more extreme than 0.5 based on the observed mean difference of -0.32, we cannot reject effects more extreme than -0.5. Therefore, we cannot completely reject the presence of meaningful mood effects. As the data does not allow us to claim the effect is different from 0, nor that the effect is, if anything, too small to matter (based on an equivalence range from -0.5 to 0.5), the data are **inconclusive**. We cannot distinguish between a Type 2 error (there is an effect, but in this study we just did not detect it) or a true negative (there really is no effect large enough to matter).

我们还看到TOST Lower指示的检验。这是第一次单侧检验，检验我们是否可以拒绝低于-0.5的效应。从检验结果来看，情况并非如此：*t*=0.46，*p*=0.33。这是一个普通的*t*检验*，*只是针对-0.5的效应*。*因为我们不能拒绝比-0.5更极端的差异，所以可能存在我们认为有意义的差异（例如，-0.60的差异）。当我们观察等价范围上限（0.5）的单侧检验时，我们可以从统计学上拒绝大于0.5的情绪效应的存在，正如在TOST upper行中我们看到的*t*=-2.08，*p*=0.02。因此，我们的最终结论是，即使我们可以根据观察到的-0.32的平均差异来拒绝比0.5更极端的效应，我们也不能拒绝比-0.5更极端的效应。因此，我们不能完全拒绝有意义的情绪效应的存在。由于数据不允许我们声称效应与0有所不同，也不允许我们说效应太小而无关紧要（基于-0.5到0.5的等效范围），因此数据是**不确定的**。我们无法区分Ⅱ类错误（存在效应，但在这项研究中，我们只是没有检测到它）或真正的阴性（确实没有足够大到要去重视的效应）。

Note that because we fail to reject the one-sided test against the lower equivalence bound, the possibility remains that there is a true effect size that is large enough to be considered meaningful. This statement is true, even when the effect size we have observed (-0.32) is closer to zero than to the equivalence bound of -0.5. One might think the observed effect size needs to be more extreme (i.e., < -0.5 or > 0.5) than the equivalence bound to maintain the possibility that there is an effect that is large enough to be considered meaningful. But that is not required. The 90% CI indicates that some values below -0.5 cannot be rejected. As we can expect that 90% of confidence intervals in the long run capture the true population parameter, it is perfectly possible that the true effect size is more extreme than -0.5. And, the effect might even be more extreme than the values captured by this confidence interval, as 10% of the time, the computed confidence interval is expected to not contain the true effect size. Therefore, when we fail to reject the smallest effect size of interest, we retain the possibility that an effect of interest exists. If we can reject the nil null hypothesis, but fail to reject values more extreme than the equivalence bounds, then we can claim there is an effect, and it might be large enough to be meaningful.

请注意，由于我们未能拒绝针对等效下限的单侧检验，因此仍有可能存在足够大以至于被认为是有意义的真实效应量。这种说法是正确的，即使我们观察到的效应大小（-0.32）比-0.5的等效边界更接近于零。有人可能认为，观察到的效应大小需要比等效边界更极端（即<-0.5或>0.5），以保持存在足够大的效应以至于被认为是有意义的可能性。但这并不是必须的。90%的置信区间不能拒绝低于-0.5的某些值。正如我们可以预期的那样，从长远来看，90%的置信区间捕捉到了真实的总体参数，真实的效应大小完全有可能比-0.5更极端。而且，这种效应甚至可能比这个置信区间捕获的值更极端，因为在10%的时间里，计算的置信区间预计不包含真实的效应量。因此，当我们不能拒绝关注的最小效应量时，我们保留了存在关注效应的可能性。如果我们可以拒绝0零假设，但不能拒绝比等效边界更极端的值，那么我们可以声称效应存在，并且它可能足够大，大到有意义。

One way to reduce the probability of an inconclusive effect is to collect suﬀicient data. Let’s imagine the researchers had not collected 15 participants in each condition, but 200 participants. They otherwise observe exactly the same data. As explained in the chapter on [confidence intervals](#_bookmark0), as the sample size increases, the confidence interval becomes more narrow. For a TOST equivalence test to be able to reject both the upper and lower bound of the equivalence range, the confidence interval needs to fall completely within the equivalence range. In Figure [9.3](#_bookmark2) we see the same result as in Figure [9.2](#_bookmark1), but now if we had collected 200 observations. Because of the larger sample size, the confidence is more narrow than when we collected 15 participants. We see that the 90% confidence interval around the observed mean difference now excludes both the upper and lower equivalence bound. This means that we can now reject effects outside of the equivalence range (even though barely, with a *p* = 0.048 as the one-sided test against the lower equivalence bound is only just statistically significant).

降低不确定效应概率的一种方法是收集有效的数据。让我们想象一下，研究人员并没有在每种情况下收集15名被试，而是收集了200名被试­。除此之外，他们观察到的数据完全相同。正如[置信区间一](#_bookmark0)章中所解释的，随着样本量的增加，置信区间变得越来越窄­。为了使TOST等效性检验能够拒绝等效范围的上限和下限，置信区间需要完全落在等效范围内。在图[9.3](#_bookmark2)中，我们看到了与图[9.2](#_bookmark1)相同的结果，但现在如果我们收集了200个观测结果。由于样本量较大，置信度比我们收集15名被试时更窄。我们看到，观察到的平均差周围的90%置信区间现在排除了等效上限和等效下限。这意味着我们现在可以拒绝等效范围之外的效应（尽管几乎没有，因为对等效下限的单侧检验仅具有统计学意义，*p*=0.048）。

Welch Modified Two-Sample t-Test

The equivalence test was significant, t(396.78) = 1.666, p = 4.82e-02

The null hypothesis test was significant, t(396.78) = -2.962, p = 3.24e-03

NHST: reject null significance hypothesis that the effect is equal to zero

TOST: reject null equivalence hypothesis

TOST Results

t df p.value

t-test -2.962 396.8 0.003

TOST Lower 1.666 396.8 0.048

TOST Upper -7.590 396.8 < 0.001

Effect Sizes

Estimate SE C.I. Conf. Level

Raw -0.3200 0.108 [-0.4981, 0.1419] 0.9

Hedges's g(av) -0.2956 0.104 [-0.4605, 0.1304] 0.9

Note: SMD confidence intervals are an approximation. See vignette("SMD\_calcs").

**−0.5 0.5**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mean Difference** | | | |
|  |  |  |  |
|  |

**−0.5 0.0 0.5**

Note: green indicates rejection region for null equivalence and MET hypotheses

注：绿色表示零等价和MET假设的拒绝区域

Figure 9.3: The mean difference and its confidence interval for an equivalence test with an equivalence range of -0.5 and 0.5.

图9.3：等效范围为-0.5和0.5的等效性检验的平均差及其置信区间

In Figure 9.4 we see the the same results, but now visualized as a confidence density plot (Schweder & Hjort, 2016), which is a graphical summary of the distribution of confidence. A confidence density plot allows you to see which effects can be rejected with difference confidence interval widths. We see the bounds of the green area (corresponding to a 90% confidence interval) fall inside the equivalence bounds. Thus, the equivalence test is statistically significant, and we can statistically reject the presence of effects outside the equivalence range. We can also see that the 95% confidence interval excludes 0, and therefore, a traditional null hypothesis significance test is also statistically significant.

在图9.4中，我们看到了相同的结果，但现在可视化为置信密度图(Schweder & Hjort, 2016)，这是置信度分布的图形总结。置信密度图允许你查看哪些效应可以通过不同的­置信区间宽度来拒绝。我们看到绿色区域的边界（对应于90%的置信区间）落在等效边界内。因此，等效性检验在统计学上是­显著的，我们可以在统计学上拒绝存在等效范围之外的效应。我们还可以看到，95%的置信区间排除了0，因此，传统的零假设显著性检验也具有统计学意义。

In other words, both the null hypothesis test and the equivalence test have yielded significant results. This means we can claim that the observed effect is statistically different from zero, and that the effect is statistically smaller than effects we deemed large enough to matter when we specified the equivalence range from -0.5 to 0.5. This illustrates how combining equivalence tests and nil null hypothesis tests can prevent us from mistaking statistically significant effects for practically significant effects. In this case, with 200 participants, we can reject an effect of 0, but the effect, if any, is not large enough to be meaningful.

换句话说，零假设检验和等效性检验都产生了显著的结果。这意味着我们可以声称，观察到的效应在统计上与0不同，并且在统计上，当我们指定-0.5到0.5的等效范围时，该效应小于我们认为足够大的效应。这说明了将等效性检验和0零假设检验相结合可以防止我们误将具有统计学意义的效应当成实际上显著的效应。在这种情况下，有200名被试，我们可以拒绝一个为0的效应，但这个效应（如果有的话）没有大到是有意义的。

**Confidence Interval**

**0.68**

**0.9**

**0.95**

**0.99**

**−0.5 0.5**



**Mean Difference**

**Mean** **Difference**

**−0.50 −0.25 0.00 0.25 0.50**

Figure 9.4: The mean difference and its confidence interval for an equivalence test with an equivalence range of -0.5 and 0.5.

图9.4：等效范围为-0.5和0.5的等效性检验的平均差及其置信区间

**9.2 Reporting Equivalence Tests**

# 报告等效性检验

It is common practice to only report the test yielding the higher *p*-value of the two one-sided tests when reporting an equivalence test. Because both one-sided tests need to be statistically significant to reject the null hypothesis in an equivalence test (i.e., the presence of effects large enough to matter), when the larger of the two hypothesis tests rejects the equivalence bound, so does the other test. Unlike in null hypothesis significance tests it is not common to report standardized effect sizes for equivalence tests, but there can be situations where researchers might want to discuss how far the effect is removed from the equivalence bounds on the raw scale. Prevent the erroneous interpretation to claim there is ‘no effect’, that an effect is ‘absent’, that the true effect size is ‘zero’, or vague verbal descriptions, such as that two groups yielded ‘similar’ or ‘comparable’ data. A significant equivalence test rejects effects more extreme that the equivalence bounds. Smaller true effects have not been rejected, and thus it remains possible that there is a true effect. Because a TOST procedure is a frequentist test based on a *p*-value, all other misconceptions of *p*-values should be prevented as well.

在报告等效性检验时，通常只报告两个单侧检验中产生较高*p*值的检验。因为两个单侧检验都需要具有统计学意义，才能在等效性检验中拒绝零假设（即存在足够大的效应），所以当两个假设检验中较大的一个拒绝等效边界时，另一个检验也是如此。与零假设显著性检验不同，报告等效性检验的标准化效应量并不常见，但在某些情况下，研究人员可能想讨论在原始尺度上，效应与等效边界的差距有多大。防止错误的解释，声称‘没有效应’、效应‘不存在’、真实效应量为‘0’，或模糊的口头描述，例如两组得出的数据“相似”或“可比”。显著的等效性检验拒绝比等效边界更极端的效应。较小的真实效应没有被拒绝，因此仍然有可能存在真实效应。因为TOST程序是一种基于*p*值的频率检验，所以也应该防止所有其他对*p*值的误解。

When summarizing the main result of an equivalence test, for example in an abstract, always report the equivalence range that the data is tested against. Reading ‘based on an equivalence test we concluded the absence of a meaningful effect’ means something very different if the equivalence bounds were *d* =-0.9 to 0.9 than when the bounds were *d* =-0.2 to *d* =0.2. So instead, write ‘based on an equivalence test with an equivalence range of *d* =-0.2 to 0.2, we conclude the absence of an effect we deemed meaningful’. Of course, whether peers agree you have correctly concluded the absence of a meaningful effect depends on whether they agree with your justification for a smallest effect of interest! A more neutral conclusion would be a statement such as: ‘based on an equivalence test, we rejected the presence of effects more extreme than -0.2 to 0.2, so we can act (with an error rate of alpha) as if the effect, if any, is less extreme than our equivalence range’. Here, you do not use value-laden terms such as ‘meaningful’. If both a null hypothesis test and an equivalence test are non-significant, the finding is best described as ‘inconclusive’: There is not enough data to reject the null, or the smallest effect size of interest. If both the null hypothesis test and the equivalence test are statistically significant, you can claim there is an effect, but at the same time claim the effect is too small to be of interest (given your justification for the equivalence range).

在总结等效性检验的主要结果时，例如在摘要中，始终报告数据所检验的等效范围。与边界为*d* = -0.2至*d* = 0.2时相比，如果等效边界为*d* = -0.9至0.9，阅读‘基于等效性检验，我们得出结论，有意义的效应不存在’意味着某些方面非常不同。反之，写下‘基于等效范围为*d*=-0.2至0.2的等效性检验，我们得出结论，我们认为有意义的效应不存在’。当然，同行们是否同意你正确地得出了有意义效应不存在的结论，取决于他们是否同意你对关注的最小效应的证明！一个更中性的结论是这样一种说法：“基于等效性检验，我们拒绝了比-0.2到0.2更极端效应的存在，所以我们可以采取行动（错误率为α），就好像这种效应（如果有的话）没有我们的等效范围那么极端一样”。在这里，你不使用诸如‘有意义’之类的充满价值的术语。如果零假设检验和等效性检验都是不显著的，那么这一发现最好被描述为‘不确定的’：没有足够的数据来拒绝零假设，或者关注的最小效应量。如果零假设检验和等效性检验都具有统计学意义，你可以声称效应存在，但同时声称效应太小，不值得关注（考虑到你对等效范围的证明）。

Equivalence bounds can be specified in raw effect sizes, or in standardized mean differences. It is better to specify the equivalence bounds in terms of raw effect sizes. Setting them in terms of Cohen’s *d* leads to bias in the statistical test, as the observed standard deviation has to be used to translate the specified Cohen’s *d* into a raw effect size for the equivalence test (and when you set equivalence bounds in standardized mean differences, TOSTER will warn: “Warning: setting bound type to SMD produces biased results!”). The bias is in practice not too problematic in any single equivalence test, and being able to specify the equivalence bounds in standardized mean differences lowers the threshold to perform an equivalence test when they do not know the standard deviation of their measure. But as equivalence testing becomes more popular, and fields establish smallest effect sizes of interest, they should do so in raw effect size differences, not in standardized effect size differences.

等效边界可以在原始效果量中指定，也可以在标准化平均差中指定。最好根据原始效果量来指定等效边界。根据Cohen’s *d*设置它们会导致统计检验中的偏差，因为必须使用观察到的标准差将指定的Cohen‘s *d*转换为等效性检验的原始效应量（当你在标准化平均差中设置等效边界时，TOSTER将警告：“警告：将边界类型设置为SMD会产生偏差结果！”）。在实践中，偏差在任何单一的等效性检验中都不会有太大的问题，并且能够在标准化平均差中指定等效边界，这降低了当他们不知道其度量的标准差时进行等效性检验的阈值。但是，随着等效性检验变得越来越流行，并且领域建立了关注的最小效应量，他们应该在原始效应量差异中这样做，而不是在标准化效应量差异这样做。

**9.3 Minimum Effect Tests**

# 最小效应检验

If a researcher has specified a smallest effect size of interest, and is interested in testing whether the effect in the population is larger than this smallest effect of interest, a minimum effect test can be performed. As with any hypothesis test, we can reject the smallest effect of interest whenever the confidence interval around the observed effect does not overlap with it. In the case of a minimum effect test, however, the confidence interval should be fall completely beyond the smallest effect size of interest. For example, let’s assume a researcher performs a minimum effect test with 200 observations per condition against a smallest effect size of interest of a mean difference of 0.5.

如果研究人员指定了关注的最小效应量，并且有兴趣检验群体中的效应是否大于关注的该最小效应，则可以进行最小效应检验。与任何假设检验一样，只要观察到的效应周围的置信区间与其不重叠，我们就可以拒绝关注的最小效应。然而，在最小效应检验的情况下，置信区间应该完全超过关注的最小效应量。例如，让我们假设一名研究人员对平均差异为0.5的最小效应量进行最小效应检验，每个条件有200个观察结果。

SMD cannot be plotted if type = "tnull"

Welch Modified Two-Sample t-Test

The minimal effect test was significant, t(396.78) = 12.588, p = 4.71e-04

The null hypothesis test was significant, t(396.78) = 7.960, p = 1.83e-14

NHST: reject null significance hypothesis that the effect is equal to zero

TOST: reject null MET hypothesis

TOST Results

t df p.value

t-test 7.960 396.8 < 0.001

TOST Lower 12.588 396.8 1

TOST Upper 3.332 396.8 < 0.001

Effect Sizes

Estimate SE C.I. Conf. Level

Raw 0.8600 0.108 [0.6819, 0.0381] 0.9

Hedges's g(av) 0.7945 0.125 [0.6234, 0.9646] 0.9

Note: SMD confidence intervals are an approximation. See vignette("SMD\_calcs").

Below the two curves we again see a line that represents the confidence interval ranging from 0.68 to 1.04, and a dot on the line that indicates the observed mean difference of 0.86. The entire confidence interval lies well above the minimum effect of 0.5, and we can therefore not just reject the nil null hypothesis, but also effects smaller than the minimum effect of interest. Therefore, we can claim that the effect is large enough to be not just statistically significant, but also practically significant (as long as we have justified our smallest effect size of interest well). Because we have performed a two-sided minimum effect test, the minimum effect test would also have been significant if the confidence interval had been completely on the opposite side of -0.5.

在这两条曲线下面，我们再次看到一条线，它代表的置信区间从0.68至1.04，以及表示观察到的0.86的平均差的线上的点。整个置信区间远高于0.5的最小效应，因此我们不仅可以拒绝0零假设，而且可以拒绝小于关注的最小效应的效应。因此，我们可以声称这种效应足够大，不仅在统计上具有显著性，而且在实践中也具有显著性（只要我们很好地证明了我们关注的最小效应量）。因为我们进行了双侧最小效应检验，如果置信区间完全在-0.5的相反侧，最小效应检验也会很显著。

**−0.5 0.5**

|  |  |  |
| --- | --- | --- |
| **Mean Difference** | | |
|  |  |  |

**−0.5 0.0 0.5 1.0**

Note: green indicates rejection region for null equivalence and MET hypotheses

注：绿色表示0等效和MET假设的拒绝区域

Figure 9.5: The mean difference and its confidence interval plotted below the *t*-distributions used to perform the two-one-sided tests against -0.5 and 0.5 when performing a minimum effect test.

图9.5：在进行最小效应检验时，用于对-0.5和0.5进行两次单侧检验的*t*分布下方

绘制的平均差及其置信区间

Earlier we discussed how combining traditional NHST and an equivalence test could lead to more informative results. It is also possible to combine a minimum effect test and an equivalence test. One might even say that such a combination is the most informative test of a prediction whenever a smallest effect size of interest can be specified. In principle, this is true. As long as we are able to collect enough data, we will always get an informative and straightforward answer when we combine a minimum effect test with an equivalence test: Either we can reject all effects that are too small to be of interest, or we can reject all effects that are large enough to be of interest. As we will see below in the section on power analysis for interval hypotheses, whenever the true effect size is close to the smallest effect size of interest, a large amount of observations will need to be collected. And if the true effect size happens to be identical to the smallest effect size of interest, neither the minimum effect test nor the equivalence test can be correctly rejected (and any significant test would be a Type 1 error). If a researcher can collect suﬀicient data (so that the test has high statistical power), and is relatively confident that the true effect size will be larger or smaller than the smallest effect of interest, then the combination of a minimum effect test and an equivalence test can be attractive as such a hypothesis test is likely to yield an informative answer to the research question.

早些时候，我们讨论了如何将传统的NHST和等效性检验相结合，从而获得信息更加丰富的结果。也可以将最小效应检验和等效性检验相结合。甚至可以说，无论何时可以指定关注的最小效应大小，这种组合都是对预测的信息更加丰富的检验。原则上，这是真的。只要我们能够收集到足够的数据，当我们将最小效应检验与等效性检验相结合时，我们总是会得到一个信息丰富、直截了当的答案：要么我们可以拒绝所有太小而不关注的效应，要么我们可以拒绝所有足够大而关注的效应。正如我们将在下面关于区间假设的统计检验力分析一节中看到的那样，每当真实效应量接近关注的最小效应量时，都需要收集大量的观测结果。如果真实效应量恰好与关注的最小效应量相同，则最小效应检验和等效性检验都不能被正确拒绝（任何显著的检验都将是Ⅰ型错误）。如果研究人员能够收集有效的数据（从而使检验具有很高的统计检验力），并且相对确信真实效应量将大于或小于关注的最小效应，那么最小效应检验和等效性检验的组合可能很有吸引力，因为这样的假设检验可能会为研究问题提供信息丰富的答案。