Statistische Inferenz mittels Schätzung: Empfehlungen der International Society of Physiotherapy Journal Editors

Autoren

Mark R. Elkins[1, 2], Rafael Zambelli Pinto[1, 3], Arianne Verhagen[1, 2], Monika Grygorowicz[4], Anne Söderlund[5], Matthieu Guemann[6], Antonia Gómez-Conesa[7], Sarah Blanton[8], Jean-Michel Brismée[9], Shabnam Agarwal[10], Alan Jette[11], Sven Karstens[12], Michele Harms[13], Geert Verheyden[14], Umer Sheikh[15]

Journals

1 Vorstand International Society of Physiotherapy Journal Editors
2 Journal of Physiotherapy
3 Brazilian Journal of Physical Therapy/Revista Brasileira de Fisioterapia
4 BMC Sports Science, Medicine and Rehabilitation
5 European Journal of Physiotherapy
6 European Rehabilitation Journal
7 Fisioterapia
8 Journal of Humanities in Rehabilitation
9 Journal of Manual & Manipulative Therapy
10 Journal of Society of Indian Physiotherapists
11 Physical Therapy
12 physioscience
13 Physiotherapy
14 Physiotherapy Research International
15 The Journal of Physiotherapy & Sports Medicine

Bibliografie

physioscience 2022; 18: 52–57

DOI 10.1055/a-1741-9919

ISSN 1860-3092

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Zitierweise für diesen Artikel

Elkins et al. Statistical inference through estimation: recommendations from the International Society of Physiotherapy Journal Editors. Journal of Physiotherapy; 2021; Volume 68, Issue 1, Pages 1–4

Korrespondenzadresse

Mark Elkins
Centre for Education & Workforce Development
Sydney Local Health District, Sydney, Australia
mark.elkins@sydney.edu.au

In der Gesundheitsforschung, einschließlich des Bereichs der Physiotherapie, werden häufig statistische Nullhypothesentests angewendet [1, 2]. Trotz ihres weit verbreiteten Einsatzes unterliegen statistische Nullhypothesentests jedoch bedeutenden Einschränkungen. Dieses gemeinschaftlich herausgegebene Editorial erklärt Inferenzstatistik unter Verwendung von statistischen Nullhypothesentests und die mit diesem Ansatz verbundenen Probleme. Es untersucht außerdem einen alternativen Ansatz für statistische Inferenz (der als Schätzen bezeichnet wird) und ermutigt Leser*innen physiotherapeutischer Forschung, sich mit Schätzmethoden und der Interpretation ihrer Ergebnisse vertraut zu machen. Darüber hinaus macht das Editorial Forschende darauf aufmerksam, dass einige Mitglieder der International Society of Physiotherapy Journal Editors (ISPJE) zukünftig Manuskripte erwarten, in denen Schätzmethoden anstelle statistischer Nullhypothesentests verwendet werden.[1]

1 Eine Stellungnahme der Herausgebenden der physioscience zu diesem Gasteditorial finden Sie im Editorial ab S. 49.

Was ist statistische Inferenz?

Der Begriff der statistischen Inferenz bezeichnet den Prozess, bei dem auf der Grundlage von Daten aus Stichproben Schlüsse auf die Grundgesamtheit gezogen werden [1]. Nehmen wir an, eine Gruppe von Forschenden möchte bei Personen mit Zustand nach Schlaganfall einen bestimmten Aspekt untersuchen (etwa den Effekt einer Intervention, die Prävalenz einer Komorbidität oder die Zweckmäßigkeit eines prognostischen Modells). Natürlich ist es den Forschenden in diesem Fall nicht möglich, sämtliche Überlebende nach einem Schlaganfall weltweit zu testen. Daher führen sie ihre Studie mit einer Stichprobe an Proband*innen aus der Grundgesamtheit der Schlaganfall-Überlebenden durch. In der Regel macht eine solche Stichprobe nur einen winzigen Teil der Grundgesamtheit aus. Aus diesem Grund weichen die Studienergebnisse auf Grundlage der Stichprobe wahrscheinlich von den Gegebenheiten in der Grundgesamtheit ab [3]. Forschende müssen daher eine statistische Analyse der Daten aus der Stichprobe

vornehmen, um Schlüsse auf die Gegebenheiten in der Grundgesamtheit zu ziehen.

Was sind statistische Nullhypothesentests?

Traditionell basiert die statistische Inferenz auf statistischen Nullhypothesentests. Bei solchen Tests wird eine sogenannte Nullhypothese aufgestellt, z. B. dass eine Intervention keinen Effekt auf ein Ergebnis hat, eine Exposition keinen Einfluss auf ein Risiko hat oder keine Beziehung zwischen 2 Variablen besteht. Außerdem wird bei solchen Tests ein p-Wert berechnet. Dieser quantifiziert die Wahrscheinlichkeit, dass bei vielfacher Wiederholung der Studie jedes Mal ein Effekt oder eine Beziehung im mindestens gleichen Ausmaß wie bei der Stichprobe in der Ursprungsstudie zu beobachten wäre, wenn die Nullhypothese zutrifft. Zu beachten ist, dass sich die Nullhypothese auf die Grundgesamtheit bezieht und nicht auf die Studienstichprobe.

Da sich die Überlegungen hinter solchen Tests auf eine imaginäre Wiederholung der Studie stützen, wird hier auch von einem „frequentistischen Ansatz“ gesprochen. Ein solcher Ansatz legt den Schwerpunkt darauf, wie stark das statistische Ergebnis – z. B. die mittlere Differenz, ein Anteil oder eine Korrelation – bei Wiederholungen der Studie variieren würde. Wenn die gewonnenen Daten aus der Studienstichprobe darauf hindeuten, dass das Ergebnis bei einer imaginären Wiederholung der Studie wahrscheinlich ähnlich wäre, wird dies als Hinweis darauf interpretiert, dass das Ergebnis in gewisser Hinsicht besonders glaubwürdig ist.

Ein Typus der statistischen Nullhypothesen-Testverfahren ist der von Fisher [4–6] entwickelte Signifikanztest. Ist es im Rahmen eines Signifikanztestes unwahrscheinlich, dass bei zutreffender Nullhypothese und imaginären Wiederholungen ein Effekt mit mindestens der gleichen Größe zu beobachten wäre wie in der Studie (angezeigt durch $p < 0{,}05$), dann wird dies Ergebnis als Beweis interpretiert, dass die Nullhypothese falsch ist. Ein weiterer Typus statistischer Nullhypothesentests ist der von Neyman and Pearson [4–6] entwickelte Hypothesentest. Dabei werden 2 Hypothesen aufgestellt: die Nullhypothese (z. B.: „In der Grundgesamtheit gibt es keinen Unterschied“) und die Alternativhypothese (z. B.: „In der Grundgesamtheit gibt es einen Unterschied“). Dabei zeigt der p-Wert den Forschenden an, welche Hypothese anzunehmen ist. Ist $p \geq 0{,}05$, wird die Nullhypothese beibehalten; ist $p < 0{,}05$, ist die Nullhypothese zu verwerfen und die Alternativhypothese anzunehmen.

Obwohl diese beiden Ansätze mathematisch ähnlich sind, unterscheiden sie sich dahingehend, wie sie interpretiert und berichtet werden sollten. Dennoch beachten zahlreiche Forschende deren Unterschiede nicht und analysieren ihre Daten mit einem unangemessenen Hybrid aus beiden Methoden.

Probleme von statistischen Nullhypothesentests

Unabhängig davon, ob Signifikanztests oder Hypothesentests (oder ein Hybrid aus beiden Verfahren) angewendet werden, sind statistische Nullhypothesentests mit zahlreichen Problemen verbunden [4, 5, 7]. 5 schwerwiegende Probleme werden in ▶ **Tab. 1** erklärt. Jedes einzelne dieser Probleme ist schwerwiegend genug, um statistische Nullhypothesentests als ungeeignet für eine Verwendung in der Forschung einzustufen. Das wird vermutlich zahlreiche Leser*innen überraschen, denn schließlich ist die Verwendung solcher Tests in Forschungspublikationen sehr weit verbreitet [1, 2].

Und es ist auch überraschend, dass sich die breitflächige Anwendung statistischer Nullhypothesentests so lange gehalten hat, wenn in Betracht gezogen wird, dass die in ▶ **Tab. 1** skizzierten Probleme schon seit Jahrzehnten immer wieder in Fachpublikationen des Gesundheitswesens aufgeworfen werden [8, 9], so auch in physiotherapeutischen Fachzeitschriften [10, 11]. Während es bereits Bewegungen weg von statistischen Nullhypothesentests gab, entwickelte sich die Verwendung von alternativen Methoden der statistischen Inferenz über Jahrzehnte nur langsam, wie Analysen der Gesundheitsforschung einschließlich physiotherapeutischer Studien belegen [2, 12]. Dies ungeachtet der Tatsache, dass alternative Methoden zur statistischen Inferenz nicht nur zur Verfügung standen, sondern auch in statistischen, medizinischen und physiotherapeutischen Fachzeitschriften beworben wurden [10, 13–16].

Schätzen als alternativer Ansatz der statistischen Inferenz

Obgleich es zahlreiche alternative Ansätze der statistischen Inferenz gibt [13], ist der einfachste Ansatz das Schätzen [17]. Auch das Schätzen basiert auf einem frequentistischen Ansatz, doch im Gegensatz zu statistischen Nullhypothesentests verfolgt es das Ziel, Parameter von Grundgesamtheiten auf der Grundlage von Daten aus der Studienstichprobe zu schätzen[2]. Die Unsicherheit und Ungenauigkeit solcher Schätzungen wird dabei durch Konfidenzintervalle vermittelt [10, 14].

Ein Konfidenzintervall lässt sich auf Grundlage der in der Studie beobachteten Daten, der Größe der Stichprobe, der Variabilität der Stichprobe und des Konfidenzniveaus berechnen. Das Konfidenzniveau wird durch die Forschenden bestimmt und liegt in der Regel bei 95 %. Dies bedeutet, dass bei einer hypothetisch vielfachen Wiederholung der Studie der wahre Parameter der Grundgesamtheit in 95 % der Fälle vom jeweiligen Konfidenzintervall überdeckt werden würde. In der Praxis wird ein solches Konfidenzintervall dann vereinfacht als der Bereich interpretiert, in dem sich der wahre Parameter mit einer Wahrscheinlichkeit von 95 % befindet.

2 Anmerkung *physioscience*: Dabei ist der beobachtete Wert der Statistik die sogenannte Punktschätzung. Das Konfidenzintervall ist eine Intervallschätzung.

▶ **Tab. 1** Probleme von statistischen Nullhypothesentests (modifiziert nach Herbert 2019 [26]).

Problem	Erläuterung
Ein *p*-Wert gibt nicht die Wahrscheinlichkeit an, mit der eine Hypothese (nicht) wahr ist.	Forschende müssen die Wahrscheinlichkeit kennen, mit der die Nullhypothese auf Grundlage der in ihrer Studie beobachteten Daten wahr ist. Ein *p*-Wert gibt stattdessen die Wahrscheinlichkeit an, dass die beobachteten Daten beobachtet werden, wenn die Nullhypothese wahr ist. Diese beiden Wahrscheinlichkeiten mögen austauschbar erscheinen, sind es aber nicht. Daher geben *p*-Werte keine Wahrscheinlichkeit an, die die Forschenden kennen müssen.
Ein *p*-Wert stellt keine Evidenz dar.	Wie vorstehend erläutert, gibt ein *p*-Wert die Wahrscheinlichkeit einer Beobachtung unter der Voraussetzung an, dass eine bestimmte Hypothese wahr ist.[1] Jegliche Wahrscheinlichkeit einer Beobachtung bei einer als wahr gegebenen Hypothese kann keine Evidenz für oder gegen diese Hypothese liefern. Es ist lediglich möglich, die Stärke der Evidenz für eine Hypothese zu quantifizieren, indem sie mit einer anderen Hypothese verglichen wird.
Statistisch signifikante Erkenntnisse sind nicht sonderlich reproduzierbar.	Wird eine Studie mit einer neuen zufälligen Stichprobe aus der gleichen Grundgesamtheit wiederholt, wird das Ergebnis (und damit auch der *p*-Wert) wahrscheinlich variieren. Stellen Sie sich eine Studie mit einem *p*-Wert zwischen 0,005 und 0,05 vor. Würde diese Studie mit einer neuen zufälligen Stichprobe aus der gleichen Grundgesamtheit wiederholt, läge die Wahrscheinlichkeit eines nicht signifikanten *p*-Wertes bei 33 % [27].
Bei den meisten klinischen Studien muss die Nullhypothese falsch sein.	Die Nullhypothese lautet, dass der untersuchte Effekt bei exakt Null liegt. Bei fast allen Interventionen ist davon auszugehen, dass sie einen gewissen Effekt haben, und sei dieser noch so verschwindend gering. Bei fast allen Studien (auch bei jenen mit solidester Methodik) ist von einem gewissen Bias auszugehen, und sei dieser noch so verschwindend gering. Deshalb sollten sämtliche Studien einen Effekt identifizieren (denn die Nullhypothese ist nicht wahr, d. h. der untersuchte Effekt ist nicht exakt Null). Dies impliziert, dass jedes statistisch nicht signifikante Ergebnis tatsächlich ein Versagen anzeigt, einen vorhandenen Effekt zu entdecken.
Forschende benötigen Informationen zur Effektstärke.	Forschende benötigen mehr als nur die Information, dass ein Effekt (nicht) vorhanden ist. Forschende müssen die Größe der Effektstärke kennen. Ein *p*-Wert gibt keine Informationen zur Effektgröße oder -richtung.

[1] Anmerkung *physioscience*: Ein p-Wert gibt die frequentistische Wahrscheinlichkeit an, mit der die vorliegende Statistik – oder extremere – unter Gültigkeit der Nullhypothese beobachtet würden.

Konfidenzintervalle werden häufig im Zusammenhang mit Behandlungseffekten in klinischen Studien diskutiert [18, 19]. Es ist aber möglich, ein Konfidenzintervall um jede Statistik zu konstruieren, unabhängig von ihrer Verwendung. Dazu gehören:

- Mittelwertdifferenz
- Risiko
- Chance
- relatives Risiko
- Chancenverhältnis
- Hazard Ratio
- Korrelation
- Proportion
- absolute Risikoreduktion
- relative Risikoreduktion
- Number Needed to Treat
- Sensitivität
- Spezifität
- Likelihood Ratio (LR)
- diagnostisches Chancenverhältnis
- Mediandifferenz

Interpretation der Ergebnisse der Schätzung

Um die Schätzung sinnvoll einzusetzen, reicht es nicht aus, lediglich Konfidenzintervalle zu berichten. Forschende müssen auch die Relevanz der durch die Konfidenzintervalle dargestellten Informationen interpretieren und deren Implikationen bedenken. Der Weg der Forschenden weg von statistischer Signifikanz und p-Werten hin zu Schätzmethoden ist mit Beispielen gesäumt, in denen Forschende auf Geheiß von Herausgebenden zwar Konfidenzintervalle berechnen, diese dann aber ignorieren und ihre Studienergebnisse stattdessen auf Grundlage des p-Wertes dichotom als statistisch signifikant oder nicht signifikant interpretieren [20]. Die Interpretation der berechneten Konfidenzintervalle ist jedoch unerlässlich.

Manche Autor*innen haben schon für ein Verbot sämtlicher Begriffe plädiert, die im Zusammenhang mit statistischen Nullhypothesentests stehen. Ein prominentes Beispiel ist das folgende Zitat aus dem Editorial einer Sonderausgabe der Fachzeitschrift The American Statistician [13] zum Thema statistische Inferenz:

Die Stellungnahme der American Statistical Association „Statement on P-Values and Statistical Significance" stand bereits kurz davor, die völlige Abwendung von Erläuterungen zur „statistischen Signifikanz" zu empfehlen. Wir gehen hier diesen Schritt. Basierend auf dem Überblick der in dieser Sonderausgabe erschienen Artikel

und der einschlägigen Fachliteratur lautet unser Fazit: Es ist an der Zeit, gänzlich auf die Verwendung des Begriffs „statistisch signifikant" zu verzichten. Auch Variationen wie „signifikant unterschiedlich", „p < 0,05" und „nicht signifikant" sollten von der Bildfläche verschwinden, egal ob sie mit Worten, durch Fußnoten in Tabellen oder auf sonstige Art zum Ausdruck gebracht werden.

Dieser Anspruch mag radikal und undurchführbar für Forschende erscheinen, die seit langem gewohnt sind, mit statistischen Nullhypothesentests zu arbeiten, aber viele ihrer Bedenken können zerstreut werden. Erstens würde ein solches Verbot die Forschungsergebnisse, die in den letzten Jahrzehnten unter Verwendung von statistischen Nullhypothesen berichtet wurden, nicht verwerfen. Die Daten, die in solchen Studien generiert wurden, bleiben valide – und sie wurden oft hinreichend detailliert berichtet, um Konfidenzintervalle berechnen zu können. Zweitens bedeutet eine solche Neuausrichtung des Studienziels letztlich nur eine einfache Verlagerung des Schwerpunkts von der Frage, ob das Ergebnis statistisch signifikant ist, auf die Frage, wie groß und genau die Schätzung des Parameters der Grundgesamtheit durch die Studie ist. Statt beispielsweise entscheiden zu wollen, ob eine Behandlung einen Effekt ungleich Null auf Überlebende eines Schlaganfalls hat, wäre die primäre Zielsetzung nun, die Größe des durchschnittlichen Effekts zu schätzen. Oder statt bestimmen zu wollen, ob ein prognostisches Modell prädikativ ist, wäre nun das Ziel zu schätzen, wie gut die Vorhersage durch das Modell ist. Drittens kann die statistische Ungenauigkeit solcher Schätzungen leicht berechnet werden. Es gibt bereits Statistik-Software, die Konfidenzintervalle berechnet, darunter auch kostenfreie Software wie R [21, 22]. Und schließlich ist die Interpretation von Konfidenzintervallen relativ einfach zu erlernen.

Viele Forschende und Leser*innen entwickeln beim frühen Zugang ein Verständnis für die Interpretation von Konfidenzintervallen im Zusammenhang mit Schätzungen zur Wirksamkeit von Behandlungen. In einer Studie, in der die behandelten Proband*innen mit denen einer Kontrollgruppe verglichen werden, und in der ein kontinuierlicher Endpunkt zum Einsatz kommt, ist die „beste" Schätzung des wahren Behandlungseffektes in der Regel der beobachtete Zwischengruppenunterschied. Um der Tatsache Rechnung zu tragen, dass die Schätzungen auf Basis einer Stichprobe vom wahren Zwischengruppenunterschied in der Grundgesamtheit abweichen kann, gibt das Konfidenzintervall einen Hinweis auf die Spanne von plausiblen wahren Zwischengruppenunterschieden oberhalb und unterhalb dieser Schätzung, innerhalb derer sich der wahre Zwischengruppenunterschied in der betreffenden klinischen Grundgesamtheit mit einer großen Wahrscheinlichkeit befindet.

Die Punktschätzung und das Konfidenzintervall sollte mit dem „kleinsten lohnenswerten Effekt" der Intervention auf diesen Ergebnisparameter in dieser Grundgesamtheit verglichen werden [23]. Der kleinste lohnenswerte Effekt ist der geringste Nutzen einer Intervention, bei dem Patient*innen noch das Gefühl haben, dass er die Kosten, Risiken und andere Unannehmlichkeiten überwiegt [23]. Liegt die untere Grenze des Konfidenzintervalls – und damit auch die Punktschätzung sowie die obere Grenze – oberhalb des kleinsten lohnenswerten Effektes, kann davon ausgegangen werden, dass Patient*innen aus der betroffenen klinischen Grundgesamtheit den Effekt der Behandlung in der Regel als relevant betrachten werden. Liegen hingegen sowohl die Punktschätzung als auch die Grenzwerte des Konfidenzintervalls unterhalb des kleinsten lohnenswerten Effektes, kann davon ausgegangen werden, dass Patient*innen aus der betroffenen klinischen Grundgesamtheit den Effekt der Behandlung in der Regel als unerheblich betrachten werden. Ergebnisse, deren Konfidenzintervalle den kleinsten lohnenswerten Effekt überspannen, deuten darauf hin, dass es einen Effekt gibt, dessen Relevanz unsicher ist. Ergebnisse mit einem engen Konfidenzintervall, das den Nulleffekt[3] beinhaltet, deuten darauf hin, dass der Effekt der Behandlung vernachlässigbar ist. Ergebnisse mit einem breiten Konfidenzintervall, die den Nulleffekt beinhalten, deuten hingegen darauf hin, dass der Effekt der Behandlung unsicher ist. Für Leser*innen, die mit dieser Art der Interpretation nicht vertraut sind, stehen einige klare, für Laien verständliche Artikel mit Beispielen aus der klinischen Physiotherapie zur Verfügung [10, 14, 18, 19].

Eine Interpretation von Schätzungen zu Behandlungseffekten und ihren Konfidenzintervallen baut darauf auf, dass der kleinste lohnenswerte Effekt (auch als minimaler klinisch relevanter Unterschied bezeichnet) bekannt ist [23]. Für manche Forschungsfragen wurde ein solcher Schwellenwert bislang noch gar nicht oder nur unter Verwendung ungeeigneter Methoden festgelegt. In solchen Fällen sollten Forschende erwägen, eine Studie durchzuführen, um den Schwellenwert zu bestimmen oder diesen zumindest prospektiv benennen.

Leser*innen, die Intervallschätzungen zu Behandlungseffekten interpretieren können, werden auch schnell mit Interpretationen zu Konfidenzintervallen für andere interessierende Phänomene vertraut sein. Vereinfacht ausgedrückt gibt das Konfidenzintervall die Spannbreite um die Punktschätzung einer Statistik an, innerhalb derer sich der wahre Parameter mit einer großen Wahrscheinlichkeit befindet. Um ein Konfidenzintervall zu interpretieren, beschreiben wir einfach die praktischen Implikationen sämtlicher Werte innerhalb des Intervalls [24]. In einer Studie zur Güte eines diagnostischen Tests zeigt uns zum Beispiel die Likelihood Ratio (LR) – bei vorliegendem positivem Test – das Verhältnis an, um wieviel wahrscheinlicher es ist, dass Personen mit der Krankheit positiv getestet werden, als Personen, bei denen die betroffene Krankheit nicht vorliegt, also das Verhältnis der richtig-positiv- zur falsch-positiv-Rate. Ein LR von mehr als 3 (LR > 3) ist in der Regel nützlich. Bei LR > 10 ist der Test sogar sehr nützlich [25]. Bei einer Punktschätzung von LR = 4,8 und einem Konfidenzintervall für das wahre LR von 4,1 bis 5,6 können wir davon ausgehen, dass das wahre LR nicht nur auf einen brauchbaren Test hindeutet, sondern es auch in etwa der Punktschätzung entspricht.

Wird hingegen in einer Studie geschätzt, dass die Prävalenz für eine Depression bei Personen mit Zustand nach einer Ruptur des hinteren Kreuzbandes 40 % beträgt, mit einem Konfidenzintervall zwischen 5 % und 75 %, können wir zwar annehmen, dass die Punktschätzung auf eine hohe Prävalenz hindeutet, für eine eindeutige Schlussfolgerung ist es jedoch zu ungenau.

3 Anmerkung *physioscience*: Zwischengruppenunterschied = 0

▶ **Tab. 2** Quellen mit zusätzlichen Informationen zur Beantwortung von Fragen zum Übergang von statistischen Nullhypothesentests zu Schätzmethoden.

Frage	Quellen
Wo finde ich nähere Informationen über statistische Nullhypothesentests und die damit verbundenen Probleme?	Dieser kurze Artikel gibt detaillierte Informationen über die Probleme, die mit Signifikanz- und Hypothesentests verbunden sind [25]: https://doi.org/10.1016/j.jphys.2019.05.001
Sind diese Probleme und die Notwendigkeit einer Alternative allgemein anerkannt?	Die Stellungnahme der American Statistical Association zu *p*-Werten [28] zeigt, dass ein diesbezügliches Problembewusstsein unter Statistiker*innen weit verbreitet ist. Zahlreiche Forschungsgebiete haben die Notwendigkeit erkannt, sich von Signifikanztests zu verabschieden, darunter Medizin im Allgemeinen [29], spezifische medizinische Teilgebiete [30, 31], Pflege [32], Psychologie [33], Neurowissenschaft [34], Pharmazie [35], Toxikologie [36], Anthropologie [37] und Veterinärforschung [38].
Gibt es eine Publikation, die Konfidenzintervalle von Grund auf erklärt?	Diese beiden Leitartikel erklären Konfidenzintervalle für kontinuierliche und dichotome Variablen [10, 14]: https://doi.org/10.1016/S0004-9514(14)60334-2, https://doi.org/10.1016/s0004-9514(14)60292-0
Gibt es Beispielpublikationen zur Interpretation von Konfidenzintervallen?	Diese beiden kurzen Artikel erläutern Konfidenzintervalle und geben Beispiele zu ihrer Beschreibung in Worten [18, 19]: https://doi.org/10.1016/j.bjpt.2019.01.003, https://www.jospt.org/doi/10.2519/jospt.2019.0706
Wie kann ich auf Grundlage meiner Rohdaten Konfidenzintervalle berechnen?	Es gibt Statistiksoftware, die Konfidenzintervalle berechnet, darunter kostenfreie Programme wie R [21, 22].
Wie kann ich schnell Konfidenzintervalle aus aggregierten Daten einer bereits publizierten Studie berechnen?	Ein kostenfreier Konfidenzintervall-Rechner auf Excel-Basis steht auf der Webseite von PEDro zum Download zur Verfügung: https://pedro.org.au/english/resources/confidence-interval-calculator/

Regelungen der ISPJE-Mitgliedszeitschriften zum Schätzen

Der Vorstand der ISPJE empfiehlt seinen Mitgliedern dringend, darauf hinzuwirken, dass in den Artikeln, die in den von ihnen herausgegebenen Fachzeitschriften publiziert werden, Punkt- und Intervallschätzungen verwendet werden. Im Einklang mit dieser Empfehlung weisen die Co-Autor*innen dieses Editorials Forschende darauf hin, dass sie zukünftig Manuskripte erwarten, in denen Schätzungen anstelle statistischer Nullhypothesentests verwendet werden. Wir erkennen an, dass es einige Zeit erfordern wird, bis der Übergang vollzogen ist. Daher werden die Herausgebenden den Autor*innen die Gelegenheit geben, ihre Manuskripte zu überarbeiten und Schätzmethoden einzusetzen, wenn ein Manuskript ansonsten die Voraussetzungen für eine Publikation erfüllt. Bei Bedarf könnten die Herausgebenden die Autor*innen bei der Überarbeitung ggf. unterstützen.

Leser*innen, die nähere Informationen zur Klärung der in diesem Editorial angesprochenen Fragen benötigen, verweisen wir auf die Quellen in ▶ **Tab. 2**. In dieser finden sie unter anderem einen wissenschaftlichen Beitrag zu den Problemen von Signifikanz- und Hypothesentests [25] sowie ein hervorragendes Lehrbuch zu den Themen Konfidenzintervalle und Anwendung von Schätzmethoden in Studien mit unterschiedlichen Designs, inklusive Beispiele zur praktischen Physiotherapie [26]. Diese beiden Quellen sind auch für Forschende und Praktiker*innen ohne Vorkenntnisse zu den behandelten Themen gut verständlich.

Quantitative Forschungsarbeiten zur Physiotherapie, die mittels Konfidenzintervallen analysiert und interpretiert werden, liefern validere und relevantere Informationen als jene, die mittels statistischen Nullhypothesentests analysiert und interpretiert werden. Daher bietet die Schätzmethode Forschenden, Praktiker*innen und anderen Nutzer*innen, die sich auf die physiotherapeutische Forschung verlassen, großes Potenzial. Vor diesem Hintergrund empfiehlt die ISPJE ihren Mitgliedern, ihre Anwendung in den Artikeln der von ihnen herausgegebenen Zeitschriften zu fördern.

Finanzielle Unterstützung: keine
Provenienz: auf Einladung, ohne Peer-Review
Danksagung: Wir danken Prof. Rob Herbert von Neuroscience Research Australia (NeuRA) für seine Präsentation zum Thema bei der ISPJE und für seine Anmerkungen zu einem Entwurf dieses Leitartikels.

Interessenkonflikt

Die Autorinnen/Autoren geben an, dass kein Interessenkonflikt besteht.

Zitierweise für diesen Artikel

Elkins et al. Statistical inference through estimation: recommendations from the International Society of Physiotherapy Journal Editors. Journal of Physiotherapy; 2021; Volume 68, Issue 1, Pages 1–4

Literatur

[1] Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods 2000; 5: 241–301. doi:10.1037/1082-989x.5.2.241

[2] Freire APCF, Elkins MR, Ramos EM et al. Use of 95% confidence intervals in the reporting of between-group differences in randomized controlled trials: analysis of a representative sample of 200 physical therapy trials. Braz J Phys Ther 2019; 23: 302–310. doi:10.1016/j.bjpt.2018.10.004

[3] Altman DG, Bland JM. Uncertainty and sampling error. BMJ 2014; 349: g7064. doi:10.1136/bmj.g7064

[4] Barnett V. Comparative Statistical Inference. London, New York: Wiley; 1973

[5] Royall RM. Statistical Evidence: A Likelihood Paradigm. 1st ed. London, New York: Chapman & Hall; 1997

[6] Gigerenzer G. The Empire of Chance: How Probability Changed Science and Everyday Life. Cambridge, England: Cambridge University Press; 1989

[7] Goodman SN, Royall R. Evidence and scientific research. Am J Public Health 1988; 78: 1568–1574. doi:10.2105/ajph.78.12.1568

[8] Ziliak S, McCloskey D. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. Ann Arbor, USA: University of Michigan Press; 2008

[9] Hubbard R. Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science. Thousand Oaks, USA: Sage; 2016

[10] Herbert RD. How to estimate treatment effects from reports of clinical trials. I: Continuous outcomes. Aust J Physiother 2000; 46: 229–235. doi:10.1016/S0004-9514(14)60334-2

[11] Maher CG, Sherrington C, Elkins M et al. Challenges for evidence-based physical therapy: accessing and interpreting high-quality evidence on therapy. Phys Ther 2004; 84: 644–654. doi:10.1093/ptj/84.7.644

[12] Yi D, Ma D, Li G et al. Statistical use in clinical studies: is there evidence of a methodological shift? PLoS one 2015; 10: e0140159. doi:10.1371/journal.pone.0140159

[13] Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". Am Stat 2019; 73: 1–19. doi:10.1080/00031305.2019.1583913

[14] Herbert RD. How to estimate treatment effects from reports of clinical trials. II: Dichotomous outcomes. Aust J Physiother 2000; 46: 309–313. doi:10.1016/s0004-9514(14)60292-0

[15] Sim J, Reid N. Statistical inference by confidence intervals: issues of interpretation and utilization. Phys Ther 1999; 79: 186–195. doi:10.1093/ptj/79.2.186

[16] Rothman KJ. Disengaging from statistical significance. Eur J Epidemiol 2016; 31: 443–444. doi:10.1007/s10654-016-0158-2

[17] Cumming G. Multivariate applications series. New York: Routledge; 2012

[18] Kamper SJ. Showing confidence (intervals). Braz J Phys Ther 2019; 23: 277. doi:10.1016/j.bjpt.2019.01.003

[19] Kamper SJ. Confidence intervals: linking evidence to practice. J Orthop Sports Phys Ther 2019; 49: 763–764. doi:10.2519/jospt.2019.0706

[20] Fidler F, Thomason N, Cumming G et al. Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. Psychol Sci 2004; 15: 119–126. doi:10.1111/j.0963-7214.2004.01502008.x

[21] R Core Team, Hrsg. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2020. Im Internet (Stand: 01.12.2021): www.R-project.org/

[22] RStudio Team, Hrsg. RStudio: Integrated Development for R. RStudio, Inc., Boston, USA; 2019. Im Internet (Stand: 01.12.2021): www.rstudio.com/

[23] Ferreira M. Research Note: The smallest worthwhile effect of a health intervention. J Physiother 2018; 64: 272–274. doi:10.1016/j.jphys.2018.07.008

[24] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019; 567: 305–307. doi:10.1038/d41586-019-00857-9

[25] Herbert R. Research Note: Significance testing and hypothesis testing: meaningless, misleading and mostly unnecessary. J Physiother 2019; 65: 178–181. doi:10.1016/j.jphys.2019.05.001

[26] Herbert RD, Jamtvedt G, Mead J et al. Practical Evidence-Based Physiotherapy. 2. Aufl. Oxford: Elsevier; 2011

[27] Boos DD, Stefanski LA. P-Value Precision and Reproducibility. Am Stat 2011; 65: 213–221. doi:10.1198/tas.2011.10129

[28] Wasserstein R, Lazar N. The ASA's Statement on p-Values: Context, Process, and Purpose. Am Stat 2016; 70: 129–133. doi:10.1080/00031305.2016.1154108

[29] International Committee of Medical Journal Editors, Hrsg. ICMJE recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals. 2013. Im Internet (Stand: 01.12.2021): www.icmje.org/icmje-recommendations.pdf

[30] McGough JJ, Faraone SV. Estimating the size of treatment effects: moving beyond p values. Psychiatry 2009; 6: 21

[31] Hayat MJ, Chandrasekhar R, Dietrich MS et al. Moving Otology Beyond p < 0.05. Otol Neurotol 2020; 41: 578–579

[32] Hayat MJ, Staggs VS, Schwartz TA et al. Moving nursing beyond p < .05. Res Nurs Health 2019; 42: 244–245. doi:10.1002/nur.21954

[33] Cumming G, Fidler F, Kalinowski P et al. The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. Aust J Psychol 2012; 64: 138–146

[34] Calin-Jageman RJ, Cumming G. Estimation for better inference in neuroscience. eNeuro 2019; 6. doi:10.1523/ENEURO.0205-19.2019

[35] Schreiber JB. New paradigms for considering statistical significance: A way forward for health services research journals, their authors, and their readership. Res Social Adm Pharm 2020; 16: 591–594

[36] Erickson RA, Rattner BA. Moving beyond p < 0.05 in ecotoxicology: A guide for practitioners. Environ Toxicol Chem 2020; 39: 1657–1669

[37] Smith RJ. P > .05: The incorrect interpretation of "not significant" results is a significant problem. Am J Phys Anthropol 2020; 172: 521–527

[38] Du Sert NP, Ahluwalia A, Alam S et al. Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLoS biology 2020; 18: e3000411