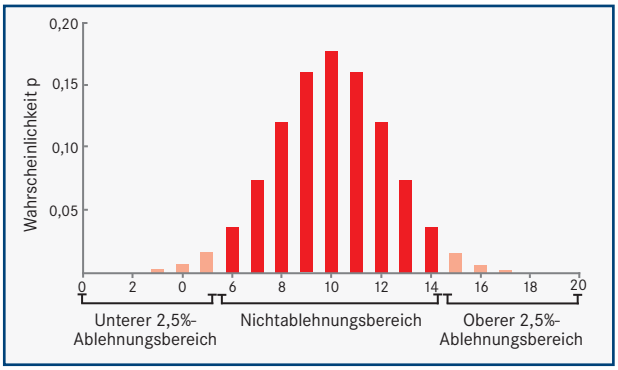
p-Wert und Konfidenzintervall richtig interpretieren

Ob die Ergebnisse einer Studie signifikant sind, lässt sich anhand des p-Werts und des Konfidenzintervalls erkennen. Doch was genau verbirgt sich hinter diesen Zahlen und wie aussagekräftig sind sie wirklich?



Dargestellt sind die Wahrscheinlichkeiten für ein Versuchsergebnis mit 20 Münzwürfen. Ergebnisse von 0 bis 5 und von 15 bis 20 sind sehr unwahrscheinlich, wenn H0 zutrifft. Deswegen würde H0 in diesem Fall abgelehnt.[1]

Forscher wollen aus Versuchsergebnissen richtige Schlüsse ziehen. Daher müssen sie in jeder Studie die Frage beantworten, ob die Ergebnisse der untersuchten Stichprobe auf die Grundgesamtheit übertragen werden können. Die Werkzeuge der beurteilenden Statistik helfen dabei, bieten aber keine absolute Sicherheit. Der p-Wert ist das Ergebnis eines Signifikanztests und gibt wieder, wie wahrscheinlich ein Versuchsergebnis ist, wenn die Nullhypothese zutrifft. Das Konfidenzintervall hingegen ist ein Wertebereich, der aus den Daten der Stichprobe errechnet wird und den wahren Wert in der Grundgesamtheit mit einer festgelegten Wahrscheinlichkeit enthält.

Häufig missverstanden: der p-Wert

Der p-Wert ist eine Maßzahl dafür, wie wahrscheinlich ein Versuchsergebnis ist, wenn die Hypothese H0 zutrifft. Auch wenn er oft so verstanden wird, ist er kein Beweis für eine der beiden Hypothesen.

Ein Beispiel: Möchte man herausfinden, ob eine Münze bei einem Wurf zufällig Kopf oder Zahl ergibt, stellt man zwei Hypothesen auf: H0 besagt, dass das Ergebnis zufällig ist, H1 besagt, dass das Ergebnis nicht zufällig ist. Anschließend werfen wir die Münze 20 Mal und zählen Kopf als 1 und Zahl als 0, die Summe ist das Versuchsergebnis.

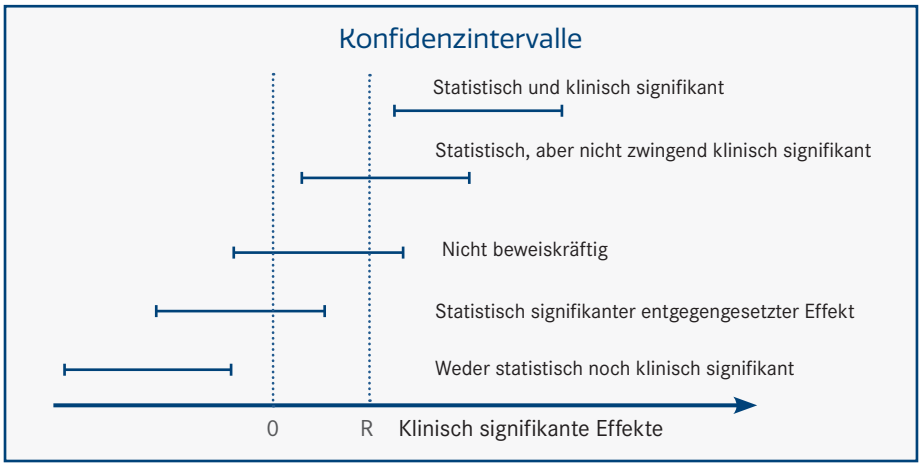
Um das Ergebnis beurteilen zu können, müssen wir uns die Wahrscheinlichkeiten aller möglichen Ergebnisse für den Fall vorstellen, dass H0 zutrifft: Sie decken die Skala von 0 (nur Zahl) bis 20 (nur Kopf) ab und entsprechen einer Normalverteilung (Abb.1). Dabei werden die Ergebnisse zu den Rändern hin unwahrscheinlicher. Der p-Wert nun gibt an, wo in dieser Grafik das Versuchsergebnis liegt: Liegt es im unteren oder im oberen Ablehnungsbereich, gilt es als nicht signifikant, da es mit großer Wahrscheinlichkeit nicht zufällig entstanden ist, wenn H0 zutrifft; daher würde dann H0 abgelehnt.

Trotzdem kann das Ergebnis mit der Wahrscheinlichkeit des p-Wertes entstanden sein, auch wenn H0 zutrifft. Anders ausgedrückt: Das Risiko, H0 fälschlich abzulehnen, ist so hoch wie der p-Wert. Die Signifikanzschwelle von 5% ist ein mehr oder weniger willkürlich gewählter Wert, der je nach den Umständen (wie wichtig ist es, H0 nicht fälschlich abzulehnen) geändert werden kann.

Nach diesem (hier stark vereinfachten) Prinzip werden auch klinische Studien ausgewertet: Zunächst werden H0 und H1 aufgestellt, dann wird aus der Grundgesamtheit eine Stichprobe ausgewählt und in der Stichprobe werden Daten erhoben. Ist der errechnete p-Wert kleiner als 5%, geht man davon aus, dass die Alternativhypothese zutrifft und H0 als zurückgewiesen gilt. Unter diesen Bedingungen geht man davon aus, dass H0 nicht auf die Grundgesamtheit übertragen werden kann.

Häufig wird ein p-Wert kleiner als 5% als Beweis für H1 verstanden, der er nicht ist: Der p-Wert gibt die Wahrscheinlichkeit an, mit der das erhaltene Versuchsergebnis eintritt, wenn H0 zutrifft. Liegt diese Wahrscheinlichkeit unter 5%, wird H0 abgelehnt und damit H1 angenommen.

Der Informationsgehalt eines reinen p-Wertes ist verhältnismäßig gering, da er nicht wiedergibt, in welche Richtung die Messwerte zwischen 2 Gruppen (z. B. Verum und Placebo) abweichen, und weil er nicht zeigt, wie groß die Differenz der Messwerte zwischen beiden Gruppen ist (klinische Relevanz des Ergebnisses). Daher muss der p-Wert immer um eine Maßzahl der Effektstärke ergänzt werden. Darüber hinaus fehlen Informationen über die Präzision der Messungen.



Gezeigt ist, wie statistische Signifikanz und klinische Relevanz eines Versuchsergebnisses am Konfidenzintervall beurteilt werden kann. R ist die Relevanzgrenze, der Null-Effektwert ist 0.[3]

Häufig unterschätzt: das Konfidenzintervall

Anders als der p-Wert geben Vertrauensbereiche (Konfidenzintervalle) die Ergebnisse auf der Ebene der Messwerte wieder: Ein Konfidenzintervall schließt bei unendlicher Wiederholung eines Experiments die wahre Lage des Parameters mit einer festgelegten Wahrscheinlichkeit (Konfidenzniveau) ein.

Ein Konfidenzintervall hat zwei wichtige Eigenschaften: Genauigkeit (Breite des Konfidenzintervalls) und Sicherheit (Konfidenzniveau). Je höher bei gleicher Stichprobengröße die Sicherheit sein soll, desto geringer wird die Genauigkeit. Die einzige Möglichkeit, ein festgelegtes Konfidenzniveau zu erreichen und gleichzeitig Konfidenzintervalle mit vorbestimmter Breite zu erhalten, ist eine ausreichend große Stichprobe.

Ein breites (ungenaues) Konfidenzintervall kann entstehen bei einer kleinen Stichprobe, bei großer Streuung der Messwerte und hohem Konfidenzniveau. Enthält das Konfidenzintervall den Messwert für „keinen Effekt“, kann man davon ausgehen, dass das Ergebnis nicht statistisch signifikant ist – und umgekehrt.

Ein Konfidenzintervall beinhaltet mehrere Informationen: Werte außerhalb des Intervalls sind möglich, aber unwahrscheinlich. Werte, die innerhalb des Intervalls liegen, sind umso wahrscheinlicher, je näher sie dem Mittelwert liegen. Gerade bei Ergebnissen, die nur knapp unter der Signifikanzgrenze liegen, können diese Informationen wertvolle Hinweise darauf geben, ob nicht bei einer größeren Stichprobe ein signifikantes Ergebnis erzielt worden wäre.

Die Beurteilung von klinischen Studienergebnissen mit Hilfe von Konfidenzintervallen und Signifikanz ist an einem schematischen Beispiel (Abb. oben) erklärt: Dargestellt ist der Effekt eines Medikaments. Ergebnisse, die über der Relevanzgrenze R liegen, sind klinisch relevant; Konfidenzintervalle, welche die 0 (Null-Effekt) mit einschließen, sind statistisch nicht signifikant. Ist das Studienergebnis statistisch nicht signifikant und klinisch nicht relevant oder statistisch signifikant und klinisch relevant, ist es leicht zu interpretieren. Ist der statistisch signifikante Effekt zu gering, um klinisch relevant zu sein, ist das Ergebnis klinisch bedeutungslos. Ist aber ein statistisch nicht signifikanter Unterschied groß genug, um klinisch relevant zu sein, kann das Ergebnis klinisch bedeutsam sein. ■ *Roland Müller-Waldeck*

1) Lorenz RJ. Grundbegriffe der Biometrie, 2. Aufl. 1988, Gustav Fischer Verlag. 2) du Prel JB, Hommel G, Röhrig B, Blettner M. Konfidenzintervall oder P-Wert?; Dtsch Arztebl 2009; 106(19):335-9. 3) Ranstam J Why the P-value culture is bad and confidence intervals a better alternative. Osteoarthritis Cartilage. 2012 Aug; 20(8):805-8. doi: 10.1016/j.joca.2012.04.001. Epub 2012 Apr 11. 4) Lee DK. Alternatives to P value: confidence interval and effect size.Korean J Anesthesiol. 2016 Dec;69(6):555-562. Epub 2016 Oct 25.