Über die Signifikanz eines statistischen Tests und die zugehörigen Fehlentscheidungen

In diesem EbM-Splitter wollen wir die prinzipiellen Zusammenhänge des statistischen Testens erläutern und besonders auf die Interpretationsmöglichkeiten in der Anwendung eingehen. Zur Illustration wählen wir ein hypothetisches Beispiel aus der zahnärztlichen Prothetik. Die Wahl einer erdachten Situation mag dem Leser auf den ersten Blick befremdlich erscheinen; im Vorteil gegenüber realen Daten ermöglicht sie jedoch eine gewisse Flexibilität hinsichtlich der Problemstellung und der äußeren Bedingungen, inklusive des Stichprobenumfangs.

Nehmen wir also an, dass ein neuer Zement für die Befestigung von Kronen im Seitenzahnbereich klinisch erprobt werden soll. Wir geben diesem Zement den Namen COLLE (französisch für Kleber). Für die angestrebte Darstellung benötigen wir zudem einen Gegenspieler oder Goldstandard. Letzterer ist in der Regel ein Zement, der sich bereits in der Praxis bewährt hat, und den wir im Folgenden einfach OR (französisch für Gold) nennen.

Wir beginnen mit dem klassischen Test auf Unterschied, dessen Null-Hypothese lautet:

H_0: COLLE klebt genauso gut wie OR. (1)

Eine solche Aussage wird natürlich erst lebendig, wenn die Güte des Klebens irgendwie quantifiziert werden kann. Wir betrachten dafür als maßgeblich das Herausfallen durch selbständiges Lösen der Krone binnen der ersten 5 Jahre und setzen H_0 gleich mit der Aussage:
Mit COLLE und OR befestigte Kronen lösen sich mit derselben Wahrscheinlichkeit heraus.

Die Alternative, auch Alternativhypothese genannt, wird gewöhnlich mit H_1 bezeichnet, und ist immer das logische Gegenstück zu der Null-Hypothese im Sinne der möglichen Versuchsausgänge:

H_1: Entweder COLLE oder OR erzielt bessere Festigkeit. (2)

Weitere Überlegungen stellen wir unter die folgende Arbeitshypothese:
Das Herausfallen einer Krone kann eindeutig dem Versagen des Zements zugeordnet werden.
Unter dieser Annahme gibt es für jeden Zement einen unbekannten Parameter, der das Herausfallen der Kronen steuert. Dieser Parameter symbolisiert die Güte des Zements; man kann ihn sich zum Beispiel als chemisch-physikalischen Eigenschaft vorstellen. Es ist wichtig zu bemerken, daß unter H_1 selbst der kleinste Güte-Unterschied fällt.

Auf die Auswahl eines geeigneten statistischen Tests soll hier nicht weiter eingegangen werden, da die zu beschreibenden Zusammenhänge auf alle Tests gleichermaßen zutreffen.

Jeder statistische Test für die obige Fragestellung kennt nur zwei mögliche Ausgänge, nämlich die Entscheidung für die Null-Hypothese (H_0) oder deren Verwerfung, was gleichbedeutend mit der Entscheidung für die Alternative (H_1) ist. Wir können bereits an dieser Stelle bemerken, dass die Testentscheidung nicht die Größe des eventuellen Unterschieds beurteilt, was einen klaren Nachteil gegenüber dem direkten Schätzen des Unterschieds mit Konfidenzintervall darstellt [vgl. 1]. Das Problem wird hier stark vereinfacht, indem nur „ja, es gibt einen Unterschied“ oder „nein, es gibt keinen Unterschied“ als mögliche Antwortsätze berücksichtigt werden.

Es gibt auch nur zwei der Realität entsprechende (dem Problem zugrunde liegende) Parameterkonstellationen: Entweder H_0 ist wahr, also die unbekannten Klebeeigenschaften sind exakt gleich, oder H_1 ist wahr, dann gibt es einen – wenn auch noch so kleinen – Unterschied. Die zentrale Aufgabe der Statistik ist, mithilfe von Daten Rückschlüsse auf die unbekannten Parameter zu erhalten, die die Güte der Zemente beschreiben. Beim statistischen Testen entsprechen Rückschlüsse den möglichen Entscheidungen, also entweder für die Null-Hypothese oder für deren Alternative (siehe Tab. 1).

Parameterkonstellation	Testentscheidung	Bewertung
H_0 ist wahr	für H_0	kein Fehler
	für H_1	Fehler 1. Art (α)
H_1 ist wahr	für H_0	Fehler 2. Art (β)
	für H_1	kein Fehler

Tabelle 1 Theoretische Entscheidungsmatrix

In einer fiktiven klinischen Studie werden nun 50 Patienten mit einer Krone versorgt, wobei jeweils die Hälfte mit COLLE und die andere Hälfte mit OR eingeklebt werden. Nach fünf Jahren – Sie werden spätestens jetzt den Vorteil der erdachten Situation bemerken – ist das Ergebnis in Tabelle 2 dargestellt.

	OR	COLLE
Erfolg	22	19
Misserfolg	3	6

Tabelle 2 Vierfeldertafel (n = 50)

Zur Durchführung eines geeigneten statischen Tests – wir wollen wie gesagt nicht näher auf die Details der Auswahl eingehen – muss zuerst das so genannte Signifikanzniveau festgelegt werden. Gewöhnlich wird dafür $\alpha = 5\ \%$ gewählt. Jeder mögliche statistische Test gelangt abhängig von α zu ei-

ner Testentscheidung auf Basis der Daten. Wir entscheiden uns für Fisher's exakten Test: P-Wert = 0,4635. Es folgt die Entscheidung für H_0, weil der P-Wert größer als das Signifikanzniveau $\alpha = 0{,}05$ ist.

Per Konstruktion ist die Wahrscheinlichkeit für eine Fehlentscheidung der 1. Art (Vergleich α in Tab. 1) des Tests durch das Signifikanzniveau auf 5 % begrenzt. Allerdings ist die Fehlentscheidung der 2. Art (β in Tab. 1) in der Regel nicht so klein wie α. Insbesondere kann aus der Testentscheidung anhand der Daten von Tabelle 2 noch nicht auf Gleichheit der Festigkeit von COLLE und OR geschlossen werden. Wenn wir nämlich die Anzahl der Patienten verzehnfachen und die Ergebnisse wie in Tabelle 3 dargestellt vorfinden, entscheidet *Fisher's* exakter Test (P-Wert= 0,00068) für H_1: Es gibt einen signifikanten Unterschied zwischen COLLE und OR.

	OR	COLLE
Erfolg	220	190
Misserfolg	30	60

Tabelle 3 Vierfeldertafel (n = 500)

Die Entscheidung für H_1 aufgrund der Daten in Tabelle 2 ist nur mit der so genannten statistischen Power $(1-\beta)$ abgesichert. Diese hängt aber entscheidend von der Anzahl der Patienten (dem Stichprobenumfang) ab. Je größer die Power des Tests, desto kleiner die Wahrscheinlichkeit für eine Fehlentscheidung der 2. Art.

Schlussbemerkungen

Ein signifikanter Unterschied muss nicht groß sein, es werden nur mehr Daten benötigt, um einen kleinen Unterschied zu dem vorgegebenen Signifikanzniveau nachzuweisen. Also Vorsicht: Auch für klinisch irrelevante Unterschiede kann man statistische Signifikanz erzielen, wenn nur die Stichprobe groß genug ist. Aus diesem Grunde reicht der Blick auf die statistische Signifikanz nicht aus, sondern man muss auch immer die Größe des betrachteten Unterschieds einbeziehen.

Ob COLLE oder OR besser ist, kann nicht an der Testentscheidung – oder stellvertretend an dem P-Wert – abgelesen werden. In Tabelle 3 erkennt man jedoch sofort, dass OR die bessere Festigkeit erzielt.

Oftmals ist es hinreichend zu zeigen, dass ein neuer Zement nicht schlechter ist als der bewährte, zum Beispiel wenn der neue Zement eindeutig bessere farbliche Eignung hat, preiswerter ist oder günstigere Verarbeitungseigenschaften aufweist. Das Prinzip der beschriebenen Tests auf Unterschied ermöglicht nur dann eine befriedigende Antwort, wenn die Größenordnung von β bekannt wäre. Besser ist es, gleich einen so genannten Äquivalenztest (Test auf Gleichheit) zu benutzen. Dabei werden einfach die Hypothesen in (1) und (2) in ihrer Bedeutung vertauscht. Näheres soll in einem der folgenden EbM-Splitter aufgenommen werden.

„Diese Geschichte ist noch längst nicht vorbei, ich habe den Eindruck, jetzt, in diesem Moment, beginnt sie."

Ror Wolf. Ein Blick auf das Leben im letzten September. In: Zwei oder drei Jahre später. Frankfurter Verlagsanstalt, Frankfurt am Main 2003, S. 89

Literatur

1. Schwarzer, G., Türp, J. C., Antes, G.: EbM-Splitter: Wie liest man klinische Studien? P-Wert und Konfidenzintervall. Dtsch Zahnärztl Z 56, 702 (2001)

Thomas Gerds, Freiburg
Jens C. Türp, Basel
Gerd Antes, Freiburg