Geschäfte machen mit Statistik

Teil 2: Der Alltag

1 • Die Strategie des Kunden

Die Gründe, einen Statistiker zu konsultieren, sind natürlich nicht vergleichbar mit denjenigen, weshalb man zu einem Arzt geht. Dennoch gibt es gewisse Ähnlichkeit: in beiden Fällen sucht man eine Entscheidung [2]. Viele Leser haben es sicherlich schon erlebt, man begibt sich zum Arzt, mit ungewissen Beschwerden, also in der Rolle des Patienten, mit der Aussicht und der Hoffnung auf Klärung der Situation. Allerdings ist die Sachlage oftmals nicht eindeutig und eine einfache Diagnose nicht möglich. Zu guter Letzt schlägt der Mediziner auch noch invasive oder anderweitig unangenehme diagnostische Maßnahmen vor. In manchen Fällen wird der Patient daraufhin einen anderen Arzt konsultieren. Diesen wird der Patient wahrscheinlich gleich zu Beginn ausführlich über die erste Beurteilung des Kollegen informieren. Vielleicht auch dadurch beeinflusst, wird die neue Berurteilung in vielen Fällen zumindest leicht von der ersten abweichen. Falls jedoch beide Einschätzungen gegen die eigene Intuition sprechen, ist es nicht unwahrscheinlich, dass der Patient seine Ärzte-Tour fortsetzt. Schließlich aber wird je nach Dringlichkeit der Beschwerden doch eine Entscheidung zu fällen sein. Ein diesbezügliches Beispiel aus dem Bereich der Zahnmedizin ist der eindrucksvolle Erfahrungsbericht von *Türp* et al. [5] über die Begründungen von Behandlungsvorschlägen für die Versorgung eines frakturierten wurzelkanalgefüllten Zahnes 12.

Der Grund, warum ich hier diese vielleicht nicht ganz alltägliche, aber durchaus bekannte Situation schildere, ist folgender: Das Ziel des Patienten ist eine erfolgreiche Behandlung. Die gewählte therapeutische Entscheidung hängt aber, und das ist vielen Menschen nicht bewusst, vom Profil der gewählten Tour ab, also von der Anzahl der konsultierten Ärzte und deren menschlichen sowie fachlichen Qualitäten. Ja bereits die Entscheidung, sich auf Ärzte-Tour zu begeben, kann Einfluss auf die Diagnose und Therapie nehmen. Es könnte sogar sein, dass eine Strategie, die grundsätzlich dem ersten Arzt vertraut, zu vergleichbaren oder sogar besseren Behandlungserfolgen führt. Ein Grund dafür könnte allein der Verzug sein, der durch eine verzögerte Entscheidung entsteht.

MARQUISE: Wir machen doch diesen Morgen eine Tour, Nichtchen?
NICHTE: Wie es Ihnen gefällt.

Johann Wolfgang von Goethe: Der Großkophta. Ein Lustspiel in fünf Aufzügen. Erstdruck in: Neue Schriften. Unger, Berlin 1792, S. 41

Im Folgenden werde ich die Risiken einer ähnlichen Tour von Statistik(er) zu Statistik(er) beleuchten. Ein Doktorand der Zahnmedizin befindet sich nämlich in einer ähnlichen Situation wie der eingangs erwähnte Patient, wenn er bei der Datenanalyse, eventuell durch Selbsthilfe, eines P-Wertes ansichtig wurde, der angenehme, da leicht publizierbare Schlussfolgerungen verbietet. Tatsächlich wird sich der Doktorand auch nur dann auf den Weg machen, um einen Statistiker zu konsultieren, wenn das selbst erreichte Ergebnis unbefriedigend war. Ebenso unbefriedigend kann die Analyse des erstbesten Statistikers ausfallen, und auf einer nachfolgenden Tour von

Statistiker zu Statistiker kann es vergleichbar stark abweichende Vorschläge geben wie in der medizinischen Diagnostik. Hier jedoch lässt sich auf einfache Weise zeigen (siehe Abschnitt 3), dass man eine solche Tour geschickt planen muss, um die langfristige Qualität der Entscheidung nicht zu gefährden [2, 3].

2 • Multiples statistisches Testen

Der Alltag des Statistikers zu dieser Thematik wird sehr schön in folgendem Zitat von Eckard Sonnemann dargestellt: *„Selten findet der Statistiker die glückliche Lage, daß ein Versuch geplant und durchgeführt und ausgewertet wurde mit dem ursprünglichen Ziel und dem endgültigen Ergebnis, daß eine Frage beantwortet wurde. Zugegeben, manchmal kommt es vor, daß keine Frage gestellt ist oder beantwortet werden kann, doch sollen diese Fälle hier nicht behandelt werden, weil ihr Glücksgewinn eher negativer Ausprägung ist. Mit dem Regelfall wollen wir uns beschäftigen, und das ist der Fall, in dem die Ergebnisse eines Versuchs dazu dienen sollen, mehrere Fragen – mindestens zwei, meistens viel mehr – zu beantworten. Immer dann wollen wir von einem multiplen Testproblem sprechen.“* [4]

Illustrieren werde ich nun verschiedene Situationen des multiplen Testens anhand eines typischen Beispiels aus dem Alltag meiner Arbeit für die Abteilung für Zahnärztliche Prothetik in der Klinik für Zahn-, Mund- und Kieferheilkunde am Universitätsklinikum Freiburg. Zur Bewährungsprobe neuer Materialien oder neuer Fertigungstechniken gehört bekanntermaßen eine Bruchfestigkeitsprüfung. Im Labor werden dafür beispielsweise Kronen, Brücken oder Implantate bis zum Bruch belastet. Die Widerstandsfähigkeit hängt von der Position des Zahnersatzes im Mund ab und sollte im Schneidezahnbereich mindestens 300 Newton, im Seitenzahnbereich mindestens 500 Newton betragen. Für einen solchen Bruchtest lässt sich z. B. eine Zwickmaschine verwenden [1]. In unserem Beispiel ist es ein Ziel, Platin-Implantate durch Implantate aus Keramik zu ersetzen. Die Anzahl der Prüfkörper pro Gruppe wird in der Regel durch die finanziellen Möglichkeiten beschränkt. Multiples statistisches Testen wird z. B. dann durchgeführt, wenn drei verschiedene Keramiken sowohl untereinander also auch jeweils mit einer Platin-Kontroll-Gruppe verglichen werden sollen. Das wären in diesem Fall bereits sechs statistische Tests mit zugehörigen Entscheidungen.

Neben der multiplen Fragestellung fallen mir spontan noch zwei weitere Gründe für mehrfaches statistisches Testen ein: sequentielle Verfahren und eine Tour von Statistiker zu

	Situation 1	Situation 2
Unterschied	Ja	Nein
Mittel K1	800 Newton	800 Newton
Mittel K2	1000 Newton	800 Newton
	Anzahl signifikanter Ergebnisse bei 1000 Simulationen	
t-Test (D1)	158	49
Wilcoxon-Test (D2)	150	51
Tour (D3)	178	57

Tabelle 1 Simulation verschiedener Strategien zum Vergleich von zwei Gruppen mit jeweils acht normalverteilten Bruchfestigkeitswerten. Die Standardabweichung ist jeweils 400 Newton.

Statistiker (siehe Abschnitt 3). In beiden Fällen wird nur eine Fragestellung, z. B. der Vergleich von nur zwei Keramikgruppen, aufgrund von mehreren statistischen Testentscheidungen beantwortet. Eine sequentielle Strategie dient in erster Linie zur Schonung der finanziellen Mittel.

Am gewählten Beispiel werden zunächst nur acht Prüfkörper pro Gruppe präpariert. Der Versuch endet, wenn die Bruchfestigkeit der beiden Gruppen bei [n = 8] bereits statistisch signifikant verschieden ist. Andernfalls werden weitere acht Prüfkörper pro Gruppe angefertigt; in diesem Fall kann der statistische Test mit erhöhter statistischer Power (n = 16) durchgeführt werden. Gegebenenfalls würde die Entscheidung dann auf den Ergebnissen von zwei oder mehr statistischen Tests beruhen.

Die Problematik des multiplen Testens ist folgende: Ein geeigneter statistischer Test erkennt Unterschiede zwischen zwei Gruppen fälschlicherweise, also wenn es in Wahrheit keinen Unterschied gibt, nur in 5 % der Fälle [siehe 2]. Betrachtet man aber zwei solche Entscheidungen als zusammengehörig, dann wird, falls in beiden Fällen kein Gruppenunterschied besteht, erwartungsgemäß in zweimal 5 %, also in 10 % der Fälle, die Gesamtentscheidung mindestens einen Fehlschluss enthalten. Bei 20 zusammengehörigen Entscheidungen, bei denen es in Wahrheit keinen Gruppenunterschied gibt, wird sogar relativ sicher – aber nur durch reinen Zufall – mindestens ein statistisch signifikantes Ergebnis herauskommen.

3 • Eine Simulationsstudie

Mithilfe eines Computers kann ich die verschiedenen Strategien auf der Suche nach einem signifikanten Unterschied zwischen zwei Keramikgruppen simulieren. Nehmen wir an, Statistiker S1 wertet die Bruchfestigkeit immer mit dem t-Test aus, wohingegen Statistiker S2 immer den Wilcoxon-Rangsummentest nimmt. Der t-Test beruht auf dem Mittelwert der Bruchfestigkeitswerte, während der Wilcoxon-Rangsummentest die Bruchfestigkeitswerte nach der Größe sortiert und mit den Rangzahlen arbeitet. Beide statistischen Methoden sind hier gut vertretbar, zumindest in der vom Computer simulierten Situation. Ich möchte folgende drei Strategien vergleichen: Doktorand D1 geht nur zu Statistiker S1, Doktorand D2 geht nur zu Statistiker S2, Doktorand D3 begibt sich auf Tour: Er geht erst zu Statistiker S1 und, falls das Ergebnis nicht signifikant ist, vertraut er der Entscheidung von Statistiker S2.

Mit dem Computer simuliere ich also jeweils acht Bruchfestigkeitswerte pro Keramikgruppe mithilfe von zwei Gaußschen Glockenkurven (siehe Abbildung auf Zehnmark-Schein). In Keramikgruppe K1 liegt der höchste Punkt der Glocke immer bei 800 Newton. Konkret simuliere ich zwei verschiedene Situationen:

(1) Es gibt einen Unterschied: Der Gipfel der Glocke von Keramikgruppe K2 liegt bei 1000 Newton.
(2) Es gibt keinen Unterschied: Der Gipfel der Glocke von Keramikgruppe K2 liegt auch bei 800 Newton.

In beiden Situationen wird eine realistische Breite der „Glocken“ der zwei Gruppen durch eine Standardabweichung von 400 Newton simuliert. Das Ergebnis von 1000 Durchläufen pro Situation ist in Tabelle 1 abgebildet.

Die Ergebnisse in Tabelle 1 bei Situation 1 zeigen die Power der verschiedenen Strategien; z. B. konnte der t-Test in 158 von 1000 Simulationen den Unterschied entdecken. (Die Power beschreibt die Fähigkeit eines statistischen Tests, einen vorhandenen Unterschied zu entdecken. Sie ist hier deshalb so niedrig weil der Stichprobenumfang [n = 8] ziemlich klein ist.) Es zeigt sich, dass die meisten Treffer von D3 erzielt werden. Ein Blick auf Situation 2 jedoch zeigt, dass die Anzahl der Fehler der ersten Art, also ein Erkennen auf Unterschied, wenn es in Wirklichkeit keinen gibt, bei D3 bei 5,7 % liegt, d. h. über dem zulässigen Niveau von 5 % = 50/1000. Der Grund dafür ist das Dilemma des multiplen Testens und gleichzeitig Motivation für einen eigenständigen Bereich der Statistik, in dem geschickte Strategien studiert werden, die auch bei multiplen Testentscheidungen noch zuverlässige Entscheidungen treffen [siehe zum Beispiel 3].

4 • Persönliches Schlusswort

Aufgrund eines reizvollen Angebots aus Kopenhagen habe ich die Leitung meines Freiburger Ladens dieses Jahr an einen Kollegen abgeben müssen. An dieser Stelle möchte ich mich bei allen Beteiligten für die gute und erfolgreiche Zusammenarbeit bedanken (vgl <www.pubmed.gov> und Eingabe von „gerds t“). Ich hoffe sehr, in den vergangenen Jahren nicht allzu viele Fehlentscheidung der ersten und zweiten Art verschuldet zu haben. DZZ

Literatur

1. Att W, Kurun S, Gerds T, Strub JR: Fracture resistance of single-tooth implant-supported all-ceramic restorations after exposure to the artificial mouth. J Oral Rehabil 33, 380-386 (2006)
2. Gerds T: Geschäfte machen mit Statistik. Teil 1: Die magischen fünf Prozent. Dtsch Zahnärztl Z 61, 168-169 (2006)
3. Horn R, Vollandt R: Multiple Tests und Auswahlverfahren. Fischer, Stuttgart 1995
4. Sonnemann E: Vorwort. Biometrisches Seminar der Region Österreich-Schweiz der internationalen biometrischen Gesellschaft. Bad Ischl, Österreich 1981
5. Türp JC, Heydecke G, Krastl G, Pontius O, Antes G, Zitzmann N: Restoring the fractured root-canal-treated maxillary lateral incisor: In search of an evidence-based approach. Quintessence Int 38, 179-191 (2007) [Volltext kostenfrei unter URL: <www.cochrane.org/news/articles/tuerp_root_canal_treatmen_quintessence_march071.pdf>]

T. Gerds, Kopenhagen (Dänemark)