Hypothesentests und p-Werte helfen bei der Überprüfung von statistischen Aussagen

Ulrike Held
Horten-Zentrum, UniversitätsSpital, Zürich

In der medizinischen Forschung werden statistische Hypothesentests zur Beantwortung von vielen Fragestellungen verwendet, und fast keine wissenschaftliche Publikation kommt ohne sie aus. Nachdem verschiedene Tests bereits in früheren Artikeln kurz beschrieben wurden, möchten wir in diesem Artikel noch einmal grundlegend die wichtigsten Prinzipien erläutern und auf einen Datensatz zum Geburtsgewicht von Neugeborenen anwenden.

Neben dem Schätzen von Effekten stellt das statistische Testen *das* zentrale Instrument jeder Datenauswertung dar. Beim statistischen Testen möchte man eine bestimmte, z.B. medizinische Fragestellung in der Form eines Ja/Nein-Schemas beantworten und muss zu diesem Zweck eine sog. Nullhypothese formulieren. Man möchte z.B. untersuchen, ob der Raucherstatus der Mutter während der Schwangerschaft einen Einfluss auf das Geburtsgewicht des Kindes hat. Nun formuliert man die Nullhypothese (H_0) als eine Aussage, die man widerlegen möchte, also z.B. «der Raucherstatus der Mutter hat keinen Einfluss auf das Geburtsgewicht des Kindes». Die Gegenhypothese oder Alternative (H_1) entspricht genau der wissenschaftlichen Hypothese, die man untersuchen will, also hier «der Raucherstatus der Mutter in der Schwangerschaft hat einen Einfluss auf das Geburtsgewicht». Das Ergebnis der statistischen Testentscheidung basiert dann auf den vorliegenden Daten.

Im Zusammenhang mit dieser Testentscheidung können nun zwei Arten von Fehlern auftreten, die unterschiedlich bewertet werden: der sog. α-Fehler oder Fehler 1. Art und der β-Fehler, der Fehler 2. Art. Mit dem Fehler 1. Art bezeichnet man die fälschliche Ablehnung der Nullhypothese: In unserem Beispiel würde das also bedeuten, dass man von einem Einfluss des Raucherstatus ausgeht, obwohl in Wirklichkeit keiner vorliegt. Der Fehler 2. Art wird gemacht, wenn in Wirklichkeit ein Einfluss des Raucherstatus vorliegt, aber die Nullhypothese nicht verworfen wird. Der Fehler 1. Art wird als schwerwiegender erachtet, da man z.B. vermeiden möchte, dass die Wirksamkeit eines Medikaments fälschlicherweise postuliert wird, ein Zusammenhang fälschlicherweise demonstriert wird usw. Aus diesem Grund «kontrolliert» man den Fehler 1. Art durch die Festlegung des Testniveaus = Wahrscheinlichkeit für den Fehler 1. Art, falls H_0 zutrifft. Typischerweise setzt man das Niveau auf 5% oder 1%. Zusätzlich möchte man auch einen niedrigen Fehler 2. Art bzw. eine hohe Güte (engl. *Power*) erreichen, was zumeist durch die Grösse des Stichprobenumfangs geschieht. Beliebig klein können die beiden Fehlerwahrscheinlichkeiten nicht gewählt werden, da diese direkt voneinander abhängen, d.h., je geringer der eine Fehler ist, umso grösser wird der andere.

Wir möchten nun exemplarisch anhand eines t-Tests für den zweiseitigen Vergleich von zwei unabhängigen Gruppen das Vorgehen erläutern. Unser Datensatz beinhaltet das Geburtsgewicht von 189 Neugeborenen, welche im Jahr 1986 im Baystate Medical Centre in Springfield, Massachusetts, auf die Welt kamen. Zusätzlich wurden alle Mütter über ihr Rauchverhalten während der Schwangerschaft befragt, und wir wissen, dass 115 der Mütter nicht geraucht haben (61%).

Wir möchten nun untersuchen, ob sich das Geburtsgewicht der Neugeborenen unterscheidet, je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht. So formulieren wir also unsere Nullhypothese (die wir verwerfen möchten): Das Geburtsgewicht μ von Neugeborenen, bei denen die Mutter während der Schwangerschaft geraucht hat, ist (im Mittel) gleich dem von Neugeborenen mit nichtrauchenden Müttern oder

$$H_0: \mu_{Raucher} = \mu_{Nichtraucher}$$

oder auch

$$H_0: \mu_{Raucher} - \mu_{Nichtraucher} = 0$$

Die Alternativhypothese ist nun, dass sich das Geburtsgewicht von Neugeborenen unterscheidet, je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht, also

$$H_1: \mu_{Raucher} \neq \mu_{Nichtraucher}$$

Wir bezeichnen dies als einen zweiseitigen Test, da die Alternative ein tieferes wie auch ein höheres Geburtsgewicht beinhaltet: Der Unterschied kann also in zwei Richtungen gehen.

Diese Hypothese testen wir auf dem Niveau $\alpha = 0{,}05$, d.h., wir möchten mit einer Wahrscheinlichkeit von höchstens 5% die Nullhypothese fälschlicherweise verwerfen. Auf Basis der Daten wird eine sog. Teststatistik berechnet, die anschliessend mit der Verteilung der Teststatistik unter H_0 verglichen wird. Ist die Teststatistik

Ulrike Held

Die Autorin erklärt, dass sie keine Interessenkonflikte im Zusammenhang mit diesem Beitrag hat.

Tabelle 1. Summary-Statistiken der Geburtsgewichte (Gramm) nach Raucherstatus.

Raucherstatus der Mutter	Minimum	1. Quartil	Median	Mittelwert	3. Quartil	Maximum
Raucherin	709	2370	2776	2773	3246	4238
Nichtraucherin	1021	2509	3100	3055	3622	4990

extremer als ein bestimmter kritischer Wert, in den die Wahl von α sowie die Ein- oder Zweiseitigkeit des Testens eingeht, dann kann man die Hypothese verwerfen. Neben der Testentscheidung, nämlich ob die Nullhypothese verworfen werden konnte oder nicht, ist auch der *p-Wert* ein häufig verwendetes Mass. Vergleicht man nun den p-Wert mit dem vorgegebenen Niveau α, dann wird die Nullhypothese verworfen, falls der p-Wert $<\alpha$ ist. Zusätzlich gibt der p-Wert an, wie deutlich die Hypothese verworfen wird: Er kann in dem Sinne als ein Mass für die Evidenz für bzw. gegen die getestete Hypothese dienen.

Wir kommen nun zurück zu unserem Beispieldatensatz und stellen damit in Tabelle 1 Summary-Statistiken über das Geburtsgewicht dar. Daran können wir ablesen, wie sich dieses in den zwei Gruppen (Mutter hat während der Schwangerschaft geraucht ja/nein) vergleicht.

Man erkennt, dass sowohl der Median als auch das 1. und 3. Quartil der Geburtsgewichte von Neugeborenen mit rauchenden Müttern tiefer liegen als in der Vergleichsgruppe. Wenn wir diesen Mittelwertsvergleich nicht nur deskriptiv, sondern auch mit Hilfe eines statistischen Tests anstellen möchten, verwenden wir einen t-Test für unabhängige Stichproben. Die unabhängigen Stichproben sind in unserem Fall gerade die zwei Gruppen von Neugeborenen.

Das Ergebnis des t-Tests ist ein p-Wert von 0,007, welcher kleiner ist als unser vorgegebenes Niveau $\alpha = 0{,}05$ (oder 5%). Somit können wir die Nullhypothese verwerfen und stellen einen Unterschied im Geburtsgewicht der Neugeborenen von rauchenden und nichtrauchenden Müttern fest. Das mittlere Geburtsgewicht in der Gruppe der rauchenden Mütter beträgt 2773,2 Gramm, während es 3055,0 Gramm bei den Neugeborenen nichtrauchender Mütter ist. Die Differenz liegt also bei 281,7 Gramm, und das geschätzte 95%-Konfidenzintervall für diese Differenz beträgt (76,47; 486,96).

In unserem Beispiel sind die beiden Vergleichsgruppen *ungepaart*, unabhängig. Falls die Gruppen abhängig sind, dann ist ein modifizierter t-Test zu verwenden, der diese Abhängigkeit berücksichtigt. Ein Beispiel für eine Situation, in der ein *gepaarter t-Test* zu verwenden ist, wäre, wenn man das Geburtsgewicht der Neugeborenen mit ihrem Gewicht nach 6 Monaten vergleicht. Dann sind die beiden Vergleichsgruppen nicht unabhängig, und man ginge von einer geringeren Variabilität zwischen den Gruppen aus als bei Vorliegen der Unabhängigkeit.

Man nennt die Nullhypothese in unserem Beispiel *zweiseitig*, da sie «=» versus «≠» (Alternativhypothese) vergleicht. *Einseitige* Hypothesen kann man auch als «≥» versus «<» (oder umgekehrt) formulieren, in unserem Beispiel könnte man also auch sagen

$$H_0: \mu_{Raucher} \geq \mu_{Nichtraucher}$$

versus

$$H_1: \mu_{Raucher} < \mu_{Nichtraucher}$$

In medizinischen Fachzeitschriften werden einseitige Tests häufig nicht akzeptiert, da möglicherweise die gleichen Daten in einer zweiseitigen Testprozedur zu einem nicht-signifikanten Ergebnis geführt haben und erst nachträglich ein einseitiger Test verwendet wurde, um die Signifikanz doch noch zu erreichen.

Falls der *Stichprobenumfang* gross ist und/oder die Daten annähernd normalverteilt sind, verwendet man am besten sog. parametrische Tests, die auf einer Verteilungsannahme beruhen. Beispiele hierfür sind z.B. der oben beschriebene t-Test oder der Gauss-Test für Mittelwertsvergleiche. Falls die Stichprobenumfänge klein sind und es unklar ist, ob die Daten normalverteilt sind, sind nicht-parametrische Tests von Vorteil, die auf keiner Verteilungsannahme beruhen. Häufige nichtparametrische Testverfahren sind der Wilcoxon-Test (als Rangsummentest für unabhängige Gruppenvergleiche oder als Vorzeichen-Rangtest für abhängige Gruppen) und der Vorzeichentest.

Zuletzt soll hier noch darauf eingegangen werden, dass die statistische Signifikanz und die praktische (klinische) Relevanz nicht dasselbe sind. Führt man einen formal korrekten statistischen Test durch und erhält anschliessend ein signifikantes Ergebnis, dann sollte man immer auch die Relevanz des Ergebnisses für die Anwendung prüfen. Vielleicht findet sich z.B. ein signifikanter Gruppenunterschied in einer grossen Studie mit vielen Patienten, aber tatsächlich ist der beobachtete Unterschied zwischen den Gruppen so klein, dass er in der medizinischen Praxis keine Relevanz hat.

Glossar

Stichprobenumfang

Zur Beantwortung einer wissenschaftlichen Fragestellung ist normalerweise die Untersuchung der gesamten Population nicht durchführbar, und aus diesem Grund ist das Ziehen einer sog. Stichprobe nötig. Diese Stichprobe ist eine Teilmenge der Grundpopulation und sollte nach Möglichkeit zufällig ausgewählt werden. Um die Grösse der Stichprobe, also den Stichprobenumfang, festzulegen, muss eine formale Planung der Fallzahl durchgeführt werden, um die Studienziele zu erreichen.

Ungepaarte/gepaarte Beobachtungen

Möchte man mit einem statistischen Test zwei Gruppen vergleichen, dann muss man überlegen, ob diese zwei Gruppen auf unterschiedlichen oder denselben Indi-

viduen basieren. Im ersteren Fall hat man ungepaarte Beobachtungen vorliegen und im zweiten Fall eben gepaarte Beobachtungen. Für den statistischen Test hat das Konsequenzen, da im zweiten Fall für die (typischerweise) geringere Variabilität in den Daten adjustiert werden muss.

Zweiseitige/einseitige Hypothesen

Zweiseitige Hypothesen sind von der Form H_0: … = … versus H_1: … ≠ … Das bedeutet, dass bei Ablehnen von H_0 der Unterschied in beide Richtungen gehen kann und es deshalb schwieriger ist, ein signifikantes Ergebnis zu erreichen. Es wird nämlich jeweils die Hälfte des Testniveaus α auf jeder Seite verbraucht. Bei einseitigen Hypothesen testet man H_0: … ≤ … versus H_1: … > … bzw. H_0: … ≥ … versus H_1: … < … und hat dann auf der jeweiligen Seite das ganze Testniveau zur Verfügung. Diese werden jedoch selten verwendet.

Korrespondenz:
Dr. Ulrike Held
Horten-Zentrum
UniversitätsSpital Zürich
Postfach Nord
CH-8091 Zürich
ulrike.held@usz.ch

Empfohlene Literatur

– Held L, Rufibach C, Seifert B. Einführung in die Biostatistik. 4. Auflage. Zürich: Abteilung Biostatistik, Institut für Sozial- und Präventivmedizin der Universität Zürich; Juli 2009. http://www.biostat.uzh.ch.
– Hüsler J, Zimmermann H. Statistische Prinzipien für medizinische Projekte. 4. Auflage. Bern: Huber-Verlag; 2006.
– Kreienbrock L, Schach S. Epidemiologische Methoden. 4. Auflage. München: Elsevier-Verlag; 2005.
– R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. ISBN 3–900051–07–0. URL http://www.R-project.org.