Einige Beobachtungen zur Definition von P-Werten

von *John E. Freund*, Arizona State Univ. und *Benjamin M. Perles*, Suffolk Univ.; Übersetzung: *Manfred Borovcnik*, Klagenfurt

Kurzfassung: Viele Bücher definieren P-Werte auf verschiedene Weisen, und es wird allgemein hin angenommen, daß diese Definitionen äquivalent sind. Daß dies nicht der Fall ist, mag überraschen, aber wir werden hier zeigen, daß einige Definitionen nicht einmal mit solch einer elementaren Prozedur wie der Spezifikation einer kritischen Region im Einklang steht.

1. Einleitung

Obwohl das Konzept eines P-Werts nicht neu ist, haben viele Statistiklehrbücher dessen Diskussion erst in jüngsten Jahren eingeführt und erweitert. P-Werte sind unter anderen Namen bekannt gewesen - prob-Werte [prob values], Schwanzwahrscheinlichkeiten, oder einfach P's. [In der deutschsprachigen Literatur sind P-Werte weniger in Gebrauch gewesen, daher kaum geläufige Fachtermini.] R.A. Fisher zum Beispiel bezeichnete sie als P's, und in Fisher (1973), erstmals 1925 veröffentlicht, stellt er fest, daß für $\chi^2 = 10.48$ bei 4 Freiheitsgraden P zwischen 0.02 und 0.05 fällt. Klar, er hat das Ergebnis mit einer Chi-Quadrat-Tabelle bekommen, und das war so präzise, wie er es ohne längliche Berechnungen bekommen konnte. Heute können uns eine Fülle an statistischer Software, ja sogar einige Taschenrechner diesen P-Wert innerhalb von Sekunden mit 0.0331 ausrechnen.

Genau dadurch ist der Gebrauch von P-Werten bei Signifikanztests so in der Beliebtheit gestiegen; tatsächlich bevorzugen viele Statistiker P-Werte gegenüber der traditionellen Methode des Vergleichs von Teststatistiken mittels kritischer Werte. Schließlich sind wir durch die P-Werte nicht länger auf die üblichen 0.01- und 0.05 -Niveaus für die Signifikanz beschränkt. Wenn wir etwa $\alpha=0.035$ wollen, so könnten wir einfach den P-Wert, der zu einem beobachteten Wert der Teststatistik gehört, mit diesem Wert von α vergleichen.

Der Nachteil dieses Ansatzes besteht darin, daß das Konzept eines P-Werts nicht immer klar verstanden wird. Unterschiede in der Definition sind zahlreich, und in vielen Fällen geht nicht klar hervor, ob die Definitionen wirklich mit anderen Aspekten statistischer Methodenlehre vereinbar sind. Wie kann es denn sein, daß ein Lehrbuch uns erzählt, daß die Nullhypothese zu verwerfen ist, wenn der [beobachtete] P-Wert *kleiner oder gleich* α (dem vorgegebenem Signifikanzniveau) ist, während ein anderes Buch aussagt, daß sie nur zu verwerfen ist, wenn der P-Wert

kleiner als α ist? Wir werden, vielleicht zu unserer Überraschung, sehen, daß dies ganz eng mit der Frage zusammen hängt, ob die Grenze eines Ablehnungsbereichs noch dazu gehört oder nicht. Zum Beispiel, wenn wir die Nullhypothese $\mu=\mu_0$ für den Mittelwert einer Normalverteilung gegen die Alternativhypothese $\mu>\mu_0$ auf dem Niveau α testen, soll der Ablehnungsbereich $z>z_\alpha$ oder $z\geq z_\alpha$ sein? Hier ist z_α durch $W(Z\geq z_\alpha) = \alpha$ bestimmt.

Man möchte vielleicht spontan sagen, 'Was soll's?', weil das praktisch nicht relevant sei. Aber wir *müssen* uns darum kümmern. Grenzwerte treten auf und sie müssen korrekt und konsequent behandelt werden. Für den Theoretiker bleibt darüber hinaus immer die Frage nach der logischen Konsistenz.

2. Definitionen und Kommentar

Zur Untersuchung des Zusammenhangs zwischen Ablehnungsbereichen und der Definition von P-Werten sei der Ablehnungsbereich $z\$z_\alpha$ mit A und der Ablehnungsbereich $z>z_\alpha$ mit B bezeichnet. Die folgende Definition für P-Werte ist eine der häufigsten:

***

(a) Der P-Wert ist die Wahrscheinlichkeit, einen Wert der Teststatistik zu bekommen, der gleich extrem oder extremer ist als das beobachtete Ergebnis, die Wahrscheinlichkeiten berechnet unter der Annahme, daß die Nullhypothese zutrifft.

***

In unserem Beispiel gilt, P-Wert = W(Z$z), wobei z der beobachtete Wert der Teststatistik Z ist. Beachten Sie, daß diese Definition sowohl auf den Ablehnungsbereich A als auch auf B angewendet werden kann. Denn für A würden wir die Nullhypothese für alle jene P-Werte ablehnen, die kleiner oder gleich dem festgesetzten Signifikanzniveau α sind, für B hingegen würden wir nur für P-Werte kleiner als α ablehnen. Für den Ablehnungsbereich B wird nämlich die Nullhypothese nicht abgelehnt, wenn $z=z_\alpha$.

In beiden Fällen gibt es keine Probleme mit der Definition, aber viele Statistiker ziehen alternative Definitionen vor, um die merkwürdige Wendung 'extremer' zu vermeiden. Wie sie hier verwendet wird, bedeutet sie 'größer als', 'kleiner als' oder 'größer oder gleich', abhängig von der Alternativhypothese. Die folgende Definition ist sehr beliebt und vermeidet diese Schwierigkeit in der Wortwahl.

***

(b) Der P-Wert ist das kleinste Signifikanzniveau, bei welchem die Nullhypothese bei den gegebenen Daten abgelehnt werden könnte.

***

Im Zusammenhang mit Ablehnungsbereich A gilt: Wenn der beobachtete Wert der Teststatistik z ist und W(Z$z) = q gilt, so könnte die Nullhypothese für jedes beliebige Signifikanzniveau größer oder gleich q abgelehnt werden. Daher ist der P-Wert gleich q und wir würden die Nullhypothese ablehnen, wenn q kleiner oder gleich dem festgesetzten Signifikanzniveau ist.

Wenn wir aber den Ablehnungsbereich B betrachten, so kommen wir in Schwierigkeiten. Hier würden wir die Nullhypothese für jedes Signifikanzniveau größer als q ablehnen, und es gibt schlicht *keinen kleinsten Wert größer als q*. Das bedeutet, daß die alternative Definition von P-Werten auf den Ablehnungsbereich B nicht angewendet werden kann. Das zeigt auch, wie vorsichtig wir sein müssen, wenn wir einen neuen Begriff (hier P-Werte) definieren, damit wir nicht in Konflikt mit der vorausgehenden Terminologie kommen. Es ist dem Leser überlassen, unter welchen Bedingungen (für welche Ablehnungsbereiche) die folgenden Definitionen des P-Werts im obigen Beispiel genutzt werden können:

***

(c) Der P-Wert ist die größte untere Schranke für die Signifikanzniveaus, auf denen die Nullhypothese bei gegebenen Daten abgelehnt werden könnte.

***

(d) Der P-Wert ist das höchste Signifikanzniveau, auf welchem die Nullhypothese bei gegebenen Daten abgelehnt werden könnte.

***

Zusammenfassend, gibt keine allgemeine Übereinstimmung über die Definition von P-Werten und es gibt nicht viel, was wir dazu beitragen können. Wie wir gesehen haben, hängt jede solche Definition von anderen Aspekten statistischer Methodologie (oder Terminologie) ab, über die ebenso keine Übereinstimmung herrscht.

Literatur:

Fisher, R.A. (1973): *Statistical Methods for Research Workers*, 14. Auflage, New York: Hafner Press.