Was ist der *p*-Wert?

– Artikel Nr. 7 der Statistik-Serie in der DMW –

What is the *p*-value?

Autoren R. Bender [1] S. Lange [1]

Institut [1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln

Der *p*-Wert ist das Ergebnis eines statistischen **Signifikanztests** [5]. Mit Hilfe eines Signifikanztests kann man a priori formulierte Hypothesen überprüfen. Die „Nullhypothese“ (H_0) ist zumeist die Formulierung der Gleichheit (kein Effekt), die „Alternativhypothese“ (H_1) die Formulierung eines Unterschieds (Effekts) bezüglich einer interessierenden Fragestellung. Man kann die Hypothesen zweiseitig (Gleichheit vs. Unterschied) oder einseitig (Gleichheit vs. positiver Effekt bzw. Gleichheit vs. negativer Effekt) formulieren. In der Regel werden zweiseitige Hypothesenformulierungen verwendet [3].

Aus einer **Zufallsstichprobe** wird das für die Fragestellung relevante Effektmaß (zum Beispiel Mittelwert, Median, Differenz zweier Mittelwerte, Regressionskoeffizient, Differenz zweier Wahrscheinlichkeiten, Risk Ratio, Odds Ratio, etc.) geschätzt. Durch eine geeignete Normierung dieses Effektmaßes erhält man eine **Teststatistik**. Beim ***t***-Test beispielsweise wird das Effektmaß (Differenz zweier Mittelwerte) auf den Standardfehler dieser Differenz normiert. Der *p*-Wert ist die Wahrscheinlichkeit, dass unter der Annahme, die Nullhypothese sei wahr, die Teststatistik den beobachteten oder einen extremeren Wert annimmt. Mit anderen (mathematisch nicht ganz exakten) Worten: Der *p*-Wert ist die Wahrscheinlichkeit dafür, dass sich die Daten wie beobachtet (oder extremer) realisieren, falls in Wirklichkeit die Nullhypothese zutrifft. Wenn diese Wahrscheinlichkeit klein ist, so spricht dieses Ergebnis gegen die Nullhypothese und es ist Evidenz für die Richtigkeit der Alternativhypothese vorhanden.

Vor der Datenerhebung wird eine maximale Irrtumswahrscheinlichkeit festgelegt (**Signifikanzniveau** α), die den Fehler 1. Art, nämlich die Nullhypothese abzulehnen, obwohl sie richtig ist, begrenzt. Häufig gewählte Niveaus sind $\alpha = 0{,}05$ und $\alpha = 0{,}01$. Ist der *p*-Wert kleiner als das festgelegte Signifkanzniveau, so liegt statistische Signifikanz zum Niveau α vor.

Der *p*-Wert ist nicht die Wahrscheinlichkeit für die Richtigkeit der Nullhypothese (häufigste Fehlinterpretation). Diesem Ereignis lässt sich (aus der Sicht der „klassischen Statistik“) gar keine Wahrscheinlichkeit zuordnen [7]. Ob ein Effekt da ist oder nicht, ist zwar unbekannt, aber fix, und ist nicht das Resultat eines Zufallsexperiments.

Schlüsselwörter
- *p*-Wert
- Statistische Signifikanz
- Signifikanzniveau
- Konfidenzintervall

Key words
- *p*-value
- Statistical significance
- Significance level
- Confidence interval

Tab. 1 Ergebnisse von t -Tests auf Unterschied (Alternativhypothese) zwischen zwei Gruppen für hypothetische Daten (systolischer Blutdruck in mm Hg) mit variierender Differenz, Stichprobengröße und Variabilität.

Test	Stichprobenumfänge	Medikament Mittelwert (SD)	Placebo Mittelwert (SD)	Differenz der Mittelwerte	*p* -Wert	Signifikanz bei $\alpha = 0{,}05$
1	$n_1 = n_2 = 10$	160 (22)	180 (22)	20	0,057	n. s.
2	$n_1 = n_2 = 10$	160 (15)	180 (15)	20	0,008	s.
3	$n_1 = n_2 = 20$	160 (22)	180 (22)	20	0,007	s.
4	$n_1 = n_2 = 50$	170 (22)	180 (22)	10	0,025	s.
5	$n_1 = n_2 = 5$	140 (50)	180 (50)	40	0,242	n. s.
6	$n_1 = n_2 = 1000$	178 (12)	180 (12)	2	< 0,001	s.

SD = Standardabweichung, n. s. = nicht signifikant, s. = signifikant

Bibliografie
DOI 10.1055/s-2007-959030
Dtsch Med Wochenschr 2007; 132: e15–e16 · © Georg Thieme Verlag KG Stuttgart · New York · ISSN 0012-0472

Korrespondenz
Privatdozent Dr. rer. biol. hum. Ralf Bender
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln
eMail Ralf.Bender@iqwig.de

Statistische Signifikanz bedeutet nicht unbedingt auch praktische Relevanz [5]. Die Höhe des *p*-Werts hängt nicht nur von der Stärke des Effekts, sondern auch von der Variabilität des geschätzten Effektmaßes ab und diese wiederum von der Variabilität der Daten und der Größe der Stichprobe. Vor allem bei großen Stichproben kann man daher sehr kleine *p*-Werte erhalten (und damit statistische Signifikanz), obwohl der Effekt gering und möglicherweise sogar unbedeutend ist. In **Tab. 1** findet man die Resultate von sechs ***t***-Tests auf Unterschied (Alternativhypothese) zwischen zwei Gruppen für hypothetische Daten (systolischer Blutdruck in mm Hg). Der *p*-Wert und damit auch die Signifikanz-Entscheidung hängt von der Differenz der Mittelwerte, der Standardabweichung (SD) und den Stichprobenumfängen (n_1, n_2) ab. Ein kleiner *p*-Wert sagt aus, dass es statistische Evidenz für einen Unterschied (irgendeiner Stärke) gibt. Wie groß dieser Effekt ist, kann man am *p*-Wert nicht ablesen. Für diesen Zweck muss man die Größe des geschätzten Effektmaßes interpretieren, am besten im Zusammenhang mit einem **Konfidenzintervall** [2].

Mit Hilfe eines Signifikanztests lässt sich ein beobachtetes Ergebnis statistisch mit einem (geringen) Irrtumsvorbehalt vom Zufall abgrenzen. Ein nicht-signifikantes Ergebnis bedeutet jedoch nicht, dass man nachgewiesen hat, dass kein Unterschied da ist. Für diesen Zweck benötigt man **Äquivalenztests** [1, 6].

kurzgefasst

Der sogenannte „*p*-Wert" ist das Ergebnis eines Signifikanztests zur Prüfung einer vorab aufgestellten (Null-)Hypothese. Ist der *p*-Wert kleiner als das, ebenfalls vorab, gewählte Irrtums-(Signifikanz-)Niveau α, dann gilt das Ergebnis als statistisch signifikant. Statistische Signifikanz ist nicht gleichbedeutend mit klinischer Relevanz.

Die englischen Bezeichnungen der hier diskutierten Begriffe zeigt **Tab. 2**.

Tab. 2 Übersetzungen (deutsch – englisch)

p-Wert	*p*-value
Signifikanztest	significance test
Null- (Alternativ-) hypothese	null (alternative) hypothesis
Zufallsstichprobe	random sample
Teststatistik	test statistic
Signifikanzniveau	significance level
Konfidenzintervall	confidence interval

Dieser Beitrag ist eine überarbeitete Fassung aus dem Supplement Statistik aus dem Jahr 2001.

Literatur

1 *Altman DG, Bland JM.* Absence of evidence is not evidence of absence. BMJ 1995; 311: 485
2 *Bender R, Lange S.* Was ist ein Konfidenzintervall? Dtsch Med Wochenschr 2007; 132: e17–e18
3 *Altman DG, Bland JM.* One and two sided tests of significance. BMJ 1994; 309: 248
4 *Guyatt GH, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S.* Basic statistics for clinicians: 1. Hypothesis testing. CMAJ 1995; 152: 27–32
5 *Lange S.* Statistische Signifikanz und klinische Relevanz. Z Hautkr 2000; 75: 225–229
6 *Lange S, Bender R, Ziegler A.* Äquivalenzstudien und Nicht-Unterlegenheitsstudien. Dtsch Med Wochenschr 2007; 132: e53–e56
7 *Pollard P, Richardson JTE.* On the probability of making type I errors. Psychol Bull 1987; 102: 159–163