Der p-Wert: Standardisierte Zufallsvariable, Überschreitungswahrscheinlichkeit oder Grenzniveau des Ablehnens?

FRANK MAROHN, WÜRZBURG

Zusammenfassung: *In der vorliegenden Arbeit soll einmal mehr auf den p-Wert eingegangen werden. Dieser Wert, der von statistischen Software-Paketen, sprich vom Computer, berechnet wird und auf den sich die statistische Anwenderwelt „stürzt", dient als Entscheidungsgrundlage beim Testen von Hypothesen. Gerade weil der Computer in den Schulen Einzug gehalten hat, ist es für den Lehrenden wichtig zu wissen, was der p-Wert (nicht) ist.*

1 Einleitung

Ein statistischer Test ist eine Entscheidungsregel, die festlegt, ob man sich aufgrund der vorliegenden Daten für die Nullhypothese H_0 oder sich gegen H_0 und somit für die Alternative H_1 entscheidet.

Üblicherweise gibt man sich nach der sogenannten Neyman-Pearson-Methode ein Signifikanzniveau (Testniveau) α vor, das die Wahrscheinlichkeit für einen Fehler 1. Art (fälschliche Ablehnung von H_0) kontrolliert. Im Gegensatz zur Vorgehensweise, einen Höchstwert α für die Wahrscheinlichkeit des Fehlers 1. Art festzulegen und daraufhin den kritischen Bereich zu wählen, ist es insbesondere bei der Verwendung von statistischer Software gängige Praxis, aus einer Stichprobe einen sogenannten p-Wert auszurechnen und die Signifikanz des erhaltenen Resultates anhand dieses Wertes zu beurteilen. Ein kleiner p-Wert (üblicherweise ≤ 0.05) führt zu einer Ablehnung von H_0.

Was ist der p-Wert? Ist diese Zahl eine Realisierung einer Zufallsvariablen? Ist sie eine Wahrscheinlichkeit? Oder ein Signifikanzniveau? Auf diese Fragen soll in diesem Artikel näher eingegangen werden. In der Literatur finden sich folgende drei Definitionen:

(1) *Der p-Wert als Zufallsvariable:* „Der p-Wert ist keine Wahrscheinlichkeit" (Stahel 2008, S. 209). Der p-Wert ist eine Zufallsvariable bzw. der konkrete p-Wert (also die Zahl, die vom Computer berechnet wird) ist eine Realisierung dieser Zufallsvariablen (Stahel 2008, Abschnitt 8.3; Falk et al. 2004, Abschnitt 2.3).

(2) *Der p-Wert als Wahrscheinlichkeit:* Der p-Wert ist die Wahrscheinlichkeit, unter H_0 den beobachteten Prüfgrößenwert oder einen in Richtung der Alternative extremeren Wert zu erhalten (Freund & Perles 1996; Stahel 2008, Abschnitt 8.7; Rudolf & Kuhlisch 2008, Abschnitt 5.3.1).

(3) *Der p-Wert als Grenzniveau:* Der p-Wert zu einer Beobachtung ist das kleinste Signifikanzniveau, zu dem der Test H_0 verwirft (Falk et al. 2014, Abschnitt 2.7.1; Henze 2013, Abschnitt 29.4; Krengel 2005, Abschnitt 6.10; Stahel 2008, Abschnitt 8.3).

Auf diese drei Definitionen des p-Wertes soll im Folgenden genauer eingegangen werden. In Abschnitt 2 wird zunächst der Fall einer stetig verteilten Prüfgröße betrachtet. Als Referenz dient der Einstichproben-Gauß-Test. Der Fall einer diskret verteilten Prüfgröße wird in Abschnitt 3 behandelt. Als Referenz dient der Binomialtest. Bezüglich des Gauß-Tests und des Binomialtests sei auf Georgii 2009 und Henze 2013 verwiesen. Aspekte rund um den p-Wert werden in Abschnitt 4 diskutiert.

2 Der p-Wert im Stetigen

Wir betrachten den Fall einer stetig verteilten Prüfgröße. Man nennt eine Zufallsvariable X stetig verteilt, falls eine Funktion $f \geq 0$ existiert (die sogenannte Wahrscheinlichkeitsdichte) mit $P(a \leq X \leq b) = \int_a^b f(x)\,dx$, $a < b$. Die Verteilungsfunktion $F(x) = \int_{-\infty}^{x} f(t)\,dt$ ist dann stetig. Exemplarisch betrachten wir eine normalverteilte Prüfgröße.

2.1 Der p-Wert als Zufallsvariable

In diesem Abschnitt werden wir sehen, dass der p-Wert eine Art „vollstandardisierte" Prüfgröße ist, die auf $(0,1)$ gleichverteilt ist, vorausgesetzt, die Verteilungsfunktion der Prüfgröße ist stetig.

Eine Zufallsvariable U heißt auf $(0,1)$ *gleichverteilt* (uniformly distributed), falls U die Dichte $f(x) = 1$, $x \in (0,1)$ und $f(x) = 0$, $x \notin (0,1)$, besitzt. Für ein Intervall $(a,b) \subset (0,1)$ gilt somit

$$P(a \leq U \leq b) = \int_a^b f(x)\,dx = b - a.$$

Speziell gilt $P(U \leq 0.05) = 0.05$. Die Wahrscheinlichkeit, dass eine auf $(0,1)$ gleichverteilte Zufallsvariable einen Wert im Intervall $(0, 0.05)$ anneh-

men wird, beträgt also 0.05. Die folgende allgemeine Aussage ist für den p-Wert wichtig.

Transformation in die Gleichverteilung: Besitzt eine Zufallsvariable X eine stetige Verteilungsfunktion F, so ist $F(X)$ auf (0,1) gleichverteilt.

Der Beweis dieser Aussage, bei der man nicht auf die Stetigkeit von F verzichten kann, ist einfach, falls F streng monton wachsend ist. Denn in diesem Fall existiert die Umkehrfunktion F^{-1}. Im allgemeinen Fall hat man die sogenannte Quantiltransformation zu betrachten, die bei der Erzeugung von (Pseudo)-Zufallszahlen eine grundlegende Rolle spielt. Für Details und die Beweise sei auf Georgii 2009, Kapitel 1 und auf Henze 2013, Abschnitt 31.14, verwiesen.

Im Folgenden wollen wir den Einstichproben-Gauß-Test betrachten, gehen also von einem Normalverteilungsmodell $N(\mu,\sigma^2)$ mit unbekanntem Mittelwert $\mu\in\mathbb{R}$ und *bekannter* Varianz $\sigma^2>0$ aus. Betrachtet wird zunächst das linksseitige Testproblem

$$\text{(L)}\quad H_0:\mu=\mu_0 \quad\text{vs}\quad H_1:\mu<\mu_0.$$

Wir können genauso gut $H_0:\mu\geq\mu_0$ schreiben. Bei der Fehlerwahrscheinlichkeit 1. Art kommt es nur auf den „Randpunkt" μ_0 an, der die Nullhypothese von der Alternative trennt. Einseitigkeit bezieht sich auf die Alternative.

Basierend auf einer unabhängig und identisch verteilten Stichprobe $X_1,\ldots,X_n$, lautet die Prüfgröße

$$Z=\frac{\bar{X}_n-\mu_0}{\sigma/\sqrt{n}}.$$

Dabei bezeichnet wie üblich $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$ das Stichprobenmittel. Unter H_0 besitzt Z eine $N(0,1)$-Verteilung. Bezeichnet

$$\Phi(z)=\int_{-\infty}^{z}\varphi(x)\,dx \text{ mit } \varphi(x):=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

die Verteilungsfunktion der $N(0,1)$-Verteilung.

Der p-Wert zum Testproblem (L) ist gegeben durch die Zufallsvariable

$$p_L(Z):=\Phi(Z)=\int_{-\infty}^{Z}\varphi(x)\,dx$$

(Verknüpfung von Φ und Z), welche auf $(0,1)$ gleichverteilt ist. Speziell gilt $P_{\mu_0}(p_L(Z)\leq 0.05)=0.05$. Die Indizierung von P mit μ_0 bedeutet, dass bei der Berechnung der Wahrscheinlichkeit der Parameter μ_0 unterstellt wird.

Ist z der konkret beobachtete Prüfgrößenwert (Realisierung von Z), basierend auf Daten $x_1,\ldots,x_n$ (Realisierungen von $X_1,\ldots,X_n$), so ist $p_L(z)=\Phi(z)$ eine Realisierung der Zufallsvariablen $\Phi(Z)$. Die Zahl $p_L(z)$, die vom Computer berechnet wird, ist der p-Wert zur Beobachtung z. Da die Werte der Verteilungsfunktion Φ tabelliert sind, lässt sich im Fall des Gauß-Tests der p-Wert zur Beobachtung z aus den entsprechenden Tabellenwerken für Φ ablesen.

Betrachtet man das rechtsseitige Testproblem

$$\text{(R)}\quad H_0:\mu=\mu_0 \quad\text{vs}\quad H_1:\mu>\mu_0,$$

so ist

$$p_R(Z):=1-\Phi(Z)=\int_{Z}^{\infty}\varphi(x)\,dx$$

der p-Wert und $p_R(z)$ ist der p-Wert zur Beobachtung z. Beachte: Mit $\Phi(Z)$ ist auch $1-\Phi(Z)$ auf $(0,1)$ gleichverteilt, so dass wieder $P_{\mu_0}(p_R(Z)\leq 0.05)=0.05$ gilt.

Im Testproblem (L) sprechen kleine z-Werte und im Testproblem (R) sprechen große z-Werte für die Alternative. Zur Beurteilung, ob ein kleiner oder großer Prüfgrößenwert beobachtet worden ist, braucht es eine Bezugsgröße, da die Prüfgröße Z im Prinzip beliebig kleine oder große Werte annehmen kann. Wird ein Signifikanzniveau α vorgegeben, so bewertet man nach der Neyman-Pearson-Methode den Wert z durch den Vergleich mit einem kritischen Wert, sprich Quantil (wir kommen in Abschnitt 2.3 darauf zurück).

Eine andere Möglichkeit (und dies ist die p-Wert-Methode) besteht darin, die Prüfgröße so zu transformieren, dass sie beschränkt ist. Die Schranken dienen dann als Bezugsgrößen. Durch die Transformation $Z\mapsto\Phi(Z)$ bzw. $Z\mapsto 1-\Phi(Z)$ der Prüfgröße Z auf ihren p-Wert lassen sich auffällig kleine bzw. große Realisierungen von Z *unmittelbar* erkennen. Übliche Konvention: p-Werte kleiner oder gleich 0.05 sprechen gegen die Nullhypothese. Fazit: Der p-Wert ist eine Art „vollstandardisierte" Prüfgröße, die unter H_0 eine uniforme Verteilung besitzt.

Beim beidseitigen (oder zweiseitigen) Testproblem

$$\text{(B)}\quad H_0:\mu=\mu_0 \quad\text{vs}\quad H_1:\mu\neq\mu_0$$

ist der p-Wert zur Beobachtung z definiert durch

$$\begin{aligned} p_B(z) &:= 2\cdot\min\{(p_L(z),p_R(z)\} \\ &= 2\cdot\min\{\Phi(z),1-\Phi(z)\} \end{aligned} \qquad (1)$$

Die Zufallsvariable $p_B(Z)$ ist ebenfalls auf $(0,1)$ gleichverteilt. Aufgrund der Symmetrie der Dichte φ

können wir für den zweiseitigen p-Wert auch schreiben $p_B(z) = 2 \cdot (1 - \Phi(|z|))$.

Bemerkung 1: Auch bei nichtsymmetrischen Verteilungen (z. B. Chi-Quadrat-Verteilung) ist der beidseitige p-Wert definiert als das Doppelte des Minimums der beiden einseitigen p-Werte.

2.2 Der p-Wert als Wahrscheinlichkeit

In diesem Abschnitt wollen wir eine (die einzige?) Möglichkeit kennenlernen, den p-Wert mit dem Begriff einer Wahrscheinlichkeit in Zusammenhang zu bringen. Betrachten wir dazu das Testproblem (R). Häufig liest man dann die folgende (etwas laxe) Definition: Der p-Wert zur Beobachtung z ist die Wahrscheinlichkeit, unter der Nullhypothese einen Prüfgrößenwert zu beobachten, der größer oder gleich z ist. Formal wird in der Literatur für diese *Überschreitungswahrscheinlichkeit* $P_{\mu_0}(Z \geq z)$ geschrieben. Diese Definition des p-Wertes zur Beobachtung z ist gleichbedeutend mit der in Abschnitt 2.1, da Z unter H_0 die Verteilungsfunktion Φ besitzt, d.h. $\Phi(z) = P_{\mu_0}(Z \leq z)$, $z \in \mathbb{R}$:

$$P_{\mu_0}(Z \geq z) = 1 - \Phi(z) = p_{\mathrm{R}}(z).$$

An dieser Stelle sei zur Notation Folgendes gesagt:

1. Wir geben der ebenfalls in der Literatur üblichen Bezeichnungsweise $P(Z \geq z|\mu_0)$ gegenüber $P_{\mu_0}(Z \geq z)$ nicht den Vorzug, da diese Notation eine bedingte Wahrscheinlichkeit suggerieren würde. Hypothesen, also Modelle, haben keine Wahrscheinlichkeiten, sie legen Wahrscheinlichkeiten fest (siehe Bemerkung 3).

2. Unreflektiert scheint die Schreibweise für den p-Wert als Überschreitungswahrscheinlichkeit $P_{\mu_0}(Z \geq z)$ unproblematisch zu sein. Aber genauer betrachtet ergibt sich hierbei die folgende Schwierigkeit: Die Beobachtung z ist eine Realisierung von Z. Die Daten $x_1, \ldots, x_n$ (Realisierungen von $X_1, \ldots, X_n$), die zum Prüfgrößenwert z führen, liegen vor. Das Zufallsexperiment ist also *bereits durchgeführt* worden (Vergangenheit!). Daher macht der Ausdruck $P_{\mu_0}(Z \geq z)$ wenig Sinn, denn Wahrscheinlichkeitsaussagen können sich nur auf künftige (oder nicht bekannte) Ereignisse beziehen. Wohin die laxe Schreibweise $P_{\mu_0}(Z \geq z)$ führt wird deutlich, wenn wir versuchen, den p-Wert wie in Abschnitt 2.1 als Zufallsvariable zu schreiben: $1 - \Phi(Z) = P_{\mu_0}(Z \geq Z) = 1$. Unproblematisch ist dagegen die Schreibweise

$$p_{\mathrm{R}}(z) = P_{\mu_0}(\tilde{Z} \geq z).$$

Dabei besitzt $\tilde{Z}$ unter P_{μ_0} die gleiche Verteilung wie die Prüfgröße Z und $\tilde{Z}, Z$ sind stochastisch unabhängig. Mit anderen Worten: Wenn man das gleiche Zufallsexperiment noch einmal unabhängig wiederholen würde, so ist der p-Wert die Wahrscheinlichkeit, dass unter der Nullhypothese die Prüfgröße $\tilde{Z}$ einen Wert annehmen wird, der größer oder gleich z ist. Jetzt können wir auch den p-Wert als Zufallsvariable schreiben:

$$p_{\mathrm{R}}(Z) = 1 - \Phi(Z) = P_{\mu_0}\big(\tilde{Z} \geq Z\big). \qquad (2)$$

Bemerkung 2: Man kann den p-Wert sehen als bedingte Wahrscheinlichkeit (Übergangswahrscheinlichkeit) im Rahmen eines zweistufigen Zufallsexperimentes. Es gilt dann

$$P_{\mu_0}\big(\tilde{Z} \geq Z | Z = z\big) = P_{\mu_0}\big(\tilde{Z} \geq z\big) = p_{\mathrm{R}}(z). \qquad (3)$$

Was sich so einfach liest und einleuchtend erscheint, ist alles andere als offensichtlich. Erstens: Es handelt sich hier um ein allgemeines Konzept von bedingten Verteilungen und nicht um den Begriff einer elementaren bedingten Wahrscheinlichkeit, wie man ihn aus der Schule kennt (beachte: $P_{\mu_0}\{Z = z\} = 0$). Zweitens: Die Gültigkeit des ersten Gleichheitszeichens in (3) ist nicht selbstverständlich. Diese „Einsetzungsregel", so intuitiv klar sie auch sein mag, bedarf eines Beweises (siehe z. B. Wengenroth 2008, Satz 6.4).

Halten wir fest: Der rechtsseitige p-Wert kann als Überschreitungswahrscheinlichkeit interpretiert werden, wobei H_0 zugrunde liegt. Eine schlimme Fehlinterpretation ist die Folgende: Ein p-Wert von 0.042 besagt, dass „die Nullhypothese die Wahrscheinlichkeit 0.042 hat". Eine solche Aussage ist unsinnig. Modelle (und Parameterwerte) selber haben keine Wahrscheinlichkeiten, sie legen Wahrscheinlichkeiten (für Daten und Teststatistiken) fest!

Bemerkung 3: In der Bayes'schen Statistik haben auch Hypothesen bzw. Parameter eine Verteilung, aber da gibt es keine p-Werte. Beachte: Der p-Wert unterstellt die Gültigkeit der Nullhypothese und ist somit nicht als a-posteriori-Wahrscheinlichkeit zu interpretieren.

Entsprechendes gilt für die links- bzw. beidseitige Testsituation. Der p-Wert lässt sich dann interpretieren als Unterschreitungswahrscheinlichkeit $p_{\mathrm{L}}(z) = P_{\mu_0}\big(\tilde{Z} \leq z\big)$ bzw. als Überschreitungswahrscheinlichkeit $p_{\mathrm{B}}(z) = P_{\mu_0}\big(|\tilde{Z}| \geq |z|\big)$.

2.3 Der p-Wert als Grenzniveau

Nach der Neyman-Pearson-Methode wird durch die Vorgabe eines Signifikanzniveaus α die Wahrschein-

lichkeit eines Fehlers 1. Art kontrolliert. Bleiben wir beim Testproblem (R). Der kritische Wert ist das $(1-\alpha)$-Quantil der $N(0,1)$-Verteilung, kurz $z_{1-\alpha}$. Diese Zahl ist per Definition die Lösung der Gleichung $\Phi(x) = 1-\alpha$. Der kritische Bereich (= Ablehnungsbereich) ist somit das Intervall $K_R(\alpha) := [z_{1-\alpha}, \infty)$.

Die Wahrscheinlichkeit eines Fehlers 1. Art ist (höchstens) α:

$$P_{\mu_0}(Z \geq z_{1-\alpha}) = 1 - \Phi(z_{1-\alpha}) = \alpha$$

(für $\mu < \mu_0$ gilt $P_\mu(Z \geq z_{1-\alpha}) < \alpha$).

Die p-Wert-Methode fragt nach dem kleinsten kritischen Bereich, der bei Vorliegen des beobachteten Prüfgrößenwertes z zu einer Ablehnung von H_0 führen würde. Alle sinnvollen kritischen Bereiche sind von der Form $[c, \infty)$, da große Prüfgrößenwerte für die Alternative sprechen. Der kleinste kritische Bereich, der z enthält, ist $[z, \infty)$. Das „Signifikanzniveau" $\alpha^*(z)$, das zu diesem kritischen Bereich gehört, ist der p-Wert. Wir können also wieder schreiben

$$\alpha^*(z) = P_{\mu_0}\left(\tilde{Z} \geq z\right) = p_{\mathrm{R}}(z).$$

Der p-Wert ist ein „Grenzniveau". Es ist die *größte untere Schranke* für das Signifikanzniveau, zu dem H_0 bei Vorliegen des Prüfgrößenwertes z (also im Nachhinein betrachtet) hätte abgelehnt werden können. Es gilt

$$p_R(z) \begin{cases} < \alpha, & z > z_{1-\alpha} \\ = \alpha, & z = z_{1-\alpha} \\ > \alpha, & z < z_{1-\alpha} \end{cases}$$

Daraus folgt: Der p-Wert ist kleiner oder gleich α genau dann, wenn der Prüfgrößenwert z im kritischen Bereich liegt:

$$p_R(z) \leq \alpha \Leftrightarrow z \in K_R(\alpha).$$

Für die Testentscheidung bedeutet dies folgendes: Nach der Neyman-Pearson-Methode wird die Nullhypothese H_0 zum Signifikanzniveau α abgelehnt, wenn der Prüfgrößenwert in den kritischen Bereich fällt; dies ist gleichbedeutend damit, dass der p-Wert kleiner oder gleich α ist.

Schauen wir uns noch kurz das beidseitige Testproblem (B) an. In diesem Fall ist der kritische Bereich zum Signifikanzniveau α gegeben durch

$$K_B(\alpha) := (-\infty, -z_{1-\alpha/2}] \cup [z_{1-\alpha/2}, \infty).$$

Alle kritischen Bereiche sind von Form $(-\infty, -c] \cup [c, \infty)$, $c > 0$. Der kleinste kritische Bereich, der die Beobachtung z enthält, ist gegeben durch $(-\infty, -|z|] \cup [|z|, \infty)$. Das „Signifikanzniveau" dieses kritischen Bereiches ist dann wieder der p-Wert:

$$\begin{aligned} p_B(z) &= 2 \cdot \Phi(-|z|) \\ &= 2 \cdot (1 - \Phi(|z|) \\ &= 2 \cdot \min\{\Phi(z), 1 - \Phi(z)\} \end{aligned}$$

Wegen

$$p_B(z) \begin{cases} < \alpha, & |z| > z_{1-\alpha/2} \\ = \alpha, & |z| = z_{1-\alpha/2} \\ > \alpha, & |z| < z_{1-\alpha/2} \end{cases}$$

führt auch hier die Neyman-Pearson-Methode und die p-Wert-Methode zur gleichen Testentscheidung: Ablehnung von H_0, falls

$$p_B(z) \leq \alpha \Leftrightarrow z \in K_B(\alpha)$$

Aufgrund der Eigenschaft eines „Grenzniveaus" in Abhängigkeit von der Beobachtung wird der p-Wert in der Literatur auch als *tatsächliches*, *exaktes*, *beobachtetes* oder *empirisches* Signifikanzniveau bezeichnet.

Aber Achtung! Der p-Wert ist nicht als ein Signifikanzniveau zu interpretieren (deshalb ist es besser, beim p-Wert von einer größten untere Schranke für das Signifikanzniveau zu sprechen). Das Signifikanzniveau α charakterisiert einen statistischen Test in dem Sinne, dass *bei Unterstellung der Gültigkeit von* H_0 die Wahrscheinlichkeit für eine Ablehnung von H_0 (Fehler 1. Art) höchstens α ist. D. h., in vielen Testdurchführungen wird es unter H_0 in etwa $\alpha \cdot 100\%$ der Fälle zu einer (fälschlichen) Ablehnung von H_0 kommen.

Der p-Wert entzieht sich einer solchen *frequentistischen* Interpretation, da er von den Daten abhängt. Aus diesem Grunde kann der p-Wert auch nicht als eine Wahrscheinlichkeit für den Fehler 1. Art interpretiert werden. Die Aussage „Die Irrtumswahrscheinlichkeit ist gleich 0.042" ist also falsch. Die Irrtumswahrscheinlichkeit charakterisiert einen Test (Nullhypothese, kritischer Bereich) und hat nichts mit Daten zu tun.

3 Der p-Wert im Diskreten

In diesem Abschnitt wollen wir auf den p-Wert eingehen, wenn die Prüfgröße diskret verteilt ist. Dann kann der p-Wert als Zufallsvariable keine (stetige) Gleichverteilung auf $(0,1)$ besitzen. Wir betrachten im Folgenden das Binomialmodell $B_{n,\theta}$, $\theta \in (0,1)$, stellvertretend für diskrete Modelle. Die Binomialverteilung $B_{n,\theta}$ ist ein Wahrscheinlichkeitsmaß auf (der Potenzmenge von) $\{0,\dots,n\}$, festgelegt durch die Einzelwahrscheinlichkeiten

$$B_{n,\theta}(\{j\}) := \binom{n}{j}\theta^j(1-\theta)^{n-j}, \quad j = 0,\dots,n.$$

Dieses Wahrscheinlichkeitsmaß taucht auf als Verteilung der zufälligen Anzahl von Treffern X in einer Bernoulli-Kette der Länge n. Unter der Trefferwahrscheinlichkeit $\theta \in (0,1)$ besitzt X eine $B_{n,\theta}$-Verteilung, d. h.

$$P_\theta(X = j) = B_{n,\theta}(\{j\}), \quad j = 0,\dots,n.$$

Wir betrachten zunächst nur die einseitige Testsituation.

3.1 Der p-Wert als Zufallsvariable

Gegeben sei das linksseitige Testproblem

$$(\mathrm{L}') \quad H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta < \theta_0$$

Eine geeignete Prüfgröße ist X, die unter H_0 B_{n,θ_0}-verteilt ist. Bezeichne

$$F_{\theta_0}(x) := \sum_{j=0}^{[x]} B_{n,\theta_0}(\{j\})$$

die Verteilungsfunktion der B_{n,θ_0}-Verteilung. Dabei ist $[x]$ der ganzzahlige Teil einer reellen Zahl x. Dies ist eine Treppenfunktion mit den Sprungstellen j und Sprunghöhen $B_{n,\theta_0}(\{j\})$, $j = 0,\dots,n$.

Sei k die beobachtete Anzahl von Treffern (Realisierung von X). Dann ist der p-Wert zur Beobachtung k gegeben durch

$$p_{L'}(k) := F_{\theta_0}(k) = \sum_{j=0}^{k} B_{n,\theta_0}(\{j\}),$$

als Zufallsvariable geschrieben

$$p_{L'}(X) := F_{\theta_0}(X) = \sum_{j=0}^{X} B_{n,\theta_0}(\{j\})$$

Diese ist diskret verteilt und kann daher nicht mehr auf $(0,1)$ gleichverteilt sein. Die Verteilungsfunktion von $F_{\theta_0}(X)$ ist eine Treppenfunktion, deren Graph immer unterhalb der Diagonalen liegt. Speziell gilt $P_{\theta_0}\big(p_{L'}(X) \le 0.05\big) \le 0.05$.

Im Fall des rechtsseitigen Testproblems

$$(R') \quad H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

ist

$$p_{\mathrm{R}'}(k) := 1 - F_{\theta_0}(k-1) = \sum_{j=k}^{n} B_{n,\theta_0}(\{j\})$$

der p-Wert zur Beobachtung k.

3.2 Der p-Wert als Wahrscheinlichkeit

Der p-Wert kann wieder als Wahrscheinlichkeit interpretiert werden: Etwa beim Testproblem (R') ist der p-Wert die Überschreitungswahrscheinlichkeit $P_{\theta_0}\big(\tilde{X} \ge k\big)$, also die Summe der Einzelwahrscheinlichkeiten $\sum_{j=k}^{n} P_{\theta_0}\big(\tilde{X} = j\big)$. Dies ist die Wahrscheinlichkeit, unter θ_0 mindestens k Treffer zu beobachten. Dabei bezieht sich die zufällige Trefferanzahl $\tilde{X}$ wieder auf die unabhängige Wiederholung des gleichen Zufallsexperiments (Bernoulli-Kette der Länge n). Formal können wir den p-Wert als Übergangswahrscheinlichkeit, sprich (elementare) bedingte Wahrscheinlichkeit, auffassen:

$$p_{\mathrm{R}'}(k) = P_{\theta_0}\big(\tilde{X} \ge X | X = k\big) = P_{\theta_0}\big(\tilde{X} \ge k\big).$$

3.3 Der p-Wert als Grenzniveau

Auch als Grenzniveau ergibt sich der p-Wert. Bleiben wir beim Testproblem (R'). Alle sinnvollen kritischen Bereiche sind von der Form $\{l,\dots,n\}$, da große Trefferzahlen für die Alternative sprechen.

Ist ein Signifikanzniveu α vorgegeben, so wählt man den kritischen Bereich $K_{\mathrm{R}'}(\alpha) := \{c_\alpha,\dots,n\}$ mit dem kritischen Wert

$$c_\alpha = \min\big\{l \in \{0,\dots,n\} : P_{\theta_0}\big(X \ge l\big) \le \alpha\big\}.$$

Der kleinste kritische Bereich, der die Beobachtung k enthält, ist $\{k,\dots,n\}$. Das „Signifikanzniveau" $\alpha^*(k)$, das zu diesem kritischen Bereich gehört, ist der p-Wert. Wir können also wieder schreiben

$$\alpha^*(k) = P_{\theta_0}\big(\tilde{X} \ge k\big) = p_{\mathrm{R}'}(k).$$

Es ist die größte untere Schranke für das Signifikanzniveau, zu dem H_0 bei Vorliegen des Beobachtungswertes k abgelehnt werden kann:

$$p_{\mathrm{R}'}(k) \begin{cases} \le \alpha, & k \ge c_\alpha \\ > \alpha, & k < c_\alpha \end{cases}$$

Bezüglich der Testentscheidung stimmen p-Wert-Methode und Neyman-Pearson-Methode überein:

$$p_{\mathrm{R}'}(k) \leq \alpha \Leftrightarrow k \in K_{\mathrm{R}'}(\alpha)$$

Gilt dies auch bei zweiseitigen Fragestellungen? Die salomonische Antwort lautet: Es kommt drauf an. Die Antwort hängt davon ab, wie man den p-Wert definiert.

3.4 Festlegung von p-Werten im beidseitigen Testproblem

In der beidseitigen Testsituation

$$(\mathrm{B}') \quad H_0 : \theta = \theta_0 \quad \text{gegen} \quad H_1 : \theta \neq \theta_0$$

wollen wir zunächst den symmetrischen Fall $\theta_0 = 0.5$ behandeln. In diesem Fall gibt es nur eine sinnvolle Festlegung des p-Wertes. Sie ist gegeben durch

$$p_{\mathrm{B}'} = \begin{cases} 2 \cdot F_{\theta_0}(k), & k < n/2 \\ 2 \cdot \big(1 - F_{\theta_0}(k-1)\big), & k > n/2 \\ 1, & k = n/2 \end{cases} \qquad (4)$$

Der zugehörige kritische Bereich ist

$$\begin{aligned} \{0,\ldots,k\} \cup \{n-k,\ldots,n\}, & \quad k < n/2 \\ \{0,\ldots,n-k\} \cup \{k,\ldots,n\}, & \quad k > n/2 \\ \{0,\ldots,n\}, & \quad k = n/2 \end{aligned}$$

Im nichtsymmetrischen Fall $\theta_0 \neq 0.5$ gibt es - anders als im stetigen Fall - verschiedene, aber gleichberechtigte Festlegungen des p-Wertes. Dies kann zu unterschiedlichen Testentscheidungen führen (Abschnitt 3.5) und es kann zu Abweichungen zwischen verschiedenen Statistik-Programmen kommen (Abschnitt 4.1).

Wir erwähnen die folgenden drei Definitionen, die im symmetrischen Fall mit (4) überein stimmen:

1. Version: Man definiert den beidseitigen p-Wert als das doppelte des Minimums vom linksseitigen und rechtsseitigen p-Wert:

$$\begin{aligned} p_{\mathrm{B}'}(k) &:= 2 \cdot \min\{F_{\theta_0}(k), 1 - F_{\theta_0}(k-1), 0.5\} \\ &= 2 \cdot \min\{P_{\theta_0}(\tilde{X} \leq k), P_{\theta_0}(\tilde{X} \geq k), 0.5\} \end{aligned}$$

2. Version: Realisierungen werden als extrem bezeichnet, wenn sie betragsmäßig mehr vom Erwartungswert $E_{\theta_0}(\tilde{X}) = n\theta_0$ abweichen als die Beobachtung k. Der beidseitige p-Wert ist somit

$$\begin{aligned} & p_{\mathrm{B}'}(k) \\ &:= P_{\theta_0}\big(|\tilde{X} - n\theta_0| \geq |k - n\theta_0|\big) \\ &= \begin{cases} P_{\theta_0}(\tilde{X} \leq k) + P_{\theta_0}(\tilde{X} \geq 2n\theta_0 - k), & k < n\theta_0 \\ P_{\theta_0}(\tilde{X} \geq k) + P_{\theta_0}(\tilde{X} \leq 2n\theta_0 - k), & k > n\theta_0 \\ 1, & k = n\theta_0 \end{cases} \end{aligned}$$

3. Version: Bei dieser Festlegung gilt eine Beobachtung als extrem, wenn sie eine kleine Eintrittswahrscheinlichkeit besitzt. Man addiert alle Trefferwahrscheinlichkeiten $P_{\theta_0}\big(\tilde{X} = j\big)$ auf, die kleiner oder gleich $P_{\theta_0}\big(\tilde{X} = k\big)$ sind:

$$\begin{aligned} & p_{\mathrm{B}'}(k) \\ &:= \sum_{\substack{j \in \{0,\ldots,n\} \\ P_{\theta_0}(\tilde{X}=j) \leq P_{\theta_0}(\tilde{X}=k)}} P_{\theta_0}(\tilde{X} = j) \\ &= \begin{cases} P_{\theta_0}(\tilde{X} \leq k) + P_{\theta_0}(\tilde{X} \geq k_1), & k < n\theta_0 \\ P_{\theta_0}(\tilde{X} \geq k) + P_{\theta_0}(\tilde{X} \leq k_2), & k > n\theta_0 \\ 1, & k = n\theta_0 \end{cases} \end{aligned}$$

Dabei sind

$$\begin{aligned} k_1 &= \min\{j > n\theta_0 : P_{\theta_0}(\tilde{X} = j) \leq P_{\theta_0}(\tilde{X} = k)\} \\ k_2 &= \max\{j < n\theta_0 : P_{\theta_0}(\tilde{X} = j) \leq P_{\theta_0}(\tilde{X} = k)\} \end{aligned}$$

Allen drei Versionen ist gemeinsam, dass der p-Wert die Wahrscheinlichkeiten an den Flanken betrachtet. Die Funktion $j \mapsto P_\theta\big(\tilde{X} = j\big)$, $j = 0,\ldots,n$, ist nämlich erst streng monoton steigend, dann streng monoton fallend (für Details siehe Georgii 2009, Lemma 8.8). Die Versionen können, müssen aber nicht zum selben p-Wert führen.

3.5 Vergleich der Testentscheidungen nach NP und der p-Wert-Methode

Nach der Neyman-Pearson-Methode wird H_0 zum Niveau α abgelehnt, falls $k \in K_{\mathrm{B}'}(\alpha) := \{0,\ldots,c_{1,\alpha}\} \cup \{c_{2,\alpha},\ldots,n\}$. Dabei sind die kritischen Werte definiert durch $c_{1,\alpha} = \max\{l : P_{\theta_0}(X \leq l) \leq \alpha/2\}$ bzw. $c_{2,\alpha} = \min\{l : P_{\theta_0}(X \geq l) \leq \alpha/2\}$. Die beiden Testentscheidungen

$$\text{Ablehnung von } H_0, \text{ falls } p_{\mathrm{B}'}(k) \leq \alpha$$

und

$$\text{Ablehnung von } H_0, \text{ falls } k \in K_{\mathrm{B}'}(\alpha)$$

sind nur im Fall des p-Wertes in der Version 1 identisch. Für die Versionen 2 oder 3 gilt dies im Allgemeinen nicht mehr!

Zahlenbeispiel: (i) Sei $\theta_0 = 0.25$ und $n = 20$.

Im Fall $k = 1$ ergibt Version 1

$$p_{\mathrm{B}'}(1) = 2 \cdot P_{0.25}(\tilde{X} \leq 1) = 0.049$$

Version 2

$$p_{\mathrm{B}'}(1) = P_{0.25}(\tilde{X} \leq 1) + P_{0.25}(\tilde{X} \geq 9) = 0.065$$

und Version 3

$$p_{\mathrm{B}'}(1) = P_{0.25}(\tilde{X} \leq 1) + P_{0.25}(\tilde{X} \geq 10) = 0.038$$

Im Fall $k = 9$ ergibt Version 1

$$p_{\mathrm{B}'}(9) = 2 \cdot P_{0.25}(\tilde{X} \geq 9) = 0.082$$

Version 2 und 3 stimmen überein:

$$p_{\mathrm{B}'}(9) = P_{0.25}(\tilde{X} \geq 9) + P_{0.25}(\tilde{X} \leq 1) = 0.065$$

Im Fall $k = 10$ ergibt Version 1

$$p_{\mathrm{B}'}(10) = 2 \cdot P_{0.25}(\tilde{X} \geq 10) = 0.028$$

Die Versionen 2 und 3 stimmen wieder überein:

$$p_{\mathrm{B}'}(10) = P_{0.25}(\tilde{X} \geq 10) + P_{0.25}(\tilde{X} = 0) = 0.017$$

Wie sieht der kritische Bereich zum Signifikanzniveau $\alpha = 0.05$ aus? Der untere kritische Wert ist $c_{1,\alpha} = 1$, der obere kritische Wert ist $c_{2,\alpha} = 10$. Damit ist der kritische Bereich $\{0,1\} \cup \{10,\ldots,20\}$. Wir sehen also, dass bei $k = 1$ der p-Wert nach Version 2 von 0.065 zu keiner Ablehnung von H_0 führt. Nach der Neyman-Pearson-Methode würden wir ablehnen, da 1 im kritischen Bereich liegt.

(ii) Sei nun $\theta_0 = 0.2$ und $n = 20$. Im Fall $k = 8$ ergibt Version 1

$$p_{\mathrm{B}'}(8) = 2 \cdot P_{0.2}(\tilde{X} \geq 8) = 0.064$$

Version 2 und 3 stimmen überein:

$$p_{\mathrm{B}'}(8) = P_{0.2}(\tilde{X} \geq 8) + P_{0.25}(\tilde{X} = 0) = 0.044$$

Zum Signifikanzniveau $\alpha = 0.05$ ist $\{0\} \cup \{9,\ldots,20\}$ der kritische Bereich. Auch hier sehen wir: Verwendet man die Version 2 oder 3, so führt der p-Wert von 0.044 zu einer Ablehnung von H_0, während es nach der Neyman-Pearson-Methode zu keiner Ablehnung von H_0 kommt, da 8 nicht im kritischen Bereich liegt.

Version 1 ist wohl die gebräuchlichste (Sheskin 2011, Seite 313) und entspricht der Festlegung im stetigen Fall (siehe Bemerkung 1).

3.6 Approximativer Binomialtest

Für große Stichprobenumfänge verwendet man die Prüfgröße

$$Z = \frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}},$$

die nach dem zentralen Grenzwertsatz asymptotisch $N(0,1)$-verteilt ist. Ist k eine Realisierung von X und schreiben wir

$$z_k = \frac{k - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}$$

so gilt $p_{\mathrm{L}'}(k) \approx \Phi(z_k)$, $p_{\mathrm{R}'}(k) \approx 1 - \Phi(z_k)$, $p_{\mathrm{B}'}(k) \approx 2 \cdot (1 - \Phi(|z_k|))$.

4 Aspekte der Diskussion um p-Werte

In diesem Abschnitt werden Probleme bei der Verwendung von Software, p-Wert versus Signifikanzniveau α und Fragen des Schulunterrichts behandelt.

4.1 Der p-Wert bei statistischer Software

Gängige statistische Softwarepakete wie R, SAS und SPSS weisen den p-Wert automatisch aus. Wir wollen auf zwei Punkte aufmerksam machen, die es dabei zu beachten gilt:

1. *Einseitiger p-Wert*: Darunter verstehen die Programme (z. B. SAS) das Minimum vom rechtsseitigen und linksseitigen p-Wert. In diesem Fall muss diese Zahl nicht der p-Wert von dem tatsächlich interessierenden Testproblem sein. Liegt also im Binomialmodell das rechtsseitige Testproblem (R') zugrunde und gilt $p_{\mathrm{L}'}(k) = \min\{p_{\mathrm{L}'}(k), p_{\mathrm{R}'}(k)\}$, so ist die im Programm-Output angegebene Zahl $p_{\mathrm{L}'}(k)$ der linksseitige p-Wert und nicht der rechtsseitige p-Wert, welchen man dann so bestimmt: $p_{\mathrm{R}'}(k) = 1 - p_{\mathrm{L}'}(k) + P_{\theta_0}(\tilde{X} = k)$.

Man hat also bei der Interpretation des Outputs zusätzlich zu berücksichtigen, ob der Prüfgrößenwert k überhaupt extrem im Sinne der Alternative (R') ist. Eine Entscheidung für die Alternative $H_1 : \theta > \theta_0$ ist nur dann sinnvoll, wenn k rechts vom Erwartungswert $n\theta_0$ liegt. Bei anderen Programmen wie R gibt es die Option, links- oder rechtsseitig zu testen.

2. *Zweiseitiger p-Wert*: Im nichtsymmetrischen Fall des Binomialtests verwenden statistische Softwarepakete unterschiedliche Versionen. Beispielsweise verwenden SAS und SPSS die Version 1, R verwendet Version 3. Wird nur ein zweiseitiger p-Wert in der Version 1 angegeben (z. B. SPSS), so ist dieser durch 2 zu teilen, um den einseitigen p-Wert zu erhalten. Bei der Interpretation dieser Zahl ist Punkt 1 zu beachten.

Bei Verwendung von Software, die den p-Wert beim Binomialtest nicht berechnen (wie z. B. Excel), bleibt nur, den p-Wert über die (kumulierten) Einzelwahrscheinlichkeiten zu bestimmen.

4.2 Diskussion: p vs α?

Die bisherigen Ausführungen suggerieren, dass es keine wesentlichen Unterschiede zwischen der Vorgabe eines Signifikanzniveaus α (Neyman-Pearson-Methode) und der p-Wert-Methode (Fisher) gibt. Eine wichtige Gemeinsamkeit ist, dass beide Verfahren von der Gültigkeitsannahme der Nullhypothese ausgehen (kein Bayes'scher Ansatz!). Aber es gibt Unterschiede zwischen diesen beiden Ansätzen. Wir möchten hier nur auf einen Punkt aufmerksam machen.

Nach der Neyman-Pearson-Methode stehen sich zwei konkurrierende Hypothesen – Nullhypothese H_0 und Alternativhypothese H_1 – gegenüber mit den zentralen Begriffen Fehler 1. Art (α-Fehler) und Fehler 2. Art (β-Fehler). Fisher hat (die explizite Formulierung von) Alternativen abgelehnt und ein Verwerfen von H_0 bedeutet (noch) nicht, dass man sich für eine alternative Hypothese entscheiden soll. Eigentlich kaum zu glauben, wenn man an den p-Wert im Sinne von Fisher denkt: Ein Signifikanztest ist ein Verfahren, das eine Wahrscheinlichkeit (unter H_0) berechnet, das beobachtete Ergebnis oder noch extremere Ergebnisse zu erzielen. Was extrem bedeutet, wird letztlich durch die Alternativhypothese bestimmt, denn sie legt die *Richtung* der Abweichungen von H_0 fest (Fisher hat daher zumindest implizit an Alternativen gedacht).

Prüft man in einem Normalverteilungsmodell den Mittelwert, dann verwendet man bei bekannter Varianz den Gauß-Test (bei unbekannter Varianz ist es der t-Test), sind dagegen Aussagen über die Varianz interessant, dann verwendet man den χ^2-Test auf Varianz.

Der p-Wert wird sich auf denjenigen Test beziehen, der die Alternative möglichst gut entdeckt, der also eine möglichst hohe *Macht* (Power), *Schärfe* hat. So ist z. B. der Gauß-Test unter den getroffenen Verteilungsannahmen ein bester unverfälschter Niveau-α-Test (siehe z. B. Georgii 2009, Kapitel 10). Bezüglich der wichtigsten in der Praxis verwendeten Testverfahren sei auf das Standardwerk von Sheskin 2011 verwiesen.

Über den Konflikt „p-Wert gegen festes Niveau", der die Auseinandersetzungen zwischen den Begründern der heutigen Testtheorie R. A. Fisher (1890-1962) auf der einen Seite und J. Neyman (1894-1981) und E.S. Pearson (1895-1980) auf der anderen Seite widerspiegelt, sei auf den Artikel von Hubbard und Bayarri 2003 mit den sich anschließenden Diskussionsbeiträgen von Berk 2003 und Carlton 2003 verwiesen. Lehmann 1993 beleuchtet das Thema (erfreulicherweise) mehr aus „statistischer" als aus „philosophischer" Sicht. Eine lesenswerte historische Einordnung gibt Stute 1989, siehe auch Sheskin 2011, S. 68-74.

4.3 Der p-Wert im Schulunterricht

Zunächst darf das Konzept des p-Wertes nicht dazu (ver)führen, auf das rational-mathematische Konzept der Neyman-Pearson-Testtheorie zu verzichten. Dies bedeutet: Das Testen von Hypothesen darf nicht beim Signifikanztest (Niveau-α-Test) stehen bleiben, wo es nur um den α-Fehler geht. Dies wäre nur eine Seite der Medaille. Man muss den β-Fehler bzw. die Macht, die Power ($= 1 - \beta$) eines Tests mit ins Spiel bringen. Ohne die Macht anzusprechen ist es schwierig, folgende Fragen zu thematisieren:

- Die Wahl von $\alpha = 0.05$ als Kompromisslösung ist erst durch die Gegenläufigkeit der beiden Fehlerwahrscheinlichkeiten ($\alpha \downarrow \Rightarrow \beta \uparrow$ bzw. $\alpha \uparrow \Rightarrow \beta \downarrow$) zu begründen (warum nicht $\alpha = 10^{-6}$?)
- Die Bedeutung des Stichprobenumfangs bleibt unklar. Welchen Einfluss hat der Stichprobenumfang auf die Power bzw. den β-Fehler?
- Welche Rolle spielen Unterschiede von praktischer Relevanz (Effektgrößen)?
- Wie groß muss der Stichprobenumfang mindestens sein, um einen Effekt mit einer gewissen Wahrscheinlicheit zu entdecken?

Für den Zweistichproben-Gauß-Test siehe dazu auch Börgens 2014. Wenn die „Philosophie" der Neyman-Pearson-Methode verstanden worden ist, ihre (zugegebenermaßen nicht einfach zu verstehenden) Begriffsbildungen und Sprechweisen geklärt und Fehlinterpretationen ausgeschlossen sind (an dieser Stelle sei noch einmal auf Henze 2013, Kap. 29 verwiesen), kann der p-Wert eingeführt werden (zuerst links- bzw. rechtsseitig, dann beidseitig). Dieser soll als eine (schnelle) Entscheidungshilfe für den Niveau-α-Test gesehen werden (man erspart sich den Vergleich mit Quantilen aus Tabellenwerken): Eine Nullhypothese ist zu verwerfen, falls der p-Wert $\leq \alpha$ ist.

Im Diskreten wird bei zweiseitiger Fragestellung empfohlen, den p-Wert in der Version 1 einzuführen,

damit auch in diesem Fall die Äquivalenz der beiden Entscheidungsvorschriften (zum Niveau α) „Ablehnung von H_0, falls p-Wert $\leq \alpha$" und „Ablehnung von H_0, falls der Prüfgrößenwert im kritischen Bereich K_α liegt" gewährleistet ist. Das „Ausweichen" auf den approximativen Binomialtest ist weniger zu empfehlen.

Egal, ob der p-Wert als Wahrscheinlichkeit oder als Grenzniveau interpretiert wird: Wichtig ist es zu betonen, was der p-Wert *nicht* ist: Wahrscheinlichkeit für die Richtigkeit einer Nullhypothese bzw. Wahrscheinlichkeit für den Fehler 1. Art.

Beim p-Wert hat die Bayes-Statistik nichts zu suchen. Ein Grund mehr bei der Einführung des Hypothesentests nach der Neyman-Pearson-Methode die Bayes-Statistik außen vor zu lassen. In diesem Zusammenhang sei an Diepgen 2002 erinnert. Erst danach (wenn überhaupt) sollte das Konzept der Bayes-Statistik vorgestellt werden. Dabei ist zu beachten, dass die *frequentistische* Sichtweise und die *Bayes'sche* Sichtweise zwei völlig verschiedene Konzepte sind, deren Wahrscheinlichkeitsaussagen sich überhaupt nicht sinnvoll miteinander vergleichen lassen (man vergleicht Äpfel mit Birnen). Wie schreibt Carlton (2003) so treffend:

> „A 5% cutoff means that 5% of true hypotheses are rejected, not that 5% of rejected hypotheses are true."

Und wer die Bayes-Statistik unbedingt im Unterricht behandeln will, sei noch einmal daran erinnert: Der p-Wert ist keine inverse Wahrscheinlichkeit (a-posteriori-Wahrscheinlichkeit).

5 Schlussbemerkungen

In der Literatur findet man gelegentlich die Auffassung, dass der p-Wert ein Maß für die Verträglichkeit (measure of evidence) von Daten und Nullhypothese ist (siehe z. B. Stahel 2008, Kap. 8). Es gibt aber auch Kritik an einer solchen Auffassung (Schervish, 1996).

Auf eine Gefahr wollen wird abschließend hinweisen und zwar auf die Gefahr, dass das Signifikanzniveau an den p-Wert angepasst wird. Wenn der p-Wert angegeben wird, so hat im Prinzip jeder das Recht zu dem Niveau zu testen, das er für geeignet hält. Aber diese Meinungsfreiheit untergräbt den wahren Sinn des statistischen Testens. Angenommen, der p-Wert ist 0.042. Möchte man ein signifikantes Testresultat, dann wählt man (im Nachhinein!) $\alpha = 0.05$. Soll die Nullhypothese nicht abgelehnt werden (z. B. aus bestimmten Interessensgründen), so wählt man $\alpha = 0.01$. Daher ist die Angabe einer oberen Schranke α für den p-Wert – und zwar bevor man den p-Wert vom Computer ausrechnen lässt – unentbehrlich.

Der p-Wert hat seine Tücken. Daher ist es wichtig, ein Verständnis für diesen Begriff zu entwickeln. Auch deswegen, damit die Statistik nicht zu einer schwarzen Kiste wird. Computerprogramme berechnen nämlich den p-Wert aus den Daten per Knopfdruck.

Dem Anwender soll es nicht so gehen wie wohl den meisten Lesern des Science Fiction Romans *Per Anhalter durch die Galaxie* von Douglas Adams. Hier hatte der Supercomputer *Deep Thought* die Frage nach dem Leben, dem Universum und dem ganzen Rest nach über sieben Millionen Jahren Rechenzeit mit der Zahl „42" beantwortet. Unbefriedigend (die Antwort, weniger die Rechenzeit). Schade, dass *Deep Thought* nicht die Zahl 0.042 ausgewiesen hat. Dann hätte sich aus statistischer Sicht eine Interpretationsmöglichkeit angeboten. Und hier wird deutlich: Ob das Ergebnis 0.042 als signifikant einzustufen ist, dazu hätte man sich auf ein Niveau α einigen müssen, bevor *Deep Thought* diesen Wert ausgespuckt hätte. Zeit genug zur Einigung wäre jedenfalls gewesen.

Literatur

Berk, K. N. (2003): Discussion of a paper by Hubbard and Bayarri (Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing). In : *The American Statistician* 57(3), S. 178-179.

Börgens, M. (2014): Die Bedeutung des β-Risikos. In: *Stochastik in der Schule* 34(1), S. 8-12.

Carlton, M. A. (2003): Discussion of a paper by Hubbard and Bayarri (Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing). In : *The American Statistician* 57(3), S. 179-181.

Diepgen, R. (2002): $P(H|D)$ versus $P(D|H_0)$? Wie man das Testen von Hypothesen - lieber doch nicht - einführen sollte. In: *Stochastik in der Schule* 22(3), S. 34-38.

Falk, M., Becker, R. und Marohn, F. (2004): Angewandte Statistik. Berlin-Heidelberg: Springer.

Falk, M., Hain, J., Marohn, F., Fischer, H. und Michel, R. (2014): Statistik in Theorie und Praxis. Mathematik für das Lehramt, Springer-Spektrum, Berlin-Heidelberg.

Freund, J.E.; Perles, B.M. (1996): Einige Beobachtungen zur Definition von p-Werten: *Stochastik in der Schule* 16(2), S. 36-38.

Georgii, H.-O. (2009): Stochastik. 4. Auflage. Berlin: de Gruyter.

Henze, N. (2013): Stochastik für Einsteiger. 10. Auflage. Wiesbaden: Springer Spektrum.

Hubbard, R. und Bayarri, M. J. (2003): Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. In: *The American Statistician 57, S. 171-178*.

Krengel, U. (2005): Einführung in die Wahrscheinlichkeitstheorie und Statistik. 8. Auflage. Wiesbaden: Vieweg.

Lehmann, E. L (1993): The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two?: In: *Journal of the American Statistical Association* 88 (424), S. 1242-1249.

Rudolf, M. und Kuhlisch, W. (2008): Biostatistik. München: Pearson Studium.

Schervish, M. J. (1996): P values: What they are and what they are not. In: *The American Statistician* 50(3), S. 203-206.

Sheskin, D. J. (2011): Handbook of Parametric and Nonparametric Statistical Procedures. Fifth Edition. Boca Raton: Chapman & Hall.

Stahel, W. A. (2008): Statistische Datenanalyse. 5. Auflage. Wiesbaden: Vieweg.

Stute, W. (1989): Der historische Streit zwischen R. A. Fisher und J. Neyman oder: Ein Sittengemälde aus der Blütezeit der englischen Schule für Statistik. In: *Mathematische Semesterberichte* 36(1), S. 61-84.

Wengenroth, J. (2008): Wahrscheinlichkeitstheorie. Berlin: de Gruyter.

Anschrift des Verfassers
Frank Marohn
Institut für Mathematik
Universität Würzburg
Emil-Fischer-Str. 30
97074 Würzburg
marohn@mathematik.uni-wuerzburg.de