Statistische Hypothesenprüfung und p-Werte
Theorie und Praxis

Thorsten Dickhaus
December 1, 2009

1. DEFINITIONEN UND EIGENSCHAFTEN

Zufällige Vorgänge in der Natur oder bei Experimenten in der Wissenschaft und Technik werden mathematisch mit Hilfe von Zufallsgrößen beschrieben. Eine Zufallsvariable ist dabei eine Abbildung $X : \Omega \to \mathcal{X} \subseteq \mathbb{R}$. Dabei bezeichnet Ω den (oftmals nicht beobachtbaren) Grundraum des Experimentes. Die zufälligen Effekte, die es zu analysieren gilt, werden in einem statistischen Modell über einen Parameter ϑ formalisiert.

Definition 1.1 (Statistisches Modell)**.** *Ein statistisches Modell ist definiert als ein Tripel $(\mathcal{X}, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$. Dabei heißt $\mathcal{X}$ Stichprobenraum, $\mathcal{A}$ bezeichnet das System der messbaren Teilmengen von $\mathcal{X}$ und $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ ist eine Familie von Wahrscheinlichkeitsmaßen auf dem messbaren Raum $(\mathcal{X}, \mathcal{A})$. Ist $\vartheta \subseteq \mathbb{R}^p$ für ein $p \in \mathbb{N}$, so heißt das statistische Experiment parametrisierbar bzw. (durch ϑ) parametrisiert.*

Inferenzstatistik beschäftigt sich mit dem Prüfen von Hypothesen bezüglich des Parameters ϑ. Es wird eine Entscheidung zwischen Hypothesenpaaren der Form

$$H_0 : \{\vartheta \in \Theta_0\} \quad \text{versus} \quad H_1 : \{\vartheta \in \Theta_1 := \Theta \setminus \Theta_0\}$$

gesucht.

Genügt nun die Zufallsvariable X dem Verteilungsgesetz $\mathbb{P}_\vartheta$ (in Zeichen: $X \sim \mathbb{P}_\vartheta$), so trägt X Information über den unbekannten Parameter ϑ. Typischerweise besteht ein Zufallsexperiment aus einer sequentiellen Beobachtung von n Zufallsvariablen $X_1, \ldots, X_n$, die stochastisch unabhängig und identisch verteilt (independent and identically distributed, in Zeichen: i.i.d.) sind, wobei $X_1 \sim \mathbb{P}_\vartheta$. Durch die wiederholte Beobachtung von Zufallsvariablen aus der zu Grunde liegenden Grundgesamtheit wird der Versuch unternommen, die Information über ϑ zu präzisieren. Das Beobachtungtupel $(x_1, \ldots, x_n)$ wird als die Realisierungen von $X_1, \ldots, X_n$ bezeichnet.

Eine Teststatistik oder auch Prüfgröße ist eine Abbildung $T : \mathbb{R}^n \to \mathbb{R}$. Sie soll die Information aus dem Zufallsexperiment aggregieren und damit zur Entscheidung H_0 versus H_1 dienen. Die reelle Zahl $T(x_1, \ldots, x_n)$ ist der Wert der Prüfgröße.

Definition 1.2 (Statistischer Test)**.** *Ein (nicht-randomisierter) statistischer Test ist eine Abbildung $\varphi : \Omega \to \{0, 1\}$. Dabei gilt die folgende Konvention:*

$$\begin{aligned} \varphi(\omega) = 1 \quad &\Longleftrightarrow \quad \textit{Nullhypothese wird verworfen, Entscheidung für } H_1, \\ \varphi(\omega) = 0 \quad &\Longleftrightarrow \quad \textit{Nullhypothese wird nicht verworfen.} \end{aligned}$$

Bei der Entscheidung H_0 versus H_1 kann man Fehler machen. Arten von Fehlern zeigt die folgende Tabelle:

	Testentscheidung	
Wahrheit	$\varphi = 0$	$\varphi = 1$
H_0	kein Fehler	Fehler 1. Art
H_1	Fehler 2. Art	kein Fehler

Table 1: Entscheidungen eines statistischen Hypothesentests

Die Schwere von Fehlentscheidungen ist asymmetrischer Natur. Der Fehler 1. Art ist schwerwiegender und daher wird seine Wahrscheinlichkeit durch eine vorher vorgegebene Schranke α, das Signifikanzniveau (z.B. $\alpha = 5\%$) begrenzt. Unter dieser Maßgabe wird dann der Test φ so konstruiert, dass die Wahrscheinlichkeit für einen Fehler 2. Art minimiert wird. Aus dieser Überlegung wird klar, dass eine statistisch abgesicherte Entscheidung (mit garantierter statistischer Sicherheit von $1-\alpha$) immer nur zugunsten von H_1 erfolgen kann. Anders ausgedrückt muss das, was bewiesen werden soll, stets als Alternativhypothese H_1 formuliert werden!

Definition 1.3 (Entscheidungsregel)**.** *Eine Entscheidungsregel für das Hypothesenpaar H_0 versus H_1 basierend auf einer Teststatistik T und einem Zufallsexperiment vom Umfang n ist von der Form*

$$\begin{aligned} T(x_1,\dots,x_n) \in \Gamma_\alpha &\quad\Rightarrow\quad \varphi(\omega) = 1, \\ T(x_1,\dots,x_n) \notin \Gamma_\alpha &\quad\Rightarrow\quad \varphi(\omega) = 0. \end{aligned}$$

Die Menge $\Gamma_\alpha \subset \mathbb{R}$ heißt Ablehnbereich und wird so bestimmt, dass gilt:

$$\mathbb{P}_{H_0}(T(X_1,\dots,X_n) \in \Gamma_\alpha) \qquad \leq \qquad \alpha, \tag{1.1}$$

$$\mathbb{P}_{H_1}(T(X_1,\dots,X_n) \in \Gamma_\alpha) \qquad \textit{maximal,} \quad \textit{gegeben dass (1.1) erfüllt ist.} \tag{1.2}$$

Der Wahrscheinlichkeitsausdruck $\mathbb{P}_{H_1}(T(X_1,\dots,X_n) \in \Gamma_\alpha)$ in (1.2) heißt Güte von φ.

Bemerkung 1.1. *Falls H_1 zusammengesetzt ist (aus mehr als einem Element besteht), so muss man sich häufig entscheiden, in welchem Punkt $\vartheta_1 \in \Theta_1$ die Güte von φ maximiert werden soll, d.h., welche Alternativen man bestmöglich erkennen möchte.*

Definition 1.4 (p-Wert)**.** *Wir schreiben abkürzend $x = (x_1,\dots,x_n)$ und nehmen an, ein Test φ mit Ablehnbereichen $\Gamma_\alpha, \alpha \in (0,1)$ sei gegeben. Der p-Wert der Beobachtung x bezüglich φ ist die Zahl*

$$p(\varphi, x) = \inf_{\{\Gamma_\alpha:\ T(x)\ \in\ \Gamma_\alpha\}} \mathbb{P}^*(T \in \Gamma_\alpha) =: \alpha_{\inf},$$

wobei das Wahrscheinlichkeitsmaß $\mathbb{P}^*$ so gewählt wird, dass $\mathbb{P}^*(T \in \Gamma_\alpha) = \sup_{\vartheta \in H_0} \mathbb{P}_\vartheta(T \in \Gamma_\alpha)$ gilt, falls H_0 zusammengesetzt ist. In Worten bezeichnet $p(\varphi, x)$ also das kleinste Signifikanzniveau $\alpha_{\inf}$, zu dem die beobachtete Datenlage x zur Ablehnung von H_0 führt.

Beispiel 1.1 (Tests vom Neyman-Pearson Typ)**.** *Ein Test vom Neyman-Pearson Typ basiert auf einer Teststatistik T, die zu größeren Werten unter Alternativen tendiert. Demnach ist der Ablehnbereich Γ_α ein einseitiges Intervall $[c_\alpha, \infty)$ und φ ist von der Form*

$$\varphi_\alpha(\omega) = \begin{cases} 1, & T(x_1, \ldots, x_n) \geq c_\alpha, \\ 0, & T(x_1, \ldots, x_n) < c_\alpha. \end{cases}$$

Falls Verteilungsdichten f_{ϑ_0} und f_{ϑ_1} von $\mathbb{P}_\vartheta$ unter H_0 und H_1 existieren und der Dichtequotient $f_{\vartheta_1}(x)/f_{\vartheta_0}(x)$ isoton in (ϑ_1, x) ist, so existiert eine Konstante c_α, die die Güte von φ_α gleichmäßig für alle $\vartheta_1 \in \Theta_1$ maximiert (Fundamentallemma von Neyman und Pearson).

Lemma 1.1. *Die Familie $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ sei so beschaffen, dass $\mathbb{P}^*$ unabhängig von α ist. Für einen Test vom Neyman-Pearson Typ gilt dann für die Berechnung des p-Wertes*

$$p(\varphi, x) = \mathbb{P}^*(T(X_1, \ldots, X_n) \geq t^* := T(x_1, \ldots, x_n)).$$

Beweis: Die Ablehnbereiche Γ_α sind von der Form $\Gamma_\alpha = [c_\alpha, \infty)$. Demnach wird $\inf\{\Gamma_\alpha : T(x) \in \Gamma_\alpha\}$ offensichtlich in dem Ablehnbereich $[t^*, \infty)$ angenommen. Aufgrund der Struktur dieses Ablehnbereiches gilt ferner $\mathbb{P}^*(T \in [t^*, \infty)) = \mathbb{P}^*(T(X_1, \ldots, X_n) \geq t^*)$. ■

Lemma 1.2. *Es herrscht eine Dualität zwischen einem Test zu einem fest vorgegebenem Niveau α und der Berechnung des p-Wertes einer Beobachtung x:*

$$\varphi_\alpha(\omega) = 1 \Longleftrightarrow p(\varphi, x) \leq \alpha.$$

Beweis: Wir beweisen das Resultat nur für Tests vom Neyman-Pearson Typ. Da die Funktion $t \mapsto \mathbb{P}^*(T(X_1, \ldots, X_n) \geq t)$ monoton fallend in t ist und wegen (1.1) und (1.2) $\mathbb{P}^*(T(x_1, \ldots, x_n) \geq c_\alpha) \leq \alpha$ sowie $\mathbb{P}^*(T(x_1, \ldots, x_n) \geq c) > \alpha$ für alle $c < c_\alpha$ gelten muss, ist $p(\varphi, x) \leq \alpha$ gleichbedeutend mit $t^* \geq c_\alpha$. Das führt bei einem Test vom Neyman-Pearson Typ aber gerade zur Ablehnung von H_0. ■

Bemerkung 1.2.

1. *Der Vorteil von p-Werten ist, dass sie unabhängig von einem a priori festgesetzten Signifikanzniveau α ausgerechnet werden können. Dies ist der Grund, warum alle gängigen Statistik-Softwaresysteme statistische Hypothesentests über die Berechnung des p-Wertes implementieren. Aus puristischer Sicht birgt das jedoch Probleme, da man mit dieser Art des Testens tricksen kann. Hält man sich nämlich nicht an die gute statistische Praxis, alle Rahmenbedingungen des Experimentes (einschließlich α!) vor Erhebung der Daten festzulegen, so kann man der Versuchung erliegen, das Signifikanzniveau erst a posteriori (nach Durchführung des Experimentes und Anschauen des resultierenden p-Wertes) zu setzen, um damit zu einer intendierten Schlussfolgerung zu kommen. Deswegen lehnen viele Statistiker die in Lemma 1.2 gezeigte Art des Testens strikt ab.*

2. *Die Interpretation des p-Wertes ist zu bedenken. Der p-Wert misst grob gesagt die Kompatibilität der beobachteten Daten mit einer vorher festgesetzten Nullhypothese. Lieber hätte man jedoch oft, die Wahrscheinlichkeit für das Vorliegen der Nullhypothese zu messen, nachdem die Daten erhoben wurden. Genau dies leistet der p-Wert jedoch* **nicht**, *obschon*

viele Praktiker fälschlicherweise zu einer solchen Interpretation neigen. Der p-Wert gibt eine Antwort auf die Frage: "Wie wahrscheinlich sind die gemessenen Daten, gegeben die Nullhypothese stimmt?" und nicht auf die Frage: "Wie wahrscheinlich ist es, dass die Nullhypothese stimmt, gegeben die erhobenen Daten?".

2. BERECHNUNGSMETHODEN FÜR P-WERTE

2.1 Der probabilistische Weg

In einigen (Standard-) Fällen ist die Verteilung von T under H_0 exakt bestimmbar. Dann läßt sich $p(\varphi, x)$ natürlich auf elementarem probabilistischen Wege berechnen. Wir illustrieren das anhand eines Beispiels.

Beispiel 2.1 (Zweiseitiger Gaußtest)**.** *Wir betrachten die Familie von Wahrscheinlichkeitsmaßen $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta} = (\mathcal{N}(\vartheta, 1))_{\vartheta\in\mathbb{R}}$, d. h., Normalverteilungen mit bekannter Varianz, die ohne Beschränkung der Allgemeinheit auf den Wert 1 gesetzt wurde. Wir interessieren uns für das Hypothesenpaar*

$$H_0 : \{\vartheta = 0\} \quad \textit{versus} \quad H_1 : \{\vartheta \neq 0\}$$

und nehmen ein Zufallsexperiment basierend auf n i.i.d. Zufallsvariablen $X_1, \ldots, X_n \sim \mathbb{P}_\vartheta$ an. Die geeignete Teststatistik für das beschriebene Testproblem ist $T = T(X_1, \ldots, X_n) = \sum_{i=1}^n X_n/n = \bar{X}_n = \hat{\vartheta}$ (der gleichmäßig beste unverzerrte Schätzer für ϑ). Wegen der Linearitätssätze für Normalverteilungen ist T ebenfalls normalverteilt mit Parametern

$$\begin{aligned} \mathbb{E}(T) &= \mathbb{E}(X_1) = \vartheta, \\ Var(T) &= \sum_{i=1}^n Var(X_i/n) = n/n^2 = \frac{1}{n}. \end{aligned}$$

Damit ist $\tilde{T} := \sqrt{n}T \underset{H_0}{\sim} \mathcal{N}(0, 1)$. Abkürzend schreiben wir außerdem wieder $T(x_1, \ldots, x_n) =: t^*$ für den Wert der Prüfgröße. Nun lässt sich der p-Wert einfach wie folgt ausrechnen:

$$\begin{aligned} p(\varphi, x) &= \mathbb{P}_{H_0}(|T| > |t^*|) \\ &= 2\mathbb{P}_{H_0}(T > |t^*|) \\ &= 2(1 - \mathbb{P}_{H_0}(T \leq |t^*|)) \\ &= 2(1 - \mathbb{P}_{H_0}(\tilde{T} \leq \sqrt{n}|t^*|)) \\ &= 2(1 - \Phi(\sqrt{n}|t^*|)), \end{aligned}$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet.

2.2 Der empirische Weg

Oftmals können selbst unter der Nullhypothese keine exakten Verteilungseigenschaften von T bestimmt werden. Damit kann der probabilistische Weg nicht beschritten werden und der p-Wert wird in einem solchen Fall durch Computersimulationen unter Verwendung sogenannter Resamplingverfahren empirisch approximiert. Auch dies soll anhand eines Beispiels (des Bootstrapverfahrens für Einstichproben-Lageprobleme) erläutert werden.

Beispiel 2.2 (Einstichproben-Bootstrap)**.** *Ähnlich wie in Beispiel 2.1 betrachten wir wieder ein Lageparametermodell bei bekannter Varianz. Allerdings sei die dem statistischen Experiment zu Grunde liegende Wahrscheinlichkeitsfamilie nun von der Form* $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta} = (F_\vartheta)_{\vartheta\in\mathbb{R}_{\geq 0}}$*, wobei* ϑ *ein Lageparameter sei, d.h., für zwei Parameter* $\vartheta_1 \neq \vartheta_2$ *aus* Θ *sei* $F_{\vartheta_1}(x-\vartheta_1) = F_{\vartheta_2}(x-\vartheta_2)$ *für alle* $x \in \mathbb{R}$*. Über die genaue Gestalt von* F *sei indes nichts bekannt, außer, dass die durch* F_ϑ *induzierte Verteilung eine existierende Varianz von (ohne Einschränkung)* 1 *für alle* $\vartheta \in \Theta$ *habe. Die zu testenden Hypothesen seien*

$$H_0 : \{\vartheta = 0\} \quad \textit{versus} \quad H_1 : \{\vartheta > 0\}. \tag{2.3}$$

Erneut ist (für i.i.d. $X_1, \ldots, X_n \sim F_\vartheta$*) die Statistik* $T = T(X_1, \ldots, X_n) = \sum_{i=1}^n X_n/n = \bar{X}_n$ *eine vernünftige Teststatistik für* H_0 *versus* H_1*, da sie die Lage der Verteilung von* $X_1 \sim F_\vartheta$ *misst, allerdings kann ihre Verteilung aufgrund des mangelnden Wissens über* F *nicht probabilistisch ermittelt werden. Ferner ist eine wahrscheinlichkeitstheoretische Approximation von* F_ϑ *gerade bei kleinem Stichprobenumfang* n *oft ungenau.*

Die Idee des sogenannten "Bootstrapverfahrens" ist nun, die aus der unzureichend reichhaltigen Datenlage resultierende geringe Präzision des Wissens über F_ϑ *durch Computersimulationen in die Höhe zu treiben. Daher rührt auch der Name "Bootstrap" (englisch für Stiefelriemen), da in der Legende von Baron Münchhausen dieser sich an seinen eigenen Stiefelriemen aus einer misslichen Situation gezogen haben soll. Der Bootstrap wird daher im deutschsprachigen Raum ab und an auch als die "Münchhausen-Methode der Statistik" bezeichnet.*

Ohne Beweis ergibt sich das folgende Verfahren für eine Schätzung des p-Wertes für das Testproblem (2.3):

Resamplingschema 2.1:

(A) Führe das Zufallsexperiment mit $X_1, \ldots, X_n$ *i.i.d. durch.*

(B) Bilde die Teststatistik
$T_n = \bar{X}_n$ *der Originalvariablen.*

(C) Generiere durch Ziehen mit Zurücklegen B *bootstrap Datensätze*
$((X^*_{1,b}, \ldots, X^*_{n,b}))_{b=1,\ldots,B}$.

(D) Berechne für $b = 1, \ldots, B$ *die bootstrap Teststatistiken*
$T_{n,b} = \sqrt{\frac{n}{n-1}} \cdot (\bar{X}^*_{n,b} - \bar{X}_n)$.

(E) Ermittle den empirischen p-Wert für das Testproblem (2.3) als

$$\frac{\#\{T_{n,b} : T_{n,b} > T_n\} + 1}{B+1}.$$

Damit der Nenner $B+1$ *in der Approximation von* $p(\varphi, x)$ *glatt wird und somit eine möglichst rundungsfehlerfreie Schätzung des p-Wertes möglich wird, sind* $B = 999$*,* $B = 4999$ *oder* $B =$

9999 *beliebte Wahlen.*

Bemerkung 2.1.

1. *Für kompliziertere Testprobleme sind raffiniertere Resamplingschemata nötig. Das Bootstrapverfahren ist mittlerweile jedoch ein Standardverfahren in der statistischen Praxis. Das Buch von Efron, dem "Erfinder" der Bootstrap-Methode und Tibshirani (siehe unten), stellt viele Anwendungsfälle dar und die jeweils adäquaten Resamplingschemata bereit.*

2. *Für Zweistichprobenprobleme sind Permutationstests dem Bootstrap häufig überlegen.*

LITERATUR

Efron, B., Tibshirani, R. J. (1993). "An introduction to the bootstrap.", Monographs on Statistics and Applied Probability. Chapman & Hall, New York.

Lehmann, E. L., Romano, J. P. (2005). "Testing statistical hypotheses. 3rd ed.", Springer Texts in Statistics. Springer, New York.

Sachs, L., Hedderich, J. (2006). "Applied statistics. Collection of methods with R. (Angewandte Statistik. Methodensammlung mit R.) 12th completely revised ed.", Springer, Berlin.

Witting, H. (1974). "Mathematische Statistik. Eine Einführung in Theorie und Methoden. 2., durchges. Aufl.", Teubner Studienbücher Mathematik. Leitfäden der angewandten Mathematik und Mechanik LAMM. Band 9.