Klassische und neue statistische Anpassungstests

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik der
Universität Karlsruhe
genehmigte

DISSERTATION

von

Dipl.-Math. Bernhard Klar
aus Riedlingen

Tag der mündlichen Prüfung: 24. Juni 1998
Referent: Prof. Dr. N. Henze
Korreferent: apl. Prof. Dr. L. Baringhaus

Inhaltsverzeichnis

Einleitung

In der induktiven Statistik werden beobachtete Daten $x_1, \ldots, x_n$ als Realisierungen von Zufallsvariablen $X_1, \ldots, X_n$ aufgefaßt. Um Schlußfolgerungen aus den Beobachtungen zu ziehen, die über eine bloße Beschreibung der Daten hinausgehen, müssen Annahmen über den zugrundeliegenden Zufallsmechanismus gemacht werden. Es ist folglich ein stochastisches Modell zu erstellen, durch das die (gemeinsame) Verteilung von $X_1, \ldots, X_n$ näher beschrieben wird. Dabei ist es intuitiv naheliegend, daß im allgemeinen um so weitreichendere Schlüsse aus den Daten gezogen werden können, je genauer das Modell die Verteilung von $X_1, \ldots, X_n$ beschreibt.

Aus wahrscheinlichkeitstheoretischer Sicht wäre es deshalb wünschenswert, die zugrundeliegende Verteilung vollständig zu spezifizieren. Handelt es sich um univariate Beobachtungen, so wäre ein mögliches Modell: $X_1, \ldots,$ X_n sind unabhängig und identisch verteilt; die gemeinsame Verteilung ist eine Normalverteilung mit Erwartungswert $\mu = 0$ und Varianz $\sigma^2 = 1$. Man schreibt dafür

$$(X_j)_{j=1,\ldots,n} \quad \overset{uiv}{\sim} \quad \mathcal{N}(0,1).$$

Unter diesen Annahmen können dann alle interessierenden Größen mit wahrscheinlichkeitstheoretischen Hilfsmitteln bestimmt werden; statistische Methoden sind überflüssig.

Die Kehrseite dieser Vorgehensweise ist offensichtlich: Je genauer man ein Modell festlegt, desto größer ist die Gefahr, daß die Daten nicht mit dem Modell vereinbar sind. Folgerungen aus einem Modell, das mit den Beobachtungen nicht verträglich ist, können zumindest für die Praxis wertlos sein.

Die Möglichkeit eines Modellfehlers läßt sich ausschließen, indem man nur völlig unzweifelhafte Annahmen in das Modell einfließen läßt, etwa die Annahme von $\mathbb{N}_0$-wertigen Zufallsvariablen, wenn die Beobachtungen aus Zähldaten bestehen. Mit einem solchen Modell (das diesen Namen kaum noch verdient) wird man jedoch nicht viel anfangen können.

Der Ausweg, den die Statistik aus diesem Dilemma findet, ist ein Kompromiß: man legt einen Modell*rahmen* fest, der irgendwo zwischen den beiden oben beschriebenen Extremen liegt, und überprüft mit statistischen Methoden, ob dieser Rahmen mit den Daten verträglich ist. Wenn dies bejaht werden kann, werden aus den Daten und dem Modell Schlußfolgerungen gezogen, die über die reine Datenbeschreibung hinausgehen. Der zweite Schritt, das Überprüfen der Verträglichkeit von Modell und Beobachtungen mit Hilfe statistischer Anpassungstests, ist der Gegenstand der vorliegenden Arbeit. Dabei nehmen wir an, daß immer die Modellannahme

$$X_1, \ldots, X_n \quad \text{sind stochastisch unabhängig und identisch verteilt}$$

erfüllt ist. Diese Voraussetzung ist in der Praxis bei kontrollierten Experimenten oft gegeben; in anderen Fällen ist sie durch geeignete Auswahl der Daten zumindest approximativ erfüllbar.

In der Literatur wird unter dem Begriff *Anpassungstest* oft folgende spezielle Situation verstanden (so z.B. in Kendall and Stuart (1979), Kap. 30): den Daten liegt als Grundannahme ein parametrisches Modell zugrunde. Ist ϑ ein endlichdimensionaler Vektor und $\{P_\vartheta : \vartheta \in \Theta\}$ eine parametrische Verteilungsfamilie, so dient ein statistischer Anpassungstest dazu, die Hypothese

$$\mathcal{H}_0 \quad : \quad P \in \{P_\vartheta : \vartheta \in \Theta\}$$

zu überprüfen. Dabei bezeichnet P die den Daten zugrundeliegende unbekannte Wahrscheinlichkeitsverteilung. Das wichtigste Beispiel für eine parametrische Verteilungsfamilie ist die Normalverteilungsklasse

$$\{\mathcal{N}(\mu, \sigma^2) \ : \ \mu \in \mathbb{R}, \ \sigma^2 > 0\}.$$

Im Unterschied zum obigen Fall einer Normalverteilung mit bekannten Parametern hat man es also nicht mit einer einfachen, sondern mit einer zusammengesetzten Hypothese zu tun.

Liegt die Situation einer zusammengesetzten Hypothese vor, so muß erst ein geeigneter Schätzer $\hat{\vartheta}_n$ für ϑ definiert werden; danach benötigt man ein Vergleichskriterium, das in geeigneter Weise überprüft, ob die vorliegenden Daten aus der Verteilung $P_{\hat{\vartheta}_n}$ erzeugt sein können.

Sind die Daten univariat, so ist die Verteilung P_ϑ durch die zugehörige Verteilungsfunktion $F(\cdot;\vartheta)$ bestimmt. In diesem Fall kann z.B. mit Hilfe der Kolmogoroff-Smirnoff-Statistik

$$K_n \quad = \quad \sqrt{n} \sup_{-\infty<t<\infty} \mid F_n(t) - F(t;\hat{\vartheta}_n) \mid$$

ein geeigneter Test konstruiert werden. Hier ist $F_n(\cdot)$ die die durch

$$F_n(t) \quad = \quad \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\{X_j \le t\}, \quad t \in \mathbb{R},$$

definierte empirische Verteilungsfunktion.

Im Falle der Normalverteilung kann man als Schätzer für μ und σ^2 den Stichprobenmittelwert $\bar{X}_n$ und die empirische Varianz $\hat{\sigma}_n^2$ verwenden; K_n nimmt dann die Form

$$K_n \quad = \quad \sqrt{n} \sup_{-\infty<t<\infty} \left| F_n\left(t\right) - \Phi\left(\frac{t-\bar{X}_n}{\hat{\sigma}_n}\right)\right|$$

an, wobei $\Phi(\cdot)$ die Verteilungsfunktion der Standardnormalverteilung bezeichnet.

Eine wünschenswerte Eigenschaft, die die Kolmogoroff-Teststatistik besitzt, ist die *globale Konsistenz*: Jede feste Alternative zur hypothetischen Verteilungsklasse wird mit gegen 1 gehender Wahrscheinlichkeit erkannt, wenn der Stichprobenumfang gegen Unendlich strebt. Solche Tests werden *omnibus*-Tests genannt. Weitergehende theoretische Aussagen über das Güteverhalten sind nur schwer zu machen. Da die Alternative nichtparametrisch ist, existiert kein Test, der die größte Güte gegenüber jeder Alternative besitzt. In eingeschränkteren Situationen wie dem Testen einer einfachen

Hypothese gegen Lagealternativen läßt sich sogar zeigen, daß zu einem bestimmten Test meist eine Verteilungsfamilie existiert, für die der Test gewisse lokale asymptotische Optimalitätseigenschaften besitzt (Nikitin (1995), Kap. 6). Um so wichtiger ist es, das Güteverhalten eines Tests gegenüber einer möglichst großen Zahl von Alternativen durch Simulationsstudien bei in der Praxis auftretenden Stichprobenumfängen zu untersuchen.

Als Minimalforderung an einen Anpassungstest bleibt die globale Konsistenz bestehen. In den letzten Jahren hat sich aber gezeigt, daß auch solche Tests ihre Güte im Raum aller Alternativen sehr ungleichmäßig verteilen (siehe etwa Janssen (1995)). Es existieren nur wenige Richtungen von Abweichungen von der hypothetischen Verteilung, gegen die ein solcher Test eine hohe Güte besitzt. Insofern verhalten sich auch omnibus-Tests zumindest bei kleinen und mittleren Stichprobenumfängen ähnlich wie Tests, die von vornherein nur zur Aufdeckung bestimmter Alternativen konstruiert wurden.

Aus diesem Grunde werden in der Praxis häufig Tests verwendet, die zwar nicht global konsistent sind, aber dafür andere positive Eigenschaften besitzen. Bei Anwendern sind insbesondere sogenannte *diagnostische Tests* beliebt, denen die Eigenschaft zugeschrieben wird, nicht nur eine Hypothese ablehnen zu können, sondern gleichzeitig die Art der Abweichung zu diagnostizieren. Ein Beispiel hierfür ist der Dispersionsindex-Test, der einer der am häufigsten verwendeten Tests auf Poissonverteilung ist. Bei der Poissonverteilung ist der Quotient aus Varianz und Erwartungswert immer gleich 1. Ist nun der als Dispersionsindex bezeichnete Quotient aus empirischer Varianz und arithmetischem Mittel nicht „hinreichend nahe“ bei 1, so kann die Hypothese der Poissonverteilung (auf einem bestimmten Signifikanzniveau) abgelehnt werden. Darüber hinaus wird diagnostiziert, daß das hypothetische Modell deshalb verworfen werden muß, weil Erwartungswert und Varianz der zugrundeliegenden wahren Verteilung nicht übereinstimmen.

Daß die konventionelle Anwendung des Dispersionsindex-Tests nicht zu dieser Diagnose berechtigt, wurde in Henze und Klar (1996) gezeigt. Der Grundfehler liegt darin, daß Aussagen über nichtparametrische Verteilungsgrößen mittels einer bestimmten parametrischen Verteilungsklasse beurteilt werden. In Henze (1997a) und Henze und Klar (1996) wurde auch gezeigt, wie man diesen und ähnliche Tests modifizieren muß, um Verfahren zu erhalten, die zumindest bei großen Stichprobenumfängen diagnostische Eigenschaften besitzen. Die richtige Vorgehensweise ist, eine nichtparametrische Hypothese zu formulieren, die die gewünschten Beziehungen beinhaltet. Im Falle des Dispersionsindex-Tests wäre die Hypothese die Gleichheit von Erwartungswert und Varianz. Wird diese nichtparametrische Hypothese auf einem gewissen Signifikanzniveau abgelehnt, so kann man folgern, daß den Daten eine Verteilung zugrundeliegt, deren Erwartungswert von der Varianz verschieden ist.

Gemäß den angesprochenen beiden unterschiedlichen Problemstellungen besitzt die vorliegende Arbeit zwei Schwerpunkte. Im ersten Kapitel werden Tests untersucht, die als Anpassungstests im engeren Sinne verwendet werden können, die sich aber auch in natürlicher Weise zu Tests mit diagnostischen Eigenschaften erweitern lassen. Diese sogenannten glatten Anpassungstests (engl. „smooth tests of fit“) wurden schon 1937 von Jerzy Neyman eingeführt und erfreuen sich insbesondere in letzter Zeit wachsender Beliebtheit.

Das zweite Kapitel behandelt dagegen neue, auf der integrierten Verteilungsfunktion basierende Tests, die in den oben skizzierten engeren Rahmen passen. Neben der Annahme von unabhängig und identisch verteilten Daten setzen sie univariate Beobachtungen voraus; diese können vom diskreten oder stetigen Typ sein.

Im ersten Abschnitt von Kapitel 1 werden glatte Anpassungstests zum Testen der Hypothese $P \in \{P_\vartheta : \vartheta \in \Theta\}$ eingeführt, die auf Teststatistiken der Form

$$\hat{\Psi}^2_{n,k} = \sum_{j=s+1}^{s+k} \hat{U}^2_{n,j} \quad \text{mit} \quad \hat{U}_{n,j} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h_j(X_i; \hat{\vartheta}_n) \tag{0.1}$$

beruhen. Hierbei ist $\{h_0(\cdot;\vartheta) \equiv 1,\ h_1(\cdot;\vartheta),\ h_2(\cdot;\vartheta), \ldots\}$ ein bezüglich P_ϑ

orthonormales Polynomsystem, es gelten also die Relationen

$$\int h_j(\cdot;\vartheta)h_l(\cdot;\vartheta)\,dP_\vartheta \quad = \quad \delta_{j,l} \quad (0 \le j,l \le s+k). \tag{0.2}$$

Dabei sind die Polynome $h_j(\cdot;\vartheta)$ vom Grade j ($j \ge 0$), weshalb die Terme $\hat{U}_{n,j}$ als Komponenten j-ten Grades bezeichnet werden. Setzt man in (0.2) $l = 0$, so ergibt sich für $j \ge 1$

$$\int h_j(\cdot;\vartheta)\,dP_\vartheta \quad = \quad E_\vartheta[h_j(X_1;\vartheta))] \;=\; 0,$$

wodurch eine unter P_ϑ gültige Relation zwischen den ersten j Momenten von X_1 beschrieben wird.

Im Falle der Poissonverteilung mit Parameter ϑ bilden die Poisson-Charlier-Polynome das Orthonormalsystem; es gilt insbesondere

$$h_1(x,\vartheta) = \frac{x-\vartheta}{\sqrt{\vartheta}}, \quad h_2(x,\vartheta) = \frac{(x-\vartheta)^2 - x}{\sqrt{2}\,\vartheta}.$$

Die Gleichung $E_\vartheta[h_j(X_1;\vartheta))] \;=\; 0$ sagt für $j = 1$ aus, daß der Erwartungswert der Poissonverteilung gerade gleich ϑ ist, während sie für $j = 2$ die Gleichheit von Erwartungswert und Varianz beschreibt. Da die erste Komponente $\hat{U}_{n,1}$ wegen

$$\sum_{i=1}^{n} h_1(X_i,\hat{\vartheta}_n) \quad = \quad \frac{1}{\sqrt{\hat{\vartheta}_n}}\left(\sum_{i=1}^{n} X_i - n\bar{X}_n\right) \;=\; 0$$

verschwindet, setzt man in (0.1) $s = 1$. Weiter gilt hier

$$\hat{\Psi}^2_{n,2} \quad = \quad \frac{1}{2n}(D_n - n)^2,$$

wobei D_n den oben beschriebenen Dispersionsindex bezeichnet. Der auf der ersten nichtverschwindenden Komponente $\hat{U}_{n,2}$ beruhende glatte Anpassungstest auf Poissonverteilung und der Dispersionsindex-Test sind folglich äquivalent.

Nach der Darstellung der allgemeinen Verteilungstheorie der glatten Anpassungstests wird im ersten Abschnitt weiter gezeigt, wie diese modifiziert werden müssen, um Tests mit diagnostischen Eigenschaften zu erhalten.

Verteilungsklassen, in denen diese Modifikation besonders einfach zu bewerkstelligen ist, sind Thema des zweiten Abschnitts; dabei handelt es sich um spezielle *Exponentialfamilien*.

Der dritten Abschnitt behandelt glatte Anpassungstests für die multivariate Normalverteilung. Für diese mit weitem Abstand wichtigste Verteilung im Mehrdimensionalen wurden in der Literatur eine Vielzahl von Verteilungsgrößen und Teststatistiken wie etwa multivariate Schiefe und Wölbung eingeführt, die im Zusammenhang mit Komponenten der glatten Anpassungstests stehen und die mit Hilfe der Resultate des ersten Abschnitts einheitlich behandelt werden können. Verteilungsaussagen für die multivariate Normalverteilung können in vielen Fällen auf die nichtparametrische Klasse der elliptisch-symmetrischen Verteilungen verallgemeinert werden.

In einem Unterabschnitt werden einfache Ausdrücke für höhere Komponenten der multivariaten Normalverteilung hergeleitet, die die Anwendung der Tests in der Praxis überhaupt erst ermöglichen. In der Literatur sind solche Ausdrücke nur für auf den beiden ersten nichtverschwindenden Komponenten beruhende Tests zu finden.

Da die allgemeinen Ergebnisse im ersten Abschnitt teilweise ohne die Voraussetzung auskommen, daß die Polynome h_j ein Orthogonalsystem bilden, können auch Testgrößen behandelt werden, die eine Darstellung wie in (0.1) besitzen, wobei die h_j nicht orthogonal sind. Dieser Umstand wird in Abschnitt 1.4 benutzt, um die asymptotische Verteilung von Verallgemeinerungen der multivariaten Schiefe und Wölbung für die Klasse der elliptisch-symmetrischen Verteilungen zu bestimmen.

Diagnostische Tests für die bivariate Poissonverteilung sind Gegenstand des fünften Abschnitts, wobei auch hier die Schwierigkeit besteht, daß kein orthogonales Polynomsystem mehr definiert werden kann.

Werden glatte Anpassungstests als diagnostische Tests verwendet, so ist es außer im Falle der im zweiten Abschnitt behandelten univariaten Exponentialfamilien aufwendig, die asymptotische Verteilung der Teststatistiken konkret zu berechnen. Kapitel 1 schließt daher mit einem Abschnitt über die Bestimmung kritischer Werte für die betrachteten Tests mit Hilfe von Resampling-Verfahren; dabei werden aus der ursprünglichen Stichprobe weitere Stichproben erzeugt, mit deren Hilfe die interessierenden Größen approximativ bestimmt werden können.

Das zweite Kapitel behandelt Anpassungstests, welche auf der integrierten Verteilungsfunktion basieren. Solche Verfahren wurden unseres Wissens nach bisher nicht betrachtet. Für eine auf $\mathbb{R}_+$ konzentrierte Zufallsvariable X mit Verteilungsfunktion $F_X(\cdot)$ kann die *integrierte Verteilungsfunktion* Ψ_X durch

$$\Psi_X(t) \quad = \quad \int_t^\infty (1 - F_X(x))\, dx \quad (t \geq 0)$$

definiert werden. Es werden neue Teststatistiken vorgeschlagen, die auf der Differenz zwischen der integrierten Verteilungsfunktion und ihrem empirischen Gegenstück aufbauen.

Im ersten Abschnitt wird dabei der Fall diskreter Zufallsvariablen untersucht. Neben der asymptotischen Verteilung der Testgröße unter der Hypothese wird auch das Verhalten unter lokalen Alternativen besprochen. Es folgt ein längerer Unterabschnitt mit Simulationsergebnissen und einem Anwendungsbeispiel.

Im zweiten Abschnitt werden auf der integrierten Verteilungsfunktion basierende Anpassungstests für die beiden wichtigsten stetigen Verteilungen, die Exponential- und die Normalverteilung, vorgestellt. Auch für diese Tests wird das Güteverhalten durch Simulationsstudien illustriert.

Ich möchte mich an dieser Stelle bei Herrn Prof. Dr. N. Henze für seine Unterstützung während der Anfertigung dieser Arbeit sehr herzlich bedanken. Er hat ihren Werdegang mit zahlreichen wertvollen Hinweisen und anregenden Diskussionen begleitet.

Ebenso gilt mein Dank Herrn Prof. Dr. L. Baringhaus für die freundliche Übernahme des Korreferats.