Stochastik in der Schule 14 (1994), Nr. 3

Über Extremwerte lernen

von Stuart, G Coles, Nottingham, übersetzt (aus Teaching Statistics, volume 16, number 1, S.23-25) und bearbeitet von Gerhard König, Karlsruhe

Zusammenfassung: Bei vielen physikalischen Prozessen und Beobachtungen ist die Verteilung der Extremwerte von großer Bedeutung. Dieser Artikel führt in die Problematik ein, diskutiert Verteilungsmodelle und Grenzverteilungsfunktionen, wie z.B. die Gumbel-Verteilung.
ZDM Klassifikation: K60

1. Einleitung

Betrachte eine Stichprobe $X_1,...,X_n$ aus einer Population mit gemeinsamer Verteilungsfunktion. Wie lautet die Verteilungsfunktion des Stichprobenmittelwertes X? Bekanntermaßen ist dies unter bestimmten Bedingungen die Normalverteilung; n muß nur genügend groß sein, und die zugrundeliegende Verteilungsfunktion muß weder bekannt sein noch muß sie eine Normalverteilung sein.

Wie lautet die Antwort, wenn nach der Verteilungsfunktion von $X_{max} = \max(X_1,...,X_n)$ gefragt wird? Dieser Frage liegen praktische Problemstellungen zugrunde. Wasserstände, Temperaturen, Luftverschmutzung, Ozon-Werte, Regenhöhen in einem Ort sind Beispiele. Tägliche Messungen der Wasserstände ergeben n Beobachtungen im Jahr, genauso bei Ozonwerten oder schweren Regenfällen. Wie verteilen sich z.B. die Pegelhöchststände von Flüssen bei starken Regenfällen oder bei starker Schneeschmelze? In vielen Fällen ist das Verhalten des Extremwertes, z.B. in einem Jahr, von Interesse. Wie verteilen sich nun die Extremwerte über die einzelnen Jahre?

2. Genaue Verteilung von X_{max}

Prinzipiell ist es leicht, einen Ausdruck für die Verteilungsfunktion X_{max} mit Hilfe der zugrundeliegenden Verteilungsfunktion F aufzuschreiben:

$X_1, X_2, ..., X_n$ seien unabhängige zufällige Variable mit den Verteilungsfunktionen $F_1, F_2, ..., F_n$, und es sei $M = \max(X_1, X_2, ..., X_n)$.
Man bestimme die Verteilungsfunktionen von M.

Wir erhalten für jedes X :

$$\begin{aligned} F_{max}(x) = P(M \le x) &= P(X_1 \le x;\ X_2 \le x\ ;\ ...\ ;\ X_n \le x) \\ &= P(X_1 \le x) \cdot P(X_2 \le x) \cdots P(X_n \le x) \\ &= F_1(x) \cdot F_2(x) \cdots F_n(x). \end{aligned}$$

Wenn speziell die F_j alle gleich sind, dann gilt

$$F_{max}(x) = (F(x))^n . \qquad (1)$$

Das sieht einfach aus, ist aber schwer zu benutzen aus folgenden Gründen:

1. Der Ausdruck (1) ist i.allg. eine analytisch kompliziert dargestellte Funktion (z.B. wenn X die Normalverteilung ist).
2. F ist nicht bekannt oder ist bekannt, aber die Parameter, wie z.B. Mittelwert oder Streuung, müssen geschätzt werden.

Aus diesen beiden Gründen werden in der Praxis Schätzungen und Näherungen von X_{max} betrachtet, wie im folgenden diskutiert werden soll.

3. Grenzverteilungsfunktion von X_{max}.

Was ist die Grenzverteilung von X_{max} für $n \to \infty$? Ist es die Normalverteilung? Folgender Algorithmus beantwortet uns diese Frage.

1. Wähle eine geeignete Verteilungsfunktion F und Stichprobengröße n
2. Simuliere eine Zufallsstichprobe $X_1, ..., X_n$ aus der gewählten Verteilung
3. Bestimme den Wert X_{max} aus der Stichprobe
4. Wiederhole Schritte 2 und 3 m-mal
5. Analysiere die m simulierten Werte von X_{max}

Beispiele: Pegelstände; es gebe m Beobachtungen im Jahr und n sei die Anzahl der Jahre, deren jährliche Maxima wir analysieren wollen. Mit Hilfe von Zufallsziffern, aus Tabellen oder einem Zufallsgenerator, kann dies simuliert werden. Da m n die Gesamtzahl der simulierten Beobachtungen ist, lassen sich Zufallsziffern nur benutzen, wenn m und n klein sind. X_{max} soll aber für große n beobachtet werden. Welche Methode man auch benutzt: man geht von einer Reihe Beobachtungen $U_1,...,U_n$ aus mit gleicher Verteilung F auf[0,1]. Die Transformation $X_i = F^{-1}(U)$ gibt eine Stichprobe $X_1,...,X_n$ mit der gewünschten Verteilungsfunktion F. Dies gilt, weil:

$$P(F^{-1}(U) \le x) = P(U \le F(x)) = F(x)$$

Visualisieren wir dieses Beispiel! Abb. 1 zeigt uns Histogramme von m=100 Werten von X_{max} aus Stichproben verschiedener Größe, die mittels Normalverteilungen simuliert werden. Bei einer Stichprobengröße n = 10 läßt wenig darauf schließen, daß die Verteilung X_{max} nicht normalverteilt ist. Mit wachsendem n wird jedoch anschaulich klar, daß X_{max} sicherlich nicht normalverteilt ist, sondern rechtsschief ist. Damit soll aber auch der Glaube erschüttert werden, daß

praktisch alle stetigen Verteilungen, besonders ihre Grenzwerte, normalverteilt sind.

Noch deutlicher sieht man,, daß X_{max} nicht normalverteilt ist, wenn eine Normalverteilung graphisch dargestellt wird. Auf speziellem Papier geht man wie folgt vor: Wenn eine Stichprobe $x_1,...,x_n$ in absteigender Folge sortiert ist, dann sollte eine graphische Darstellung der x_i gegen $G^{-1}\{(i - 0.5)/m\}$, mit G als Standard Normalverteilung, ungefähr linear sein. Der Grund dafür ist, daß die empirische Verteilungsfunktion der Daten mit der Verteilungsfunktion selbst vergleichbar sein sollte. Abweichungen von der Linearität suggerieren, daß das Modell der Normalverteilung falsch ist. Weiterhin gilt, wenn sich eine Gerade gut anpassen läßt, würden die Steigung und der Achsenabschnitt der Geraden Näherungen für μ und σ geben.

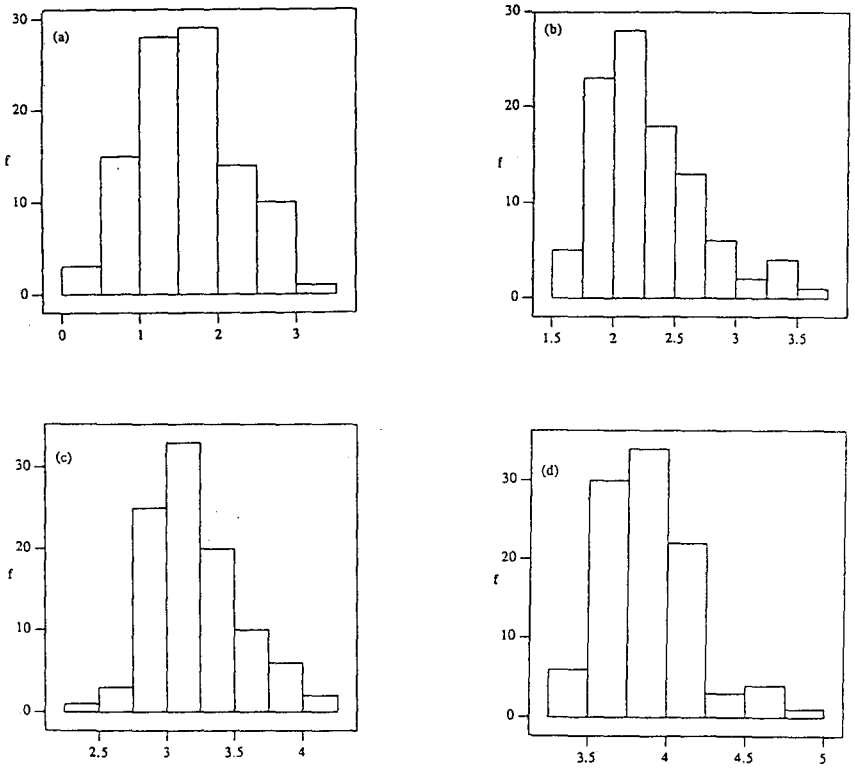

Abb. 1: Histogramme der Maxima von simulierten Stichproben einer Standard-Normalverteilung. Stichproben (a)10, (b)100, (c)1000, (d)10.000

Abb. 2 zeigt graphische Darstellungen für jede der oben diskutierten Mengen der x_{max}. Die Darstellung für n =10 scheint linear zu sein, aber für große n erscheint ein Knick. Das illustriert uns wiederum, daß die Verteilung der x_{max} bei steigender Stichprobenzahl nicht gegen die Normalverteilung strebt.

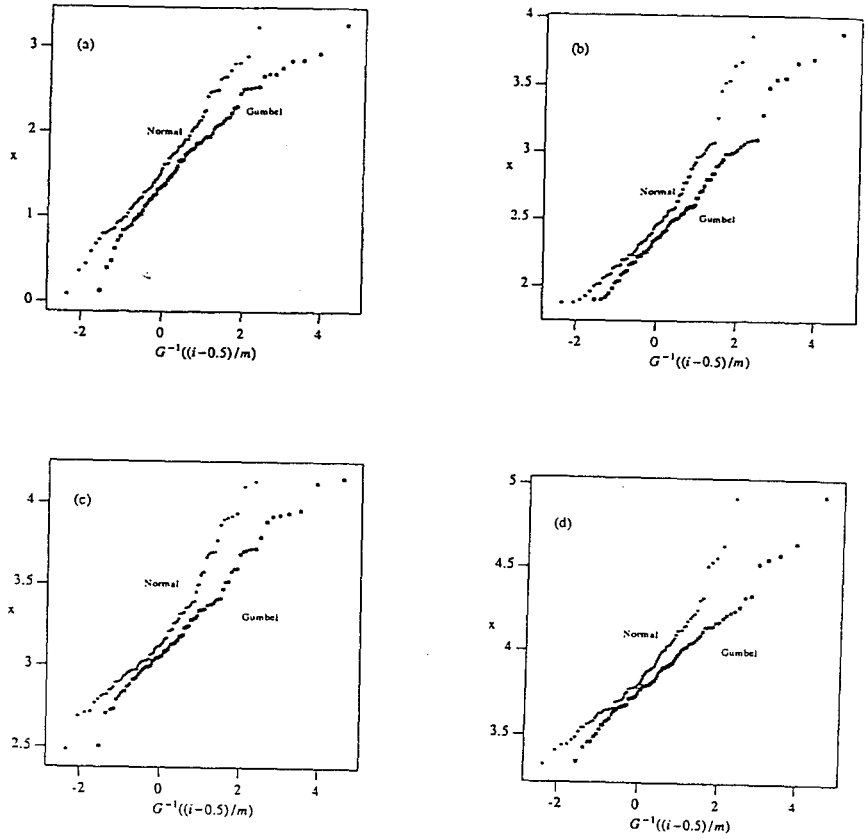

Abb.2 Wahrscheinlichkeiten der simulierten Stichprobenmaxima unter der Annahme von Normal-bzw. Gumbelverteilung. Stichprobengrößen: (a)10, (b)100, (c)1.000, (d)10.000

4. Die Gumbel Verteilung

Daß die Extremwerte X_{max} nicht normalverteilt sind, ist nicht weiter überraschend. Eine genauere mathematische Analyse von (1) zeigt, daß für große n die Verteilungsfunktion von X_{max} der sog. Gumbelverteilung gehorcht:

$$G(x) = \exp\{-\exp(-(\frac{x-a}{b}))\} \qquad (2)$$

wobei a und b Lage- bzw. Streuungsparameter sind, ähnlich μ und σ bei der Normalverteilung (s.Gumbel 1958,1962). Unterstützen die simulierten Daten diese Aussage? Betrachten wir Fig.2, wo diesmal die geordnete Stichprobe gegen $G^{-1}(i/(m+1)$ = -log(- log(i/m+1))) aufgetragen ist. Die Standard-Normalverteilung wurde ersetzt durch die standardisierte Verteilung von Gumbel (a=0, b=1 in Gleichung (2)).

Wenn die Gumbel-Verteilung zutrifft, würde diese graphische Darstellung ungefähr linear sein, und die Steigung und der Achsenabschnitt würden Schätzungen für a und b darstellen. Mit größerer Stichprobe scheint diese Annahme wahr zu werden, sodaß in der Tat die Gumbel-Verteilung eine Grenzwertfunktion des Maximums darzustellen scheint.

Eines ist aber doch zu beobachten: obwohl die Wahrscheinlichkeitsdarstellungen von Gumbel etwa linear sind, scheinen große Werte von X_{max} doch von der Geraden abzuweichen. Dies legt den Schluß nahe, daß für große Stichproben die Gumbel-Verteilung keine geeigneten Wahrscheinlichkeiten für die Extremwerte liefert. Das ist ein Problem, da gerade solche Ereignisse von großer Bedeutung für die Praxis sind.

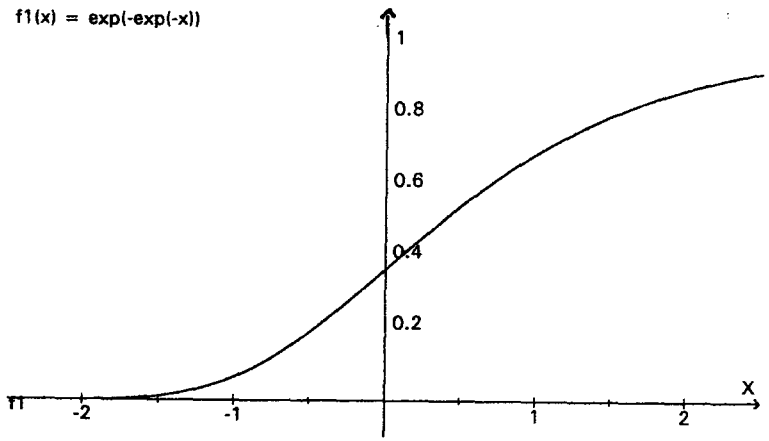

Abb. 3: Die Gumbelverteilung

Eine noch genauere mathematische Analyse zeigt, daß die Gumbel-Verteilung nur eine spezielle Grenzverteilung für Extremwerte ist. Sie gehört zur Familie der Extremwertverteilungen und Gnedenko zeigte, daß es drei Grenzverteilungen gibt, gegen die die Extremwertverteilungen konvergieren können.(s.Galambos 1987, Gnedenko 1943)

5. Ergänzung: Verteilung von X_{min}

Wie müssen die Überlegungen in 2. modifiziert werden, um die Verteilung von X_{min}, dem Minimum einer Stichprobe, zu erhalten. Auch diese Werte haben eine Bedeutung, wie die Beispiele minimale Temperaturen oder die Wasserversorgung zeigen .

Für die Behandlung des Minimums empfiehlt sich die Einführung der sogenannten Restverteilung G_j, die jedem F_j wie folgt entspricht:

$$G_j(x) = P(X_j>x) = 1-F_j(x).$$

Wenn wir nun mit $S_j = (x_j, \infty)$ eine analoge Formel anwenden, erhalten wir

$$\begin{aligned} G_{min}(x) &= P(m>x) = P(X_1>x\ ;\ X_2>x\ ;\ \ldots\ ;\ X_n>x) \\ &= P(X_1>x)\ P(X_2>x) \ldots P(X_n>x) \\ &= G_1(x)\ G_2(x) \ldots G_n(x). \end{aligned}$$

Daher ist $F_{min}(x) = 1-G_1(x)\ G_2(x) \ldots G_n(x).$

Sind die F_j alle gleich, dann wird daraus: $G_{min}(x) = G(x)^n$, $F_{min}(x) = 1-G(x)^n$.

Literatur

Buishand, T.A.:Statistics of extremes in climatology. In:Statistica Neerlandica 43(1989),nr.1, S.1-30

Galambos, J.: The asymptotic theory of extreme order statistics. Malabar, Florida: Robert E. Krieger Publishing Company, 1987

Gnedenko, B.: Sur la distribution limite du terme maximume d'une serie aleatoire. In: Annals of Mathematics, Second series 44(1943), S.389-410

Gumbel,E.: Statistics of extremes. New York: Columbia University Press: 1958,1962 (*Dies ist ein Standardwerk über die Stochastik der Extremwerte)*

Phien, H; N; Debrata, P.: A Poisson Process for maximum rainfall analysis. In.: Int.J.Math.Educ.Sci.Technol.,1982, vol.13,no.1,S.117-123