IST ALLES NORMALVERTEILT ?

nach Oliver D. Anderson

Originaltitel in "Teaching Statistics" Vol. 11 (1989) Nr. 1:
Teaching Statistics in US - An Alien's-eye View

Bearbeitung: Hans-Joachim Bentz

Kurzfassung: Während die Verteilung von beobachteten Variablen oft doch recht deutlich von der Normalverteilung abweicht, ist letztere meist für Mittelwerte in guter Näherung adäquat. Dies wird in der Beurteilenden Statistik, etwa für Vertrauensintervalle ausgenützt.

1. Normalverteilung von beobachteten Variablen

Ich kam über die Mathematik zur Angewandten Statistik. Alles wurde als normal verteilt angenommen. "Begründet" wurde dies durch einen vagen Bezug auf organisches Wachstum, das einer Fülle von zufälligen Einflüssen ausgesetzt sei, was nahe legt, daß jede Quantität in Wirklichkeit die Summe einer großen Zahl von unabhängigen Zufallsvariablen ist, und daher, wegen des Zentralen Grenzverteilungssatzes eben normal verteilt sein muß.

Dies beunruhigte mich eine Zeit lang, denn es schien doch klar, daß die meisten Objekte auf den verschiedensten Skalen gemessen werden können. Normalverteilung der Messungen auf der einen Skala schließen nun aber diese Verteilung auf einer anderen, nicht davon linear abhängigen Skala aus. Angenommen, die Körpergröße wäre normal verteilt, dann könnte es doch Gewicht nicht sein, sogar wenn man bedenkt, daß

Gewicht proportional (Körpergröße)^3

offensichtlich eine allzu vereinfachte Beziehung darstellt.

[Zusatz des Bearbeiters:] Viele Variablen, die man so gemeinhin beobachtet, sind daher gar nicht (besonders gut) normalverteilt. Das beste, was man einer realen Verteilung zusprechen kann, ist, daß sie eingipfelig und nicht allzu schief ist. Der Zentrale Grenzverteilungssatz wird, was Einzelvariable anbelangt, allzu sehr strapaziert.

2. Normalverteilung für die Verteilung von Mittelwerten

Es brauchte eine gewisse Zeit, bis ich erkannte, daß man in der Praxis häufig mit Mittelwerten aus Stichproben zu tun hat. Für ausreichend große Stichprobenumfänge können (und werden) sowohl durchschnittliche Körpergröße als auch durchschnittliches Körpergewicht normal verteilt sein - wenn man sich, nun korrekt, auf den Zentralen Grenzverteilungssatz beruft. Klar, daß damit auch sicher gestellt sein muß, daß die Einzelbeobachtungen unabhängig voneinander sind.

Die Ausgangsannahme ist daher nicht

$$X \sim N(\mu, \sigma^2),$$

d.h., die Variable X ist normal verteilt mit Parametern μ und σ^2, sondern

$$X \sim ?(\mu, \sigma^2),$$

wobei ? nun für eine unbekannte Verteilung der Variablen X steht. Das hat zur Folge, daß

$$\bar{X} = \Sigma X/n \approx N(\mu, \sigma^2/n),$$

für einen ausreichend großen Stichprobenumfang n. (Um Ausnahmefälle zu vermeiden, muß man von X annehmen, daß Erwartungswert und Varianz existieren.)

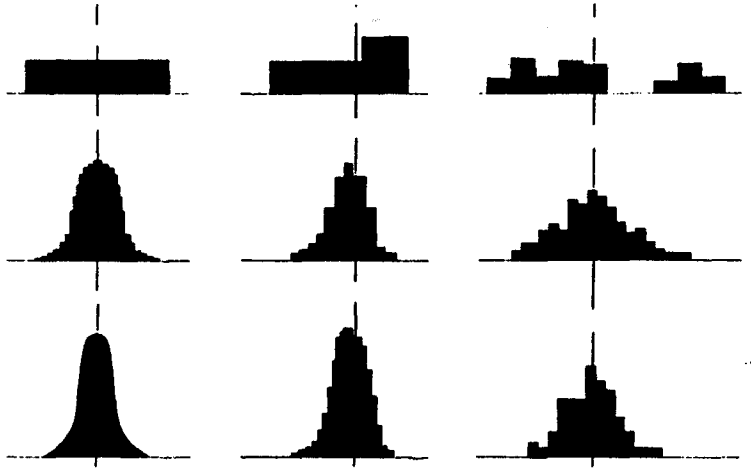
Was hinreichend großes n heißt, hängt vom Grad der Abweichung von der Normalverteilung für die Ausgangsverteilung von X ab. Diese Abweichungen werden üblicherweise grob durch Inspektion der Häufigkeitstabelle oder eines Histogramms beurteilt, die Unabhängigkeit der Beobachtungen dagegen wird durch informelles Wissen über das Zustandekommen der Daten eingeschätzt.

3. Wozu braucht man, daß Mittelwerte normalverteilt sind, und, wann sind sie dies

[Dies ist ein Zusatz des Bearbeiters.] Was die Verteilung der Mittelwerte $\bar{X}$, von der man üblicherweise einen einzigen Wert kennt, anbelangt, so kann man zu Recht sagen, daß sie, ziemlich unbhängig von der ursprünglichen Verteilung der Einzeldaten X in guter Näherung eine Normalverteilung ist. Hier "regiert" der Zentrale Grenzverteilungssatz wirklich.

Ist X eingipfelig, so ist mitunter n=5 (oder noch kleiner) schon ausreichend, es gilt: je symmetrischer X, desto kleiner n; zerfällt die Verteilung von X in mehrere Teile, so wird die Normalisierung der Verteilung von $\bar{X}$ erst bei viel größerem n eintreten.

Fig.: Normalisierung der Verteilung der Mittelwerte bei unterschiedlichen Ausgangsverteilungen



Man beachte, daß man für gewöhnlich wohl über die Verteilung von X mehrere Daten verfügt, aus der Verteilung von $\bar{X}$ aus dem vorhandenen Datensatz über nur einen Wert verfügt. Es ist für gewöhnlich sehr aufschlußreich, mit einem PC die Datensätze von X mit dem jeweiligen Mittelwert $\bar{x}$ und somit die Verteilung von

$\bar{X}$ zu simulieren. Die unterschiedliche Normalisierung von $\bar{X}$ kann daraus schön beobachtet werden. Ferner erhält man durch die Simulationsstudie Einblick in die Zuverlässigkeit der Information, die im Stichprobenmittel steckt. Das würde, formalisiert, zur Methode der Vertrauensintervalle führen (siehe Borovcnik, 1987).

Mittelwerte sind also weitgehend normalverteilt und das ist gut so. Sonst würde die Methode der Vertrauensintervalle nicht so einfach sein. (Die Komplexität der Berechnungen würde das Verständnis des Konzepts verschleiern.) Was die Vertrauensintervalle in der Regel jedoch unzuverlässig macht, ist die Schätzung der Standardabweichung σ durch s aus der Stichprobe. Diese ist sehr empfindlich hinsichtlich Abweichungen der ursprünglichen Verteilung, der Verteilung der Daten von X, von der Normalverteilung: Ausreißer, größere Ausläufer, Zerfallen der Verteilung in mehrere Teile etc. (siehe Österreicher, 1988). Eine solche Abweichung kann dazu führen, daß der Vertrauensgrad der berechneten Intervalle zu einer schlichten Nominalzahl degeneriert.

Die Prüfung der Normalität von X erfolgt mit Vorteil mit Hilfe des sogenannten Wahrscheinlichkeitsnetzes, in welchem die empirische Verteilungsfunktion (die Summenfunktion) im Falle einer idealen Normalverteilung als eine Gerade erscheint.

Die vorstehende Bearbeitung betrifft lediglich einen kleinen Ausschnitt aus dem englischen Originalartikel.

Literatur:

Borovcnik, M.: Zum Schätzen von Mittelwerten - Ein intuitiver Zugang zu Vertrauensintervallen. In: Didaktik der Mathematik 29 (1987), 399-404.

Österreicher, F.: Die Normalverteilung in Wort und Bild. In: Ausflüge in die Mathematik. 21 Jahre Institut für Mathematik. Salzburg: Abakus Verlag 1988, 93-104.

Pfanzagl, J.: Allgemeine Methodenlehre der Statistik, Bd. 2. Berlin - New York: de Gruyter 1974.