Eine Grundfrage der Datenanalyse: Addiert oder multipliziert die Natur?

Eckhard Limpert[1)], Georg Ohmayer[2)], Werner A. Stahel[3)]

[1)]ELI-o-Research, Life Sciences, Zurich
[2)]Hochschule Weihenstephan-Triesdorf (HSWT)
[3)]Seminar for Statistics, ETH Zürich
georg.ohmayer@hswt.de

Abstract: Bei der Datenanalyse landwirtschaftlicher Versuche wird bislang fast standardmäßig von der Normalverteilung ausgegangen. Hier zeigen wir, dass die alternative, ähnlich einfach handhabbare logarithmische Normalverteilung vielfach die Daten besser beschreibt und effizienter ist. Das ist im Einklang mit den häufig multiplikativen Effekten als Ursache von quantitativer Variation.

1. Einleitung

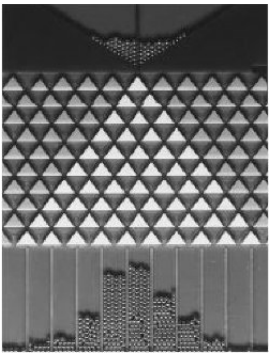
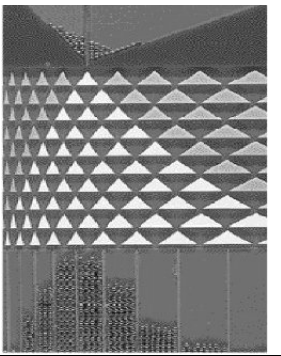
Diese Frage nach der „Rechenart“ der Natur ist so zu verstehen: Wie entstehen typische biologische Zufallseffekte bzw. wie kumulieren sie sich, additiv oder multiplikativ? Wenn beispielsweise verschieden hohe Pflanzen zufällig alle mehr Dünger bekommen, kann man fragen, ob dies eine Wachstumszunahme um etwa gleich viele cm oder um etwa den gleichen Prozentsatz bei allen bewirkt. Im einen Fall ist der Wachstumseffekt additiv, $h \rightarrow h+x$ [cm], im anderen Fall multiplikativ, $h \rightarrow h \cdot (1+y/100)$ [cm].

Eine Betrachtung der Naturgesetze zeigt, dass multiplikative Verknüpfungen deutlich wichtiger sind als additive. Das zeigt z.B. das Gravitationsgesetz $F = G \cdot m_1 \cdot m_2 / r^2$. In der Chemie als Grundlage des Lebens sind u.a. Reaktionsgeschwindigkeiten proportional dem *Produkt* der Konzentrationen der Reaktionspartner. Auch Wachstum und Vermehrung sind im Wesentlichen multiplikativ. Die Folgen aus dieser Erkenntnis werden anhand von Beispielen sowie der Literatur gezeigt.

2. Wahrscheinlichkeitstheoretischer Hintergrund

Der Zentrale Grenzwertsatz besagt, dass die Summe von n beliebig verteilten, voneinander unabhängigen Zufallsgrößen für $n \rightarrow \infty$ normalverteilt ist [Fi58, Sa04]. Ganz analog hat auch das Produkt von unabhängigen Zufallseffekten eine Grenzverteilung, und zwar die logarithmische oder *multiplikative* Normalverteilung [AB57, LSA01] (Tab. 1). Wenn in der Natur multiplikative Zufallseffekte vorherrschen, dann passt die letzte

Verteilung besser als die übliche Gauß-Normalverteilung. Das kann empirisch anhand publizierter Datensätze auch festgestellt werden.

Physikalische Simulations-modelle: Galton [Ga89] nutzte für sein Brett Nägel - statt der symmetrischen Dreiecke hier. Beim multiplikativen Brett sind die Dreiecke asymmetrisch [LSA01]; siehe auch: - http://www.inf.ethz.ch/personal/gut/lognormal/brochure.html - http://f1000.com/1020726#evaluations	**additiv** 	**multiplikativ** 
Theoretische Behandlung: Zentraler Grenzwertsatz ⇒ Art der Verteilung	in der additiven Variante ⇒ Gauß-Normalverteilung $NV(\mu,\sigma)$	in der multiplikativen Var. ⇒ Log-Normalverteilung $LNV(\mu,\sigma)$

Tab. 1: Unterschiede bei additivem und multiplikativem Kumulieren von Zufälligkeiten

Abb. 1 zeigt Ergebnisse von 471 Blättern eines Schmetterlingsflieders (*Buddleja davidii*). Die Verteilung von Länge und Breite passt zur Gauß-NV, die der Fläche zur LNV. Das ist im Einklang mit Überlegungen zu Durchmessern und Querschnitts- und Oberflächen, wie auch Volumina und Gewicht bei kugelförmigen Früchten [Ka03].

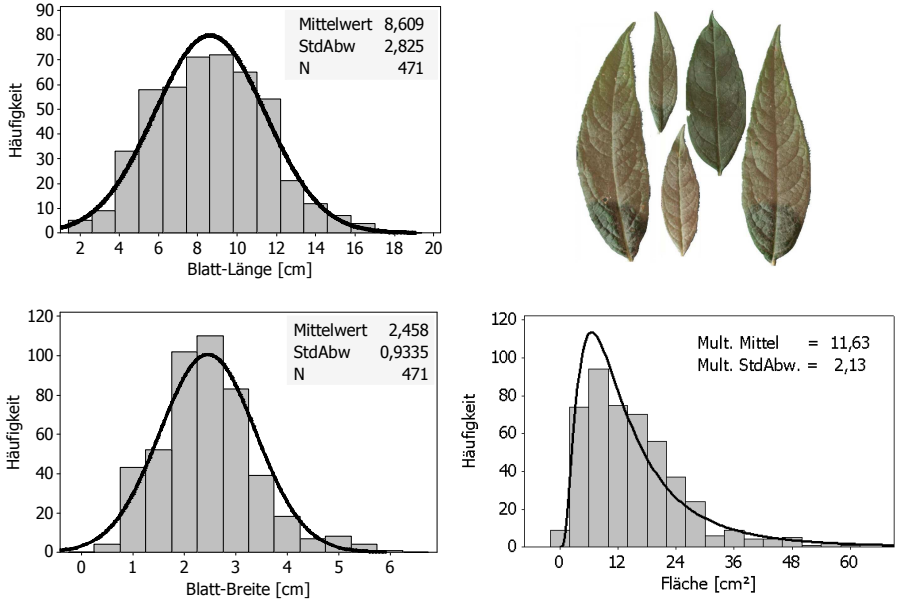


Abb. 1: Verteilung von Länge, Breite, Fläche der Blätter eines Schmetterlingsflieders

3. Beispiele für Datenanalysen

Schon bei sehr einfachen Versuchsdaten kann die Frage nach der richtigen Verteilungsannahme relevant sein. Abb. 2a zeigt Daten eines studentischen Versuches, in dem zu prüfen war, ob sich die Erträge für 4 Spinatsorten bei Berücksichtigung des unterschiedlichen Aufganges [% der ausgebrachten Samen] statistisch unterscheiden. Bei Durchführung einer Kovarianzanalyse (Faktor: Sorten, Kovariable: Aufgang) unterscheiden sich die Sorten nicht signifikant (p = 0.057), wobei die Verteilung der Residuen deutlich rechtsschief ist (Abb. 2b). Die alternative Auswertung der logarithmierten Daten ergibt dagegen nicht nur statistisch signifikante Unterschiede der Sorten (p = 0.048), sondern auch eine symmetrische Residuenverteilung.

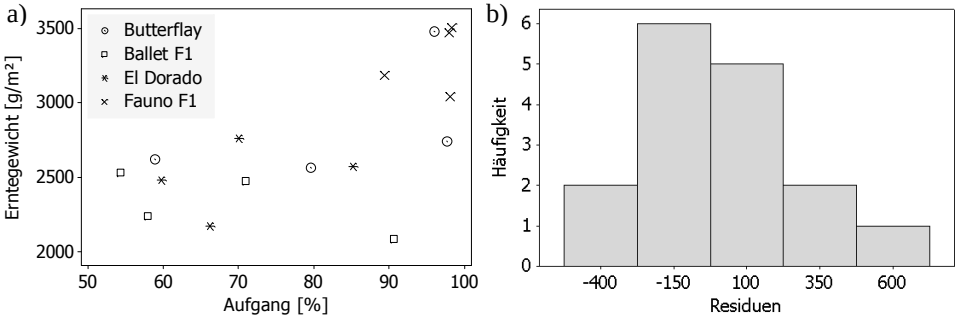


Abb. 2: Daten eines Sorten-Versuches bei Spinat (a) und Residuenverteilung (b) nach einer praxisüblichen einfaktoriellen Kovarianzanalyse

Im Vortrag werden weitere Beispiele für Datenanalysen gezeigt [siehe auch LS11]. Natürlich gibt es auch Datensätze, für die die Modellierung mit der Normalverteilung ebenso gut passt wie mit der Lognormalverteilung - besonders, wenn Variationskoeffizienten klein sind – ebenso wie solche, für die sich weder die Gauß‘sche noch die logarithmische Normalverteilung gut eignet.

4. Konsequenzen für die Praxis

Die Beispiele zeigen, dass oft präzisere Aussagen möglich sind, wenn statt der Normal-Verteilung die Lognormal-Verteilung genutzt wird. Vertrauensintervalle werden im Mittel kürzer und Tests haben mehr Macht. Wenn zwei Stichproben vom Umfang *n* auf gleiche Lage getestet werden, kann man mit einer Simulation zeigen, um wie viel effizienter die adäquate Auswertung ist - unter der Voraussetzung, dass die Lognormal-Verteilung passt. Anschaulich gesagt: Wie viele zusätzliche Beobachtungen müssen gemacht werden, um mit dem üblichen t-Test die gleiche statistische Macht zu erhalten wie mit der Auswertung der logarithmierten Daten [LS11]. Die Ergebnisse zeigt Abb. 3. Diese Aussage gilt natürlich in ähnlicher Weise für Varianzanalyse und Regression [FM23, LS03].

Hinweis:
n ist die Anzahl notwendiger Beobachtungen in jeder von 2 Gruppen, um mit dem Test für untransformierte Daten die gleiche Macht (90%) zu erreichen wie für n_0 Beobachtungen und den adäquaten Test

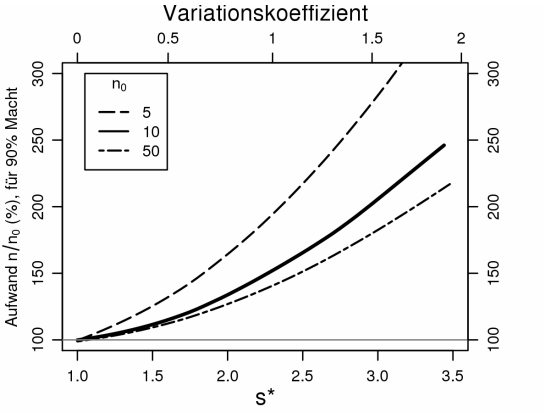


Abb. 3: Effizienz des t-Tests für logarithmierte Daten gegenüber dem t-Test für untransformierte Daten für den Fall, dass lognormale Daten vorliegen.

Die Erkenntnis, dass multiplikative Gesetze in der Natur vorherrschen, führt aber auch zu statistischen Modellen, die in der Regel besser auf die Daten passen, und in der Folge zu neuen Erkenntnissen in der Theorie. Schließlich sollte die Erkenntnis, dass Messdaten oft eine schiefe Verteilung aufweisen, sich auch in den üblichen grafischen Darstellungen von Unsicherheiten durch Fehlerbalken zeigen. Wenn sie mit multiplikativem Mittel und Standardabweichung gemäß $\bar{x} * {}^{x}/ s*$ charakterisiert werden, geben sie ein klareres Bild der Streuung von Daten als mit dem klassischen $\bar{x} \pm s$ [LS11].

Da die meisten Naturgesetze multiplikativer Art sind, ist ein Modell für eine Verteilung von zufälligen Abweichungen von einem Idealwert plausibel, das dieser Operation entspricht. Das ist auf Grund der entsprechenden Variante des Zentralen Grenzwertsatzes die Lognormal-Verteilung. Es gibt viele deskriptive Methoden und Modelle, die vor allem dann gerechtfertigt sind, wenn die Daten normalverteilt sind. Die entsprechenden Methoden, die auf der Lognormal-Verteilung beruhen, sind im Wesentlichen ebenso einfach anzuwenden.

Literaturverzeichnis

[Fi58] Fisher RA: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh,1958.

[AB57] Aitchison J., Brown JAC: The Lognormal Distribution. Cambridge Univ. Press, 1957.

[Ga89] Galton, F: Natural Inheritance. London: Macmillan, 1889.

[Ka03] Kapteyn JC: Skew frequency curves in biology and statistics. Astronomical Laboratory, Groningen: Noordhoff, 1903.

[LSA01] Limpert E, Stahel WA, Abbt M: Log-normal distributions across the sciences - keys and clues. BioScience 51, 341-352, 2001.

[LS03] Limpert E, Stahel WA: Das Leben ist multiplikativ - neue Aspekte zur Verteilung von Daten. Bericht 53. Int. Züchtertagung, Gumpenstein, 15-21.

[LS11] Limpert E, Stahel WA: Problems with using the normal distribution – and ways to improve quality and efficiency of data analysis. PloS ONE 6, 2011.

[Sa04] Sachs L: Angewandte Statistik. 11. Aufl., Springer, Berlin, 2004.

[FM23] Fisher RA, Mackenzie WA: Studies in crop variation II. J Agric Sci 13, 311-320, 1923.