Zur Quantifizierung des Normalverteilungsgrades

Christian Langesberg*† Uwe Ligges* Claus Weihs*

25. April 2018

Zusammenfassung

Ob die Werte einer Stichprobe aus einer Normalverteilung stammen ist in der Statistik eine häufig gestellte Frage. Gebräuchliche Werkzeuge zur Beantwortung dieser Frage sind häufig entweder nicht automatisierbar oder nicht in der Lage, Abstufungen zu erkennen. Der vorliegende Aufsatz stellt aktuell verwendbare Ansätze vor, welche keinen dieser Nachteile aufweisen. Mit theoretischen Überlegungen und einer Simulationsstudie, sowie der Berücksichtigung von Stichprobengrößen und der Schätzung von Normalverteilungsparametern, werden diese Ansätze verglichen. Als beste Verfahren stellen sich Abwandlungen der Metriken nach Kolmogorov und Lévy, sowie eine Transformation der Teststatistik von Jarque und Bera heraus.

Stichworte Normalverteilung, Tests auf Vorliegen einer Verteilung, Metriken und Distanzmaße für Verteilungen, Simulationsstudie

*TU Dortmund
†E-Mail: clangesberg@statistik.tu-dortmund.de

Inhalt

1 Einleitung **3**

2 Ausgewählte Methoden **4**

2.1 Anforderungen 4

2.2 Distanzen 5

2.3 Teststatistiken 10

2.4 Umsetzung 14

2.5 Ausgewählte Verteilungen 15

3 Allgemeine Vergleiche **16**

3.1 Vergleich der Konzepte 16

3.2 Vergleiche für theoretische Verteilungen 18

3.3 Folgerungen 20

4 Simulationsstudie **22**

4.1 Simulationsdesign 22

4.2 Ergebnisse 22

4.3 Vergleich mit theoretischen Größen 29

4.4 Folgerungen 31

5 Fazit **32**

6 Ergänzungen **34**

6.1 Zur Stichprobengröße 34

6.2 Zur Optimierung 36

7 Zusammenfassung **40**

A Ergänzende Grafiken **42**

Literatur **46**

1 Einleitung

Die Normalverteilung ist eines der wichtigsten Werkzeuge eines Statistikers oder einer Statistikerin. So kommt kein umfassendes Lehrbuch ohne die Definition der „Gauß'schen Glockenkurve“ aus, und beispielsweise auch Hartung (2005) oder Genschel und Becker (2005) heben die Normvalverteilung als „eine der wichtigsten statistischen Verteilungen“ hervor (Seite 143 f. bzw. Seite 42). Bei zahlreichen weit verbreiteten Methoden wie Regressionsmodellen oder Maximum-Likelihood-Schätzern findet sich die Normalverteilung wieder, etwa als unterstellte Verteilung von Modellfehlern oder als Limes der Verteilung einer Schätzfunktion. Auch und insbesondere der Zentrale Grenzwertsatz verdeutlicht die Bedeutung der Normalverteilung. Wie Hartung (2005) in der Einführung ausführt, stammt ein großer Teil der Fundamente heutiger statistischer Verfahren erst aus dem frühen 20. Jahrhundert - die Normalverteilung hingegen wird bereits 1718 von Abraham de Moivre in einer theoretischen Arbeit, und im 19. Jahrhundert von Adolphe Quetelet in praktischer Anwendung (nämlich der Messung des Brustumfangs von Soldaten) verwendet.

Aus der Bedeutung der Normalverteilung erwächst die für die praktische Anwendung zahlreicher statistischer Verfahren wichtige Frage, ob die jeweils vorliegenden Daten aus einer Normalverteilung stammen - oder zumindest nicht allzu stark von dieser abweichen. Letzteres trägt dem Umstand Rechnung, dass in vielen Fällen nur eine Konvergenz gegen die Normalverteilung vorliegt. So etwa beim Satz vom zentralen Grenzwert (ZGWS) für eine steigende Anzahl von Summanden, oder bei Dichtfunktionen, welche mit passender Parameterwahl gegen die Dichte der Normalverteilung konvergieren. Hier seien als Beispiele die Konvergenz der t_n-Verteilung gegen die Standardnormalverteilung (für $n \to \infty$), oder die durch das Galtonbrett gut veranschaulichte Konvergenz der Binomialverteilung genannt.

In den meisten Fällen ist nun jedoch unklar, was eine *nicht zu starke Abweichung* bedeutet. Zur Beurteilung der Normalität eines Datenvektors stehen zwar zahlreiche Methoden zur Verfügung. Diese führen jedoch zumeist entweder nur zur binären Entscheidung normal- oder nicht normalverteilt, was insbesondere die Vielzahl an statistischen Tests auf Vorliegen einer Normalverteilung betrifft, oder sie beruhen auf einer grafischen Beurteilung, wie der Form eines Histogramms oder eines Boxplots. Während beide Varianten im Einzelfall praktikabel sind und sich sogar ergänzen können, haben sie bei der Auswertung einer großen Anzahl an Datensätzen klare Mängel: Die Beurteilung von Grafiken ist nicht uneingeschränkt objektiv möglich, erfordert zumeist sogar eine gewisse Vorbildung des Analysten und ist nicht ohne Weiteres automatisierbar. Denkbare Ansätze, beispielsweise ein Histogramm einer statistischen Verteilung automatisiert zuzuordnen, sind aber im Bereich künstlicher Intelligenz gut vorstellbar - müssten dann aber auch mit Verfahren konkurrieren, denen statt dem Histogramm die Daten selbst zur Verfügung stehen. Die Idee einer objektiven Beurteilung von Q-Q-Plots findet sich bei der Gruppe von Teststatistiken wieder, welche auf Regression und Korrelation beruhen (siehe unten, Teil 3.1).

Bei der Durchführung von Tests ist es hingegen nicht möglich, graduelle Unterschiede festzustellen. Wird etwa die Normalität einer Summe in Abhängigkeit von der Anzahl der Summanden untersucht, so können Tests trotz stetiger Konvergenz (vgl. ZGWS) nur die „Türschwelle“ zwischen Nicht-Normalität und Normalität erkennen. Über Abstände oder Näherungsgeschwindigkeiten können keine jedoch keine Aussagen getroffen werden.

Die vorliegende Ausarbeitung beschreibt Möglichkeiten, den Abstand einer Stichprobe reellwertiger Zahlen zur Normalverteilungsfamilie zu messen. Diese Abstandsbestimmungen sollen leicht automatisierbar sein, sowie graduelle Unterschiede bezüglich der Nähe zur Normalverteilung angeben können. Eine genauere Beschreibung der Anforderungen folgt als Kapitel 2. Daran anschließend werden die ausgewählten

Diese Arbeit wurde von der Deutschen Forschungsgemeinschaft im Rahmen der Forschergruppe FOR 1511 gefördert.

„Methodik-Kandidaten“ zur Messung des Abstands zur Normalverteilung vorgestellt. Es bieten sich insbesondere Metriken für Dichte- oder Verteilungsfunktionen an; auch Teststatistiken erscheinen als sinnvolle Möglichkeit. Neben inhaltlichen Erwägungen wird bei der Auswahl auch die Verfügbarkeit oder einfache Umsetzbarkeit in der statistischen Programmiersprache R berücksichtigt. Eine umfassendere Sicht auf die Verfügbarkeit von Teststatistiken in verschiedenen Programmen geben Yap und Sim (2011) in ihrer Tabelle 1.

Nach einer Gegenüberstellung der verschiedenen Konzepte und Vergleichen theoretischer Eigenschaften der Verfahren im dritten Kapitel folgt die Beschreibung einer durchgeführten Simulationsstudie. Dessen Ergebnisse werden für sich genommen betrachtet, aber auch den theoretisch zu erwartenden Resultaten gegenübergestellt. Theoretische und praktische Eigenschaften werden im Kapitel 5 zusammengefasst und auf dieser Basis dann - soweit möglich - eine Empfehlung für das oder die beste/n Verfahren gegeben. Vor einer abschließenden Zusammenfassung erfolgen im Kapitel 6 ergänzende Betrachtungen der als am besten eingestuften Methoden. So sind Abhängigkeiten von der Stichprobengröße sowie von der Bestimmung der jeweiligen Normalverteilungsparameter zu berücksichtigen.

2 Ausgewählte Methoden

2.1 Anforderungen

Für die Beurteilung des Grades der Annäherung an eine Normalverteilung sind verschiedene Herangehensweisen denkbar. Da eine Normalverteilung für einen Erwartungswert $\mu \in \mathbb{R}$ und eine Varianz $\sigma^2 \in \mathbb{R}_{>0}$ vollständig durch ihre Dichtefunktion $f : \mathbb{R} \to \mathbb{R}$ mit

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

definiert ist, können alle Charakteristiken dieser Verteilung auch aus dieser Funktion abgeleitet werden. Dabei kann die Dichtefunktion offenbar als Ganzes verwendet werden, es können aber auch einzelne Charakteristiken wie beispielsweise die Symmetrie um den Erwartungswert ($f(\mu + x) = f(\mu - x)$) oder Transformationen wie die Verteilungsfunktion als Integral der Dichte ($F(x) = \int_{-\infty}^{x} f(t)\ dt$) verwendet werden.

Liegt nun eine Stichprobe $y = (y_1, y_2, \ldots, y_n)' \in \mathbb{R}^n$ der Größe $n \in \mathbb{N}$ vor, deren Normalverteilungsgrad beurteilt werden soll, so ist also eine Distanzfunktion $d : \mathbb{R}^n \to \mathbb{R}$ gesucht. Wie im Folgenden dargestellt, ist ein Teil der möglichen Varianten nur zur Beurteilung der Abstands der Daten zu einer *bestimmten* Normalverteilung geeignet, unter Berücksichtigung gegebener Parameter μ und σ^2 handelt es sich dabei also um eine Abbildung $\mathbb{R}^{n+2} \to \mathbb{R}$.

Wird im Folgenden das Wort „Maß“ verwendet, ist dies nicht nur im mathematischen Sinne einer σ-additiven Abbildung zu verstehen, vielmehr wird es auch als allgemeiner Begriff etwa wie ein Längenmaß verwendet. Als erstrebenswert für ein solches „gutes“ Abstandmaß können die folgenden Punkte festgehalten werden.

1. Es können sowohl Abstände zwischen einem Datenvektor und einer Verteilung, als auch zwischen zwei Datenvektoren oder zwei Verteilungen gemessen werden.

2. Es handelt sich bei d um eine stetige Funktion, so dass beispielsweise eine kleine Änderung der Daten nur zu einer kleinen Änderung des Abstands führt.

3. Das Maß ist nach oben und unten durch zwei Größen $a,\ b \in \mathbb{R}$ beschränkt, so dass ein Funktionswert für sich alleinstehend interpretiert werden kann.

4. Symmetrie: Sind x, y zwei Datenvektoren oder zwei Verteilungen, so sollte $d(x, y) = d(y, x)$ gelten.

5. Die Zuordnung normal/nicht-normal soll mit möglichst großem Anteil korrekt getroffen werden, sofern eindeutig möglich.

Eine Erfüllung der ersten Forderung würde nicht nur einen praktikablen Umgang mit dem Maß ermöglichen, es könnten auch weitere Konsistenzüberprüfungen stattfinden: Werden zwei Verteilungen als nah zueinander beurteilt, sollte das auch für zwei Vektoren aus diesen Verteilungen gezogener Zufallszahlen zu erwarten sein.

Eine zunächst naheliegende Variante zur Beurteilung des Grades der Normal- oder auch einer anderen Verteilung stellen Signifikanztests dar. Dabei wird jedoch nur eine binäre Entscheidung ja oder nein erzeugt. Feinere Abstufungen sind nicht möglich, womit beispielsweise keine gute Untersuchung des Konvergenzverhaltens von Datenvarianten möglich ist. Außerdem führen die Stellschrauben des Signifikanzniveaus und der Teststärke (oder des Niveaus und der Stichprobengröße) zu nicht eindeutigen Ergebnissen. Alternativ oder parallel zur Entscheidung eines Tests kann auch der zugehörige p-Wert herangezogen werden, welcher die Forderung nach Abstufung und Begrenzung erfüllt. Auch die mitunter angewandte rein deskriptive Betrachtung es p-Werts ist aber nicht praktikabel: Trifft die Nullhypothese einer bestimmten Verteilung oder Verteilungseigenschaft zu, so ist der p-Wert eine im Intervall (0,1) gleichverteilte Zufallsvariable. Damit kann ein kleiner p-Wert nicht sicher als große Abweichung von einer Normalverteilung aufgefasst werden.

Beeinflusst wird der p-Wert direkt von der jeweiligen Teststatistik. Diese hat deutliche Vorteile gegenüber dem p-Wert: Zwar folgt auch sie einer zufälligen Verteilung, welche abhängig von der Richtigkeit der Nullhypothese ist. Dabei ist es aber nicht ausgeschlossen, dass die Verteilungen der Teststatistik unter Null- und Alternativhypothese zu disjunkten Bereichen führen, oder zumindest weniger Überdeckung aufweisen als bei p-Werten. Des Weiteren liegt einer Teststatistik zumeist eine sachlogische Idee zugrunde, so dass ein großer oder kleiner Wert oft direkt interpretiert werden kann. So wird beispielsweise beim bekannten Kolmogorov-Smirnov-Test der maximale Abstand von empirischer und unterstellter Verteilungsfunktion angegeben - ein Wert, welcher inhaltlich klar und eindeutig verständlich ist. Zudem ist eine Teststatistik - im Gegensatz zum p-Wert oder zur Testentscheidung - frei von Einflüssen der Konvergenzgeschwindigkeit bei approximativ verteilten Teststatistiken.

Neben Signifikanztests bestehen weitere Konzepte, um den Abstand von empirischen oder theoretischen Verteilungen zu beurteilen. Diese entstammen im Allgemeinen der mathematischen Statistik und werden im folgenden Abschnitt vorgestellt. Die Auswahl berücksichtigt dabei, ob eine Umsetzung in der statistischen Programmierung R vorliegt, welche das jeweilige Maß zwischen Daten und Verteilung berechnet.

2.2 Distanzen

Zur Beurteilung des Abstands zwischen zwei Verteilungen bestehen verschiedene Möglichkeiten. Oft liegen per definitionem zwei theoretische (nicht empirische) Verteilungen vor, dann müssen die Berechnungsweisen für Datenvektoren entsprechend modifiziert werden (s. u.). Berücksichtigt werden die folgenden Distanzen, welche zum Teil Dichte- und zum Teil Verteilungsfunktionen verwenden. Dabei wäre es jeweils auch möglich, jeden Abstand nur für Dichten oder Verteilungsfunktionen zu definieren, da sich beide eindeutig ineinander übertragen lassen.

Totalvariationsabstand

Im Teil 3.1.5 „Abstandsmaße und Konvergenzarten für Verteilungen“ führt Rüger (2002) auf den Seiten 41 ff. verschiedene Varianten für Distanzmaße an. Dabei werden P und Q als zwei Verteilungen über dem Borel'schen Messraum $(\mathbb{R}, \mathfrak{B})$ vorausgesetzt. Eine Metrik wird also über der Menge $\mathfrak{M}$ aller Verteilungen dieses Raums gebildet. Der sogenannte Totalvariationsabstand ist dann definiert als

$$TV(P,Q) = \sup_{B \in \mathfrak{B}} |P(B) - Q(B)| .$$

In alternativer Darstellung mittels eines Maßes[2] ν und Dichten f und g von P und Q,

$$TV(P,Q) = \frac{1}{2} \int |f - g| \, d\nu,$$

wird klar, dass sich der Totalvariationsabstand aus der Fläche von zwei Dichten berechnet: Angegeben wird der Anteil der Fläche unter einer (beliebigen) der Dichtefunktionen, welcher schnittmengenfrei mit der Fläche der anderen Dichte ist, sinngemäß und kurz ist $TV(P,Q) = 1 -$ Schnittfläche der Dichten. In der Abbildung 2 auf der Seite 9 ist dies gemeinsam mit den Ideen anderer Verfahren veranschaulicht. Ist im Folgenden jeweils eindeutig, welche Verteilungen beziehungsweise Dichten gemeint sind, wird zumeist statt $TV(P,Q)$ kurz nur TV geschrieben.

Hellinger-Distanz

Ein ähnliches Maß wie der Totalvariationsabstand stellt der Hellinger-Abstand dar. Dieser beruht auf dem geometrischen Mittel zweier Dichten und gewichtet damit ebenso Bereiche stark, in denen die Dichten beide „groß“ sind. Er berechnet sich durch

$$H(P,Q) = \sqrt{\frac{1}{2} \int \left(\sqrt{f} - \sqrt{g} \right)^2 d\nu} = \sqrt{1 - \int \sqrt{fg} \, d\nu}.$$

Wie Witting (1985) auf Seite 136 darlegt, können die beiden genannten Distanzen auch allgemeiner als

$$d_r(P,Q) = \left(\int \left| f^{1/r} - g^{1/r} \right|^r \right)^{1/r}$$

formuliert werden, wobei $TV(P,Q) = \frac{1}{2} d_1(P,Q)$ und $H(P,Q) = \frac{1}{\sqrt{2}} d_2(P,Q)$ ist. Beide Distanzmaße sind durch die Null und die Eins begrenzt, stellen Metriken über $\mathfrak{M}$ dar und es gilt

$$TV(P,Q) = 1 \Leftrightarrow H(P,Q) = 1 \Leftrightarrow P \perp Q :\Leftrightarrow \exists A \in \mathfrak{B} : P(A) = 1 \wedge Q(A) = 0.$$

Lévy-Metrik

Weiter mit Verweis auf Rüger (2002) ist die Metrik nach Paul Lévy anzuführen, welche dort als „das älteste Abstandsmaß für Verteilungen“ (Seite 52) vorgestellt wird. Dabei wird um eine Verteilungsfunktion

[2] genauer: mittels eines σ-finiten Maßes ν, „das P und Q dominiert (ein solches existiert stets, z. B. $\nu = P + Q$)“ (Seite 44).

F eine Lévy-Umgebung so gebildet, dass eine zweite Verteilungsfunktion G komplett in dieser enthalten ist. Die kleinstmögliche Größe dieser Umgebung bildet dann den Abstand zwischen diesen beiden Verteilungen:

$$L(F,G) = \inf\{\varepsilon > 0 : G \in \mathcal{U}_\varepsilon(F)\} \, (= \inf\{\varepsilon > 0 : F \in \mathcal{U}_\varepsilon(G)\})$$

wobei die Lévy-Umgebung mit der Menge aller eindimensionalen Verteilungsfunktionen $\mathcal{F}$ durch

$$\mathcal{U}_\varepsilon(F) = \{G \in \mathcal{F} : F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon \; \forall \; x \in \mathbb{R}\}$$

definiert wird.

Die Grundidee des Abstands zwischen zwei Verteilungsfunktionen findet sich auch in der durch den Kolmogorov-Smirnov-Test bekannten und unten dargestellten Kolmogorov-Metrik wieder. Dabei ist anzumerken, dass die Kolmogorov-Variante jeweils nur die Abstände an einer Position x betrachtet. Bei den von Lévy verwendeten Umgebungen wird durch die Betrachtung an den Stellen $x \pm \varepsilon$ eine zusätzliche Einschränkung gemacht, welche eine Stetigkeitsbedingung darstellt. Die Idee der Lévy-Distanz ist in der folgenden Abbildung 1 genauer dargestellt.

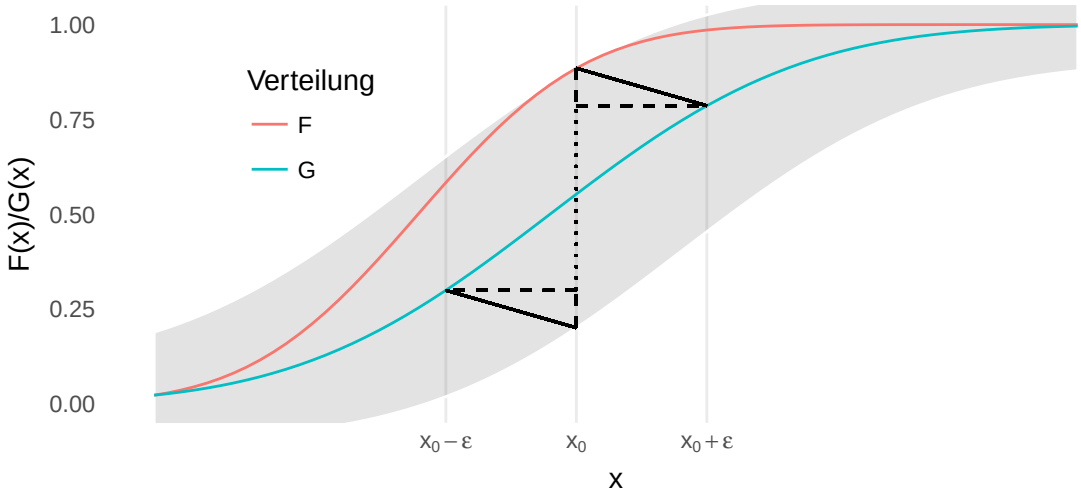

Abbildung 1: Die Lévy-Distanz wird gemäß Definition durch das Infimum der Menge $\{\varepsilon > 0 : F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon \; \forall \; x \in \mathbb{R}\}$ berechnet. Diese Menge ist hier für das entsprechende Infimum grau hinterlegt. Gemessen wird der Abstand zwischen den beiden Verteilungsfunktionen in 45-Grad-Richtung (bei gleicher Skalierung der Koordinatenachsen): Die beiden Katheten (gestrichelte Linien) des Dreiecks der Punkte mit den Koordinaten $(x_0, G(x_0 + \varepsilon))$, $(x_0, G(x_0 + \varepsilon) + \varepsilon)$ und $(x_0 + \varepsilon, G(x_0 + \varepsilon)$ haben jeweils die Länge ε, die Hypotenuse (durchgezogene Linie) als diagonaler Abstand der beiden Verteilungsfunktionen damit die Länge $\sqrt{2}\varepsilon$. Entsprechendes gilt für das mittels Punktspiegelung erzeugte zweite Dreieck.

Die Lévy-Prohorov-Metrik beruht auf dem gleichen Gedanken wie die Lévy-Metrik und ist auch für mehrdimensionale Verteilungen definiert. An dieser Stelle führt sie aber nur zu einer weniger anschaulichen Definition, so dass die einfachere, ältere Variante vorgezogen wird.

Kolmogorov-Metrik

Der Kolmogorov-Smirnov-Test dürfte zu den bekannteren statistischen Tests zählen und beruht auf dem maximalen Abstand zwischen zwei Verteilungsfunktionen. Dieser Abstand erfüllt alle Eigenschaften einer Metrik und ist zudem nach oben durch die 1 begrenzt. Damit kann ein Maß für die Unterschiedlichkeit zweier Verteilungen gebildet werden durch:

$$K(F, G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|.$$

Die Kolmogorov-Metrik findet sich etwa bei Rüger (2002) auf der Seite 50. Der guten Interpretierbarkeit und hohen Vielseitigkeit dieser Kennzahl steht die Reduzierung der in F und G vorliegenden Informationen auf einen Abstand an einem einzelnen Punkt entgegen. Wie gut diese Verdichtung geeignet ist, wird im Laufe der folgenden Auswertungen untersucht.

Cramér-von-Mises-Metrik

Die Idee der Kolmogorov-Metrik lässt sich leicht auf zwei Weisen weiterentwickeln: Zum einen kann statt der größten Abweichung der beiden Verteilungen auch die kumulative Abweichung der beiden Verteilungsfunktionen verwendet werden. Zum anderen haben sich in der Statistik an vielen Stellen quadratische Abstände als praktikabel erwiesen (etwa bei der Varianz oder dem Kleinste-Quadrate-Schätzer für Regressionsmodelle), so dass sie hier auch als naheliegend erscheinen.

Beide Ideen berücksichtigt die Cramér-von-Mises(CvM)-Distanz, welche nach Harald Cramér und Richard von Mises benannt ist und durch

$$CM(P, Q) = \int (F(x) - G(x))^2 \, dF(x)$$

definiert ist. In Rieder (1994) findet sich zudem der Hinweis auf die Möglichkeit der Gewichtung durch eine Funktion w, das heißt durch Integration von $(F(x) - G(x))^2 w(x)$. Diese Idee findet etwa beim Anderson-Darling-Test Anwendung, der unten vorgestellt wird.

Vergleich der Ideen

Alle fünf der vorgestellten Metriken lassen sich in Abhängigkeit von Dichte- oder Verteilungsfunktion darstellen. Beispielhaft zeigt die folgende Abbildung 2 die jeweiligen „Kernelemente“ der Metriken. Dabei wird auf die zusammenfassenden Elemente wie Integral oder Supremum verzichtet, um die zugrundeliegenden Ideen zu verdeutlichen.

Die Grafik zeigt die Dichte- und Verteilungsfunktionen zweier Normalverteilungen und die entsprechenden Werte der Distanzmaße. Dabei ist festzustellen, dass die Verwendung von Dichten oder Verteilungsfunktionen zu gegenläufigen punktuellen Resultaten führt: Bei etwa 0.9 weisen die beiden Dichten einen Schnittpunkt auf. Dass die beiden Funktionen hier gleich sind bewerten die Hellinger- und TV-Elemente mit dem Minimalwert von 0, während durch den dortigen maximalen Abstand der Verteilungsfunktionen die Methoden von Kolmogorov, Lévy und CvM um 0.9 ihren jeweils größten Wert annehmen.

Weitere Distanzmaße

Es existiert ein Vielzahl weiterer Distanzmaße, welche hier nicht berücksichtigt werden. Dabei fußt die vorgenommene Auswahl nicht zuletzt auf der Auswahl der angegebenen Literatur. Des Weiteren

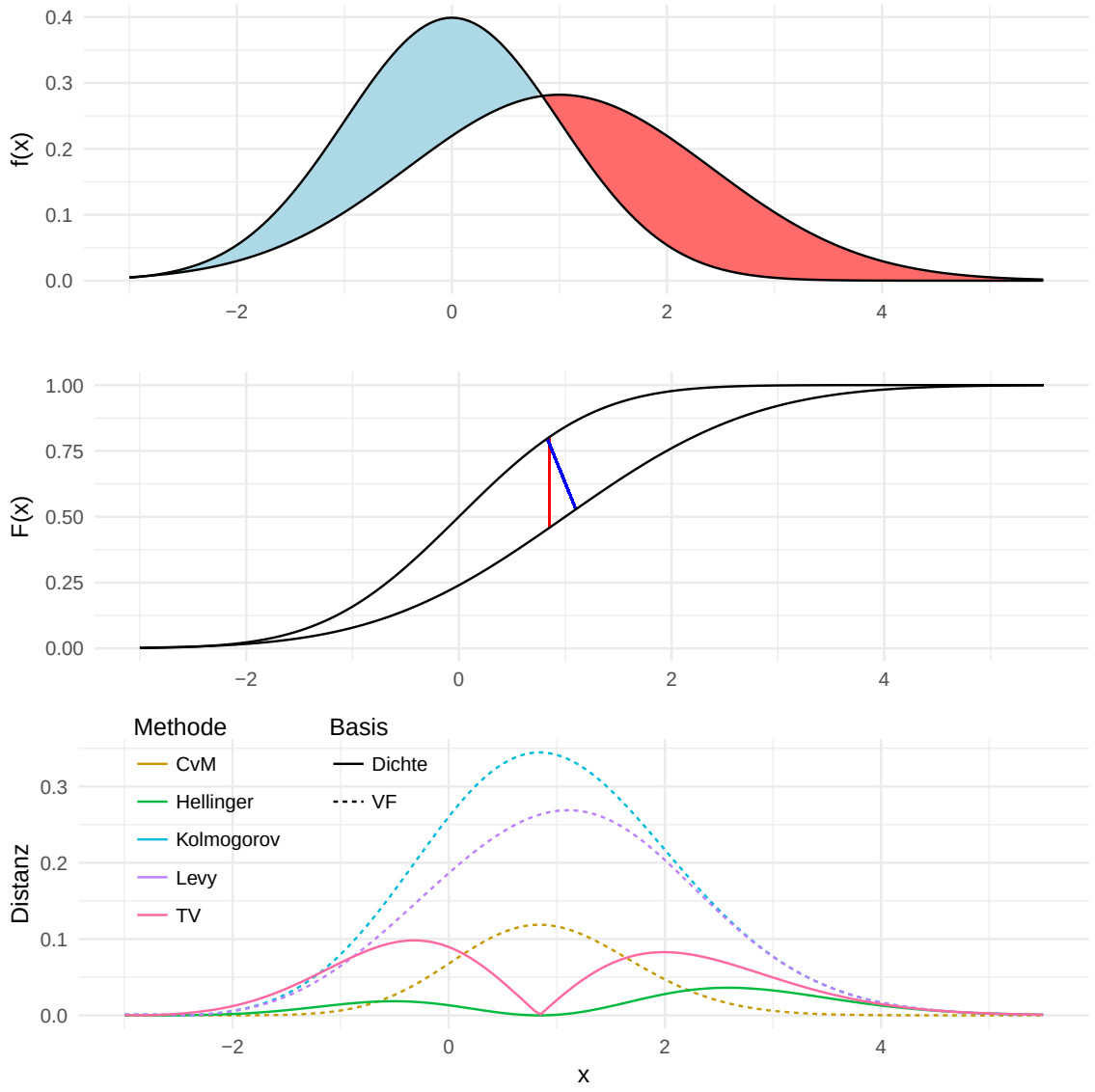

Abbildung 2: Dichte, Verteilungsfunktion und punktueller Abstand von zwei Normalverteilungen mit den Parametern (0,1) und (1,$\sigma^2 = 2$). Die blaue *oder* gleichwertig die rote Fläche ergibt den Totalvariationsabstand, also das Integral der Hälfte der entsprechenden Funktion aus der unteren Grafik. Die vertikale rote Linie zwischen den beiden Verteilungsfunktionen kennzeichnet den Abstand nach Kolmogorov, die diagonale blaue Linie den Abstand nach Lévy. Die Werte der unteren Grafiken werden für die Bildung der Maße in allen Fällen noch zu einer einzelnen Größe verdichtet, etwa durch Integralbildung oder Extremwertbestimmung.

sollen keine weiteren Maße hinzugezogen werden, sofern deren Konzept bereits in den einbezogenen Maßen enthalten ist. So beruhen mehrere Maße wie die Bhattacharyya-Distanz auf dem Bhattacharyya-Koeffizienten $BC = \int \sqrt{p(x)q(x)}dx$. Ob dieser Wert groß oder klein ausfällt kann aber stellvertretend an der Hellinger-Distanz durch die Beziehung $H = \sqrt{1 - BC}$ erkannt werden.

Denkbar ist auch, Kombinationen der Ideen von Kolmogorov- und CvM-Metrik zu verwenden, das hieße zusätzlich auch die maximale quadratische Abweichung und die kumulative absolute Distanz zu betrachten (letztere bezeichnet Rieder (1994) als L_1-Distanz, siehe Seite 125). Da hier jedoch keine wesentlichen weiteren Erkenntnisse zu erwarten sind, wird auf diese Varianten aus Gründen der Übersichtlichkeit verzichtet.

Auf andere Maße wurde verzichtet weil sie per definitionem nachteilige Eigenschaften haben, so ist etwa die Kullback-Leibler-Divergenz nicht symmetrisch in P und Q. Die Energy-Distanz andererseits ist allgemein für Verteilungsfunktionen mehrdimensionaler Beobachtungen konstruiert; die Definition ergibt im eindimensionalen Fall gerade die CvM-Metrik.

Eine bekannte Distanz stellte auch die Mahalanobis-Distanz dar, welche unter anderem in der Diskriminanzanalyse Anwendung findet. Wird ein Datenvektor $(x_1, x_2, \ldots, x_n)'$ als Realisierung unabhängig identisch verteilter Größen aufgefasst, ergibt sich die Mahalanobis-Distanz als Summe von zentrierten, standardisierten und insbesondere wiederum unabhängig identisch verteilten Größen. Diese Summe konvergiert gegen eine normalverteilte Zufallsvariable, und das unabhängig von der Ausgangsverteilung. Unterschiede ergeben sich allenfalls bei kleinen Stichprobengrößen, auf welche an dieser Stelle aber nicht eingeschränkt werden soll.

2.3 Teststatistiken

Neben den angeführten Metriken werden insgesamt vier Teststatistiken in Betracht gezogen, um den Grad der Normalverteilung einer Stichprobe zu quantifizieren. Diese Statistiken verwenden verschiedene Eigenschaften einer Verteilung, beispielsweise die Form der Dichte oder die erwartete Häufigkeit bestimmter Wertebereiche. Ein Teil der vorgestellten Verfahren kann zur Untersuchung von verschiedenen Verteilungsannahmen verwendet werden, an dieser Stelle wird die Anwendung jedoch jeweils auf die Normalverteilung begrenzt.

Die folgende Auswahl der Tests soll verschiedene Konzepte der Abweichung von einer Normalverteilung abdecken, zudem sollen die gängigen Methoden enthalten sein. Insbesondere die Einordnung als „gängig“ ist dabei ohne Weiteres sicher als subjektiv zu bewerten. Berücksichtigt werden die Statistiken der folgenden Tests.

χ^2-Anpassungstest

Die allgemeine Chi-Quadrat-Statistik vergleicht für gegebene Klassen jeweils die beobachtete und die unter einer Nullhypothese erwartete Häufigkeit an Beobachtungen. Für einen Test auf eine stetige Verteilung muss, wie Hartung (2005) auf Seite 182 f. beschreibt, das Intervall $(-\infty, \infty)$ in k disjunkte Abschnitte unterteilt werden. Sind für eine bestimmte Normalverteilung $E_i \in \mathbb{N}$ Beobachtungen in der Klasse $1 \leq i \leq k$ zu erwarten, und werden $O_i \in \mathbb{N}$ tatsächlich beobachtet, so wird die Teststatistik berechnet durch

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

es wird also die relative quadratische Abweichung der absoluten Häufigkeiten von den erwarteten summiert. Bei der Durchführung des Tests ist zu berücksichtigen, wie die Parameter der Normalverteilung hergeleitet wurden, mit denen die E_i berechnet wurden (vgl. ebd.). Das kann hier vernachlässigt werden. Zu berücksichtigen ist jedoch, dass die Statistik eben diese Parameter benötigt, also für μ und σ^2 Schätzer einzusetzten sind. Dazu werden hier das arithmetische Mittel und die Stichprobenvarianz verwendet.

Für die Anzahl der Klassen k wird der Standardwert der verwendeten Software übernommen, das bedeutet $k = \lceil 2n^{2/5} \rceil$. Bei einer perfekt normalverteilten Stichprobe ist zwar von $O_i = E_i \; \forall \; i$ auszugehen, durch die Quadrierung des Abstandes sind Abweichungen der Summanden von der Null jedoch nur in positiver Richtung möglich und der Erwartungswert von χ^2 ist somit ebenso größer als Null. Die Vergleichsgröße (der „kritische Wert“) stammt bei einem Test aus der χ^2_{k-3}-Verteilung, der Erwartungswert beträgt damit $k-3$. Werden die beiden Parameter einer Normalverteilung nicht geschätzt sondern vorgegeben, sind für die Orientierungsgröße nur $k-1$ Freiheitsgrade anzusetzen. Nach oben ist diese Statistik offenbar nicht begrenzt.

Anderson-Darling-Test

Der Test von Theodore W. Anderson und Donald A. Darling nutzt aus, dass der Wert der Verteilungsfunktion einer Zufallsvariablen einer Gleichverteilung folgt, was bei der Inversionsmethode zur Erzeugung von Zufallszahlen einer bestimmten Verteilung genau andersherum ausgenutzt wird.

Für die Berechnung der Teststatistik wird die Stichprobe $x_g = (x_{(1)}, x_{(2)}, \ldots, x_{(n)})'$ als aufsteigend geordnet vorausgesetzt. Diese wird für einen Test auf Normalverteilung durch die Transformation

$$z_{(i)} = (x_{(i)} - \hat{\mu})/\hat{\sigma}$$

mit Schätzern für Erwartungswert und Standardabweichung auf eine (0,1)-Verteilung skaliert. Diese empirische Verteilung würde sich bei Vorliegen einer Normalverteilung für die x_i nicht stark von einer N(0,1)-Verteilung unterscheiden, in der Theorie ergibt sich exakt die Standardnormalverteilung.

Die Statistik A^2 des Anderson-Darling-Tests bildet sich daher durch die Verteilungsfunktion Φ der Standardnormalverteilung, ausgewertet an den Stellen $z_{(i)}$. Ist die Anzahl der Datenpunkte bekannt, so können auch die zu erwartenden Werte bestimmt werden. Hinzu kommt nun eine Gewichtung. Die Statistik A^2 wird hier in zwei Varianten angegeben, wobei die zweite einfacher zu interpretieren ist:

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left(\log(\Phi(z_{(i)})) + \log(1 - \Phi(z_{(n+1-i)})) \right).$$

Diese Form findet sich unter anderem bei Lewis (1961) (Seite 1 119 oben mit abweichender Notation der Indizes), wird dort aber auch nur als „äquivalente Form“ angeführt. In der Grundform wird für empirische und theoretische Verteilungsfunktionen F_n und Φ die Variante

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - \Phi(x))^2}{\Phi(x)\,(1 - \Phi(x))} \, d\Phi(x)$$

angegeben. Es handelt sich also wie bei der Distanz nach Cramér und von Mises im Wesentlichen um die Beurteilung des quadratischen Abstands von empirischer und unterstellter Verteilungsfunktion. Dabei wird hier jedoch eine Gewichtung $(\Phi(x)(1 - \Phi(x)))^{-1}$ verwendet, welche bei Cramér und von

Mises konstant gleich Eins ist. Dieses Gewicht wird beim Anderson-Darling-Test an den Rändern groß, welchen damit mehr Bedeutung zukommt.

Wie der gleichen Quelle in Form der dortigen Tabelle 2 sowie der Gleichung (6) entnommen werden kann, ist die Verteilung von A^2 auch unter der Nullhypothese abhängig von der Stichprobengröße. Auch der kleinste mögliche Wert der Statistik ist nur asymptotisch gleich Null. Erst für 118 Beobachtungen wird beispielsweise die Schranke 0.01 für den kleinsten *möglichen* Wert unterschritten.

Shapiro-Wilk-Test

Im Jahr 1965 veröffentlichten Samuel S. Shapiro und Martin Wilk den später nach ihnen benannten „analysis of variance test for normality“ (Shapiro und Wilk (1965)). Dabei wird das Verhältnis von zwei Streuungsschätzern verglichen, wobei einer durch die gewöhnliche Stichprobenvarianz gebildet wird, also durch $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Für den zweiten Schätzer wird zunächst angenommen, dass die Beobachtungen x_i einer Normalverteilung mit Erwartungswert μ und Varianz σ^2 entstammen. Dann können die Werte auch durch

$$x_i = \mu + \sigma y_i$$

dargestellt werden, wobei die y_i aus einer Standardnormalverteilung stammen. Auf Basis der Methode der verallgemeinerten kleinsten Quadrate kann nun ein weiterer Schätzer für σ ermittelt werden. Dazu werden die geordneten y_i betrachtet, notiert als $y_{(i)}$. Dann ist für einen Vektor $m = (m_1, m_2, \ldots, m_n)$ der Erwartungswerte $m_i = E(y_{(i)})$ und eine Matrix $V = (v_{ij})_{1 \leq i,j \leq n}$ der Kovarianzen $v_{ij} = \mathrm{cov}(y_{(i)}, y_{(j)})$ auch das Quadrat von

$$b = \frac{mV^{-1}}{mV^{-1}m'}x$$

ein Schätzer für die Streuung. Die Teststatistik wird dann gebildet als

$$W = \frac{b^2}{(n-1)S^2}.$$

Für die Berechnung von m und V wird auf die Literatur verwiesen. Wichtig ist an dieser Stelle lediglich, dass bei einer Normalverteilung der x_i ein W von 1 zu erwarten ist. Des Weiteren müssen keine Schätzer für μ und σ herangezogen werden, so dass hier ohne Ergänzungen eine Abbildung der Art $\mathbb{R}^n \to \mathbb{R}$ vorliegt.

Als einzige Variante der im Rahmen der vorliegenden Ausarbeitung verwendeten Distanzen ist für diese Teststatistik bei Vorliegen einer Normalverteilung der größte Wert zu erwarten. Zu interpretieren ist W damit wie ein Ähnlichkeitsmaß. Um die Interpretationsrichtung „Je größer, desto entfernter von der Normalverteilung“ konsistent anzubieten, wird im Weiteren auch der Wert

$$W^* = 1 - W \in [0, 1]$$

verwendet. Eine kleines W^* deutet demnach auf normalverteilte Daten hin.

Jarque-Bera-Test

Der Jarque-Bera-Test beruht auf der Idee, dass das dritte und das vierte Moment der Normalverteilung unabhängig von den Parametern konstant sind: Alle Vertreter der Normalverteilungsfamilie sind symme-

trisch um den Erwartungswert, weisen also eine Schiefe von 0 auf. Die Wölbung, also das vierte zentrale Moment, berechnet sich zur Konstanten 3.

Mit der empirischen Schiefe M_3 und der empirischen Kurtosis M_4, das heißt mit

$$M_i := \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^i}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{i}{2}}},$$

berechnet sich die Teststatistik dann durch

$$JB = \frac{n}{6}\left(M_3^2 + \frac{(M_4 - 3)^2}{4}\right).$$

Die Normalverteilung wird also auf zwei Eigenschaften ihrer Dichtefunktion reduziert, und es werden die entsprechenden quadratischen Abstände betrachtet (wobei offenbar $M_3^2 = (M_3 - 0)^2$ ist). Diese Abstände werden durch die Stichprobengröße n normiert; je größer die Fallzahl ist, desto geringer dürfen die Abweichungen ausfallen, um JB nicht größer werden zu lassen.

Mit den Faktoren $1/6$ beziehungsweise $1/24$ werden die Abstädnde so skaliert, dass der Vergleich mit den Quantilen der χ_2^2-Verteilung zu einem Signifikanztest führt. Diese Verteilung ergibt sich bei einer Normalverteilung der Daten approximativ für die Größe JB (vgl. Jarque und Bera (1987), insbesondere Seite 165).

Wie auch in einer Diplomarbeit der Universität Würzburg ausführlich beschrieben (vgl. Hain (2010), Teile 3.2.1 und 3.2.2), bilden die beiden Summanden bereits für sich genommen und ohne Quadrierung die Möglichkeit, Tests durchzuführen: Es gilt für normalverteilte x_i , dass

$$\sqrt{\frac{n}{6}}M_3 \xrightarrow[n\to\infty]{D} N(0,1) \text{ und } \sqrt{\frac{n}{24}}(M_4 - 3) \xrightarrow[n\to\infty]{D} N(0,1).$$

Die Summe aus quadrierten standardnormalverteilten Zufallsvariablen konvergiert dann gegen eine χ_2^2-Verteilung.

Weitere Tests

Offenbar bilden die vorgestellten Verfahren nur eine Teilmenge aller veröffentlichten und insbesondere aller möglichen Testverfahren ab. Für einzelne Tests sind aber die grundlegenden Ideen bereits durch andere dargestellte Methoden abgebildet. Beispielsweise für die Tests nach Kolmogorov/Smirnov und Cramér/von Mises ist dies durch die gleichnamigen Distanzen der Fall. Der der Test nach Lilliefors andererseits stellt nur eine Spezialisierung des Kolmogorov-Smirnov-Tests dar, welche sich bezüglich der Teststatistik aber nicht unterscheidet.

Wie der Jarque-Bera-Test basieren auch die Tests von Anscombe/Glynn und von D'Agostino auf der Schiefe beziehungsweise auf der Schiefe und der Wölbung der zu beurteilenden Daten. Während die Grundidee dieses Vorgehens bereits durch die oben dargestellte Größe JB abgedeckt ist, und die Alternative wie bei D'Agostino et al. (1990) dargestellt zudem deutlich komplexer und damit schwerer zu interpretieren ist, wird die Variante von Jarque und Bera hier bevorzugt.

An zahlreichen Stellen finden sich weitere Hinweise auf Konzepte, welche zu weiteren Normalverteilungsmaßen führen. So widmen Patel und Read (1996) den „Characterizations" der Normalverteilung ein ganzes Kapitel. Beispielsweise unter der Nummer 4.2.5 (a) findet sich dort der Sachverhalt, dass für X_i

aus einer symmetrischen Verteilung die χ^2_{n-1}-Verteilung von $\sum_{i=1}^{n} {(X_i-\bar{X})}/{\sigma^2}$ äquivalent zur Normalverteilung der X_i ist. Um dies verwendbar zu machen, muss aber zunächst die Verknüpfung mit einer Beurteilung der Symmetrie vorgenommen werden (wobei ein entsprechender Test auch oben bereits genannt ist).

2.4 Umsetzung

Für den Großteil der vorgestellten Verfahren werden feste Werte für die Normalverteilungsparameter μ und σ^2 benötigt. Für Schätzungen stehen bekanntlich zahlreiche Methoden bereit, von denen das arithmetische Mittel und die Stichprobenvarianz als Standardmethodik bezeichnet werden können. An dieser Stelle stehen jedoch die Abstände eines Datenvektors zur Normalverteilung an sich, das heißt zur Familie der Normalverteilungen, im Vordergrund. Damit sollen die Distanzen und Statistiken ausdrücklich auch für nicht-normale Datenlagen berechnet werden. Inwiefern sich Mittelwert und empirische Varianz für diese Verwendung eignen, ist nicht klar. Statt der genannten Schätzer kann auch jeweils der naheste Vertreter der Normalverteilung für die Abstandsberechnung herangezogen werden, das heißt eine Distanz d wird für eine Dichte f der Normalverteilung und eine aus den Daten geschätzte Dichtefunktion $\hat{f}_n$ angegeben als

$$d(x) = \min_{\mu,\sigma} \; d\left(\hat{f}_n(x), f(x|\mu,\sigma)\right).$$

Dieses Vorgehen entspricht der minimum distance estimation, wie sie unter anderem bei Boos (1982) für Anderson-Darling-Abstände, bei Rüger (2002) (Seite 211) für χ^2-Abstände oder bei Beran (1977) für Hellinger-Distanzen beschrieben wird. In Rieder (1994) wird das Vorgehen für Kolmogorov- und Cramér-von-Mises-Distanzen besprochen (Seite 232 ff.). Mittelwert und empirische Varianz können dabei als Startwerte für Optimierungen verwendet werden. Eine Untersuchung zu den Vorteilen oder sogar der Notwendigkeit der Optimierung erfolgt später.

Offen ist, auf welche Weise der Schätzer $\hat{f}_n$ gebildet wird. Prinzipiell kommen dabei insbesondere alle Varianten der Kerndichteschätzung in Betracht. An dieser Stelle wird dazu das Vorgehen des R-Pakets distrEx verwendet (Ruckdeschel et al. (2006), siehe dazu auch Kohl (2005)). Für die Gegenüberstellung einer Dichtefunktion und einer diskreten und endlichen Menge an Datenpunkten wird dort, wie am Programmcode ersichtlich ist, die stetige Dichte „diskretisiert“ (bei Einstellung des Parameters asis.smooth.discretize auf ‚discretize‘). Damit wird das Integral der Distanzfunktionen zu einer Summe. Die Kernidee beruht dabei auf einem Gitter des Bereichs zwischen dem 0.001- und dem 99.999-Prozent-Quantil der zu f gehörenden Verteilung, welches dann den Träger der diskretisierten Verteilung bildet. Für zwei benachbarte Gitterpunkte $y_k < y_{k+1}$ sowie für $y_{k'} = {(y_k+y_{k+1})}/{2}$ ist dann $\hat{f}_n(y_{k'}) = F(y_{k+1}) - F(y_k)$.

Das distrEx-Paket stellt Funktionen für vier der fünf angeführten Metriken bereit, jedoch keine für die nach Lévy. Dabei ist dieses Paket nicht das einzige mit den jeweiligen Distanzfunktionen, hebt sich aber in einem Punkt ab: Beispielsweise die Funktion HellingerDist() akzeptiert als Argumente sowohl zwei Verteilungen, als auch eine Verteilung und einen numerischen Vektor. Für die Anwendung zur Messung der Normalität des Datenvektors wird genau die zweite Varianten benötigt - während die erste für theoretische Überlegungen hilfreich ist.

Für den Lévy-Abstand scheint aktuell keine Implementierung in der Programmiersprache R zur Verfügung zu stehen, somit war hier eine entsprechende Funktion zu erstellen. Es wird dabei auf Funktionen des distrEx-Pakets zurückgegriffen. Zwischen diskretisierten theoretischen und empirischen Verteilungsfunktionen wird kein Unterschied gemacht. Die Forderung der Definition „für alle reellwertigen x“ wird dabei nur durch die Auswertung auf einem Gitter realisiert. Dazu werden zwischen den Punkten, an

denen jeweils mindestens eine der beiden Verteilungsfunktionen größer als 10^{-5} und kleiner als $1-10^{-5}$ ist, äquidistante Gitter bestehend aus 5 000 Punkten verwendet. Für jeden dieser Punkte muss dann die Mengendefinition $F(x-\varepsilon)-\varepsilon \leq G(x) \leq F(x+\varepsilon)+\varepsilon$ erfüllt sein. Die Minimierung des Parameters ε wird dann durch einen Bisektionsalgorithmus realisiert. Dabei wird für die folgenden Anwendungen eine Intervallbreite von maximal 10^{-4} als Stopp-Kriterium verendet. Bei den Startwerten kann ausgenutzt werden, dass stets $0 \leq L \leq K \ (\leq 1)$ gilt.[3]

Für die Teststatistiken wird auf die R-Pakete nortest (für χ^2- und AD-Test) und moments (JB-Test) zurückgegriffen, der Shapiro-Wilk-Test wird durch das Basispaket stats bereitgestellt. Genaueres ist angegeben bei R Core Team (2017), Groß und Ligges (2015) sowie bei Komsta und Novomestky (2015). Optimierungen der Parameter werden mit dem Algorithmus von Nelder und Mead durchgeführt, wobei für die Varianz eine Restriktion auf den positiven Zahlenraum zu berücksichtigen ist.

Abschließend zu diesem Teil sei daraufhin gewiesen, dass alle bisher genannten Verfahren im Folgenden als „Distanzen", „Abstände" oder „Maße" bezeichnet werden. Dem liegt, neben dem einfacheren Sprachgebrauchs, die Idee zugrunde, dass hier alle Methoden als genau solches verwendet werden: Als Werkzeug, *Abstände* beziehungsweise *Distanzen* zu *messen*. Soll zwischen den oben ersichtlichen zwei Gruppen unterschieden werden und gehen diese nicht aus dem Zusammenhang hervor, wird sprachlich zwischen *Metriken* und *Teststatistiken* unterschieden, wobei jeweils der herkömmliche mathematisch-statistische Sinn gemeint ist.

2.5 Ausgewählte Verteilungen

Für einen Eindruck der angeführten Kandidaten zur Messung der Normalverteilung werden Simulationen mit verschiedenen Verteilungen durchgeführt. Die Notation $N(\mu, \sigma)$ führt dabei als zweite Größe stets die Standardabweichung, nicht die Varianz auf.

Als Beispiel wird zunächst eine Standardnormalverteilung mit Erwartungswert 0 und Standardabweichung 1 betrachtet. Von dieser ausgehend werden die Abstände zu den Normalerteilungen N(0,3), N(3,1) und N(3,3), sowie zu den Chi-Quadrat-Verteilungen χ^2_2 und χ^2_5 betrachtet. Weiter werden mit den t-Verteilungen t_5 und t_{20} sowie den Gleichverteilungen $U(0,6)$ und $U(-1.5,1.5)$ zwei symmetrische Verteilungsfamilien einbezogen. Die Poisson-Verteilung $P(2)$ stellt einen Vertreter diskreter Verteilungen dar.

Eine Reduzierung auf einzelne Verteilungen kann nicht allumfassend sein. Diese elf Vertreter sollten aber die Bandbreite der intuitiven Nähe (beziehungsweise des intuitiven Abstands) zur Normalverteilung abbilden. Die folgende Abbildung 3 zeigt die Dichten dieser elf Verteilungen, wobei sie aus Gründen der Übersichtlichkeit in Hälften geteilt wurden. Die Verteilungen werden hier und auch im Weiteren kurz durch einen einzelnen Buchstaben oder einen Buchstaben und die Parameter abgekürzt, so wird etwa kurz „die N-Verteilung" für die Normalverteilung oder „die t_5" für die t-Verteilung mit 5 Freiheitsgraden geschrieben.

Offenbar ist nun von einem guten Maß eine klare Trennung zwischen der Normal- und der Nicht-Normal-Verteilung eines Datensatzes zu erwarten. Dabei sind Abstufungen erwünscht, so sollte die diskrete Verteilung intuitiv den größten Abstand zur Normalverteilung haben. Die Gleichverteilung teilt mit der Symmetrie immerhin eine zentrale Eigenschaft der Normalverteilung, so auch die t-Verteilung. Durch die Verteilungskonvergenz der t_n-Verteilung gegen die $N(0,1)$-Verteilung (für $n \to \infty$) ist *eine irgendwie geartete Nähe* der t- zur N-Verteilung zu erwarten, wobei die Distanz für t_{20} zumindest nicht größer als für t_5 sein sollte. Schließlich kann gefordert werden, dass die vier Normalverteilungen alle zu Werten

[3] Dass $0 \leq L$ und $0 \leq K \leq 1$ ist, folgt direkt aus den Definitionen und aus der Begrenzung von Verteilungsfunktionen auf das Intervall [0,1]; wegen $L \leq K$ folgt auch $L \leq 1$. Die Ungleichung von Lévy- und Kolmogorov-Distanz findet sich mit Beweis bei Huber und Ronchetti (2009) auf Seite 36 als Ungleichung (2.25).

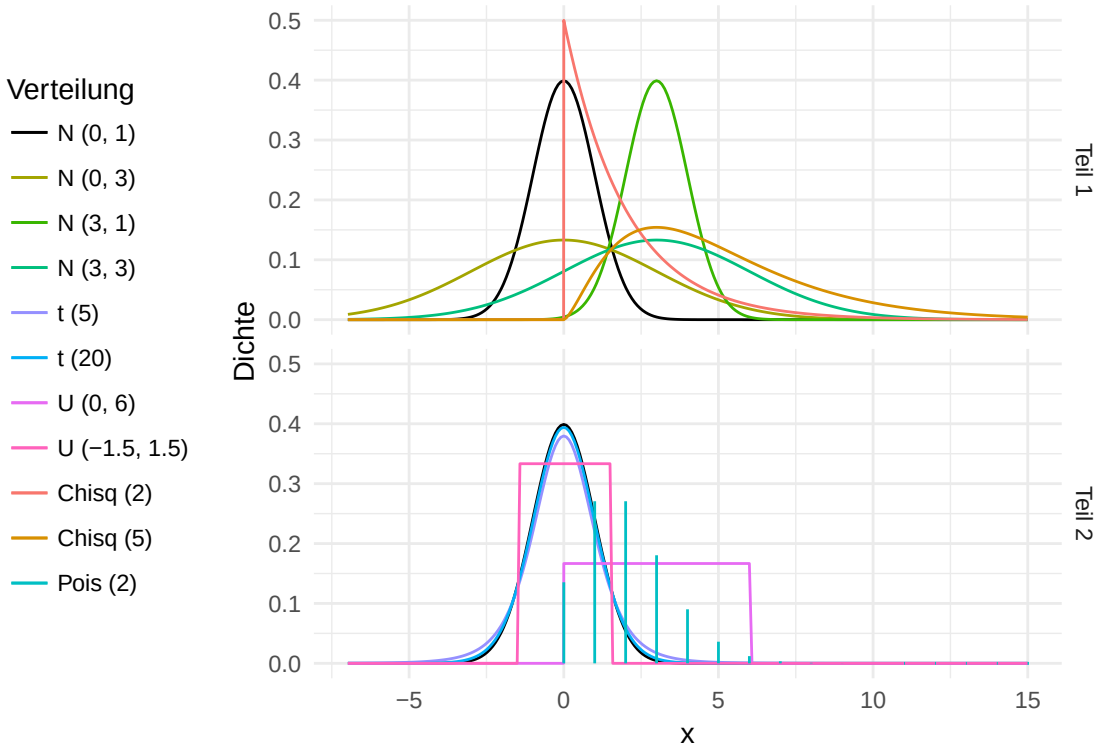

Abbildung 3: Die elf im Folgenden betrachteten Verteilungen. Betrachtet werden zunächst beispielhaft die jeweiligen Abstände zur schwarz dargestellten N(0,1). Diese ist in beiden Teilen eingezeichnet.

gleicher Größenordnung führen, wenn lediglich die Distanz zu „irgendeiner“ Normalverteilung, also zur Normalverteilungsfamilie insgesamt, untersucht wird.

3 Allgemeine Vergleiche

3.1 Vergleich der Konzepte

Vor Bewertungen sollen hier die verschiedenen Konzepte der neun oben vorgestellten „Kandidaten“ gegenübergestellt werden. In der Literatur wird beim Vergleich der Methoden im Allgemeinen die Teststärke (Power) in den Mittelpunkt gestellt, so etwa bei Hain (2010), Sürücu (2008), Thadewald und Büning (2007) oder Yap und Sim (2011). Allein diese ist aber, wie oben angeführt, hier nicht von Interesse. Resultate, welche bestimmte Gruppen von Alternativhypothesen betrachten, können aber Hinweise liefern, welche Statistiken gute Indikatoren für Abweichungen von der Normalverteilung sind.

Solche Gruppierungen verwendet Arghami (2011) bei der Konzeption einer Simulationsstudie. Dabei ergeben sich unterschiedliche beste Verfahren, je nach Verteilung der Daten gemäß Alternativhypothese. Symmetrische, nicht-normale Verteilungen mit dem ganzen Raum der reellen Zahlen als Träger werden am besten vom Jarque-Bera-Test erkannt. Das deutet darauf hin, dass die Teststatistik bereits recht sensibel auf von 3 abweichende Wölbungen reagiert. Bei asymmetrischen Verteilungen mit gleichem Träger fallen die Ergebnisse des Shapiro-Wilk-Tests am besten aus, allerdings werden hier auch ausschließlich Gumbel-Verteilungen betrachtet. Diese weist eine konstante Schiefe von etwa 1.14 und eine konstante Wölbung von 5.4 auf. Dass der JB-Test hier nur (mit wenig Abstand) den zweiten Platz belegt könnte darauf zurückzuführen sein, dass der Shapiro-Wilk-Test ebenfalls gut auf die Schiefe reagiert: Schiefe

Verteilungen produzieren Ausreißer, welche wiederum die Varianzschätzer stark beeinflussen. Ist dieser Einfluss für die beiden in die Teststatistik W eingehenden Schätzer unterschiedlich, zeigt diese entsprechend den Unterschied an. Für Verteilungen mit dem Träger $(0, \infty)$ weisen der Shapiro-Wilk-Test und der (hier nicht weiter besprochene) Test von Vasicek die besten Werte auf, für solche mit Träger $(0, 1)$ nur letzterer. Es ist jedoch ohne Weiteres nicht klar, wie sich diese Träger auf die Qualität der Tests auswirken.

Auch Monografien und Handbücher liefern neben Kommentaren zur Power zumeist, wie beispielsweise Hartung (2005), lediglich eine Beschreibung einzelner Tests, ohne allgemeine Vor- und Nachteile eingehender und vergleichend zu diskutieren. Bei Judge et al. (1985) findet sich zwar der Verweis auf auf die „enorme“ Literatur zum Testen auf Normalität (Seite 826), die Autoren geben dann aber nur die Vorschrift zur Berechnung des Shapiro-Wilk-Tests an: Dieser habe demnach „performed reasonably well in a wide variety of circumstances“ (ebenda).

Eine positive Ausnahme stellt Groß (2004) dar: Hier wird eine Einteilung von Tests auf Normalverteilung in drei Gruppen vorgenommen: Die Gruppe der χ^2-Tests, die Gruppe der auf der empirischen Verteilungsfunktion beruhenden Tests und die Gruppe der auf Regression und Korrelation beruhenden Tests (vgl. Seite 106 f.). Die dritte Gruppe enthält Methoden, welche auf der Gegenüberstellung der geordneten Datenreihe mit der bei Normalität zu erwartenden Reihe beruht. Mit dieser Klassierung werden dann der Chi-Quadrat-Test (erstgenannte Gruppe), der Kolmogorov-(Smirnov-)Test, der Cramér-von-Mises-Test, der Anderson-Darling-Test (zweite Gruppe) und der Shapiro-Wilk-Test (dritte Gruppe) zugeordnet. Die Lévy-Metrik lässt sich zudem klar der zweiten Gruppe zuordnen.

Nicht berücksichtigt werden im Vergleich mit den hier vorgestellten Verfahren die Hellinger-Distanz, der Totalvariationsabstand und der Jarque-Bera-Test. Wie die obige Abbildung 2 bereits verdeutlichte, lassen sich die Distanzmaße in Dichte- und Verteilungsfunktion-basierte Methoden aufteilen. Somit kann eine ergänzende, vierte Gruppe vorgeschlagen werden: Die der Dichte-basierten Methoden. Dieser kann auch der Test von Jarque und Bera zugeordnet werden, welcher mit Schiefe und Wölbung gerade Eigenschaften der Dichtefunktion untersucht.

Eine Verteilungsfunktion ist bekanntlich stets monoton nichtfallend, rechtsseitig stetig und konvergiert für $x \to \mp\infty$ gegen 0 beziehungsweise 1 (vgl. bspw. Hartung (2005), Seite 106). Damit ist zumindest die optische Form recht stark eingegrenzt, für die Dichtefunktion bestehen mehr geometrische Möglichkeiten: So fällt beispielsweise der Unterschied von Dichtefunktionen stetiger und diskreter Verteilungen, oder uni- und bimodaler Verteilungen direkt ins Auge. Ob diese Mehrzahl an Charakteristika aber auch von den Methoden gewinnbringend berücksichtigt wird, ist zu untersuchen.

Wieder bei Groß (2004) im Kapitel 4 finden sich Empfehlungen, welche Eigenschaften der Verfahren beinhalten. So wird vom χ^2-Test abgeraten, nicht zuletzt da das Resultat von der Wahl der Klassen abhängt. Weitere Argumente behandeln dann im Wesentlichen die Teststärke. Bemerkenswert ist das auf Seite 113 angeführte Zitat von R. D'Agostino, nach dem der Kolmogorov-Smirnov-Test nur eine „historische Kuriosität“ sei, welche „niemals verwendet werden sollte“. Als gut geeignet werden hingegen die Tests nach Anderson/Darling, Cramér/von Mises, Shapiro/Francia und Shapiro/Wilk empfohlen. Inwieweit sich diese Ergebnisse übertragen lassen, wenn die Teststatistiken selbst betrachtet werden, weitere Verfahren hinzugenommen werden und die angeführten Verteilungen als Beispiele verwendet werden, wird die folgende Analyse zeigen.

Es lassen sich weitere Merkmale der Verfahren ausmachen, so wurde die Unterscheidung zwischen Metriken und Teststatistiken bereits deutlich. Auch fällt die Spezialisierung auf die Normalverteilung ins Auge. So sind mit Ausnahme des Shapiro-Wilk- und des Jarque-Bera-Tests alle vorgestellten Verfahren nicht an die Normalverteilung gebunden, sondern lassen sich auch auf andere Verteilungen übertragen. Es ist zu erwarten, dass die Spezialisten bei ausschließlicher Betrachtung der Anwendung für Normalverteilungen bessere Resultate liefern.

Rüger (2002) führt in Teil 3.1.5 (ab Seite 41) Zusammenhänge von Verteilungskonvergenzen an, wenn diese für verschiedene Metriken definiert werden. Eine Folge von Verteilungen P_n ist demnach gegen eine Verteilung P konvergent bezüglich einer Metrik d , wenn $d(P_n, P) \to 0$ $(n \to \infty)$ gilt. Dies wird als $P_n \xrightarrow{d} P$ notiert. Es bestehen die Zusammenhänge

$$P_n \xrightarrow{TV} P \Leftrightarrow P_n \xrightarrow{H} P \Rightarrow P_n \xrightarrow{K} P \Rightarrow P_n \xrightarrow{L} P,$$

(vgl. a. a. O. (3.41) sowie Satz 3.6). Wenn sich also zwei Verteilungen im Sinne von Hellinger oder dem Totalvariationsabstand „nah" beieinander befinden, dann auch im jeweils anderen Sinne und im Sinne von Kolmogorov und Lévy. Zeigt die Lévy-Metrik die Nähe zweier Verteilungen an, muss dies aber offenbar nicht für die anderen gelten. Es wird also nicht überraschen, wenn das Lévy-Maß tendenziell häufiger Nähe zur Normalverteilung anzeigt als die anderen Maße. Ob das förderlich oder hinderlich ist, und ob eventuell die Kolmogorov-Metrik als „mittleres" Maß der dargestellten Implikationen die ausgewogensten Zuordnungen trifft, ist zu untersuchen.

3.2 Vergleiche für theoretische Verteilungen

Die Frage nach der Distanz zur Normalverteilungsfamilie kann auf zwei Weisen gestellt werden: Zum einen für eine Verteilung als Ganzes, zum anderen für eine Realisierung in Form eines Datenvektors. Die erste Variante wird in diesem Kapitel betrachtet, die Übertragung auf die Datensituation folgt unten im Kapitel 4. Abstände zwischen jeweils zwei Verteilungen können dabei durch die Metriken per definitionem berechnet werden, also zwischen jeweils zwei theoretischen Verteilungen in ihrer Darstellung durch eine Dichte- oder Verteilungsfunktion. Für die im Teil 2.5 angeführten Verteilungen ergeben sich dann Werte wie weiter unten angegeben.

Da die Teststatistiken in ihrer Grundform von einem Datenvektor ausgehen, müssen Anpassungen vorgenommen werden, um theoretische Verteilungen miteinander vergleichen zu können. Die Chi-Quadrat-Statistik etwa vergleicht die erwarteten und beobachteten Häufigkeiten in jeweils einer Klasse miteinander. Für gegebene Klassengrenzen können auch die erwarteten Häufigkeiten zwei theoretischer Verteilungen verglichen werden. Der Vergleich von relativen Häufigkeiten ermöglicht die Unabhängigkeit von der Stichprobengröße n, wobei dann statt der Größe χ^2 der Wert χ^2/n betrachtet wird.

Die Festlegung der Klassengrenzen wird hier anhand einer der beiden zu vergleichenden Verteilungen so vorgenommen, dass für jedes Intervall i ein Anteil $E_i \equiv 1/k$ zu erwarten ist. Über Quantils- und Verteilungsfunktionen können dann die erwarteten relativen Häufigkeiten für die zweite Verteilung bestimmt werden. Wird nun außerdem jeweils statt χ^2/n der Wert $1/(k-1) \cdot 1/n \cdot \chi^2$ verwendet, das heißt die durchschnittliche relative Abweichung pro Klasse, ist auch eine Unabhängigkeit von der Anzahl der Klassen gegeben. Dabei sollte k offenbar nicht zu klein gewählt werden, um Streuungseinflüsse zu begrenzen.

Der Anderson-Darling-Test stellt in der Definitions-Variante über das Integral eine empirische und eine theoretische Verteilungsfunktion gegenüber. Dabei kann zwar prinzipiell einfach die empirische Funktion gegen eine theoretische Funktion ausgetauscht und das Integral berechnet werden - je nach Verteilung ändert sich hier aber das (nach Lebesgue–Stieltjes definierte) Integral. Dessen recht aufwändige Berechnung würde im Aufwand den Nutzen vermutlich deutlich überschreiten, so dass der AD-Test hier nicht berücksichtigt wird.

Das Gleiche gilt für den Test nach Shapiro und Wilk. Dabei werden, wie oben beschrieben, zwei Varianzschätzer verglichen, wobei einer der Schätzer auf Erwartungswerten und Kovarianzen von sortierten und transformierten Beobachtungen beruht. Deren Herleitung für alle der hier verwendeten nicht-normalen Verteilungen gestaltet sich recht aufwendig.

Die Variante nach Jarque und Bera ist nur abhängig von Schiefe und Wölbung, beide Kennzeichen sind für die theoretischen Verteilungen bekannt. Einzige Variable in der obigen Definition von JB ist dann die Stichprobengröße n, welche für den Vergleich zweier theoretischer Verteilungen jedoch ohne Interpretationsverlust ausgelassen werden kann. Betrachtet wird dann die Größe $1/nJB$.

Zu beachten ist sowohl für die χ^2- als auch für die JB-Variante: Bei einer perfekten Übereinstimmung zweier Verteilungen, und simplem Einsetzen der Parameter (Schiefe ist gleich, usw.), ergeben sich Werte von $\chi^2 = JB = 0$. Da aber in beiden Fällen keine symmetrische Verteilung der Teststatistik vorliegt (folgend aus dem Betrachten jeweils quadratischer Abstände), beträgt der Erwartungswert für die Untersuchung eines Datenvektors jeweils nicht Null.

Die Werte für die vier Metriken und die beiden berücksichtigten Teststatistiken finden sich in der folgenden Tabelle 1. Dabei wird jeweils gemessen, wie nah die genannten Verteilungen an der Standardnormalverteilung sind.

	Kolm.	TV	Hell.	Lévy	CvM	$\frac{1}{n(k-1)}\chi^2$	$\frac{1}{n}JB$
N (0, 1)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
N (0, 3)	0.242	0.484	0.475	0.166	0.218	0.092	0.000
N (3, 1)	0.866	0.866	0.822	0.552	0.741	0.567	0.000
N (3, 3)	0.625	0.653	0.618	0.391	0.557	0.345	0.000
t (20)	0.008	0.016	0.031	0.005	0.007	0.000	0.006
t (5)	0.030	0.061	0.109	0.018	0.028	0.001	1.500
U (-1.5, 1.5)	0.067	0.185	0.284	0.025	0.059	0.003	0.060
U (0, 6)	0.687	0.687	0.737	0.483	0.588	0.374	0.060
Chisq (2)	0.500	0.500	0.625	0.379	0.359	0.100	1.767
Chisq (5)	0.847	0.847	0.844	0.549	0.748	0.643	0.907
Pois (2)	0.706	1.000	1.000	0.312	0.537	0.262	0.094

Tabelle 1: Die fünf Metriken und die zwei angepassten Teststatistiken als Beurteilungsgrundlage des Abstands verschiedener Verteilungen von der N(0,1).

Alle Maße erkennen die Identität, wie die erste Zeile zeigt. Die veränderte Streuungskomponente (N(0,1) verglichen mit der N(0,3), zweite Zeile) bewerten die TV- und H-Metriken nahezu gleich, während die K- und L-Messungen hier weniger stark ausschlagen. Die vergleichsweise geringe Schnittmenge der N(0,1)- und der N(3,1)-Dichten erkennen die fünf Metriken als solche mit Werten zwischen 0.55 und 0.87. Bei der Variante in der vierten Zeile, also beim Vergleich von N(0,1) und N(3,3), erkennen alle vier Metriken einen mittelgroßen Abstand. Numerisch fällt die Lévy-Variante dabei etwas ab.

Die durchschnittliche χ^2-Statistik schwankt numerisch relativ stark innerhalb der Vergleiche von Normalverteilungen, bewertet aber ebenfalls die vorgenommene Streuungs-Änderung deutlich schwächer als die Verschiebung. Der Jarque-Bera-Test hingegen kann hier konstruktionsgemäß keine Unterschiede feststellen, da alle Normalverteilungen die gleiche Schiefe und die gleiche Wölbung aufweisen

Für die t-Verteilungen werden wie zu erwarten von fast allen Varianten eher kleine Abstände ausgegeben. Dass etwa die Dichte der t_{20} näher an der N(0,1) ist als die der N(3,3), ist nach Betrachtung der Funktionsverläufe auch eine sinnvolle Interpretation (die Näherung an *irgendeine* Normalverteilung wird erst später betrachtet). Einzig der JB-Abstand erkennt für die t_5-Verteilung eine recht klare Abweichung von der N(0,1), welche auf die unterschiedliche Wölbung zurückzuführen ist (9 bei der t_5, 3 bei der N-Verteilung).

Für die Gleichverteilungen hingegen hat der JB-Abstand kaum Indizien auf Abweichungen von der N(0,1): Schiefe und Kurtosis sind Parameter-unabhängig konstant, nämlich 0 und 1.8, so dass nur der Unterschied in der Wölbung zum Tragen kommt. Die Totalvariations- und Hellinger-Metriken erkennen die

Unterschiede von N(0,1)- und Gleichverteilungs-Dichten bzw. -Verteilungsfunktionen relativ gut, insbesondere wenn dazu die Lageverschiebung (U (0, 6)) kommt. Die Methoden nach Kolmogorov, Lévy und CvM, sowie der χ^2-Wert, scheinen genau diese Lageverschiebung für eine große Sensibilität zu benötigen.

Die Dichten der χ^2-Verteilungen sind ausschließlich echt positiv, wenn auch das Argument größer als Null ist, und mit Schiefen von 2 bzw. 1.3 weichen sie recht gut sichtbar von der N(0,1)-Dichte ab. Gerade für die Trennung der Standardnormalverteilung von der Chi-Quadrat-Verteilung mit zwei Freiheitsgraden scheint die Jarque-Bera-Methode am besten geeignet. Auffällig ist, dass sie als einzige Methode den Abstand von N(0,1) und χ^2_2 größer bewertet, als den von N(0,1) und χ^2_5 (vgl. neunte und zehnte Zeile). Hier zeigen sich die unterschiedlichen Konzepte von Form- und Flächendifferenzen. Innerhalb der Metriken fällt auf, dass hier die Kolmogorov-Variante numerisch näher an TV und Hellinger ist als an der Lévy-Methode, der sie konstruktiv näher ist. Ein solches Paar-Verhalten findet sich auch bei den χ^2-Verteilungen wieder: Gleiche (gerundete) Werte bei K- und T-Metrik, abweichende Beurteilung durch die H-Metrik. Der Unterschied fällt bei der χ^2_5-Verteilung jedoch geringer aus.

Die intuitive Forderung nach einem maximal großen Abstand der stetigen N(0,1) zur diskreten Poisson-Verteilung erfüllen der T- und der H-Abstand ohne Einschränkung. Das ist plausibel, das sich anhand von Dichtefunktionen der Unterschied zwischen stetigen und diskreten Trägern auch grafisch sehr gut verdeutlichen lässt. Die K-, L und CvM-Metriken erkennen den Abstand wesentlich weniger gut (die diskrete Verteilungsfunktion bildet eine Treppenfunktion), zumindest aber deutlich. Nur moderate Abweichungen von der N(0,1) erkennt die χ^2-Variante. Der Abstand nach Jarque und Bera kann kaum Unterschiede aufzeigen: Offenbar weichen Schiefe und Wölbung der Poisson (2)-Verteilung mit $1/\sqrt{2}$ und $1/2$ nicht stark genug von dener der N(0,1) ab.

Insgesamt erscheinen mit diesen Eindrücken alle sechs Varianten als grundsätzlich geeignet: Insgesamt ergeben sie jeweils ein schlüssiges Bild, und Schwächen gegenüber anderen Methoden bei einzelnen Vergleichen werden durch Stärken bei anderen Vergleichen kompensiert. Auch inakzeptables Verhalten, wie beispielsweise das Nicht-Erkennen der Gleichheit zweier Verteilungen, fällt nicht auf.

In der Qualität abfallend gegenüber den Alternativen ist aber der χ^2-Abstand, welcher bei Vorliegen eines anderen der Maße keine neuen Information liefern würde. Dieses Ergebnis ist einerseits bemerkenswert, da dieser Methode ein sehr großer Bekanntheitsgrad attestiert werden kann - welcher sich aber vermutlich vor allem in der vergleichsweise einfachen Idee und der Flexibilität der Methode begründet. Zudem ist zu berücksichtigen, dass hier für Vergleiche von stetigen Verteilungen Klassen gebildet wurden. Dass sich dieser Informationsverlust letztlich in der Qualität niederschlägt, ist nicht verwunderlich.

3.3 Folgerungen

Die abgeleiteten Eigenschaften der einzelnen Verfahren werden nun zusammengefasst. Eine Übersicht findet sich anschließend am Ende dieses Teils auf der Seite 22 in Form der Tabelle 2. Diese ordnet die im Folgenden dargelegten Charakteristiken jeweils als „erfüllt“, „nicht erfüllt“ oder „teilweise erfüllt“ ein.

Die Gruppierung der Methoden in Metriken und Teststatistiken bei deren Vorstellung im Kapitel 2 war durch die die grundsätzliche Herangehensweise motiviert: Ein Teil untersucht die Abstände zweier Funktionen, der andere Teil wurde zur Beurteilung eines Datenvektors konstruiert. Wünschenswert ist es hingegen, die gleiche Methode für beide Fälle anwenden zu können. Es ist, wie oben ausgeführt, prinzipiell auch in allen Fällen möglich, diesen Übergang zu gewährleisten. Durchgeführt wurde dies jedoch nur teilweise durch allgemeine Verfahren wie Diskretisierungen stetiger Verteilungen. Eine effiziente und gut untersuchte, sowie direkte Übertragung der jeweilige Idee, ist zum jetzigen Zeitpunkt nur für die Methoden nach Anderson-Darling, Cramér-von Mises und Kolmogorov bekannt (und zwar in Form der jeweils gleichnamigen Distanzen und Teststatistiken).

Für die anderen Methoden ist diese Forderung nur als teilweise oder gar nicht erfüllt anzusehen: Etwa der Gedanke des Jarque-Bera-Tests, der Vergleich von Schiefe und Wölbung zweier Verteilungen also, ist auf alle Verteilungen und empirischen Werte übertragbar. Wirklich untersucht, inklusive der Verwendung einer zweckdienlichen Gewichtung der Abstände, sind solche Adaptionen bisher aber nicht. Der SW-Test beruht auf dem Vergleich zweier Varianzschätzer. Dieser Vergleich ergibt für theoretische Verteilungen keinen Sinn, da dann die Varianzen eindeutig bekannt sind und nicht geschätzt werden müssen. Werden ersatzweise zwei Varianzen verglichen, um zwei Verteilungen zu vergleichen, hat sich die Vorgehensweise schon weit vom Shapiro-Wilk-Grundgedanken gelöst, und würde auch eher dem F-Test auf Varianzgleichheit als empirischem Analogon entsprechen.

Die oben in Teil 2.1 gestellte Forderung nach Symmetrie kann nur für den Vergleich von zwei Verteilungen oder zwei Datenvektoren sinnvoll formuliert werden. Da hier jeweils der Abstand zu einer (Normal-)Verteilung von Interesse ist, kann die Forderung weiter auf den Vergleich von zwei Verteilungen abgeschwächt werden. Die Gruppe der *Metriken* erfüllt eine Definition, welche die Symmetrie-Eigenschaft enthält. Die Teststatistiken hingegen mussten für Betrachtungen zwischen Verteilungen angepasst werden. Dabei läuft mit den Gewichtungen durch die erwartete Klassenhäufigkeit für eine der Verteilungen beim χ^2-Test direkt eine zentrale Eigenschaft der Symmetrie-Forderung entgegen. Auch der Anderson-Darling-Test verwendet eine bei Integral-Schreibweise leicht erkennbar asymmetrische Vorgehensweise. Beim Jarque-Bera-Test werden die quadratischen Abstände von Schiefe und Wölbung zweier Verteilungen verglichen. Ein Vertauschen der Argumente führt offenbar zu gleichen Resultaten, so dass hier Symmetrie vorliegt. Bei Shapiro-Wilk-Test besteht keine Variante für zwei Verteilungen, und damit insbesondere auch keine symmetrische.

Wichtig für die Interpretierbarkeit der numerischen Resultate sind absolute Bezugsgrößen. So ist etwa für die fünf Metriken klar, dass Werte nahe der Null für zwei ähnliche Verteilungen stehen, Werte nahe Eins entsprechend für sehr unterschiedliche. Mit welchen Zahlenbereichen der Begriff „nahe“ belegt wird ist ohne weitere Betrachtungen zwar nicht klar, ohne beidseitige Beschränkungen der Werte ist aber zumindest einer der Begriffe „klein“ und „groß“ gar nicht fassbar. Bei der Verwendung im Signifikanztest werden die Größen χ^2 und JB mit (H_0-)χ^2-Verteilungen verglichen. Das verdeutlicht, dass keine oberen Schranken angegeben werden können. Die Normierung durch ein großes Quantil könnte zwar ein praktikables Vorgehen sein, wäre aber weiter zu untersuchen. Für die Chi-Quadrat-Statistik wäre dabei auch der Einfluss der Klassenzahl zu betrachten, diese beeinflusst beim Chi-Quadrat-Test die Verteilung der H_0-Teststatistik über die Anzahl der Freiheitsgrade.

Auch für den AD-Test kann keine obere Schranke angegeben werden. Das wird an der hier verwendeten Summen-Schreibweise klar: Neben den Gewichten beinhaltet jeder der Summanden einen Term $-\log(\Phi(z_{(i)})) - \log(1 - \Phi(z_{(n+1-i)})$. Da keine Anforderungen an die z- oder die zugrunde liegenden $x-$ Werte gestellt werden, kann dieser Term jede beliebige Grenze überschreiten. Einzig der Shapiro-Wilk-Test führt zu einer beschränkten Größe, wobei W wie bereits erwähnt in das Intervall von 0 bis 1 fällt (und folglich ebenso die Variante $W^* = 1 - W$).

Nicht weiter verglichen werden an dieser Stelle die Resultate der theoretischen Abstände verschiedener Verteilungen zur Standardnormalverteilung, wie sie oben in der zugehörigen Tabelle 1 (Seite 19) dargestellt wurden: Diese dienen vor allem einem ersten Eindruck der Maße. Auch sollten alle Schlussfolgerungen, die mit Rücksicht auf die Ergebnisse zu ziehen sind, sich auch in der Simulationsstudie wiederfinden - andernfalls wäre gerade die Tatsache dieses nicht konsistenten Bildes zu bewerten.

Eine Zusammenfassung der Eigenschaften folgt nun als Tabelle 2. Für weitere Schlussfolgerungen sind die Resultate der im nächsten Kapitel beschriebenen Simulationsstudie hinzuzuziehen. Eine gemeinsame Betrachtung erfolgt dann im Kapitel 5.

Eigenschaft	Kolm.	TV	Hell.	Lévy	CvM	χ^2	JB	AD	W
Theorie und Daten	●	◐	◐	◐	●	◐	◐	●	○
Symmetrie	●	●	●	●	●	○	●	○	○
Beschränktheit	●	●	●	●	●	○	○	○	●

Tabelle 2: Die im Text genannten Eigenschaften in Zusammenfassung. Die geforderten Eigenschaften sind erfüllt oder nicht, was durch die Symbole ● und ○ gekennzeichnet wird. Die teilweise Erfüllung, markiert mit ◐, wird ebenfalls jeweils im Text erläutert.

4 Simulationsstudie

Die insgesamt neun Varianten zur Messung der Normalität werden nun nicht nur bezüglich einiger theoretischer Eigenschaften, sondern auch durch eine Simulationsstudie verglichen. Die dazu ausgewählte Vorgehensweise und die Ergebnisse werden in diesem Kapitel vorgestellt.

4.1 Simulationsdesign

Statt des Abstands zu einer bestimmten Verteilung soll der Abstand von empirischen Daten zur Normalverteilungsfamilie insgesamt, das heißt praktisch zu *irgendeiner* oder *der nahesten* Normalverteilung, beurteilt werden. Das bedeutet, dass statt zwei Verteilungen nun Realisationen der einen und die Familie der anderen Verteilung vorliegen. Wie beschrieben werden dazu mittels Minimierung der Maße Parameterschätzer ermittelt und die Normalverteilungsfamilie somit auf einen Vertreter reduziert. Weiter wird dessen Dichte dann an abzählbar und endlich vielen Stellen in den Vergleich mit den Realisationen gestellt.

Für die Simulation werden Zufallsstichproben aus den oben angeführten elf Verteilungen gezogen und dafür jeweils die Abstandsmessung zur Normalverteilung durchgeführt. Dabei werden Stichproben der Größen 10, 100 und 1 000 gewählt. Dabei könnten - gerade bei den verwendeten Teststatistiken, aber auch für die Metriken - für den kleinsten Wert 10 gegebenenfalls Probleme durch unterstellte Konvergenzen der Maße aufgezeigt werden. Bei den Varianten 100 und 1 000 sollten diese Probleme allenfalls nur noch in zu vernachlässigender Größenordnung bestehen. Dann kann durch den Faktor 10 gut geprüft werden, ob weitere Abhängigkeiten von der Stichprobengröße bestehen. Diese können offenbar nicht nur in den Maßen selbst bestehen, auch die Diskretisierung der Verteilungen zum Vergleich mit einer Stichprobe können hiervon betroffen sein.

Es werden für jede Stichprobengröße und jede Verteilung 100 mal Zufallszahlen gezogen, und damit insbesondere die Maße für die jeweils gleichen Daten verglichen. Dass diese Anzahl an Wiederholungen ausreicht zeigt die Streuung im Folgenden präsentierten Ergebnisse. Für die Ziehung der Zufallszahlen wurden die in den „Standard"-Funktionen des R-Paktes stats hinterlegten Methoden verwendet, also die Funktionen rnorm(), rchisq(), usw.

Die Resultate ergeben sich dann für ein Gitter bestehend aus neun Distanzmaßen, elf Verteilungen, drei Stichprobengrößen und einhundert Wiederholungen, das Ergebnis der Simulation besteht also aus $9 \cdot 11 \cdot 3 \cdot 100 = 29\,700$ reellwertigen Zahlen. Diese werden nun bezüglich der verschiedenen Einstellungen verglichen.

4.2 Ergebnisse

Die Abbildung 4 auf der Seite 24 zeigt einen beispielhaften Einzel-Lauf der Simulation, wobei nur ein Teil der elf Verteilungen einbezogen ist. Dargestellt sind jeweils Histogramme von Zufallsstichproben

und eine Kerndichteschätzung. Der Fokus wird auf die Parameterschätzung gelegt, daher sind die für die fünf Metriken optimalen Normalverteilungen ergänzt. Außerdem ist die Normalverteilung für Stichprobenmittel und -varianz eingezeichnet, was dem Startwert der Optimierungen entspricht. Diese Schätzer werden auch für die Berechnung der Teststatistiken verwendet, soweit nötig.

Zu erkennen ist, dass der Unterschied zwischen den Schätzern erwartungsgemäß abhängig von der Verteilung der Daten ist: Für die Normal- und die eher „ähnliche“ t-Verteilung ergeben sich praktisch keine Unterschiede zwischen den angepassten Normalverteilungen. Insbesondere für die recht schiefe χ^2_2-Verteilung hingegen ergeben sich Unterschiede in den angepassten Dichten. Dass hier keine dieser Anpassungen der Kerndichte entspricht, ist zu erwarten: Die χ^2_2-Verteilung ist eben insbesondere keine Normalverteilung. Geringer, aber nicht weniger offensichtlich, fallen die Abweichungen bei der Gleichverteilung aus. Bei der diskreten Poisson-Verteilung ergeben sich zwei Gruppen, wobei hier auch der gewählte Gauß-Kernel nicht gut geeignet ist, die Dichte der Daten durch eine Kerndichte zu repräsentieren.

Insgesamt ergibt sich der Eindruck, dass sich die Optimierung vor allem für Daten aus besonders Normalverteilungs-fernen Strukturen in nennenswerter Größenordnung auswirkt. Gerade dann aber ist fraglich, ob diese Unterschiede noch von praktischem Interesse sind. Für die Simulation bedeuten die Erkenntnisse aus der Abbildung, dass die verschiedenen Varianten durch die Optimierung Distanzen zu verschiedenen Normalverteilungen berechnen können. Dabei ist unklar, ob diese Unterschiede auch immer relevant sind. Weitere Betrachtungen zur Optimierung werden unten in Teil 6.2 folgen.

Die Abbildungen 5 und 6 (sowie die Abbildung 13 im Anhang, siehe Seite 42) stellen die Ergebnisse aus der Simulation dar, wobei zwischen den Stichprobengrößen unterschieden wird. Es wird jeweils eine Teilgrafik für jedes Distanzmaß verwendet, die Skalierungen sind entsprechend unterschiedlich und orientieren sich insbesondere an den realisierten Werten, nicht an den gesamten möglichen Intervallen. Eingezeichnet ist jeweils eine horizontale Linie. Diese bildet das 95 %-Quantil ab, welches sich für jeweils alle $4 \cdot 100$ Werte der Normalverteilungen ergibt. Damit soll eine Orientierungshilfe zur „Trennschärfe“ der Methoden gegeben werden: Idealerweise wären offenbar alle Werte nicht-normaler Daten oberhalb einer solchen Linie.

In der Abbildung 5 mit den Resultaten für die Stichprobengröße 10 ist zu erkennen, dass die Unterschiede zwischen einem Datenvektor und der Normalverteilungsfamilie zumindest für kleine Fallzahlen kaum festzustellen sind. Bei derartig kleinen Stichproben sind auch die zwei Freiheitsgrade in Form der zu schätzenden Verteilungsparameter noch relevant. Zunächst fällt auf, dass alle Maße für alle vier Normalverteilungen erwartungs- und forderungsgemäß gleiche Verteilungen bilden, die jeweils vier linken Boxplots lassen keine systematischen Unterschiede zueinander erkennen.

Während die horizontale Anordnung der Verteilungen tendenziell steigende Boxplots erwarten lässt, sind hier kaum Unterschiede zu notieren. Allein die Poisson- und zumeist auch die χ^2_2-Verteilung fallen augenscheinlich „aus der Reihe“ der anderen grafisch zusammengefassten Ergebnisse. Dabei ist einzig die Hellinger-Distanz in der Lage, den Poisson-Beispielen einen Abstand zuzuordnen, welcher durchweg den für Normalverteilungen üblichen Bereich übersteigt - im Gegenzug erkennt diese Distanz aber kaum anders geartete Abweichungen.

Auffällig sind die Cramér-von-Mises und die Totalvariations-Distanz, welche den P(2)-verteilten Daten die kleinsten Werte zuweisen. Bei der zweitgenannten Methode offenbaren sich zudem weitere Probleme: Während mit Überlegungen zu mathematischen Eigenschaften (s. o.) nur Werte zwischen der Null und der Eins möglich sein sollten, werden hier zu großen Teilen Werte oberhalb dieses Intervalls ausgegeben - die Gründe sind unklar, jedoch in der Diskretisierung bei kleinen Stichproben zu vermuten. Weitere Unterschiede der Maße (oder der zugehörigen Implementierungen) ergeben sich in der Lage der Werte: So sind diese auch für normalverteilte Daten nur teilweise nahe der Null, etwa für Hellinger-Distanzen ist dies gar nicht der Fall (auch wenn hier das Intervall $[0, 1]$ eingehalten wird). Beide Auffälligkeiten

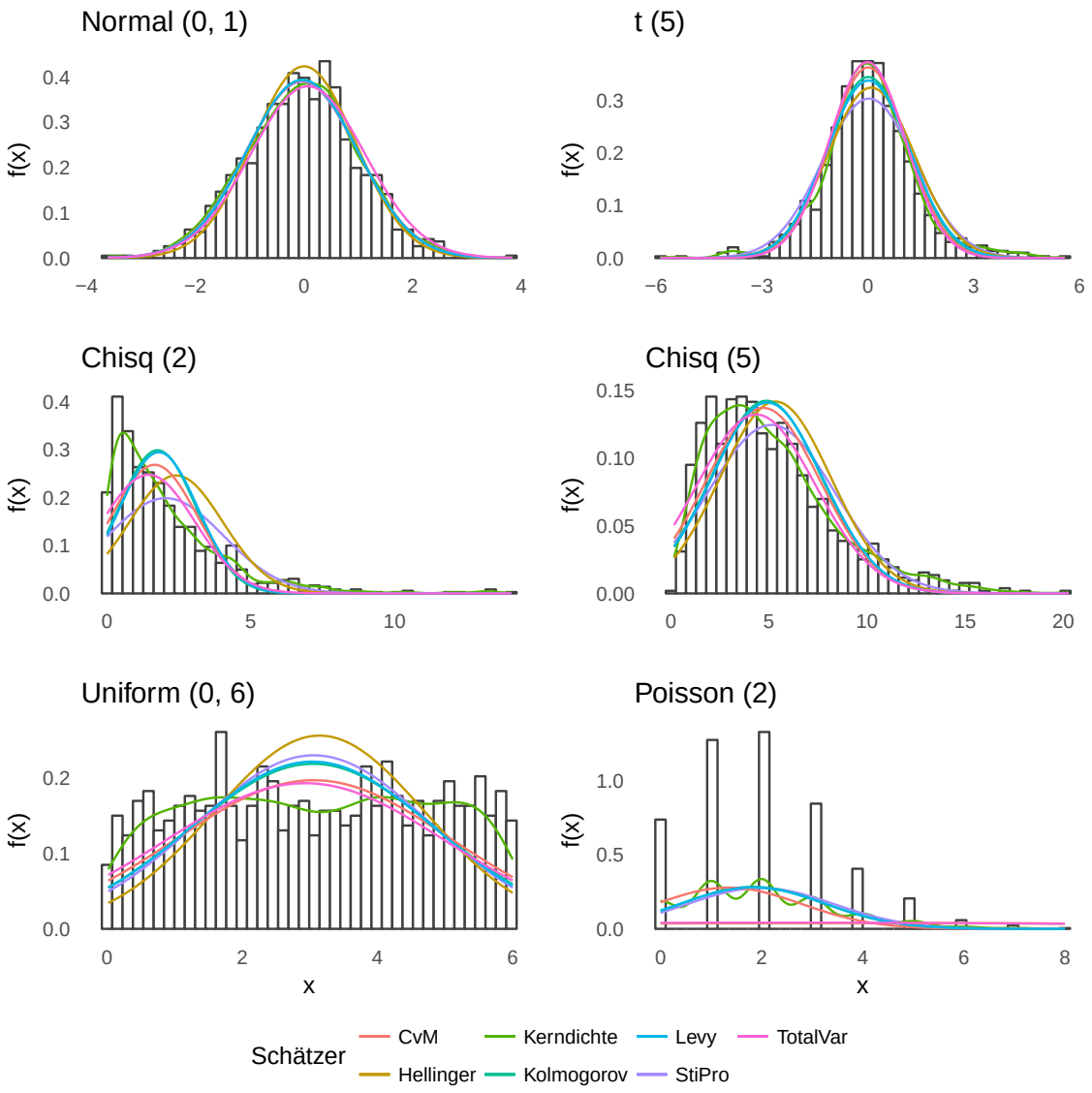

Abbildung 4: Sind die Distanzen zur Normalverteilung allgemein zu bestimmen, muss die Verteilungsfamilie für die meisten Varianten auf einen einzelnen Vertreter reduziert werden. Dabei können die Stichprobenschätzer (d. h. hier Mittelwert und Varianz) verwendet werden, oder das jeweilige Maß wird optimiert. Alle Varianten sind hier für die fünf Metriken dargestellt, ergänzt um eine Kerndichteschätzung. Gezeigt werden nur sechs der elf Verteilungen der Simulation, hier für eine Stichprobengröße von 1 000 Beobachtungen

.

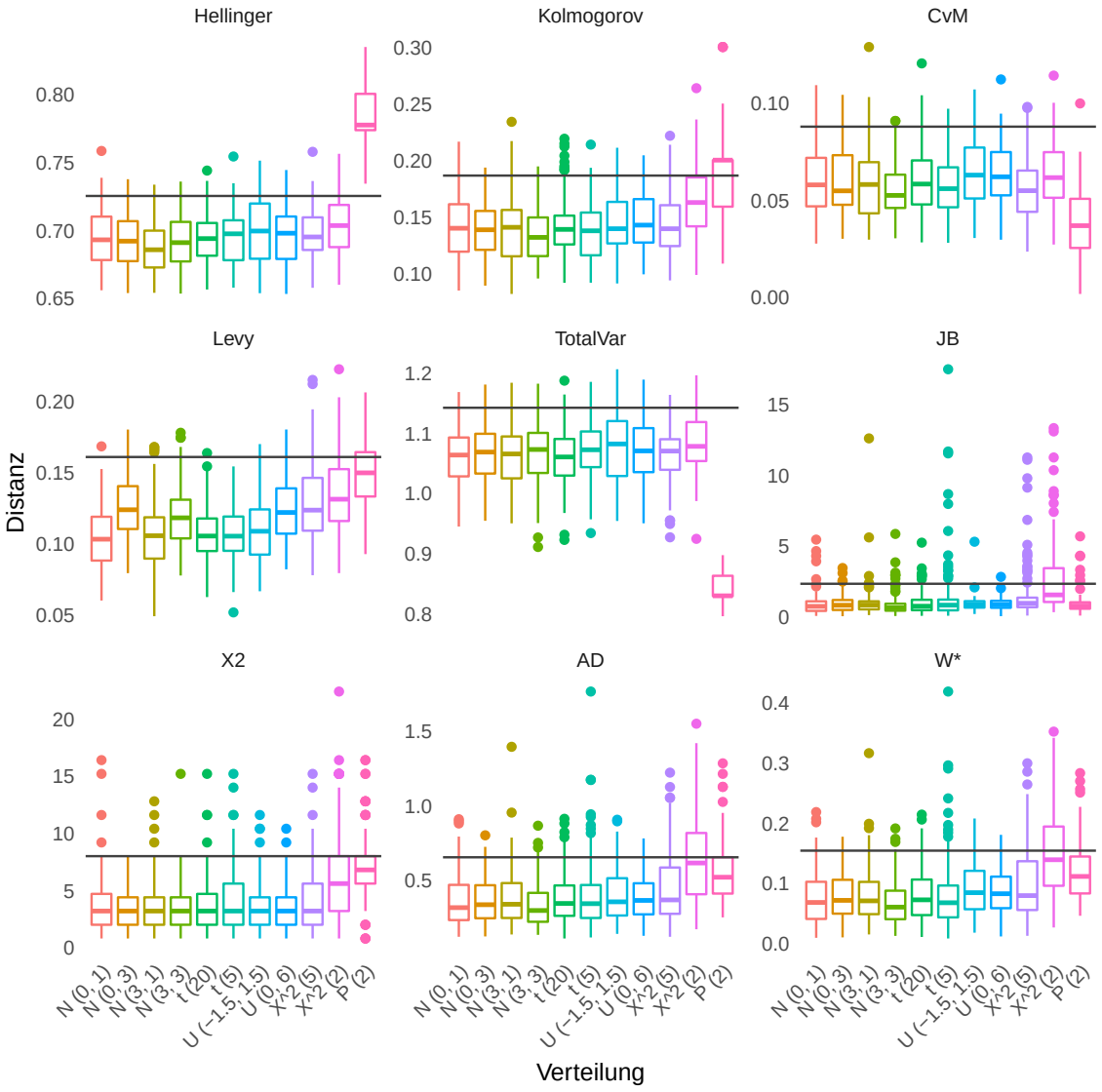

Abbildung 5: Die Simulationsergebnisse für die Stichprobengröße 10: Es sind mit allen Methoden kaum Unterschiede auszumachen.

könnten jedoch ignoriert werden, wenn sie konsequent auftreten - und sich so dennoch eine Unterscheidungsmöglichkeit zwischen den Verteilungen ergäbe.

Insgesamt ist für kleine Stichproben die offenbar grundsätzlich schwierige Trennung der Verteilungen zu notieren. Dabei fällt keines der Verfahren durchweg positiv auf. Einzig die Hellinger-Distanz trennt die Poisson-Verteilung gut von der Normalverteilung. Die Abweichungen der χ^2_2- Verteilung erkennen die Teststatistiken tendenziell besser, insbesondere die Variante nach Anderson und Darling. Die Ergebnisse des Totalvariationsabstandes sind zumindest für kleine Stichproben nicht vertrauenswürdig.

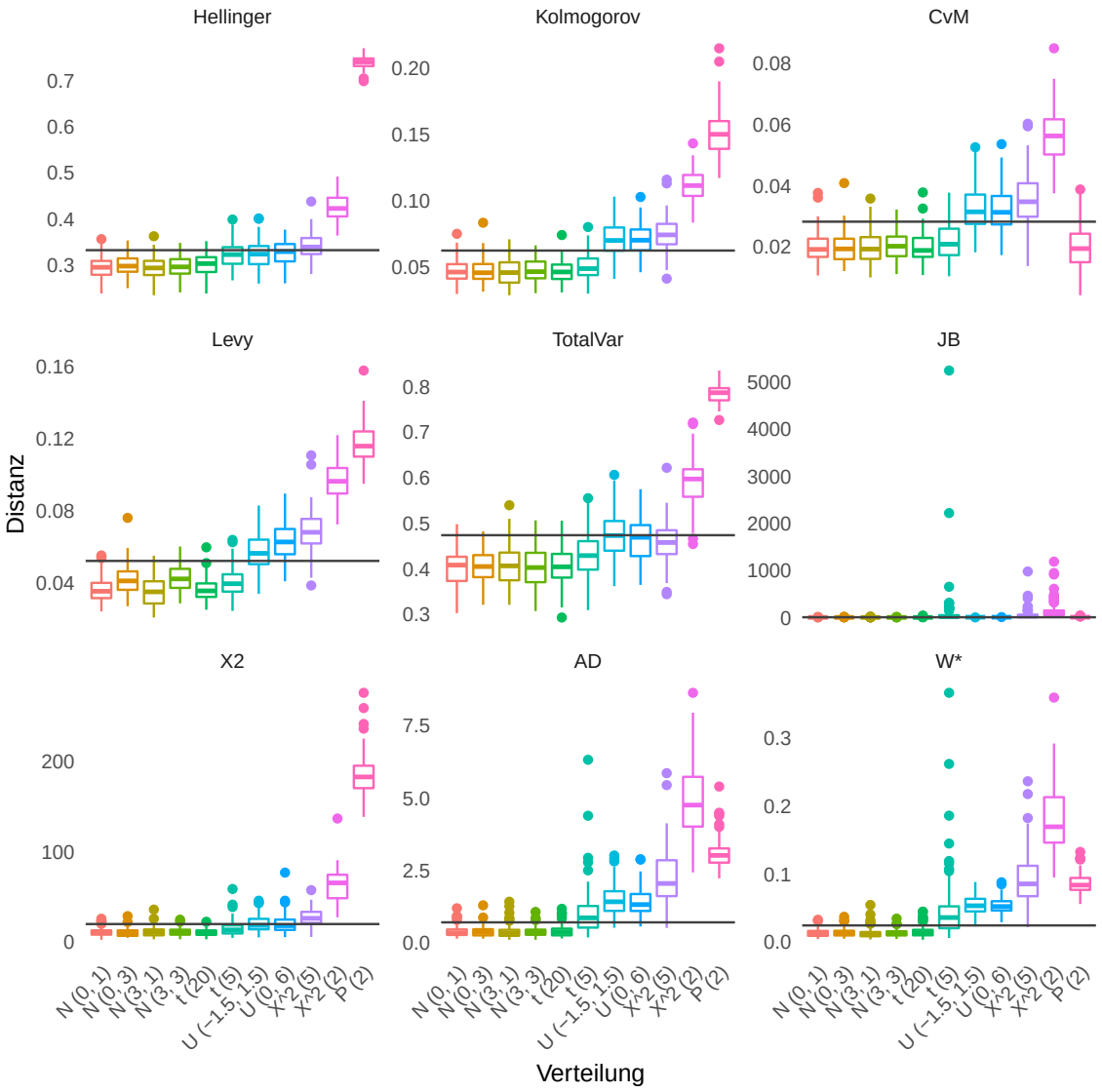

Abbildung 6: Die Simulationsergebnisse für die Stichprobengröße 100: Die Maße trennen die Verteilungen unterschiedlich gut.

Aus der statistischen Testtheorie ist bekannt, dass sich die Power (oder auch Güte, Trennschärfe) eines Verfahrens im Allgemeinen mit steigender Stichprobengröße verbessert. Dieses Prinzip findet sich auch im Vergleich der Abbildung 6 für die Ergebnisse der Simulationen zur Stichprobengröße 100 mit der vorigen Grafik. Je nach Methode können nur nicht nur die Daten aus Poisson- und χ^2_2-Verteilungen relativ sicher von der Normalverteilung getrennt werden: Auch für die anderen Verteilungen können klare Abwei-

chungen notiert werden, so trennen Kolmogorov, CvM und Lévy auch die zweite χ^2-Verteilung sowie die Gleichverteilungen recht gut von den normalen Verteilungen. Den Teststatistiken nach Anderson/Darling und Shapiro/Wilk gelingt das sogar sehr gut.

Insgesamt hat sich auch die Skala der jeweiligen Größen verändert: Dabei sind für die Metriken insbesondere die aus den Normalverteilungen gewonnenen „Baselines“ nach unten verschoben, zum Teil deutlich. Für die Teststatistiken zeigt sich ein anderer Effekt: Hier führen einzelne bis wenige Datenbeispiele nicht-normaler Verteilungen zu - auch im Vergleich mit dem Bild oben - extremen Abweichungen von der Null.

Für die Simulation mit der Stichprobengröße 1 000 ergeben sich keine qualitativen Unterschiede zu bisherigen Erkenntnissen. Die zugehörige Grafik ist der Vollständigkeit halber im Anhang auf Seite 42 als Abbildung 13 zu finden. Die Werte für die Metriken nähern sich hier bei tatsächlichem Vorliegen einer Normalverteilung weiter der Null an; und für feste Verteilungen und feste Maße sinkt die Streuung der Ergebnisse. Der wesentliche Unterschied zu den Resultaten der Stichprobengröße 100 besteht darin, dass für 1 000 Beobachtungen auch die Ergebnisse der t_5-Verteilung von der Normalverteilungsfamilie getrennt werden können. Im Rahmen der Metriken ist hier die Hellinger-Distanz positiv hervorzuheben.

Insbesondere für die Jarque-Bera-Methode (100 Beobachtungen), aber auch mit Blick auf alle Teststatistiken (1 000 Beobachtungen) ergeben sich mitunter recht schief verteilte Distanzmaße. Das trifft vor allem innerhalb einer Verteilung, zum Teil aber auch für alle Ergebnisse zu. Aus diesem Grund werden in der Abbildung 7 die Grafiken der Teststatistiken wiederholt, wobei jeweils eine Vergrößerung des Bereichs geringerer Distanzen enthalten ist. Das Äquivalent für die Metriken ist im Anhang auf Seite 43 als Abbildung 14 zu finden, bietet aber nur im Detail neue Erkenntnisse.

Wie mit den vergrößerten Bereichen sofort zu erkennen ist, steigt die Trennschärfe für die Entscheidung Normalverteilung ja/nein mit der Stichprobengröße klar an. Für die Stichproben der Größe 1 000 (rechte Seite der Grafik) werden die beiden t-Verteilungen zum Prüfstein: Die Werte der vier Normalverteilungen weichen nicht systematisch voneinander ab, alle anderen Verteilungen werden sicher von der Normalverteilung getrennt. Dabei kann eine klare Rangfolge der Verfahren abgeleitet werden.

Die χ^2-Statistik kann die Nicht-Normalität der t_{20}-Daten nicht erkennen, bei den t_5-Daten ist die Trefferquote ebenfalls recht gering. Wird weiter nur die t_{20}-Verteilung betrachtet, so folgen auf den Plätzen drei bis eins der AD-, der W^*-Abstand und der JB-Abstand. Die Jarque-Bera-Variante ist dabei für gut die Hälfte der 100 Wiederholungen größer als 95 Prozent der Abstände normalverteilter Daten. Dabei ist zu berücksichtigen, dass der Vergleich anhand der Hilfslinie lediglich punktuell ist, was für den Vergleich der gesamten Verteilung (also hier der Boxplots) nicht gilt.

Für die kleinere Stichprobengröße 100 fällt die JB-Methode jedoch durch eine negative Eigenschaft auf, so kann offenbar die diskrete Verteilung nicht sicher von der Normalverteilung getrennt werden. Für die am schwierigsten zu trennende Verteilung t_{20} liefern alle Methoden qualitativ ähnliche, schlechte Resultate. Insgesamt erscheint die χ^2-Statistik auch hier im Vergleich als wenig geeignet. Insbesondere bei den Gleichverteilungen, aber auch bei den t_5-Daten lassen sich leichte Vorteile der W^*-Statistik ausmachen, wobei die Unterschiede besonders zu Anderson-Darling gering sind.

Für größere Stichproben sind damit auf Seiten der Teststatistiken JB, AD und insbesondere W^* als gut zu bewerten. Dabei ist weiter anzumerken, dass nur die Shapiro-Wilk-Variante begrenzt ist, die Werte also alleinstehend am einfachsten interpretierbar sind. Auch ist hier die Skalenverschiebung durch Änderungen der Stichprobengrößen am geringsten.

Diese Liste der „besten“ Methoden ist auf Seiten der Metriken um Kolmogorov/Lévy und Hellinger zu erweitern. Dabei ist die Wahl der Methode abhängig vom inhaltlichen Abstand der Verteilungen: Die Hellinger-Distanz reagiert offenbar am sensibelsten bei den t-Verteilungen, für die Gleichverteilungen fallen die Ergebnisse der Verteilungsfunktion-basierten Methoden wiederum besser aus.

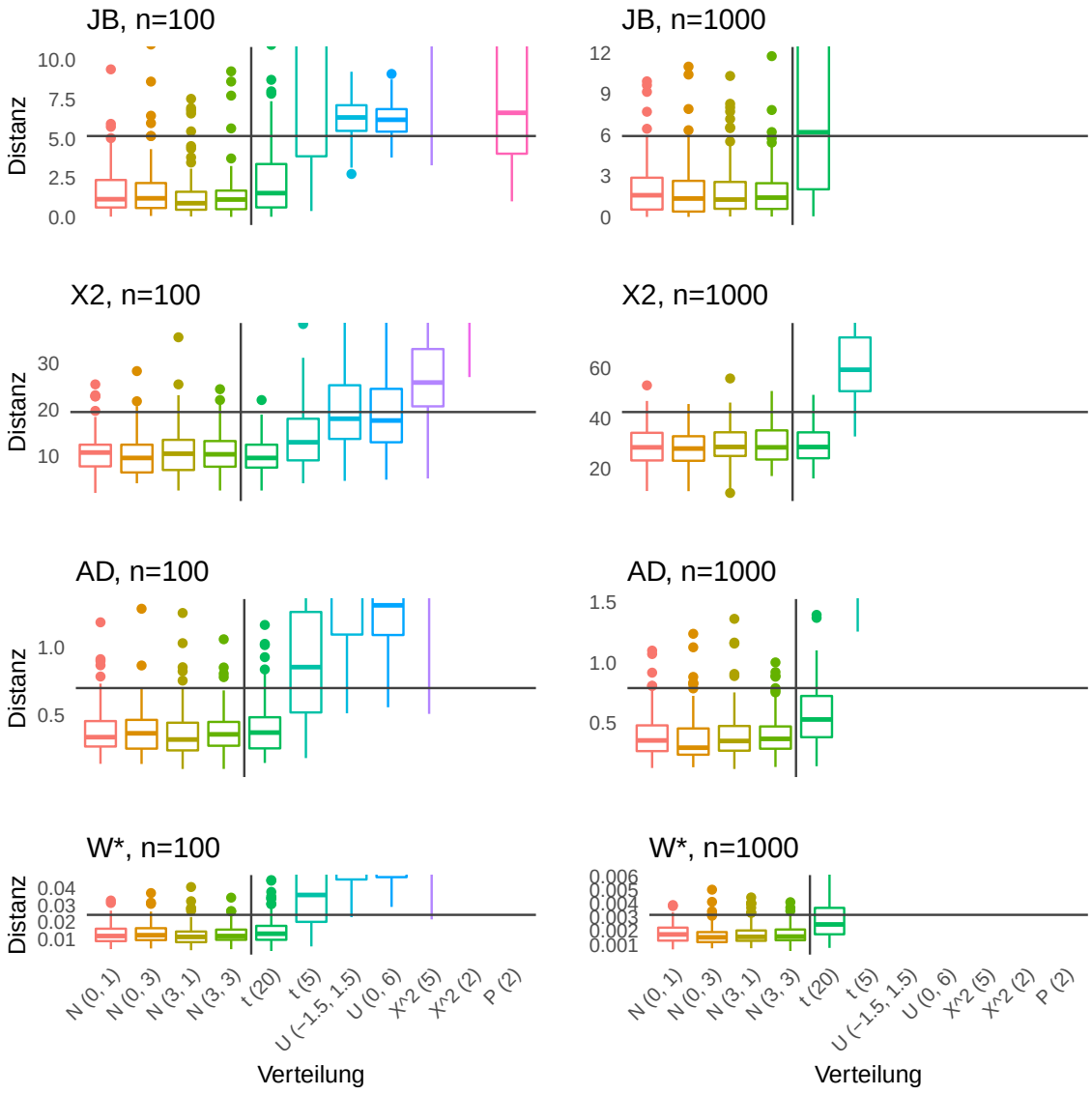

Abbildung 7: Auf den vorigen Abbildungen waren die Bereiche entlang der Entscheidungsgrenze „normal oder nicht normal“ mitunter recht klein skaliert. Hier wird eine Vergrößerung gegeben. Das hat zur Folge, dass einige Verteilung-Distanz-Kombinationen hier keine Datenpunkte mehr im Bildausschnitt aufweisen. Dann sind diese aber immer im numerisch größeren „nördlichen“, Teil der Skala zu verorten.

4.3 Vergleich mit theoretischen Größen

Als weitere Eigenschaft der Distanzmaße wird der Abstand der realisierten Werte vom jeweils theoretisch zu erwartenden betrachtet: Dies ist hier möglich, da die den Zufallszahlen zugrunde liegenden Verteilungen bekannt sind. Mit diesen kann nun jeweils die Normalverteilung mit gleichem Erwartungswert und gleicher Varianz der Verteilung gegenübergestellt und so die Distanz zu einer „passenden" Normalverteilung bestimmt werden. Das bedeutet beispielsweise: Die t-Verteilung mit n Freiheitsgraden hat einen Erwartungswert von 0 und eine Varianz von $n/(n-2)$. Für die t_5-Verteilung werden die Distanzen zur $N(0, 5/3)$-Verteilung ermittelt.

Während die Metriken hier per definitionem berechnet werden können, sind bei den Teststatistiken weitere Überlegungen nötig. Insbesondere bei der Anderson-Darling- und der Shapiro-Wilk-Methode sind die jeweiligen Werte nicht trivial zu berechnen (mit Ausnahme der erfüllten Nullhypothese, in diesem Fall ist die Verteilung bekannt oder es liegen zumindest tabelliert kritische Werte vor), der Aufwand würde den Nutzen hier übersteigen. Bei der Jarque-Bera-Variante fällt die Berechnung einfacher, da jeweils nur die bekannten Schiefen, Wölbungen und Stichprobengrößen zu verwenden sind. Bei fester Stichprobengröße und fester Regel zur Anzahl der Klassen (s. o.) sind auch die zu berücksichtigenden Einflüsse für die χ^2-Statistik bekannt.

Die Abbildung 8 stellt die zu erwartenden Werte den Realisierungen gegenüber. Dabei findet zum einen die Analyse der Distanzen selbst und implizit auch beispielsweise ihrer Robustheit gegenüber Ausreißern statt. Zum anderen werden hier aber auch und insbesondere gleichzeitig die zur Verfügung stehenden Diskretisierungen und Implementierungen beurteilt. In der Grafik sind jeweils die bereits bekannten Boxplots für einzelne Distanz-Verteilungs-Kombinationen aufgeführt. Die theoretisch zu erwartenden Werte sind jeweils durch ein rotes Kreuz markiert. Die Simulationsdaten werden hier auf die Stichprobengröße 1 000 reduziert. Dass die Distanzschätzer für diesen Wert „gute" Eigenschaften aufweisen, kann als Minimalforderung angesehen werden: Die Bezeichnung einer Stichprobe als „groß" ist sicher abhängig vom Bereich der Anwendung. In vielen Gebieten aber ist der Wert 1 000 schon ein aus Kosten- und Zeitgründen nicht zu erreichender Wert, etwaige Approximationen sollten auch für kleinere Werte schon ausreichen.

Bei Interpretation der Grafik ist berücksichtigen, dass die Simulationsergebnisse Optimierungen der Parameter beinhalten. Damit sind leichte Abweichungen unterhalb der Erwartungen kein Widerspruch zu passenden Werten. Zunächst fällt auf, dass in allen Fällen für die Normalverteilungen eine Null (bzw. für die Chi-Quadrat- und JB-Statistiken ein Wert relativ nahe bei Null) zu erwarten ist. Dieser wird von den Methoden unterschiedlich gut erreicht, was zudem - wie oben gezeigt - abhängig von der Stichprobengröße ist. Augenscheinlich erreichen die beiden Teststatistiken die Null hier am besten bezüglich des relativen Abstandes zum Minimalwert, was aber nicht zuletzt auch auf die relativ großen Skalierungen zurückzuführen ist. Auch insgesamt ergibt sich aber der Eindruck, dass die Realisierungen dieser Varianten am besten den theoretischen Größen gegenübergestellt werden können.

In den mittleren Bereichen (was hier für die Abbildungen jeweils gleichermaßen im horizontalen wie vertikalen Sinne zu verstehen ist) scheinen Hellinger, Lévy und, mit Abstrichen, Kolmogorov gute Resultate zu liefern. Bei der CvM-Methode könnte durch die Optimierungen der Unterschied zwischen einer bestimmten und allen Normalverteilungen hier in der Grafik enthalten sein. Die niedrigen Werte bei der Poisson-Verteilung weichen dennoch auch von inhaltlichen Erwartungen ab. Der Grad der Normalität der χ^2_2-Verteilung wird bezüglich der Metriken einzig von der Hellinger-Distanz passend beurteilt.

Auffällig ist hier wiederum der Totalvariationsabstand, dessen Abweichungen nach oben nicht durch die Optimierungen erklärt werden können: Hier ist der gesamte zu erwartende Verlauf nach oben verschoben, mit Ausnahme der Ergebnisse zur Poisson-Verteilung. Dabei ist theoretisch ein Abstand von 1 zu erwarten, die Dichten überdecken sich schließlich nur punktweise. Das wird aber auch vom Hellinger-Abstand nur im Groben erfasst, was bei der Verwendung von Diskretisierungen nicht ganz zu umgehen ist.

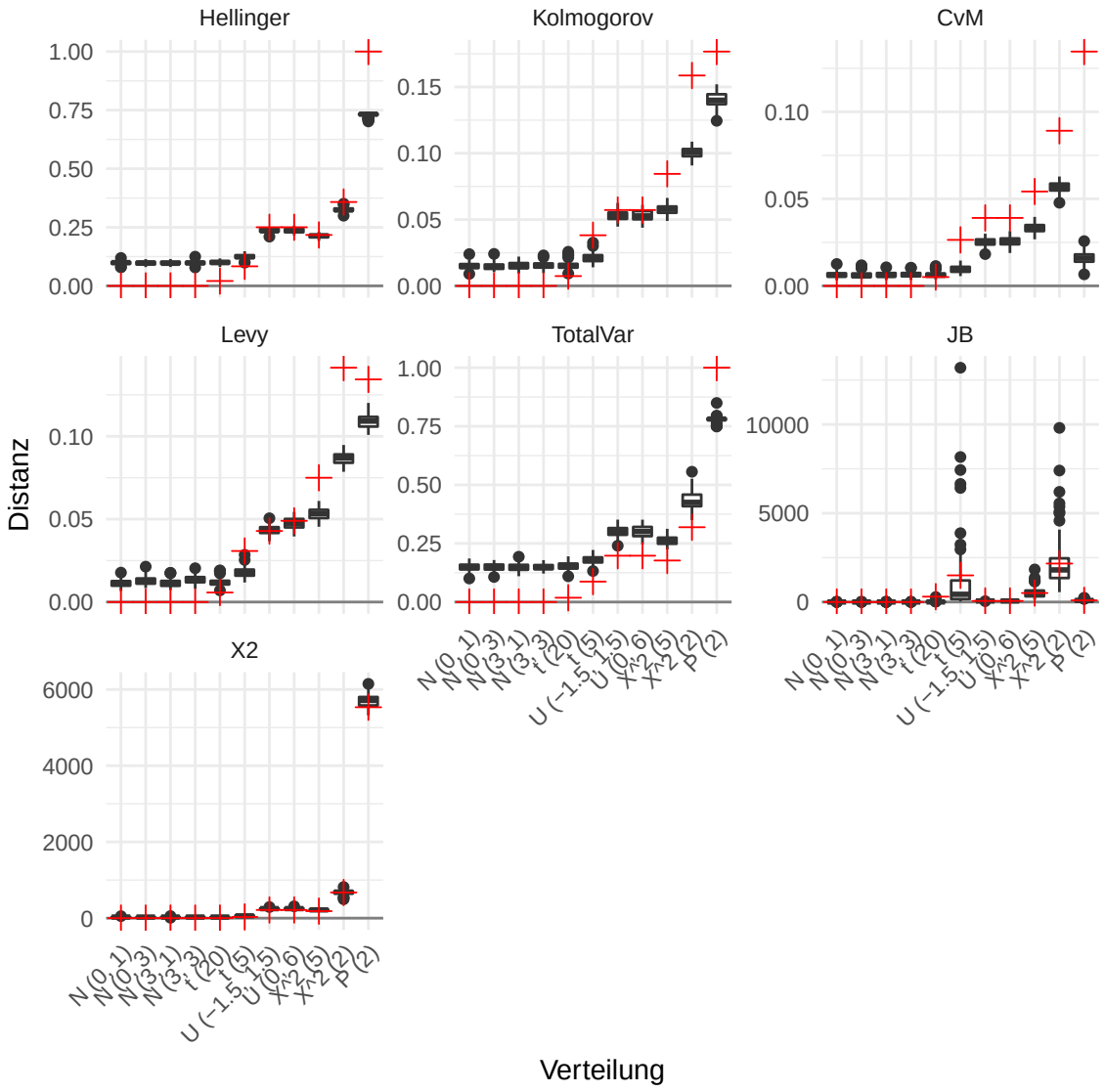

Abbildung 8: Die Ergebnisse der Simulation für 1 000 Beobachtungen, ergänzt um die theoretisch zu erwartenden Werten (rot), ohne AD und W^*.

Bezogen auf die zu erwartenden Werte weisen alle fünf Metriken eine plausible Rangordnung auf, auch wenn diese nicht immer exakt gleich ist. Für die JB-Methode gilt das nicht unbedingt, so werden die t-Verteilungen hier als weiter von der Normalverteilung entfernt als die Gleichverteilungen eingeordnet. Das ist zurückzuführen auf die Symmetrie beider Verteilungstypen und die kleinere Wölbung der Gleichverteilung, ein Effekt der oben bereits kommentiert wurde. Ebenso findet sich bei JB die weniger starke Wölbung der χ^2_5- gegenüber der t_5-Verteilung wieder (bei Schiefen von 1.23 und 0).

Die Eigenschaft der Methoden, ihre Werte den theoretisch zu erwartenden anzunähernden, kann auch aus einem anderen Blickwinkel betrachtet werden: Unabhängig davon, wie gut der theoretische Wert für eine bestimmte Stichprobengröße realer oder simulierter Daten angenähert wird, kann gefordert werden, dass die Näherung mit steigender Stichprobengröße besser wird. Dieses Prinzip ist als Konsistenz bekannt und in der Schätztheorie elementar. Zur Entscheidung, ob die Verfahren solch ein Verhalten zeigen, wird wiederum die Abbildung 8 herangezogen. Für die beiden kleineren Stichprobengrößen sind die Resultate als entsprechende Abbildungen im Anhang unter den Nummern 15 und 16 finden (auf den Seiten 44 und 45). Für alle der sieben diesbezüglich untersuchten Verfahren ergibt sich der Eindruck konsistenten Verhaltens. So bringt die Vergrößerung der Stichproben jeweils eine Verschiebung der Boxplots mit sich: Die Ergebnisse werden differenzierter zwischen den Verteilungen, die Skalierung ändert sich indes. Zugleich ergeben sich dabei „Bewegungen“, wie sie für Konsistenzen zu erwarten sind. Die Distanz nach Cramér und von Mises entfernt sich für die Poisson-Verteilung zwar zunehmend vom Soll, das aber konsequent. Das wird hier als in sich stimmige Eigenschaft aufgefasst, während die große Abweichung von Ist- und Soll-Wert selbst an anderer Stelle berücksichtigt wird.

4.4 Folgerungen

Die dargestellten Erkenntnisse sollen nun zusammenfassend bewertet werden. Wie bereits in Teil 3.3 erfolgt zunächst eine Diskussion der Eigenschaften der neun Verfahren, anschließend wird eine tabellarische Zusammenfassung gegeben (siehe die Tabelle 3).

Die Ergebnisse der Simulationen zeigten für alle Varianten eine Abhängigkeit von der Stichprobengröße, was sowohl die numerische Lage der Ergebnisse als auch die Trennschärfe der Methoden betrifft. Für sehr kleine Stichproben ist keines der Verfahren geeignet. Da für kleine Stichproben Abweichungen von einer Verteilung offenkundig auch nur sehr schwer zu detektieren sind stellt diese Tatsache noch kein grundsätzliches Hindernis zur Verwendung der Methoden dar, zu notieren ist es jedoch. Gegenteilig fällt das Urteil für den Vergleich auf Basis der 1 000-er Stichprobe auf: Hier können alle Verfahren die nicht-normalen Verteilungen angeben, mit Abweichungen für die t_{20}-Daten. Dabei sind die Unterschiede jedoch nur gradueller Natur.

Am besten zur Unterscheidung der Maße eigenen sich die Daten mit der Stichprobengröße 100. Die Forderung nach einer guten Trennbarkeit für diese Stichprobengröße kann ohne Weiteres formuliert werden, um ein Maß möglichst allgemein als „gut“ bezeichnen zu können. Das Kriterium der Trennbarkeit von der Normalverteilungsfamilie ist dabei nur vor dem Hintergrund der jeweiligen Verteilungen zu beantworten, wie sich an den Grafiken in Teil 4.2 zeigte. Die vier Gruppen Poisson, χ^2_k, $U(a, b)$ und t_k werden daher jeweils für sich betrachtet. Die Einordnung in eine gute, schlechte oder mittelmäßige erzielte Trennbarkeit erfolgt auf Basis der Abbildungen und kann der folgenden Tabelle entnommen werden.

Im vorigen Teil wurden Gleichheit und Unterschiedlichkeit der realisierten und der erwarteten Werte untersucht. Dabei stellte sich heraus, dass auch bei 1 000 Beobachtungen nicht alle Maße zum theoretischen Wert passen. Weiter wurde neben der Konvergenz auch die Konsistenz kommentiert, auch diese beiden Eigenschaften sind zusammenfassend in der Tabelle notiert.

Eigenschaft	Kolm.	TV	Hell.	Lévy	CvM	χ^2	JB	AD	W
Eignung StiPro 10	○	○	○	○	○	○	○	○	○
Eignung StiPro 100									
Poisson-Vtlg.	●	●	●	●	○	●	◐	●	●
χ^2_k-Vtlg.	●	◐	◐	●	●	●	●	●	●
$U(a,b)$-Vtlg.	●	◐	◐	●	●	◐	●	●	●
t_k-Vtlg.	○	○	○	○	○	○	◐	◐	◐
Eignung StiPro 1 000	●	●	●	●	●	●	●	●	●
Konvergenz	◐	○	●	●	○	●	●	x	x
Konsistenz	●	●	●	●	●	●	●	x	x
Implementierung	●	◐	●	●	●	●	●	●	●

Tabelle 3: Die Folgerungen aus den Simulationsergebnissen in Zusammenfassung. Die geforderten Eigenschaften sind erfüllt oder nicht, was durch die Symbole ● und ○ gekennzeichnet wird. Die teilweise Erfüllung, markiert mit ◐, wird jeweils im Text erläutert; Eigenschaften mit der Kennzeichnung x sind nicht untersucht worden.

Die Eigenschaften der verwendeten Implementierungen enthalten wie bereits erwähnt jeweils zwei bis drei Quellen: Zum einen die Eingenschaften der Methoden selbst, zum anderen die Eigenschaften der gewählten Form der Implementierung. Für einen Teil der Methoden kommen die Eigenschaften der notwendigen Diskretisierung hinzu. Auffälligkeiten ergaben sich beim Totalvariationsabstand, für welchen bei kleinen Stichproben entgegen der Definition Werte oberhalb der Eins ausgegeben werden. Auf welcher Ebene der angeführten Quellen der Eigenschaften sich dies begründet, wird an dieser Stelle nicht weiter untersucht. Für die anderen Methoden waren im Rahmen der hier vorgestellten Arbeiten keine Auffälligkeiten zu notieren, es wird daher - jedoch ohne weitere Prüfung und nach dem Prinzip in dubio pro reo - von geeigneten Implementierungen ausgegangen.

5 Fazit

In den beiden vorigen Kapiteln wurden die neun ausgewählten Verfahren für einige beispielhafte Verteilungen mittels verschiedener Kriterien verglichen. Dabei wurde deutlich, dass keines der Abstandsmaße durchweg die beste Wahl zur Behandlung einer spezifischen Fragestellung ist - eine Art *gleichmäßig* bestes Verfahren kann nicht herausgestellt werden. Sowohl die Diskussion der theoretischen Eigenschaften also auch die Ergebnisse aus der Simulationsstudie zeigen, dass die Wahl eines optimalen Verfahrens auch von der Anwendung abhängen kann, etwa wenn verschiedene Formen der Nicht-Normalität verschiedene Auswirkungen haben. Dennoch kann zusammenfassend diskutiert werden, welche Methode zumindest die *durchschnittlich* beste darstellt, welche eventuell gänzlich ungeeignet ist oder welche Auswahl von Maßen die Menge von neun Methoden verlustfrei auf eine praktikablere Größe reduzieren kann.

Eine Zusammenstellung der Ergebnisse bildet die Tabelle 4. Diese stellt als Kombination der beiden vorigen Tabellen 2 und 3 die bisherigen Erkenntnisse zusammen. Es wird dabei auf die Charakteristiken verzichtet, welche keine Unterschiede zwischen den Verfahren aufzeigen. Die letzten beiden Zeilen enthalten eine Gesamtbewertung der neun Methoden, welche durch simples Auszählen der vorher vergebenen „Punkte“ gebildet wurde. Die ideale Gewichtung der einzelnen Aspekte ist dabei jedoch nicht eindeutig und je nach Anwendungsfall zu berücksichtigen, hier wurde eine gleiche Gewichtung aller Merkmale angesetzt.

Offenbar ist kein Verfahren durchweg besser als alle anderen: Zwar weisen die beiden Metriken nach Kolmogorov und Lévy, also zwei der auf Verteilungsfunktionen beruhenden Methoden, die meisten erfüllten

Eigenschaft	Kolm.	TV	Hell.	Lévy	CvM	χ^2	JB	AD	W
Theorie und Daten	●	◐	◐	◐	●	◐	◐	●	○
Symmetrie	●	●	●	●	●	○	●	○	○
Beschränktheit	●	●	●	●	●	○	○	○	●
Eignung StiPro 100									
Poisson-Vtlg.	●	●	●	●	○	●	◐	●	●
χ^2_k-Vtlg.	●	◐	◐	●	●	●	●	●	●
$U(a,b)$-Vtlg.	●	◐	◐	●	●	◐	●	●	●
t_k-Vtlg.	○	○	○	○	○	○	◐	◐	◐
Konvergenz	◐	○	●	●	○	●	●	×	×
Implementierung	●	◐	●	●	●	●	●	●	●
Insgesamt ●	7	3	5	7	6	4	5	5	5
Insgesamt ○	1	2	1	1	3	3	1	2	2

Tabelle 4: Die Kombination der nicht-konstanten Zeilen aus den beiden Tabellen 2 und 3 führt zu einer stark verdichteten, aber übersichtlichen Zusammenfassung der erkannten Unterschiede. Die unteren Zeilen enthalten die Häufigkeiten der Eigenschaften, was eine gleiche Gewichtung aller Kriterien impliziert.

Eigenschaften auf. Die Bereiche, in denen diese beiden Verfahren jedoch Schwächen aufweisen, werden von anderen Methoden zumindest teilweise erfüllt.

Dominanz kann hier zweckmäßig folgendermaßen definiert werden: Beim Vergleich zweier Verfahren wurde das dominierende Verfahren in allen Bereichen mindestens genau so gut bewertet das dominierte Verfahren, sowie in mindestens einem Bereich besser. Dann dominiert dem vorigen Absatz gemäß keine Methode alle anderen. Jedoch werden einige Methoden dominiert und sind damit obsolet. Das betrifft den Totalvariationsabstand (dominiert durch Kolmogorov, Hellinger und Lévy), den Hellinger-Abstand (durch Lévy), den Cramér-von-Mises-Abstand (durch Kolmogorov) und die χ^2-Statistik (durch Lévy und Anderson-Darling).

Für die übrigen fünf Varianten stellt sich die Frage nach der Gewichtung der spezifischen Vor- und Nachteile: Ist etwa die Sensibilität gegenüber der t-Verteilung von besonderem Interesse, sollte eine der Teststatistiken gewählt werden. Können diskrete Verteilungen von vornherein ausgeschlossen werden, verlieren die CvM- und JB-Kennzahlen einen Nachteil, und so weiter. Insgesamt erscheinen aber Kolmogorov und Lévy als die besten Varianten, gefolgt von JB. Mit dem Kolmogorov- und dem Jarque-Bera-Verfahren weisen zwei dieser Methoden zudem die positiven Eigenschaften einer anschaulichen Grundidee und einer vergleichsweise einfachen Berechnungsweise auf.

Statt nur einer Größe kann auch die Verwendung von zwei Methoden in Betracht gezogen werden, was ebenfalls als praktikabel erscheint. Werden die beiden letztgenannten Methoden Kolmogorov und Jarque-Bera gemeinsam betrachtet, so ist jede der tabellierten Forderungen maximal erfüllt („maximal“, da kein Verfahren für die t-Verteilung mehr als die Einordnung „teilweise erfüllt“ erreicht). Von allen 36 Möglichkeiten, zwei der Verfahren zu kombinieren, wird nur für drei Paare dieses maximal mögliche Ergebnis erzielt: Hellinger und Anderson-Darling, Lévy und Anderson-Darling sowie, bereits genannt, Kolmogorov und Jarque-Bera (und entsprechend dominieren dieses drei Paarungen alle anderen). Da die Kombination der Kolmogorov-Distanz und des JB-Tests auch für die einzelnen Verfahren die besten Werte dieser drei Möglichkeiten aufweist, kann sie als die eindeutig beste bezeichnet werden.

6 Ergänzungen

Für die Vereinfachung der praktischen Anwendung dieser beiden ausgewählten Maße sollen an dieser Stelle noch zwei Punkte untersucht werden: Zum einen wurden oben Abhängigkeiten von der Stichprobengröße notiert, welche eine stand-alone-Interpretation der Werte verhindern. Zum anderen enthalten die vorgestellten Ergebnisse Parameteroptimierungen. Diese setzen eine vielmalige Auswertung des Distanzmaßes voraus und erhöhen den Rechenaufwand damit im Vergleich zur einmaligen Berechnung der Abstände für die $N(\hat{\mu}, \hat{\sigma}^2)$-Verteilung enorm. Es ist zu klären, ob dieser Mehraufwand lohnenswert ist.

6.1 Zur Stichprobengröße

Die Interpretationen der Werte ist wie gezeigt nur bei Kenntnis des Stichprobenumfangs ohne Weiteres möglich, da auch die Werte für normalverteilte Daten von diesem abhängen. Für die Kolmogorov-Distanz und die Jarque-Bera-Statistik wird die Abhängigkeit an dieser Stelle weiter untersucht. Die Abbildung 14 wiederholt die Ergebnisse der Simulationen bei Gegenüberstellung der Stichprobengrößen. Da keines der Maße - den Forderungen und Erwartungen entsprechend - Unterschiede zwischen den vier gewählten Normalverteilungen angezeigt hat, werden diese im Weiteren der Übersichtlichkeit halber nur durch die Standardnormalverteilung repräsentiert.

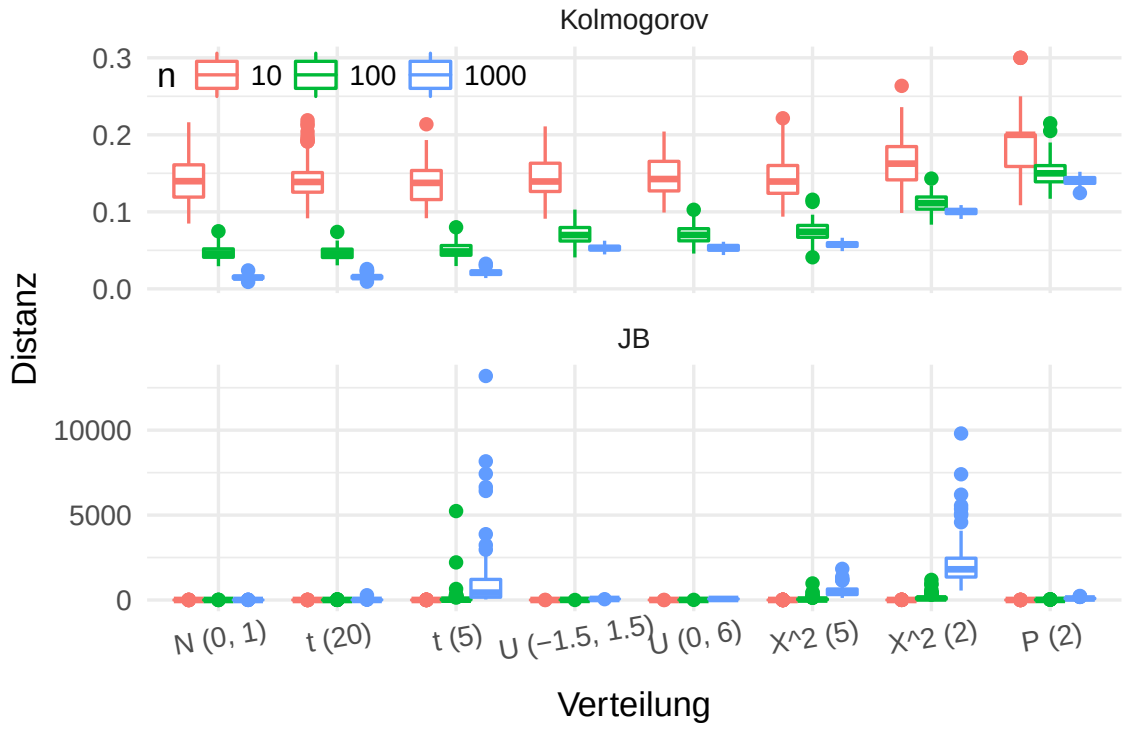

Abbildung 9: Die Kolmogorov-Distanzen und JB-Statistiken aus den Simulationen, für verschiedene Verteilungen und verschiedene Stichprobengrößen n.

Auf der Basis dieser Abbildung und auch der vorherigen Erkenntnisse wird die Teststatistik von Jarque und Bera logarithmiert betrachtet: Die mitunter extrem schiefen Verteilungen der Größe führen zu Skalierungen, welche neben einzelnen Ausreißern kaum Systematiken erkennen lassen (vgl. dazu bspw. auch Abbildung 6 auf Seite 26). Da der Logarithmus eine streng monotone Funktion ist, führt dies bezüglich

der Ordnung mehrerer Distanzen nicht zu Verzerrungen, die Skalierungen werden jedoch praktikabler. Da sich für perfekt normalverteilte Daten ein Wert von Null ergibt, der Logarithmus aber dort nicht definiert ist, wird die JB-Statistik zunächst verschoben:

$$LJB = \log(JB + 1).$$

Diese Verschiebung führt auch dazu, dass die Null wieder den minimalen Abstand bildet, da $\log(0+1) = 0$ ist. Alternativ wurde auch die Verwendung der radizierten JB-Statistik in Betracht gezogen, welche aber nicht zielführend ist.

Die Abbildung 10 zeigt bei (hier nicht dargestellter Vergrößerung der unteren LJB-Skala) ein gegenläufiges Verhalten der beiden Größen an: Während die optimierte Kolmogorov-Distanz gerade bei einer Normal- und ähnlichen Verteilungen große Unterschiede zwischen den Stichprobengrößen aufweist, sind die LJB-Werte hier am nahesten beieinander. Das folgt aus der Eigenschaft als Teststatistik, welche unter der Nullhypothese approximativ einer Chi-Quadrat-Verteilung mit zwei Freiheitsgraden folgt. Unterschiede zwischen den drei Boxplots zur Normalverteilung können dieser Approximation sowie Zufallstreuungen zugeschlagen werden.

Bei der Kolmogogrov-Distanz führt die Multiplikation der Ergebnisse mit $\sqrt{n}$ zu gleichen Werten für die normalverteilten Stichproben. Dieser Faktor wird auch beim Kolmogorov-Smirnov-Test genutzt, wie Hartung (2005) auf Seite 184 entnommen werden kann. Der Nachteil der Betrachtung von

$$K^* = \sqrt{n}\ K = \sqrt{n}\ \sup_x |F_n(x) - G_n(x)|$$

besteht in der Verlagerung der Stichprobenabhängigkeit auf die Schranken: Während $K \in [0, 1]$ ist, ist $K^* \in [0, \sqrt{n}]$. Das wird aber als leichter zu interpretierende Variante aufgefasst. Zur Orientierung: Es ist $\sqrt{10} \approx 3.16$, $\sqrt{100} = 10$ und $\sqrt{1000} = 10\sqrt{10} \approx 31.6$. Die Abstände sind also durchweg recht weit entfernt von dieser Schranke, was entsprechend auch bei den Auswertungen der Simulationsstudie im Kapitel 4 bereits zu beobachten war. Die beiden transformierten Größen sind in der Abbildung 10 dargestellt.

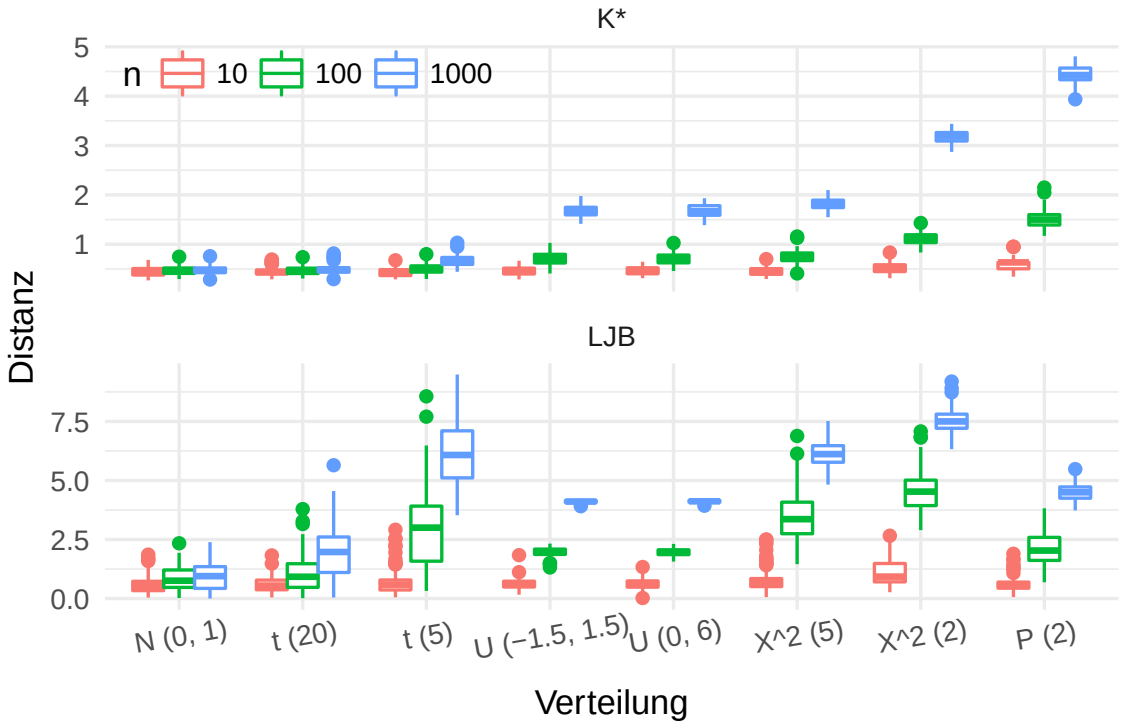

Abbildung 10: Mit den beiden angegebenen Transformationen sind zum einen die Skalierungen praktikabler angelegt, zum anderen bestehen keine systematischen Unterschiede für die Werte unterschiedlich großer, normaler Stichproben.

6.2 Zur Optimierung

Bei den bisher betrachteten simulierten Kolmogorov-Distanzen wurden die Parameter der Normalverteilung als Variablen aufgefasst. Damit wurden die Metrik dann optimiert. Von Interesse ist nun jedoch die Notwendigkeit dieser Optimierung, beziehungsweise gegebenenfalls die Größenordnung des Vorteils gegenüber der Verwendung geschlossen darstellbarer Parameterschätzer. Für die Jarque-Bera-Variante stellt sich diese Frage nicht, da hier gerade ein von den ersten beiden Momenten unabhängiges Vorgehen gewählt wird.

In einer kleinen Simulationsstudie wurden aus verschiedenen Verteilungen Zufallsstichproben vom Umfang 100 gezogen und die K^*-Distanz jeweils sowohl für die Startwerte der Optimierung (das heißt hier Mittelwert und Stichprobenvarianz), als auch für die optimierte Variante festgehalten. Dieses Vorgehen wird 50 mal wiederholt. Die folgenden beiden Abbildungen 11 und 12 verdeutlichen die Unterschiede, welche sich durch die Optimierung im Vergleich zu den Startwerten („direkte Bestimmung“) ergeben. In der ersten Grafik ist dabei der Unterschied für jede der Stichproben dargestellt, wobei die Distanzen für die Start- und die optimierten Werte jeweils durch eine Gerade verbunden sind (jede Teilabbildung enthält also $3 \cdot 50$ Geraden). Verwendet wurden hier die oben bereits genutzten Verteilungen, wobei wiederum nur Ergebnisse für eine statt für vier Normalverteilungen gezeigt werden.

Auch unter Berücksichtigung der unterschiedlichen Intervallbreiten ($[0, \sqrt{n}]$, s. o.) fallen Verbesserungen der Abstände ins Auge, wobei diese recht unterschiedlich stark ausfallen: Das Optimierungspotential steigt mit der Distanz, welche sich für die Startparameter berechnet. Damit können vor allem bei den χ^2- und Poisson-Verteilungen recht starke Verbesserungen erreicht werden, teilweise wird der Distanzwert durch die Optimierungen um ein Drittel und mehr gesenkt. Ebenso steigt der K^*-Wert mit der

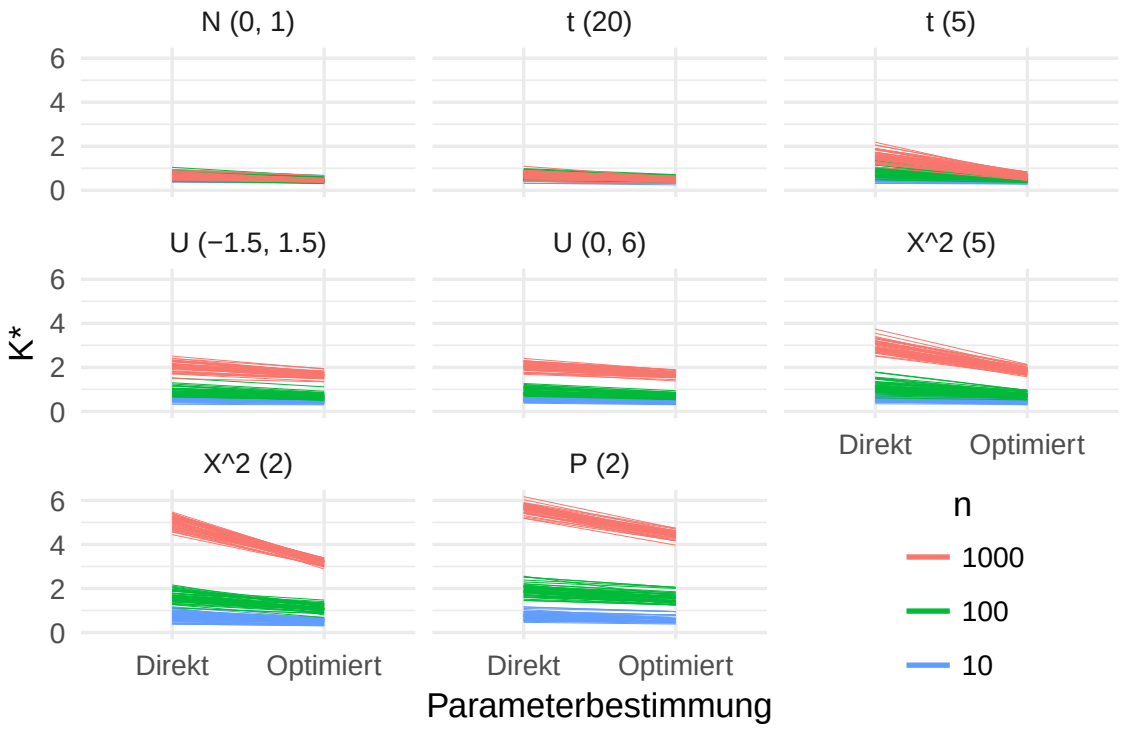

Abbildung 11: Die K^*-Distanzen zur Normalverteilung für jeweils 50 Wiederholungen verschiedener Stichprobengrößen. Dargestellt sind die Werte, welche sich für die $N(\hat{\mu}, \hat{\sigma}^2)$-Verteilung ergeben (geschlossene oder „direkte“ Parameterbestimmung), sowie die Distanzen für die nach Optimierung naheste Normalverteilung. Jede Gerade verbindet diese beiden Werte für die jeweils gleiche Zufallsstichprobe.

Stichprobengröße, sofern keine Normalverteilung vorliegt - und mit der Entfernung vom Minimalwert 0 dann ebenfalls das Optimierungspotential. Dieser Effekt verwundert nicht: Zum einen sind die kleineren Distanzen bereits näher am minimal möglichen Wert von 0. Zum anderen weisen Mittelwert und Stichprobenvarianz gute Eigenschaften wie etwa Erwartungstreue als Schätzer für Erwartungswert und Varianz auf, sofern tatsächlich normalverteilte Daten vorliegen - gerade deswegen sind sie gebräuchliche Schätzer. Steigt der Abstand von gegebenen Daten zur Normalverteilung, fällt auch die Argumentation für deren Verwendung entsprechend schwerer: Es sollen dann optimale Parameter für die Anpassung einer Normalverteilung gefunden werden, obwohl keine Normalverteilung vorliegt. Begriffe wie Erwartungstreue oder Konsistenz sind dann nicht sinnvoll definiert.

Klar ist also, dass die Optimierungen in viele Fällen zu numerisch anderen Resultaten führen; der beste Vertreter der Normalverteilungsfamilie also nicht immer durch Parameter nahe bei den hier verwendeten geschlossenen Schätzern gekennzeichnet ist. Offen ist jedoch, ob die Unterschiede auch von Relevanz sind: Sollte die Optimierung der Distanzen eine streng monotone Abbildung darstellen, wären die Ergebnisse mit und ohne Optimierung quantitativ zwar unterschiedlich, qualitativ könnten aber keine Unterschiede ausgemacht werden. Die Abbildung 12 zeigt aus den Simulationen resultierende Boxplots, wobei wie oben die Linien für das jeweilige 95 %-Prozent-Quantil der Normalverteilungs-Distanzen eingezeichnet sind (auch wenn diese Orientierungsgröße bei nur 50 Beobachtungen jedoch einer recht großen Varianz unterliegt).

Wie bereits vorher klar wurde, zeigt auch diese Darstellung ein teilweise deutliches Absinken der Distanzmaße durch die Optimierungen. Das beutetet, dass die Optimierung einen sichtbaren Effekt erzielt hat, und nicht nur die Startwerte oder Werte sehr nahe bei diesen wieder ausgegeben werden. Andererseits wird hier aber auch klar, dass die Ordnungen aus der direkten Parameterschätzung erhalten bleiben: Zwar auf anderem Niveau, ist auch nach den Optimierungen das gleiche „Muster“ wir vorher zu erkennen. Die Boxplots beispielsweise zu den t_5-Daten heben sich vor den Optimierungen sogar besser von denen der Normalverteilung ab (für $n \geq 100$), eine Beobachtung, welche an dieser Stelle jedoch nicht von zufälligen Einflüssen getrennt werden kann und daher nicht interpretiert werden sollte.

Insgesamt kann mit diesen Ergebnissen festgehalten werden, dass die Optimierung für theoretische Betrachtungen durchaus von Interesse sein kann: Die beste Annäherung einer Stichprobe durch eine Normalverteilung wird nicht ohne Weiteres durch die genannten Schätzer erreicht. In der Praxis sind diese Unterschiede jedoch nicht relevant, so dass auf die Optimierungen verzichtet werden kann. Damit kann Rechenzeit sowie eine mögliche Fehlerquelle eingespart werden.

In diesem Zusammenhang interessant ist eine Erkenntnis von A. Kagan, welche Patel und Read (1996) auf der Seite 99 folgendermaßen zusammenfassen:

> „It turns out that, among all continuous distributions depending on location parameter, and having certain regularity conditiones, the location parameter ist estimated worst of all in the case of normal samples.“

Ob dieser Sachverhalt einen Einfluss auf den Wert der Maße hat, und wie groß dieser gegebenenfalls ausfällt, kann in weiteren Untersuchgen betrachtet werden. Alternative, geschlossene Parameterschätzer können also in Betrachtet gezogen werden, um die Berechnung der Maße zu verbessern, ohne den Rechenaufwand deutlich erhöhen zu müssen.

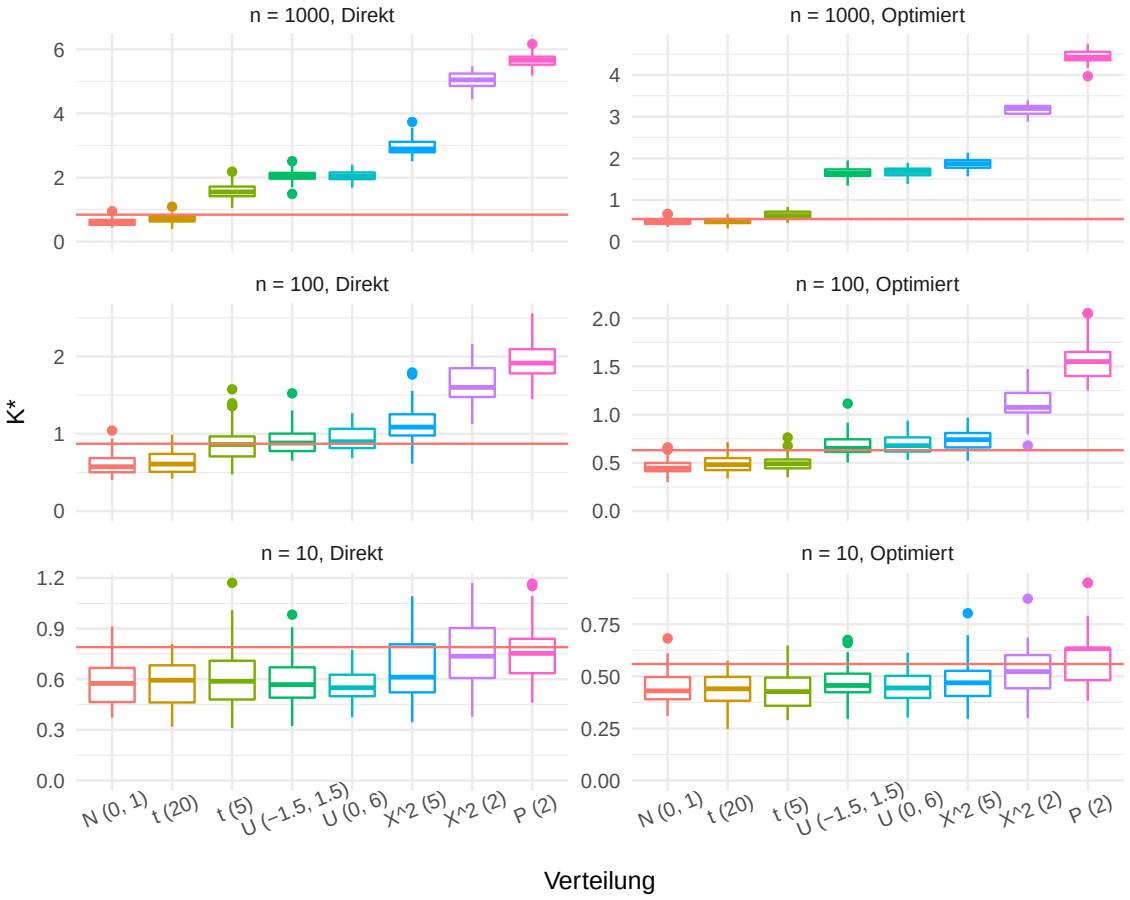

Abbildung 12: Die bereits aus vorangegangenen Abbildung bekannten K^*-Distanzen in anderer Darstellung: Auf der linken Seite die Distanzen zur Normalverteilungsfamilie in Repräsentation durch die Parameterschätzer Mittelwert und Stichprobenvarianz, auf der rechten Situation in Repräsentation durch optimierte Vertreter. Die horizontale Linie markiert jeweils das 95 %-Quantil der Distanzen der N(0,1). Zu beachten sind die unterschiedlichen Skalierungen der Ordinatenachsen bei sonst (zeilenweise) sehr ähnlichen Mustern.

7 Zusammenfassung

Die in den vorangegangenen Kapiteln beschriebenen Untersuchungen hatten das Ziel, eine möglichst gute Methode zur Messung der Normalverteilungsnähe einer Stichprobe zu identifizieren. Der Bedarf einer solchen Methode wurde im ersten Kapitel skizziert. Anschließend wurde die Problemstellung genauer ausgeführt, nämlich die Anforderungen an eine Methode herausgearbeitet. Eigenschaften wie Verwendbarkeit für die Verteilung selbst in ihrer Form als Dichte- oder Verteilungsfunktion, als auch für eine gegebene Stichprobe stellten sich dabei unter anderem als wünschenswerte Eigenschaften heraus.

Die neun hier ausgewählten Kandidaten bestehen aus fünf Metriken und vier Teststatistiken, wobei zwei der Metriken auf der Dichte- und drei auf der Verteilungsfunktion beruhen. Ein nicht zu vernachlässigender Punkt bei der Auswahl und Anwendung von Metriken stellt die Übertragung auf die Verarbeitung von Stichprobenwerten dar. Hierzu wurden die im verwendeten R-Paket implementierten Diskretisierungen der stetigen Verteilungsfunktionen verwendet. Für nahezu alle Varianten (das heißt für Metriken und Teststatistiken) muss die Familie der Normalverteilungen auf einen einzelnen Vertreter reduziert werden. Dazu wurden die Distanzen zunächst als Funktionen der Normalverteilungsparameter aufgefasst und optimiert.

Theoretische Eigenschaften wie die Gemeinsamkeiten und Unterscheide der Konzepte sowie die zu erwartenden Werte für beispielhafte statistische Verteilungen wurden im Kapitel 3 diskutiert. Neben Minimal-Forderungen wie der Ausgabe des kleinstmöglichen Werts bei tatsächlichem Vorliegen einer Normalverteilung, sind Beschränktheit und Symmetrie der Maße anzustrebende Eigenschaften.

Als zentrale Entscheidungshilfe bei der qualitativen Unterscheidung der Methoden dient eine Simulationsstudie, welche im Kapitel 4 vorgestellt wurde. Für Zufallsstichproben verschiedener Verteilungen und verschiedener Größen wurden die Maße unter verschiedenen Gesichtspunkten verglichen. Zusammengefasst mit vorherigen Erkenntnissen ergibt sich schließlich kein eindeutiges Bild: Keine Methode ist allen anderen in allen Kriterien gleichwertig oder überlegen. Damit stellt sich für die Benennung eines Verfahrens als *das beste* die Frage nach der Gewichtung der einzelnen Kriterien, welche sich je nach Anwendungsfall unterscheiden kann: So kann beispielsweise die gute Trennung der Normalverteilung von diskreten Verteilungen in einigen Anwendungen elementar sein, während in anderen Fällen diskrete Stichproben a priori ausgeschlossen werden können.

In jedem Fall als nicht relevant können der Totalvariations-, der Hellinger- und der Cramér-von-Mises-Abstand sowie die Teststatistik des Chi-Quadrat-Tests angesehen werden: Für diese vier Verfahren (in der hier verwendeten Form inklusive Implementierung) findet sich in jedem Fall ein besseres anderes Vorgehen. Während die Anderson-Darling- und Shapiro-Wilk-Teststatistiken auch als geeignet eingeordnet werden können, scheinen die Distanzen von Kolmogorov und Lévy sowie die Teststatistik nach Jarque und Bera am besten geeignet zu sein. Als Empfehlung wird schließlich die gemeinsame Verwendung von Kolmogorov-Distanz und JB-Teststatistik herausgestellt.

Diese beiden Methoden wurden im Kapitel 6 eingehender betrachtet. Dabei konnten die Vorteile von Transformationen der Art $LJB = \log(JB + 1)$ und $K^* = \sqrt{n}\ K$ herausgestellt werden. Weiter konnte für K^* gezeigt werden, dass die Optimierungen der Normalverteilungsparameter zwar zu quantitativen Unterschieden der Abstandsmessungen führen, diese qualitativ aber nicht relevant sind. Für die Berechnung von LJB stellt sich diese Frage konstruktionsbedingt nicht.

Die Bindung der Vorgehensweise an eine schnelle Verfügbarkeit in der Programmiersprache R war für die Durchführung der dargestellten Untersuchungen Voraussetzung. Dabei ist diese Voraussetzung, trotz der großen Verbreitung von R gerade im wissenschaftlichen Bereich, offenbar nicht objektiven, sondern rein praktischen Gründen anzulasten. Dies gilt auch für die Diskretisierungen der Metriken. Weitergehende Untersuchungen können demnach weitere Methoden, oder die hier vorgestellten Methoden in variierter Umsetzung betrachten.

Ebenfalls für weitergehende Betrachtungen bieten sich zwei naheliegende Problemstellungen an: Die hier untersuchte Frage nach dem Abstand zur Normalverteilung lässt sich leicht auf die verallgemeinerte Situation eines Abstands zu einer beliebigen Verteilung übertragen. Je nach Methode ist die Übertragung der Grundidee dabei ohne Weiteres möglich (bspw. χ^2) oder erscheint zunächst eher aufwendig (Shapiro-Wilk). Eine Erweiterung auf *beliebige* Verteilungen könnte insbesondere auch mehrdimensionale Verteilungen umfassen, für welche sich die hier zugrunde liegende Frage nach dem Grad der Nähe zu einer bestimmten Verteilung oder einer Verteilungsfamilie ebenfalls formulieren lässt.

A Ergänzende Grafiken

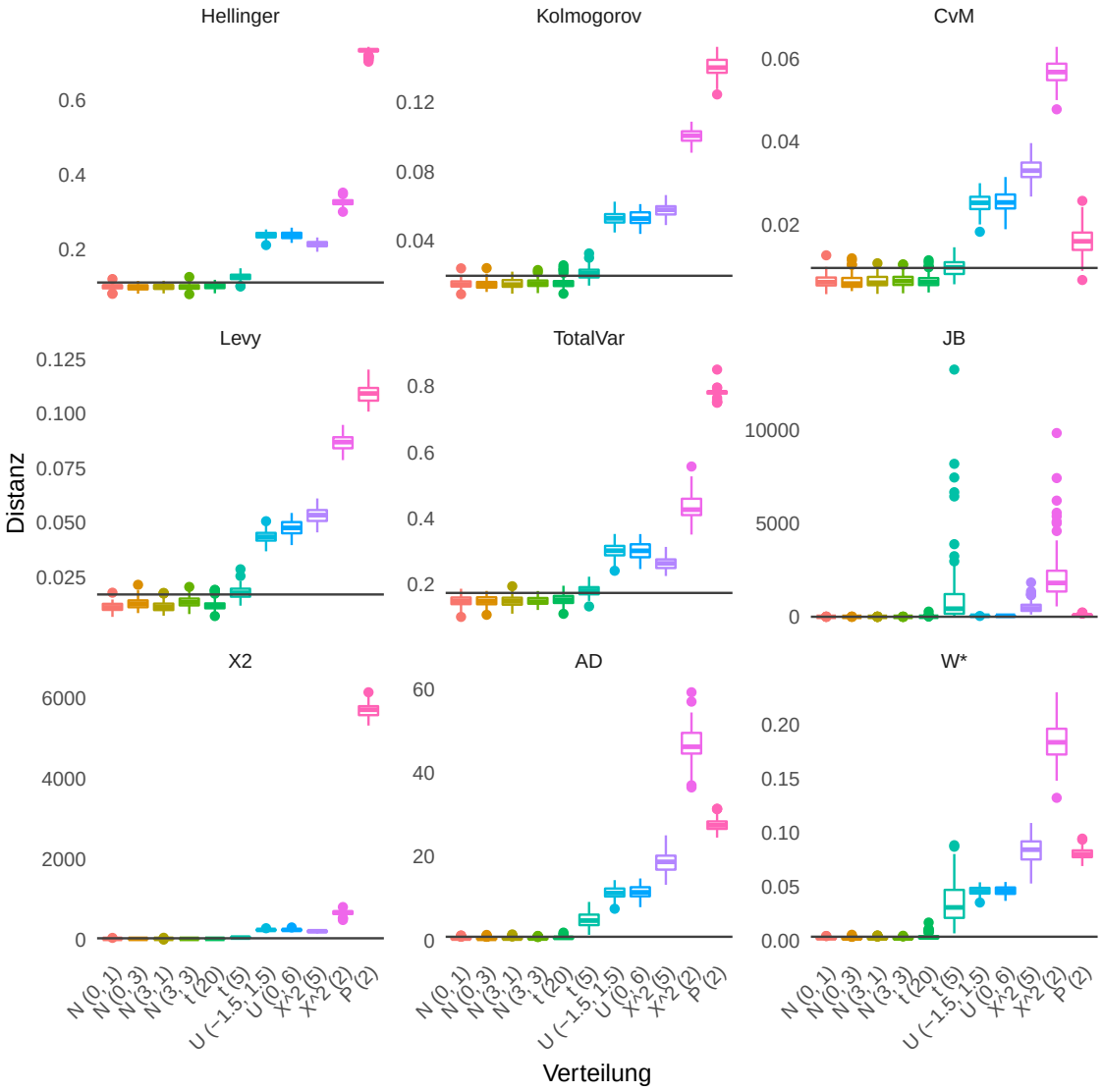

Abbildung 13: Ergänzend zu Teil 4.2 und den Abbildungen 5 und 6, hier die Simulationsergebnisse für die Stichprobengröße 1 000: Trotz der recht großen Stichprobe sind grafisch nicht alle Abweichungen der Daten von Normalverteilungen durch die Distanzen zu erkennen.

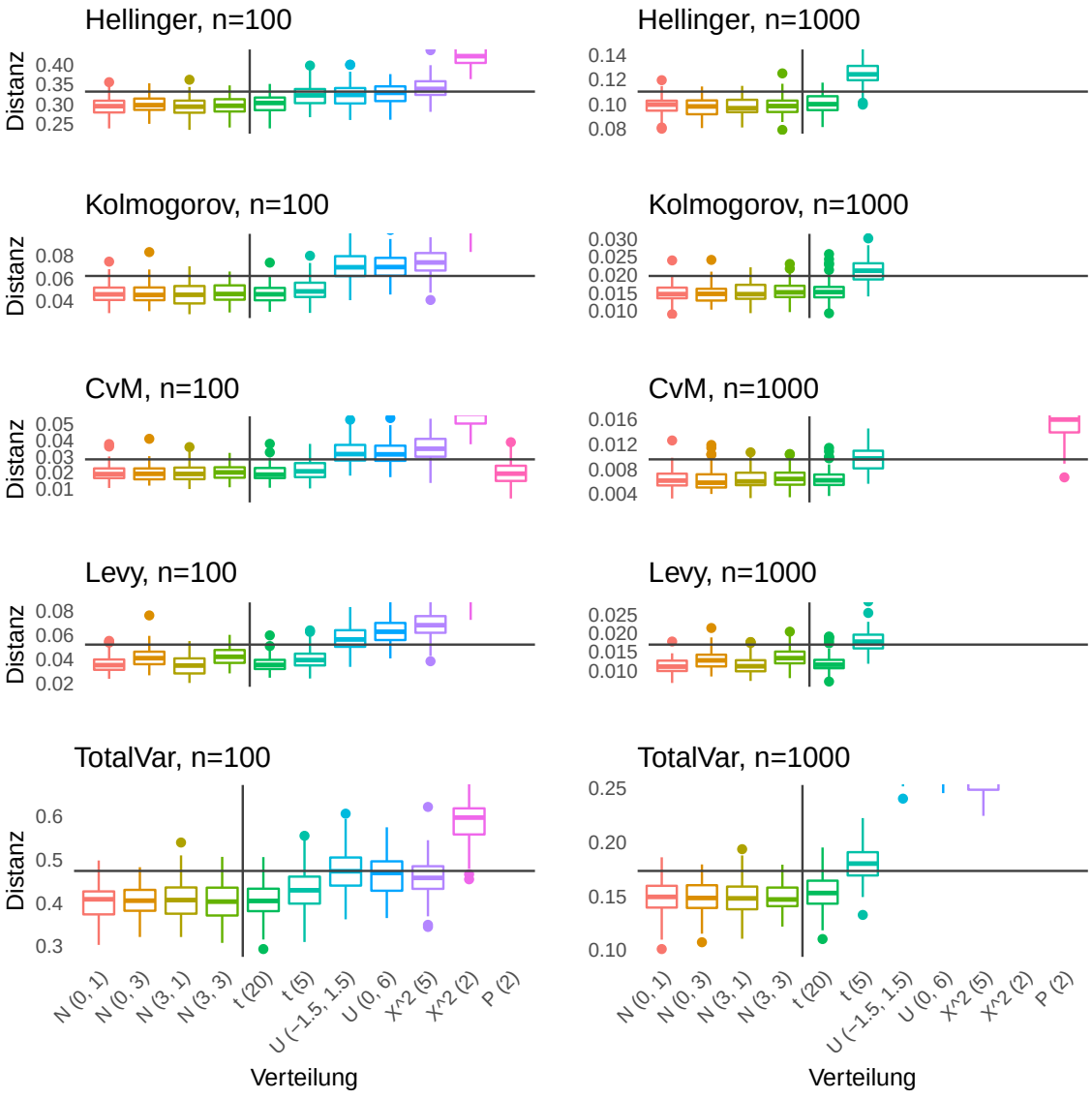

Abbildung 14: Ergänzend zu Teil 4.2 und Abbildung 7, die Simulationsergebnisse als „Zoom“ für kleine Distanzen und die Metriken.

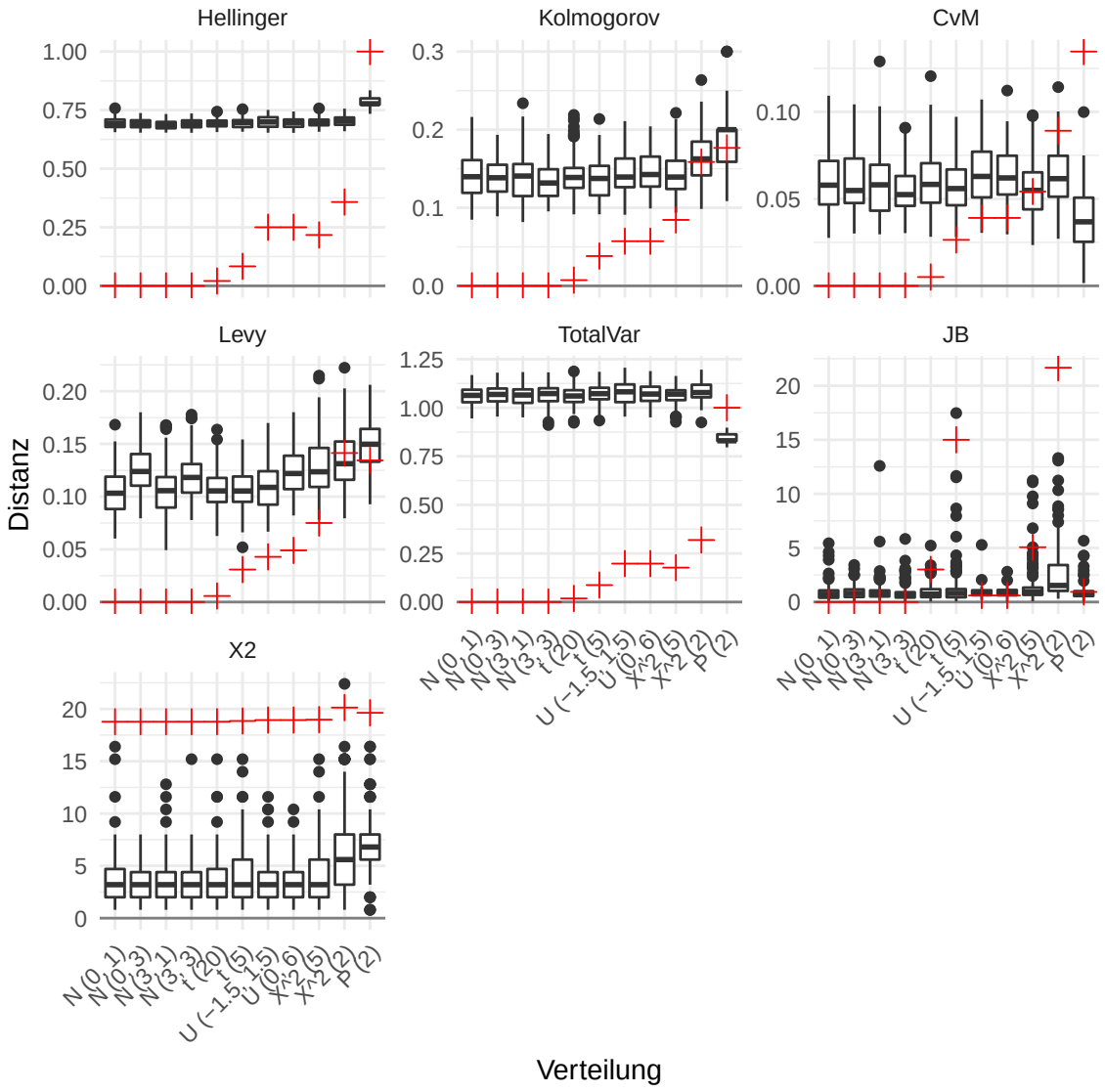

Abbildung 15: Ergänzend zu Teil 4.3 und Abbildung 8: Die Ergebnisse der Simulation mit den theoretisch zu erwartenden Werten (rot). Hier für die Stichprobengröße 10 und ohne AD und W^*.

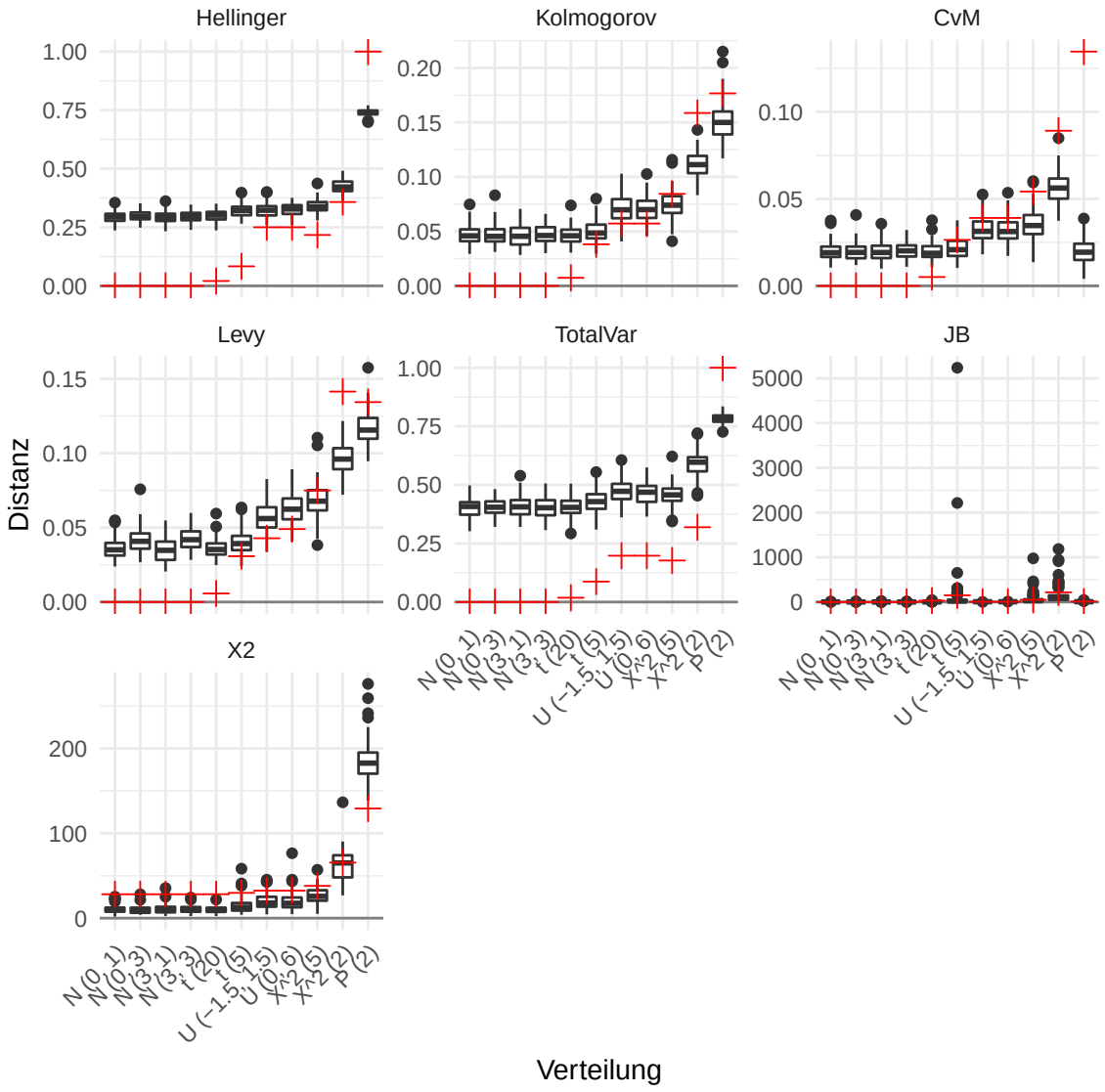

Abbildung 16: Ergänzend zu Teil 4.3 und Abbildung 8: Die Ergebnisse der Simulation mit den theoretisch zu erwartenden Werten (rot). Hier für die Stichprobengröße 100 und ohne AD und W^*.

Literatur

Naser Reza Arghami, 2011: *Monte Carlo comparison of seven normality tests*. Journal of statistical computation and simulation. Jahrgang 11, Nummer 81, Seite 965 - 972. Talyor & Francis, Abington.

Rudof Beran, 1977: *Minimum Hellinger Distance Estimates for Parametric Models*. The Annals of Statistics, Jahrgang 5, Nummer 3, Seite 445-463. Institute of Mathematical Statistics.

Dennis D. Boos, 1982: *Minimum anderson-darling estimation*. Communications in Statistics - Theory and Methods, Jahrgang 11, Nummer 24, Seite 2 747-2 774. Taylor and Francis.

Ralph B. D'Agostino, Albert Belanger und Ralph B. D'Agostino, 1990: *A suggestion for using powerful and informative tests of normality*. The American Statistician, Jahrgang 44, Nummer 4, Seite 316-321. American Statistical Society.

Ulrike Genschel und Claudia Becker, 2005: *Schließende Statistik. Grundlegende Methoden*. Springer-Verlag Heidelberg.

Jürgen Groß, 2004: *A Normal Distribution Course*. Peter Lang Verlag, Frankfurt am Main.

Jürgen Groß and Uwe Ligges, 2015: *nortest: Tests for Normality*. R-Paket, Version 1.0-4.

Johannes Hain, 2010: *Comparison of Common Tests for Normality*. Diplomarbeit am Lehrstuhl für Mathematik VIII (Statistik) der Julius-Maximilians Universität Würzburg, Institut für Mathematik und Informatik.

Joachim Hartung, Bärbel Elpelt und Karl-Heinz Köster, 2005: *Lehr- und Handbuch der angewandten Statistik*. Oldenbourg Wissenschaftsverlag, München, 14. Auflage.

Peter J. Huber und Elvezio M. Ronchetti, 2009: *Robust Statistics*. John Wiley & Sons, Hoboken, New Jersey. Zweite Auflage.

Carlos M. Jarque und Anil K. Bera, 1987: *A Test for Normality of Observations and Regression Residuals*. International Statistical Review, Jahrgang 55, Nummer 2, Seite 163 - 172. International Statistical Institute.

George G. Judge, W. E. Griffiths, R. Carter Hill, Helmut Lütkepohl und Tsoung-Chao Lee, 1985: *The Theory and practice of econometrics*. John Wiley & Sons, New York. Zweite Auflage.

Matthias Kohl, 2005: *Numerical Contributions to the Asymptotic Theory of Robustness*. Dissertation, Universität Bayreuth

Lukasz Komsta und Frederick Novomestky, 2015: *moments: Moments, cumulants, skewness, kurtosis and related tests*. R-Paket, Version 0.14.

Peter A. W. Lewis, 1961: *Distribution of the Anderson-Darling Statistic*. The Annals of Mathematical Statistics, Jahrgang 32, Nummer 4, Seite 1 118-1 124. Institute of Mathematical Statistics.

Jagdish K. Patel und Campbell B. Read, 1996 *Handbook of the normal distribution*. Marcel Dekker, New York. Zweite Auflage.

R Core Team, 2017: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien.

Helmut Rieder, 1994: *Robust Asymptotic Statistics*. Springer-Verlag, New York.

P. Ruckdeschel, M. Kohl, T. Stabla und F. Camphausen, 2006: *S4 Classes for Distributions*. R News, Jahrgang 6, Nummer 2. R Foundation for Statistical Computing, Wien.

Bernhard Rüger, 2002: *Test- und Schätztheorie, Band II: Statistische Tests*. Oldenbourg Wissenschaftsverlag, München.

S. S. Shapiro und M. Wilk, 1965: *An analysis of variance test for normality (complete samp1es)*. Biometrika, Jahrgang 52, Nummer 3 und 4, Seite 591-611. Biometrika Trust.

Bariş Sürücü, 2008: *A power comparison and simulation study of goodness-of-fit tests*. Computers and Mathematics with Applications, Jahrgang 56, Seite 1 617 - 1 625. Elsevier.

Thorsten Thadewald und Herbert Büning, 2007: *Jarque – Bera Test and its Competitors for Testing Normality – A Power Comparison*. Journal of Applied Statistics, Jahrgang 34, Nummer 1, Seite 87 - 105. Taylor & Francis.

Hermann Witting, 1985: *Mathematische Statistik I: Parametrische Verfahren bei festem Stichprobenumfang*. B. G. Teubner, Stuttgart.

B. W. Yap und C. H. Sim, 2011: *Comparisons of various types of normality tests*. Journal of Statistical Computation and Simulation, Jahrgang 81, Nummer 12, Seite 2 141 - 2 155.