Wie man das Einschätzen der Normalverteilung mit großen und einfach hergestellten Datensätzen lehrt[1]

CHRISTOPHER W. KULP & GENE D. SPRECHINI, WILLIAMSPORT

[1] Original: „Teaching the assessment of normality using large easily-generated real data sets" in *Teaching Statistics* 38 (2016) 2, S. 56–62. Übersetzung: ANDREAS EICHLER, Kassel

Zusammenfassung: *Es wird ein konkreter Unterrichtsvorschlag mit einem einfach herzustellenden, großen und realen Datensatz für den Statistikunterricht präsentiert. Der Vorschlag besteht darin, die Videoaufzeichnung eines beliebigen Objekts zu analysieren. Die Farbdaten des aufgenommenen Objekts dienen als Datensatz, dessen grafische Darstellung mit statistischen Graphen wie Histogrammen, Quantil-Quantil-Plots oder Punktwolken die Untersuchung der Streuung ermöglichen Ebenso sind vertiefte formale Auswertung möglich.*

1 Einleitung

Eine Vielzahl von Aufgaben zur statistischen Datenanalyse wird momentan im Unterricht verwendet. Manche dieser Übungen umfassen reale Daten, wobei der interessierte Leser auf Levine und Roos (1998) verwiesen wird. Andere, eher traditionelle Übungen, die ähnlich zu denen in Levine und Roos (1998) sind, fordern Schülerinnen und Schüler auf, Objekte zu zählen oder Messwerte zu diesen Objekten zu ermitteln. Solch ein Vorgehen kann allerdings erheblich Unterrichtszeit verbrauchen. Außerdem scheint es für größere Datensätze notwendig zu sein, die Daten mehrerer Klassen zu vereinen. Daher will dieser Beitrag ein Verfahren präsentieren, das das Erzeugen und das Analysieren von Daten ermöglicht, die schnell und einfach gewonnen werden können. Das Verfahren basiert darauf, ein Handy oder eine Kamera dafür zu verwenden, ein beliebiges Objekt mit Video aufzunehmen. Das Video wird dann in eine geeignete Software importiert (etwa Mathematica, wobei unser Vorschlag nicht auf dieses Programm beschränkt sein muss) und die Farbinformation für die Pixel vom eigentlichen Video separiert. Digitale Kameras erzeugen neben dem eigentlichen Farb-Signal auch ein Rauschen, das als Messfehler gedeutet werden kann. Man kann annehmen, dass diesem Messfehler die Normalverteilung zugrunde liegt. Das wollen wir aber im Folgenden noch näher untersuchen. Die Methoden für diese Art Untersuchung haben wir so entwickelt, dass sie effizient in der Schule eingesetzt werden können. Zudem soll hier betont werden, dass wir auf eine formale Analyse des Rauschens einer Kamera verzichten. Interessierte seien hier auf Irie et al. (2008) verwiesen.

Unser Vorgehen hat Vorteile gegenüber der traditionellen Datenanalyse. Zunächst verwenden wir große und reale Datensätzen, die sehr schnell erhoben werden können. Camcorder nehmen 30 Bilder pro Sekunde auf, das heißt, in nur wenigen Minuten können Tausende von Datenpunkten erzeugt werden. Verwendet man also Camcorder, so erhalten Schülerinnen und Schüler Erfahrungen im Umgang mit großen Datensätzen. Dabei werden sie auch erfahren, dass reale Datensätze nicht so sauber sind, wie es die Schulbuchbeispiele vorgeben. Weiterhin können Schülerinnen und Schüler den Unterschied von großen und kleinen Datensätzen erfahren. Schließlich ermöglicht die potentiell sehr schnelle Datenerhebung eine Zeitersparnis, die zugunsten des Verständnisses der statistischen Methoden verwendet werden kann.

Ein weiteres wesentliches Argument für den Einsatz von Camcordern ist, dass die Verwendung der Technik motivierend für Schülerinnen und Schüler ist. Sie selbst könnten etwa mit ihren Handys auch Videos aufzunehmen. Camcorder bieten aber zu Handys noch eine relativ günstige und leicht zugängliche Alternative. Die Camcorder können dann auch hinsichtlich ihrer Qualität beurteilt werden, sich in der Qualität der Daten bemisst. Schülerinnen und Schüler erfahren damit also auch unterschiedliche Qualitäten von Datensätzen.

Weiterhin lernen Schülerinnen und Schüler mit unserem Vorgehen auch, den Rechner bzw. Technologie für die Erforschung und Analyse von Daten einzusetzen (insbesondere ältere Schüler). Gerade weil die Datensätze groß sind, müssen Schülerinnen und Schüler einen Rechner verwenden. Während ambitionierte Schülerinnen und Schüler versucht sein könnten, eigene Algorithmen für die Datenauswertung zu schreiben, macht das Software wie Maple oder R überflüssig. Lernen müssen Schülerinnen und Schüler dagegen auf jeden Fall, ein file (hier ein Video) in eine Software zu importieren.

Mathematica ist ein Computeralgebra-System (CAS), das enorme numerische Möglichkeiten bietet. Es muss aber noch einmal betont werden, dass es nicht zwingend notwendig ist, Mathematica zu nutzen. Stattdessen ermöglicht auch andere Software unseren Vorschlag auszuführen. Tatsächlich ist aber die Verwendung von Mathematica für das Lehren von Statistik nicht neu. Hogg und Tanis (2010) haben beispielsweise Mathematica verwendet; eine Aufgabe in diesem Beitrag wird den von Hogg und Tanis propagierte Nutzung des Kolmogorov-Smirnoff-Test sinnvoll ergänzen. Mathematica ermöglicht es weiterhin, mit eigenen Prozeduren eine Datenanalyse schnell auszuführen und eine Vielzahl von Analysemethoden effizient einzusetzen. Auch können Schülerinnen und Schüler Videos importieren, einen Datensatz aus einem Video zu generieren und dann im gleichen Programm die Daten auch auswerten, was ein Vorteil gegenüber Lösungen mit verschiedener Software ist. In diesem Beitrag gehen wir zwar auf Mathematica-Befehle ein, aber nicht so sehr auf die eigentliche Verwendung dieser Software, um auch Anwendern anderer Software die Übersetzung der Befehle zu ermöglichen. Die geeigneten Befehle kann man zwar häufig im Internet finden, eine Basis von Grundbefehlen ist aber hilfreich. Wir hoffen, dass wir mit den Mathematics-Befehlen diese Basis schaffen. Mathematica kann einerseits eher traditionellen Software-Lösungen wie etwa SPSS überlegen sein, da es beispielsweise einfach ist, ein Video zu importieren. Andererseits mag die stetige Entwicklung von Zusatzpaketen für Software wie SPSS es irgendwann einfacher machen, Daten ohne zusätzliche Programmierung zu importieren.

2 Die Unterrichtseinheit

Die Einheit beginnt damit, ein Video von einem beliebigen Objekt aufzunehmen. Einzige Einschränkung für das Objekt ist, dass es in der Farbe konstant ist, d. h. man sollte beispielsweise kein Blinklicht aufnehmen. Für diesen Beitrag wurde ein rotes Spielzeug verwendet. Für die Aufnahme kann wie gesagt ein Handy oder ein Camcorder verwendet werden. Dabei muss man nicht besonders auf ausreichende Beleuchtung achten, sofern die Beleuchtung nur einigermaßen konstant ist. Sich ändernde Lichtverhältnisse könnten für sehr gute Schüler allein die Möglichkeit bieten, zu lernen, wie man Daten von einem Trend bereinigt. Wir empfehlen allerdings, ein Stativ für die Kamera zu verwenden, damit das gewählte Objekt nicht innerhalb des Bildes wandert. Wir gehen später darauf ein, warum das wichtig ist. Eine letzte Überlegung betrifft das Format der Aufnahme. Mathematica kann zwei Formate importieren, .mov und .avi. Solche Formate sind aber auch über die Wahl der Software hinaus zu empfehlen, da die Dateien nicht zu groß werden und dadurch weniger Berechnungszeit verlangen. Die Länge der Aufnahme entscheidet dann schließlich über die Anzahl der Datenpunkte. Auch hier muss man wiederum bedenken, dass große Dateien höhere Bearbeitungszeiten erfordern. Camcorder nehmen beispielsweise 30 Bilder pro Sekunde auf. Jedes Bild entspricht dann hier einem Datenpunkt. Für diese Arbeit haben wir ein 50-Sekunden-Video erstellt und damit 1500 Datenpunkte. Allgemein ist es hier sinnvoll zu wissen, wie viele Bilder ein Camcorder pro Sekunde aufnimmt, um die Größe der Datensätze abschätzen zu können.

Nachdem das Video aufgenommen ist, muss es in die ausgewählte Software importiert werden. Mathematica importiert ein Video automatisch als Folge von Bildern, wobei es notwendig ist, die Anzahl dieser Bilder zu wissen, bevor man das Video importiert.

Sobald das Video importiert ist, kann man den Farb-Kanal für das Pixel, für das man sich interessiert, aufrufen. Am Einfachsten ist es, dazu ein Pixel in einem Bild auszuwählen und mit einem Klick der rechten Maustaste „GetIndices" zu wählen. Die Reihen- und Spaltennummer für ein Pixel wird angezeigt, wenn mit der Maus über diese Pixel gefahren wird. Wählt man ein Pixel aus, dann kann der Farb-Kanal für jedes Bild mit dem Befehl „ImageData" gefunden werden. Die Farbe jedes Pixels ist mit einem Tripel RGB angegeben. Jeder Farbkanal (R,G,B) hat dabei einen Wert zwischen 0 und 255. Die Verwendung eines Stativs wird an dieser Stelle klar: Wenn das gefilmte Objekt in der Bild wandert, dann stammen die Daten nicht vom gleichen Punkt in dem Objekt, d. h. die Daten stammen nicht aus der gleichen Quelle. Abbildung 1 zeigt die Daten aus den Farbkanälen von unserem Video.

Das aufgenommene Objekt ist rot. Daher hat der Farbkanal für R die höchsten Werte. Ebenso wird ersichtlich, dass die Kamera auch ein Rauschen produziert. Die Farbe des Objekts ändert sich in der Aufnahme nicht, aber die Elektronik der Kamera verzeichnet

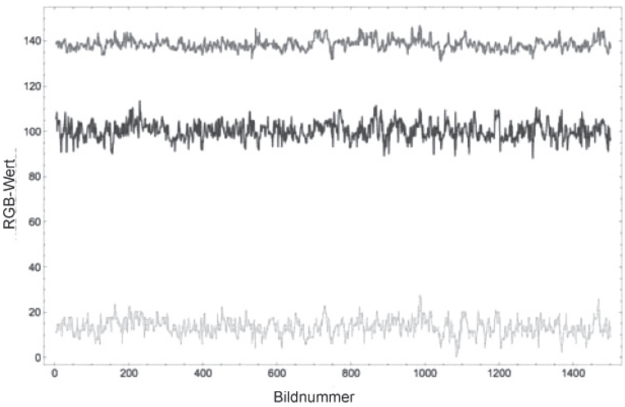

Abb. 1: Farbkanäle des Videos für R oben, B in der Mitte und G unten.

dennoch ein Rauschen, sei es aufgrund natürlicher Lichtschwankungen, sei es aus anderen Gründen.

Herkömmliche Datenanalyse basiert typischerweise darauf, Histogramme der Daten zu zeichnen und zu interpretieren. Das kann man auch in Mathematica leicht mit dem Befehl „Histogram" ausführen. Abbildung 2 zeigt das Histogramm unserer Aufnahme mit Säulenbreite 1. Die Lehrkraft könnte solch ein Histogramm präsentieren und dabei die Beschränkungen der grafischen Darstellung ansprechen. Für kleine Datensätze ist beispielsweise die Darstellung sehr stark abhängig von der Säulenbreite und auch vom Startpunkt der Säulen. Das ist für große Datensätze nicht der Fall, so dass sich Raum für andere Betrachtungen eröffnet.

Drei Aspekte in Abbildung 2 sind diskussionswürdig. Die erste betrifft die Frage, ob die Farbe eine stetige Zufallsgröße darstellt. Stetige Zufallsgrößen werden in einer bestimmten Einheit gemessen. Die Genauigkeit, mit der hinsichtlich dieser Einheit gemessen wird, sollte aber während der gesamten Beobachtung gleich bleiben. D. h., wenn Höhen in Zentimetern gemessen werden, dann ist es nicht angemessen, die eine Messung mit 168 cm zu machen und die andere mit 171,5 cm. Schülerinnen und Schüler sind das auch gewohnt, wenn sie Größen wie Höhen, Gewichte oder Zeiten messen. In gleicher Weise sollte, wenn einmal Geldbeträge (etwa Stundenlöhne) mit zwei Dezimalen gemessen werden, alle weiteren Beträge auch mit diesen beiden Dezimalen gemessen werden. In unserem Beispiel soll die Größe Farbe mit einem Wert von 0 bis 255 angegeben werden. Diese Größen haben keine explizite Einheit und sind ungewohnt für Schülerinnen und Schüler. Dennoch sollte diesen bewusst sein, dass es sich hier um eine stetige Größe handelt.

Bisher haben wir also Zufallsgrößen, die theoretisch Werte zwischen 0 und 255 annehmen können. In Abbildung 2 wird allerdings deutlich, dass die Spannweite der Werte für jede Farbe deutlich kleiner ist. Das ist auch nicht überraschend, die Streuung der Werte, die als Rauschen bezeichnet wird, ist durch den Camcorder bedingt. Weiterhin kann man der Abbildung 2 entnehmen, dass manche Werte nicht gemessen wurden. Das kann Schülerinnen und Schülern den Unterschied zwischen realen Daten und solchen in Schulbüchern verdeutlichen. Interessant ist dabei, dass die Löcher nur im blauen und grünen Kanal zu sehen sind, und diese auch eine größere Streuung zu

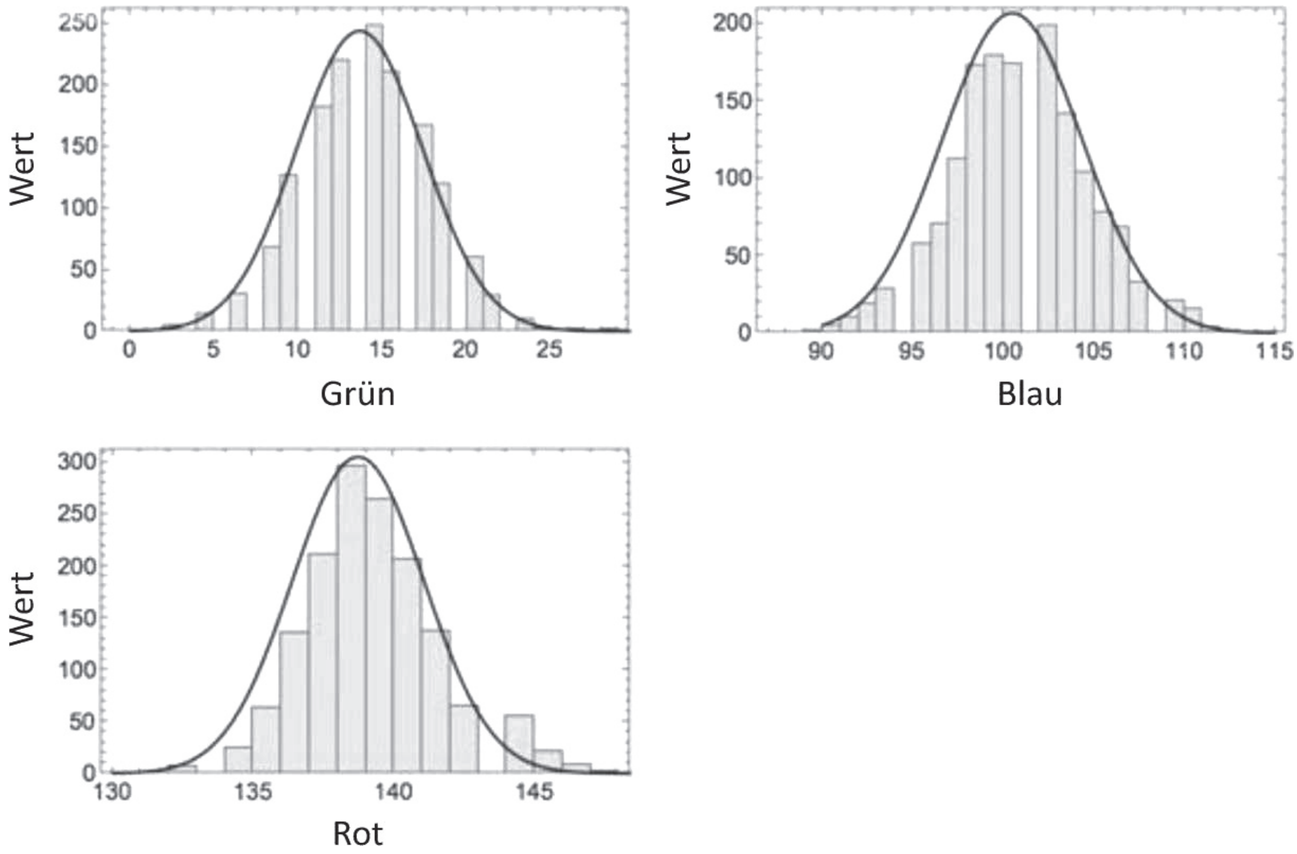

Abb. 2: Histogramme für jeden Farbkanal. Die schwarze Linie repräsentiert die zugehörige Normalverteilung mit gleichem Mittelwert und Standardabweichung wie die Daten des Farbkanals.

haben scheinen als der rote Kanal. Da das Objekt rot war, eröffnet das einige Anschlussfragen einschließlich der Qualität des Camcorders. Camcorder mit hoher Qualität haben hoch empfindliche Sensoren und nehmen daher auch mit weniger Rauschen auf und produzieren schließlich auch Histogramme mit weniger Löchern. Natürlich würden auch größere Balkenbreiten im Histogramm die Löcher verschwinden lassen. Ob aber so ein Vorgehen Vor- oder Nachteile birgt, könnten Schülerinnen und Schüler untersuchen bzw. diskutieren. Ebenso ergäbe ein längeres Video größere Streuung in den beobachteten Werten (unser Video ist 50 Sekunden lang). Dabei müsste man aber bedenken, dass längere Videos auch größere Dateien und Bearbeitungszeiten nach sich ziehen.

Es kann auch thematisiert werden, dass größer nicht gleich besser ist. Obwohl wir bereits einen Datensatz mit 1500 Werten haben, hätten wir die Größe leicht auf Zehntausende Werte vergrößern können. Die Vorteile solch einer Größe sind allerdings stark durch die Güte des hier eingesetzten Camcorders beschränkt. Was deutlich größere Vorteile hätte, wäre der Einsatz eines Camcorders mit höherer Qualität und damit einer höheren Bildpräzision. Mit anderen Worten ist die Qualität der Messung einem vergrößerten Datensatz vorzuziehen.

Wir wollen nun die Frage untersuchen, inwieweit die Abweichungen in den Farbwerten, die als durch die Elektronik verursachtes Rauschen gelten können, normalverteilt sind. Die schwarzen Linien in Abbildung 2 repräsentieren Normalverteilungen mit dem gleichen Mittelwert und der gleichen Standardabweichung wie die zugehörigen Verteilungen. Schülerinnen und Schüler sollten hier erinnern, dass die schwarzen Glockenkurven ein theoretisches Modell darstellen. Tatsächlich sehen wir selbst bei 1500 Daten kein glattes Histogramm, auch dann nicht, wenn man einen Camcorder mit höherer Qualität eingesetzt hätte. Schülerinnen und Schüler könnten hier die Kurven als Beleg nehmen, dass das Rauschen normalverteilt ist, könnten aber auch Zweifel haben. Wie auch immer, es sollte herausgestellt werden, dass der Vergleich eines Histogramms mit einer theoretischen Verteilung insbesondere dann fehlgedeutet werden könnte, wenn wenige Beobachtungen mit kleinen Wahrscheinlichkeiten des Auftretens vorliegen.

Für Lehrkräfte, die über grafische Darstellung hinaus gehen möchten, könnten die Farbwerte als Ergebnis der Simulation einer normalverteilten Zufallsgröße mit dem Mittelwert und der Standardabweichung der beobachteten Daten verstanden werden. Allerdings sollte man bedenken, dass keine Notwendigkeit für und vielleicht sogar deutliche Bedenken gegen eine allzu vertiefte theoretische Diskussion bestehen. Streift man aufgrund der altersgemäßen Fähigkeiten der Schülerinnen und Schüler nur den Kern der mathematischen Konzepte, so könnte der Sinn der Betrachtung verloren gehen. Man kann aber die theoretischen Werte schlicht als das beschreiben, was für den kleinsten Farbwert, den zweitkleinsten Farbwert usw. zu erwarten ist und zwar in einer Stichprobe mit 1500 normalverteilten Farbwerten.

Der Befehl „RandomReal" in Mathematica erzeugt ein Histogramm die auf der Simulation normalverteilter Zufallsgrößen basieren. In diesem Sinne ist Abbildung 3 bezogen auf die ersten beiden Histogrammen entstanden und zwar jeweils basierend auf dem Mittelwert und der Standardabweichung der roten Farbwerte. Beide Datensätze in Abbildung 3 enthalten 1500 Daten, die quasi in Echtzeit erzeugt werden können. Herausgestellt werden sollte dabei, dass die Menge der Erwartungswerte das theoretische Modell darstellen, während die simulierten Daten nur einen möglichen empirischen Datensatz bilden. Als Zusatz könnten Schülerinnen und Schüler hier überlegen, wie unterschiedlich verschiedene empirische Datensätze sein könnten.

Diskutiert man die Schwierigkeit, die Histogramme in Abbildung 3 per Augenmaß zu vergleichen, so kann als Lösung ein Quantil-Quantil-Plot motiviert werden, um mit diesem die empirischen Daten mit den theoretischen Verteilungen zu vergleichen. Es könnte den Schülerinnen und Schüler erklärt werden, dass die experimentellen Daten im Histogramm von kleinsten bis hin zum größten Wert geordnet werden können. Das Histogramm links in Abbildung 3 zeigt dagegen, was theoretisch zu erwarten ist bei 1500 normalverteilten Daten mit gegebenen Mittelwert und Standardabweichung. Auch hier besteht die Ordnung vom kleinsten bis hin zum größten Wert. Weiter müssen Schülerinnen und Schüler erinnern, dass eine Gerade mit Steigung 1 und Nullstelle im Ursprung die Punkte enthält, bei denen die x- und y-Koordinaten gleich sind. Anschließend ist es für Schülerinnen und Schüler keine Schwierigkeit Folgendes zu erkennen: Liegen zwei gleich große Datensätze von Beobachtungen vor und sind beide Datensätze von kleinsten zum größten Wert geordnet und zeichnet man nun jeweils die beiden kleinsten Werten, die beiden zweitkleinsten Werten usw., dann sind genau die Werte gleich, die auf der oben beschriebenen Gerade liegen. Wir führen das mit den empirischen sowie den theoretischen Daten aus. Eine Untersuchung, wie dicht die Punkte an der genannten Geraden liegen, zeigt an, wie weit die empirischen Daten einer Normalverteilung entsprechen. Diese grafische Darstellung nennt man Q-Q-Plot.

Für das Zeichnen eines Q-Q-Plots kann der Rechner verwendet werden so wie hier für die Darstellung der Farbkanäle in Abbildung 4. Dabei ist auf der x-Achse jeweils der empirisch gemessene Wert des Farbkanals gegeben, auf der y-Achse der auf der Basis der Normalverteilung theoretisch erwartete Wert, wobei die Normalverteilung denselben Mittelwert und dieselbe Standardabweichung hat wie die empirischen Daten.

In Abbildung 4 ist für jeden Farbkanal auffällig, dass die Punkte nahe an der Ursprungsgeraden mit Steigung 1 liegen und dabei zwei Muster zeigen: Zunächst scheinen die Punkte Geradenstücke in Treppenform zu bilden und keine glatte Kurve. Während bei den theoretischen Werten keine Wiederholungen vorkommen, ist das bei den experimentellen Daten

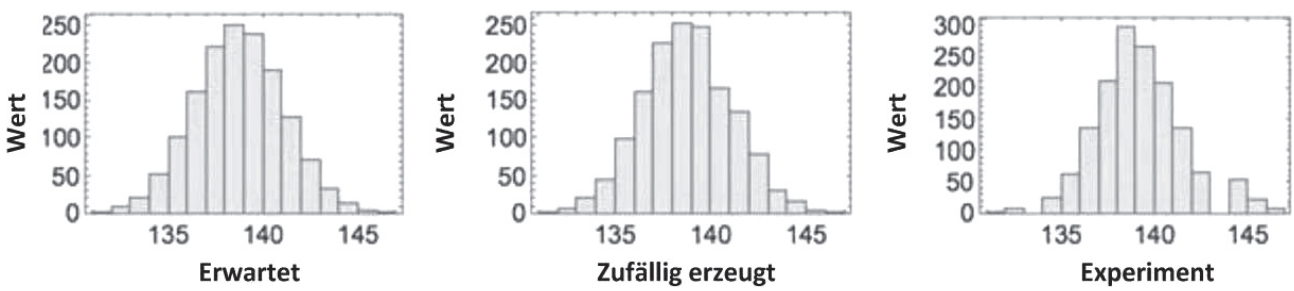

Abb. 3: Histogramme zu den Erwartungswerten auf der Basis der Normalverteilungsannahme (links), simulierte Daten mit Mathematica (Mitte) und experimentelle Daten (rechts)

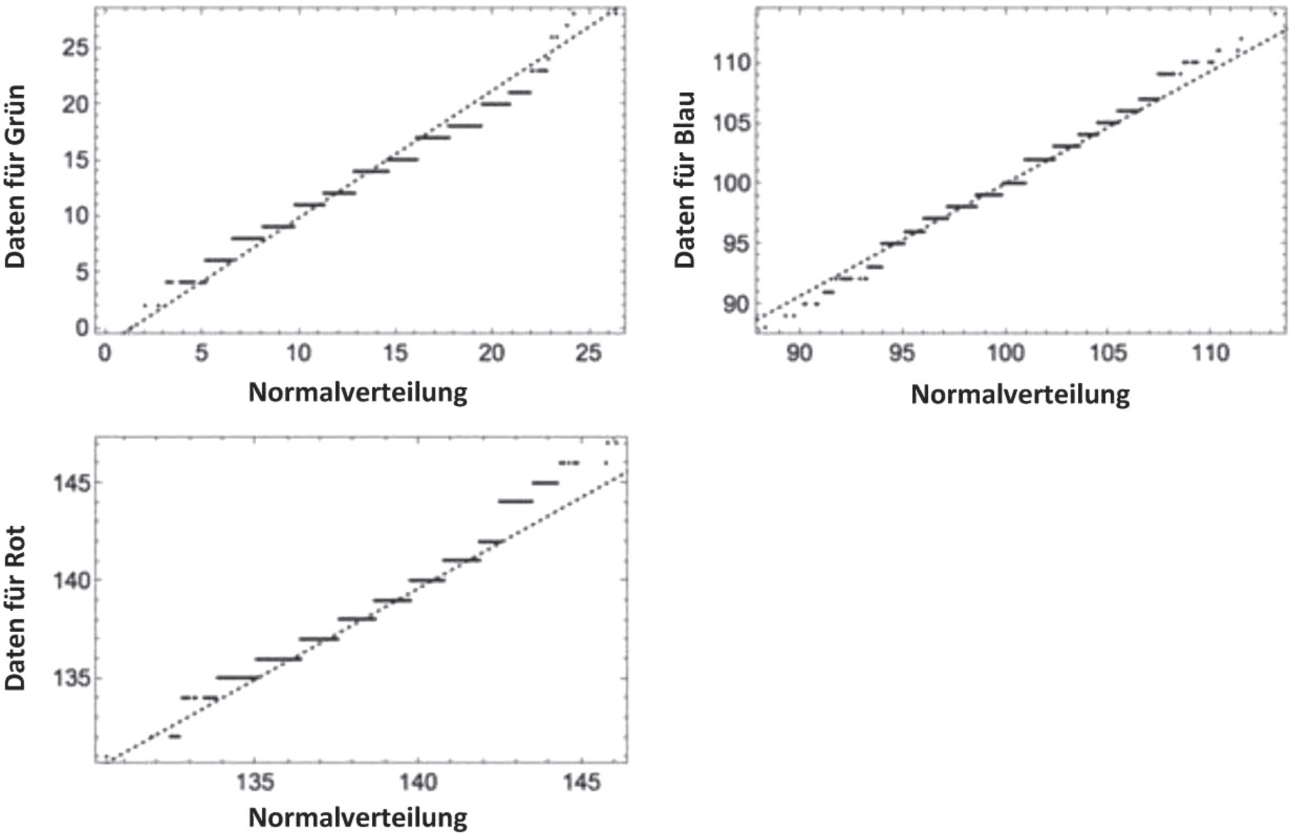

Abb. 4: Quantil-Quantil-Plots für jeden Farbkanal. Die gestrichelte Linie ist die Ursprungsgerade mit Steigung 1.

der Fall wie Schülerinnen und Schüler leicht erkennen können.

Ein weiteres Muster betrifft den Unterschied an den beiden Enden des Graphen, der besonders ausgeprägt am rechten Ende des roten Farbkanals ist. Der Q-Q-Plot gibt also vertiefte Erkenntnisse über die Ränder der Daten, die so im Histogramm nicht enthalten sind. So sind die experimentellen Daten offenbar an den Enden deutlicher gestreut als man das auf der Basis der Normalverteilung erwartet hätte. Die Aufgabe des Betrachters ist es nun zu entscheiden, ob die Daten noch als normalverteilt bzw. zumindest annähernd normalverteilt eingeschätzt werden können oder nicht. Dabei kann auch diskutiert werden, was denn „zumindest annähernd" bedeuten könnte. Für diejenigen, die eine Datenanalyse auch jenseits der grafischen Darstellungen anstreben, wird dazu ein Hypothesentest im nächsten Abschnitt diskutiert. Diesen Abschnitt möchten wir dagegen mit einer Fragestellung beenden, die eine weitere wertvolle Erkenntnis für Schülerinnen und Schüler darstellen kann.

In den Daten für die Farbkanäle R, G und B ist Rauschen vorhanden, das die Elektronik der Kamera erzeugt hat. Die Beobachtungen selbst stellen eine Zeitreihe dar, also Werte, die über einen bestimmten Zeitabschnitt gemessen wurden. Kann man hier annehmen, dass die Beobachtungen unabhängig sind? Vielleicht sind die Beobachtungen ja auch noch von anderen Umgebungseinflüssen (neben der Elektronik der Kamera) bestimmt. Untersuchen kann man die Frage der Unabhängigkeit bezogen auf die Zeit, indem man Punktwolken verwendet. Jede Punktwolke besteht aus den Daten zweier Farbkanäle, wobei die beiden Farbwerte die Koordinaten eines Punktes in der Wolke bilden. Abbildung 5 zeigt diese Punktwolken für alle Paare von Farbkanälen. Allein durch die Anschauung wird die große Streuung ersichtlich. Dennoch scheint eine positive lineare Abhängigkeit in den Daten zu stecken. Das ist für jedes Paar von Farbkanälen so, aber besonders offenbar für die Punktwolken, in denen auch der rote Farbkanal eingeht. Die Analyse über die grafische Darstellung hinaus nehmen wir wie für die Q-Q-Plots im nächsten Abschnitt vor.

Hypothesentests

Die Diskussion in diesem Abschnitt geht über den normalen Stoff für Schülerinnen und Schüler im Alter von 9 bis 19 hinaus. Allerdings glauben wir, dass diese Diskussion für die Leser, die mit den relevanten statistischen Methoden und zudem mit Mathematica vertraut sind, dennoch interessant sein kann. Darüber hinaus sind die folgenden Überlegungen auch als Zusatzaufgabe für sehr gute Schülerinnen und Schüler möglich. Der Abschnitt kann bei Bedarf aber auch übersprungen werden, ohne dass dadurch ein Bruch erzeugt wird.

Der Vergleich der Reihenfolge der experimentellen Daten mit den normalverteilten ist die Basis des Shapiro-Wilk-Tests (Shapiro & Wilk 1965).

Die Nullhypothese des Tests ist, dass die experimentellen Daten normalverteilt sind. Lehrkräfte, die diese vertiefte Analyse anstreben, können den Mathematica-Befehl „ShapiroWilkTest" verwenden. Wenn man zusätzlich die Option „TestDataTable" verwendet, erhält man die Teststatistik und den *p*-Wert. Natürlich bietet auch andere Software diesen Test an.

Tabelle 1 zeigt die Teststatistik und den *p*-Wert für den Shapiro-Wilk-Test für jeden Farbkanal. Der kleine *p*-Wert in allen drei Fällen zeigt an, dass die Daten ebenso in allen drei Fällen den Shapiro-Test auf Normalverteilung nicht bestehen, die Nullhypothese wird stets abgelehnt. Das kann überraschend für Schülerinnen und Schüler sein, da ja die Histogramme gut zu den Glockenkurven zu passen schienen. Diese Überraschung kann von Lehrenden dazu verwendet werden, um ein bekanntes Phänomen bei Hypothesentests zu thematisieren: Auch wenn die Daten der Nullhypothese nur marginal zu widersprechen scheinen, wird die Nullhypothese bei sehr großen Datensätzen abgelehnt. Die Aufgabe ist nun daher zu untersuchen, welche Abweichungen von der Normalverteilung bestehen und ob diese praktisch relevant sind. Das führt potentiell zu der Frage der praktischen Signifikanz (oder Effektstärke) versus der statistischen Signifikanz. Für die Diskussion der verschiedenen Ursachen des Rauschens selbst, verweisen wir den interessierten Leser auf Irie et al. (2008).

Farbe	Teststatistik	*p*-Wert
Rot	0,961951	$7{,}38975 \cdot 10^{-19}$
Grün	0,98057	$2{,}21175 \cdot 10^{-13}$
Blau	0,986389	$1{,}14697 \cdot 10^{-10}$

Tab. 1: Ergebnisse des Shapiro-Wilk-Tests

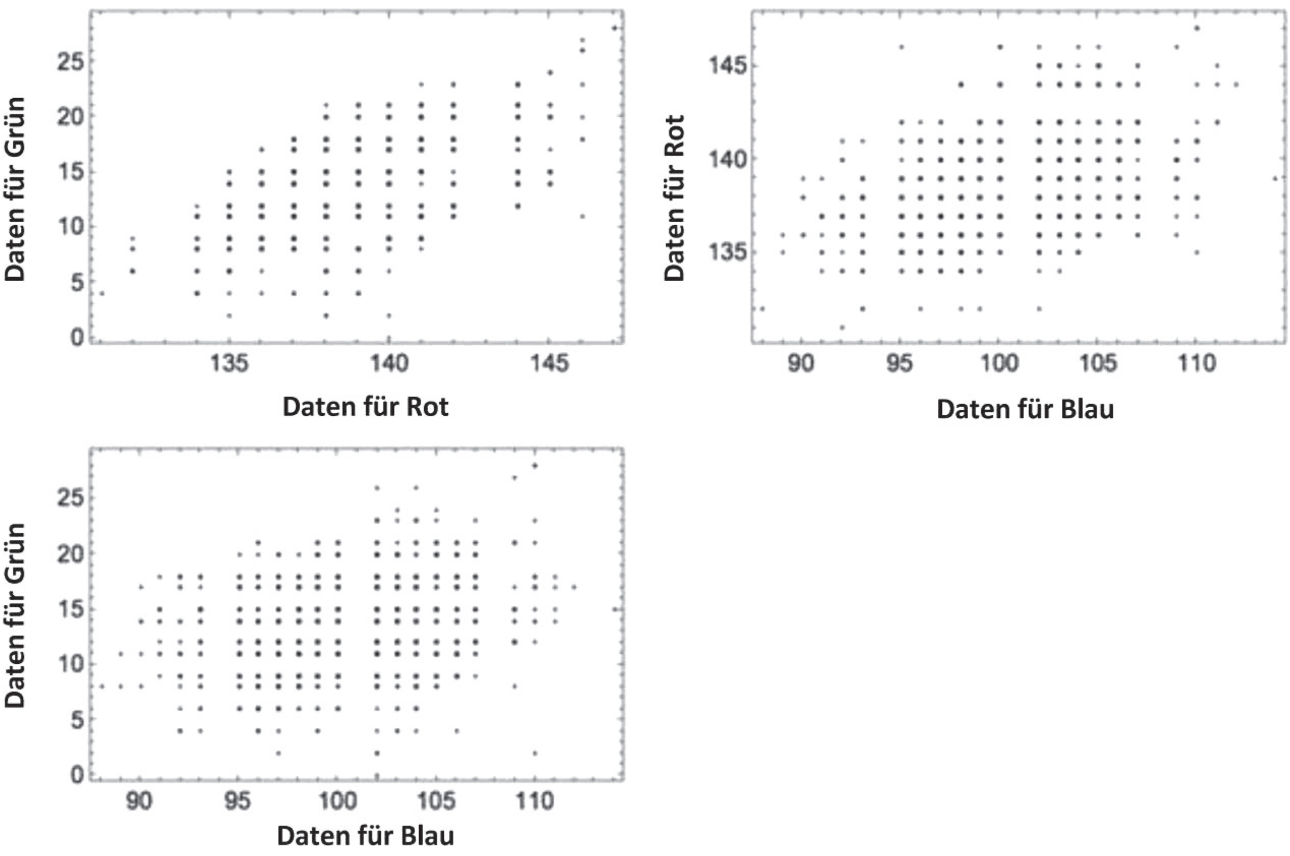

Abb. 5: Punktwolken zu allen Paaren von Farbkanälen

Die Überlegungen oben können auch auf andere Arten von Anpassungstests wie etwa dem Chi-Quadrat-Anpassungstest übertragen werden und für Schülerinnen und Schüler den Horizont bezogen auf p-Werte und Hypothesentests erweitern. Mathematica umfasst verschiedene Anpassungstests, die für Lernende wie Lehrende interessant sein können. Da diese Tests in das Programm implementiert sind, umgeht man Schwierigkeit des Programmierens oder Erstellens von Tabellen. Man kann sich vielmehr voll auf den Prozess des Hypothesentestens konzentrieren.

Zu guter Letzt kann es eine interessante Anfangsüberlegung sein, was die Werte des Korrelationskoeffizienten in den Punktwolken bedeuten. Diese Koeffizienten sind 0,560 ($p < 0{,}0001$) für den roten und grünen Farbkanal, 0,422 ($p < 0{,}0001$) für den roten und blauen Farbkanal und 0,253 ($p < 0{,}0001$) für den grünen und blauen Farbkanal, wobei sich alle Koeffizienten statistisch signifikant von 0 unterscheiden. Wiederum kann die Lehrkraft thematisieren, dass man praktisch alle Nullhypothesen verwerfen kann sofern der Datensatz nur groß genug ist und dabei das Verhältnis von praktischer Signifikanz (Effektstärke) und statistischer Signifikanz beleuchten.

Zusammenfassung

Verwendet man einen Camcorder in Verbindung mit einem Rechner, so ist es möglich traditionelle Wege der Datenerhebung gewinnbringend zu ersetzen. Insbesondere erlaubt dieser Weg, dass Schülerinnen und Schüler mit einem großen, realen Datensatz Erfahrungen machen. Außerdem kann die Zeit, die man bei der reinen Datenerhebung spart, genutzt werden, um Fertigkeiten wie etwa die Arbeit mit Datensätzen mit Technologieunterstützung (etwa durch Mathematica) zu fördern. Daten können so auf der Basis verschiedener grafischer Darstellungen untersucht und diskutiert werden, da diese Darstellungen mit dem Rechner ungeheuer einfach werden. Für den Einstieg reicht es dabei, auf die Daten und deren Interpretation zu fokussieren, ohne technische Details zu betrachten. Weiter ermöglicht unser Weg es Schülerinnen und Schülern, über die Qualität von Daten nachzudenken. Eine Erweiterung dazu könnte etwa der Vergleich von Camcordern im High-End-Bereich und digitalen Spiegelreflex-Kameras sein. In diesem Beitrag wollten wir allerdings zeigen, was ohne größere Kosten und Aufwand alles möglich ist.

Literatur

Levine, J.; Roos, T. B. (1998): http://www.math.dartmouth.edu/~matc/DataAnalysis/DataLabs.html [Letzter Zugriff: 2014].

Irie, K.; McKinnon, A. E.; Unsworth, K.; Woodhead, I. M. (2008): A technique for evaluation of CCD video-camera noise. IEEE Transactions on Circuits and Systems for Video Technology, 18(2), S. 280–284.

Hogg, R. V.; Tanis, E. A. (2010): Probability and Statistical Inference. London: Pearson Prentice Hall.

Shapiro, S. S.; Wilk, M. B. (1965): An analysis of variance test for normality (complete samples). *Biometrika, 52*, 591–611.

Anschrift der Verfasser

Christopher W. Kulp
Lycoming College,
Williamsport,
PA USA
Kulp@lycoming.edu

Gene D. Sprechini
Lycoming College,
Williamsport,
PA USA
Sprgene@lycoming.edu