Statistik
für Anfänger

Entwurf

von
Norbert Kessel
www.verlagkessel.de

Erster Teil, Einführung

Eine Einführung mit Kartoffeln, Taschenrechner und R

Viele Schüler und Studenten kommen irgendwann einmal mit Statistik in Berührung. Das sorgt nicht immer für Freude, denn zum einen sind die Methoden der Statistik oft schwer zu verstehen, zum anderen ist das Datenmaterial häufig uninteressant: immer wieder werden Daten aus Fragebögen ausgewertet, entstanden meist bei Umfragen, das ist zwar manchmal informativ, aber spannend ist es selten.

In diesem Buch wird ein anderer Ansatz verfolgt: einfache Daten, die man entweder selbst erheben oder aus dem Internet herunterladen kann, sollen auf den Stoff neugierig machen. Was kann man aus Daten schlussfolgern? Was ist der Unteschied zwischen einer Vermutung und einem statistisch signifikanten Ergebnis?

Ohne Neugier gäbe es keine Wissenschaft, und niemand hat behauptet, dass das langweilig sein muss.

Wieso Kartoffeln?

Diese Einführung wird sich mit Kartoffeln beschäftigen, um anhand des Gewichtes der einzelnen Kartoffeln ein paar Fragen zu beantworten, die wir uns alle schon einmal gestellt haben (naja, mehr oder weniger):

- Was wiegt eine Kartoffel?
- Wieviele Kartoffeln sind in einer Packung?
- Wie schwer ist die leichteste, wie schwer die schwerste Kartoffel?
- Was ergibt ein Vergleich zwischen Bio- und herkömmlich erzeugten Kartoffeln?

Falls Sie sich diese Fragen noch nie gestellt hatten, dann tun Sie einfach so, als würde es Sie interessieren. Falls das nicht möglich sein sollte, dann stellen Sie sich einfach etwas anderes vor: anstelle des Gewichts einzelner Kartoffel können Sie sich auch den Umfang von Bäumen oder die Körpergröße von neu eingeschulten Kindern vorstellen. Auch industriell hergestellte Gegenstände können Sie betrachten: die Länge von Streichhölzern oder das Gewicht von einzelnen Schrauben, den Durchmesser von Röhren, die Füllhöhe von Biergläsern, überall lässt sich irgendwas messen. Ansonsten können Sie ja eine Umfrage machen.

Kartoffeln

Taschenrechner

Mit dem Taschenrechner rechnen, wenn es Computer gibt, wieso tun wir uns sowas an? Weil man ein Gefühl für die Daten bekommt und sich danach umso mehr freut, wenn man mit dem kleinen Gerät die gleichen Ergebnisse erhält, wie mit dem großen Computer.

Das bedeutet aber auch: bleiben Sie misstrauisch, denn der Computer findet immer irgendein Ergebnis, aber zu prüfen, ob das sinnvoll ist, das bleibt dem Menschen überlassen: Computer können Unsinn nicht erkennen. Und es schadet nicht, wenn man bei komplexten Auswertungen auf dem Computer gelegentlich den Taschenrechner mit Papier und Bleistift benutzt, um ein Ergebnis zu überprüfen.

R

R ist ein Computerprogramm, das man gratis aus dem Internet herunterladen kann. Für alle, die schon mal Geld für Software ausgegeben haben (SPSS, Statgraph, PlotIt, SAS, Statistica) ist es kaum zu glauben, was man da umsonst bekommen kann. Und dass es so einfach zu bedienen ist, ist geradezu ein Märchen. Im Anhang ist beschrieben, wo man es herunterladen kann und welche Pakete man installieren sollte.

Stichprobe und Grundgesamtheit

Meistens interessiert man sich für Eigenschaften der Grundgesamtheit, das wäre – beim Kartoffel-Beispiel – die Menge aller Kartoffeln, die von allen Bauern in Deutschland in einem Jahr geerntet werden. Aber es ist gar nicht möglich, so eine Auswertung zu machen. Deshalb hilft man sich mit einer Stichprobe. Da man nicht unbedingt zum Bauer auf das Feld möchte, wenn der die Kartoffeln aus dem Boden holt, kauft man sich beispielsweise ein 2,5 kg-Paket im Laden.

Man versucht anhand dieser doch relativ kleinen Menge auf die dahinter liegende Menge aller Kartoffeln zu schließen. Ob das möglich und sinnvoll ist, werden die ersten Seiten im Buch zeigen. Dabei leuchtet es sofort ein, dass die Größe der Stichprobe von großer Bedeutung ist: man kann aus fünf Kartoffeln keine vernünftigen Schlüsse ziehen, 30 Stück sind schon besser, 100 sind noch besser.

Wenn Sie im Radio oder in der Zeitung von Umfagen hören, dann sollten Sie die Ohren spitzen: bei den wenigsten Umfragen wird mitgeteilt, wieviele Menschen befragt wurden. Das ist aber extrem wichtig! Stellen Sie sich vor, Sie befragen 20 Menschen tagsüber in der Fußgängerzone nach Ihrer Meinung zu einem bestimmten Thema. Das kann nicht gut gehen, denn alle berufstätigen Menschen werden sich um diese Zeit gar nicht dort aufhalten. Genauso werden Sie vor einem Fabriktor immer andere Ergebnisse erhalten, als in einer Fußgängerzone. Also ist die Auswahl (wie eine Stichprobe erhoben wird) von großer Bedeutung.

Mittelwerte und Streuung

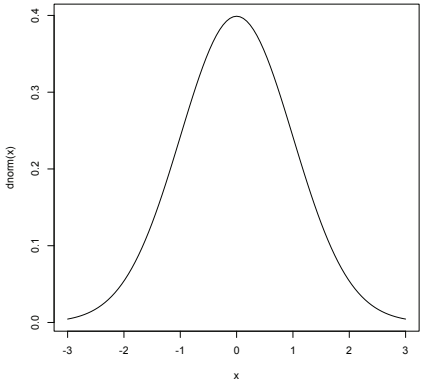
Die ersten Abschnitte im Buch zeigen, wie man Daten beschreiben kann, hierzu werden verschiedene *Mittelwerte* und *Streuungsmaße* berechnet. Das ist im Grun-

de immer der Anfang jeder Auswertung. Danach lassen lassen sich mehrere Stichproben miteinander vergleichen, zum Beispiel Bio-Kartoffeln mit herkömmlich erzeugten Kartoffeln.

Die Normalverteilung

Bei allem, was auf Feldern wächst oder in Fabriken hergestellt wird, gibt es Streuung. Selbst wir erzeugen Streuung, wenn wir z.B. nacheinander 20 Linien mit einer Länge von 5 cm zeichnen. Da ist vermutlich keine einzige genau 5 cm lang, irgendwie immer zu kurz oder zu lang werden sie sein.

Im Normalfall streuen die Werte um den arithmetischen Mittelwert in der Form, dass es nur wenige ganz kleine und nur wenige ganz große (Kartoffeln) gibt, die meisten finden sich in der Mitte. Würde man Brötchen oder Brote wiegen, die in Bäckereien hergestellt werden, dann würde man das dort ebenfalls entdecken. Die folgende Abbildung zeigt das.



Eine Normalverteilung, erstellt mit R
(die einzige dazu nötige Anweisung in R lautet: ‚curve(dnorm, -3, 3)‘

Ist das nicht eine schöne Kurve? Carl Friedrich Gauß hat das dankenswerterweise zu Papier gebracht (im Jahr 1809) und dazu auch eine Funktionsgleichung entwickelt (die uns aber hier noch nicht interessiert).

Allerdings hat sich Gauß nicht mit Kartoffeln beschäftigt, sondern mit Planetenbahnen, und seine Ankündigung, dass der Zwergplanet Ceres zu einem bestimmten, von Gauß berechneten Zeitpunkt erneut am Himmel erscheinen werde, hat den jungen Mann mit einem Schlag weltberühmt gemacht (der Titel des Buches ist: „Theorie der Bewegung der Himmelskörper ...“). Natürlich ging die Welt damals nicht unter, wie viele andere vermutet hatten, als Ceres erstmals erschien und plötzlich wieder verschwand. Zum Dank kam Gauß auf den alten 10-DM-Schein in Deutschland. Eigentlich hätte er eine wertvollere Banknote verdient.

Und warum ist diese Normalverteilung wichtig? Zum einen können wir herausfinden, ob die uns vorliegenden Werte überhaupt normalverteilt sind (hierzu folgt später ein interessantes Beispiel mit Brötchen, die von einem misstrauischen Kun-

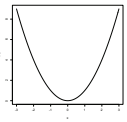
Einschub: was ist eine Funktion?

Viele kennen den Begriff ‚Funktion' aus dem Mathematikunterricht in der Schule, dort war er ungefähr so definiert: eine Funktion wird durch ihren Namen aufgerufen und ausgeführt. Beim Aufruf übergibt man der Funktion eine Zahl, die Funktion liefert ein Ergebnis zurück (die beim Aufruf übergebene Zahl wird auch Parameter genannt).

Zum Beispiel die Funktion $y = x^2$: man übergibt ihr eine Zahl (3), sie liefert ein Ergebnis zurück (9). Man kann die Ergebnisse in mit einer kleinen Tabelle anschaulich machen:

x	-3	-2	-1	0	1	2	3
$y=x^2$	9	4	1	0	1	4	9

Zeichnet man die Datenpaare auf, erhält man eine Parabel.



Zum Taschenrechner: dort befindet sich eine Taste mit dem Ausdruck ‚x^2'. Gibt man eine Zahl ein und drückt anschließend diese Taste, wird ein kleines Computerprogramm ausgeführt, das die Quadratzahl berechnet und anschließend anzeigt (der Taschenrechner hat *keine* Tabelle, aus der er das Ergebnis ablesen könnte).

In der Informatik ist der Begriff Funktion etwas anders definiert: die dort verwendeten Funktionen können auch mehr als einen Wert zurückliefern oder aber sie tun etwas im Verborgenen, zum Beispiel das Laden von Daten in den Hauptspeicher (denn erst nachdem Daten von der Festplatte oder von der Tastatur eingelesen und im Hauptspeicher stehen, kann mit ihnen gearbeitet werden, das gilt auch für die Textverarbeitung).

In R gibt es sehr viele Funktionen, einige werden hier im Buch oft verwendet, sie sind in der folgenden Übersicht zusammengefasst:

Funktion in R	liefert	Beispiel	Ergebnis
min()	die kleinste Zahl	min(1,2,3)	1
max()	die größte Zahl	max(1,2,3)	3
range()	Spannweite	range(1,2,3)	1 3
c()	kopiert Daten in eine Variable (hier mit dem Namen ‚meinedaten')	meinedaten<-c(1,2,3)	(es wird nichts weiter angezeigt, aber im Hauptspeicher ist nun ein Teil mit den Daten gefüllt)
mean()	arithmetischer Mittelwert	mean(meinedaten)	2
sd()	Standardabweichung	sd(meinedaten)	1
hist()	zeichnet ein Histogramm	hist(meinedaten)	Histogramm
curve()	eine Zeichnung, im Beispiel eine Parabel	curve(x^2, from=-3, to=3)	Parabel

den gewogen wurden). Zum anderen verlangen viele Tests, die wir später berechnen, dass die Daten normalverteilt sind; sind sie es nicht, dann muss man mit anderen Tests rechnen, die aber oft weniger aussagekräftig sind. Außerdem kann man – sofern Normalverteilung vorliegt – angeben, wieviele Messwerte innerhalb bestimmter Wertebereiche liegen, das ist bei der industriellen Produktion enorm wichtig.

Auch bei der Herstellung von Lebensmitteln wird darauf geachtet, dass die verarbeiteten Rohstoffe (Tomaten, Spargel, Kartoffeln, Blumenkohl) nicht zu sehr streuen, genauer: dass die Größen nicht zu sehr vom arithmetischen Mittelwert abweichen.

Beschreibende Statistik ...

Im ersten Teil dieses Büchleins werden Mittelwerte und Streuungswerte ermittelt: das ist das, was man am Anfang einer Auswertung immer macht, man erhält damit einen ersten Eindruck. Der schon genannte arithmetische Mittelwert ist sicher der bekannteste (jeder, der in der Schule mal den Durchschnitt einer Klassenarbeit berechnet hat, kennt ihn). Es gibt aber auch noch andere Mittelwerte, die einem helfen, die Daten zu verstehen.

Von großer Bedeutung sind die Streuungswerte, die man sich ganz einfach mit zwei Schulkindern vorstellen kann: das eine Kind schreibt in zwei Klausuren die Noten 1 und 5. Der Mittelwert ist

$$(1 + 5) / 2 = 3$$

Ein anderes Kind schreibt zweimal die Note 3, auch hier ist der Mittelwert 3. Beim ersten Kind streuen die Werte sehr, beim zweiten Kind streuen sie überhaupt nicht.

Allerdings ist das Beispiel mit den Schulnoten – so eingängig und leicht verständlich es auch ist – leider äußerst schlecht. In einem folgenden Kapitel wird über die „Skalierung von Werten" berichtet und da wird es sich herausstellen, dass Schulnoten „ordinal skaliert" sind, im Grunde eine willkürliche Zuweisung von Zahlen an eine Leistung (sehr gut = 1, gut = 2 usw.). Und eigentlich darf man mit solchen Daten überhaupt keine Mittelwerte berechnen – aber die ganze Welt tut es.

Übertragen auf die Kartoffel könnten wir feststellen: wenn der Sack 25 Kartoffeln enthält, dann gibt es typischerweise darin (wenige) ganz kleine und (wenige) ganz große Kartoffeln die meisten sind ungefähr so schwer, wie die ‚mittlere' Kartoffel, also die, die durch den arithmetischen Mittelwert berechnet wurde (insofern ist sie also eine ‚theoretische' Kartoffel, die nicht unbedingt im Sack enthalten sein muss).

... und schließende Statistik

Typischerweise hat man eine Vermutung, wie zum Beispiel: sind die Kartoffeln vom Bio-Bauer kleiner als die herkömmlich erzeugten? Deswegen sammelt man Daten. Liegen die dann vor, kann man zur Überprüfung der Vermutung sogenannte Tests rechnen. Abhängig von den Test-Ergbnissen kann man dann *fundierte* Aussagen machen, die viel sicherer sind als bloße Vermutungen.

Über Skalen

Skalen begegnen uns an vielen Orten: Fieberthermometer, Waage, Lautstärkeregler, Kaffeemaschine aber auch Schulnoten folgen verschiedenen Skalen, wie im Text schon erwähnt (sehr gut = 1 ...). Im Grunde geht es ja darum, etwas zu messen und die Werte anschließend zu vergleichen. Nun gibt es Skalen, die sind für statistische Aussagen ideal (Zollstock oder Schieblehre zur Längen- oder Durchmessermessung); werden auf diese Art Messwerte gewonnen, dann steht das komplette Instrumentarium der Statistik zur Verfügung, um die Daetn auszuwerten und letztendlich fundierte Aussagen machen zu können. Andere Messwerte dagegen stammen von Skalen, mit denennicht alle Tests gerechnet werden dürfen. Man unterscheidet üblicherweise vier Skalen, sie haben besondere Eigenschaften (welche Tests später gerechnet werden dürfen).

Die Computer können eine Postleitzahl nicht von einem codierten Schlüssel oder dem Gewicht eines Versuchsobjekts unterscheiden, deshalb werden manchmal unsinnige Ergebnisse angezeigt.

Zu den Skalen:

- Die *schwächste* Skala ist die Nominalskala, in ihr werden Merkmale beschrieben, anschließend wird gezählt und festgelegt, wieviele Elemente in eine Gruppe (oder Klasse) gehören.
- Die *stärkste* Skala ist die Verhältnisskala, sie hat alles, was man für Auswertungen braucht: einen ‚richtigen' Nullpunkt (der nicht willkürlich festgelegt wurde, so wie bei unserem Thermometer mit Celsius-Skala) und Messungen, die alle Auswertungen gestatten.

Hier die Einteilung der Skalen

Art	Erläuterung, Beispiel
Nominalskala	eine Bezeichnung (oder Ziffer), zum Beispiel für das Geschlecht einer Person (weiblich=1, männlich=2). Man kann mit solchen Daten keinen arithmetischen Mittelwert berechnen. Es können nur Häufigkeiten angegeben werden (12 Damen, 8 Herren) und man kann Aussagen darüber machen, wovon es mehr oder weniger gibt. Manchmal ist es verwirrend, denn es gibt Werte, die nach einer Messung ausschauen, aber in Wahrheit nur willkürlich festgelegte ‚Schlüssel' sind: Postleitzahlen zum Beispiel oder Codierungen, zum Beispiel für Haarfarben: 1=schwarz, 2=blond). Auch eine Sortierung ist nicht zulässig. Die beim Würfeln erzeugten Daten sind nominal skaliert, ebenso die Häufigkeiten bei der Vererbung von Merkmalen (Mendel, Beispiele folgen). Und schließlich sind die Ergebnisse von Umfragen häufig nominal skaliert.
Ordinalskala	Schulnoten: sie erlauben nur Aussagen darüber, was größer oder kleiner ist (bzw. besser oder schlechter). Anders als Schulnoten suggerieren, sind die Abstände zwischen den einzelnen Schulnoten (von 1 zu 2 und von 5 zu 6) nicht festgelegt.
Intervallskala	Datumswerte (Jahreszahlen usw.), Grad Kelvin (nicht: Grad Celsius),
Verhältnisskala	Hier nun endlich eine Skala mit Nullpunkt: Das Alter von Menschen, Tieren oder Pflanzen, der Durchmesser eines Baumes, das Gewicht einer Kartoffel. Hat man zur Auswertung solche Daten, dann kann das komplette Instrumentarium zur Versuchsauswertung verwendet werden.

Man stellt also eine Frage, berechnet mit den Daten einige Kenngrößen und entscheidet dann, wie die Frage beantwortet werden kann. Das kann durchaus spannend sein, wie die folgenden Fälle zeigen:

- In einem späteren Beispiel geht es zum Beispiel um die Frage, welche Volksgruppe in den USA mehr Alkohol trinkt als die anderen. Hierzu wurden um das Jahr 1980 herum Menschen mit einem Fragebogen befragt, die darüber Auskunft gegeben haben. Eine Vermutung hat jeder, der sich mit diesem Thema beschäftigt, aber mit Hilfe eines Tests kann man die Frage nach den unterschiedlichen Trinkgewohnheiten *eindeutig* beantworten, die Ergebnisse sind dann gesichert (oder ‚signifikant'); ohne diese Tests sind Aussagen wie zum Beispiel: ‚die Iren in den USA trinken am meisten' nicht sinnvoll, nicht fundiert und gehören somit eher in das Reich der Märchen.
- Auch die immer wieder zitierte höhere Herzinfarkt-Rate von Männern an Montagen, kann man mit einem Test überprüfen. So kann man die Aussage: ‚an Montagen bekommen Männer häufiger einen Infarkt' untermauern oder verwerfen.
- Schließlich die Frage: werden bei Vollmond tatsächlich mehr Kinder geboren als an anderen Tagen, was immer wieder zu hören ist. Auch diese Frage lässt sich mit einem Test beantworten.

Kartoffeln – ein erster Kontakt

Für dieses einführende Kapitel haben wir drei Sorten Kartoffel gekauft:

- Bio-Kartoffeln aus dem Laden (1,5 kg)
- herkömmliche Kartoffeln aus dem Laden (2,5 kg)
- herkömmliche Kartoffel direkt vom Bauer (10 kg)

Die Fragen, die uns dazu gebracht haben sind hier genannt:

- sind Bio-Kartoffeln kleiner als herkömmliche Kartoffeln?
- sind die Kartoffeln vom Bauern größer oder kleiner als die Supermarkt-Kartoffeln?

Hier könnten natürlich weitere Fragen folgen (Inhaltsstoffe der Kartoffel, Rückstände von Pestiziden, Fungiziden usw.), aber zur Beantwortung solcher Fragen muss man immer zuerst Daten erheben, für die man oft ein Labor braucht; für unsere Zwecke reichte eine einfache Waage, auf die wir die Kartoffeln gelegt und gewogen haben. Notiert wurde zunächst von Hand: Papier und Bleistift sind an dieser Stelle unschlagbar, die Originaldaten sollten auf jeden Fall aufbewahrt werden.

Die Rohdaten (alle Zahlen in Gramm)

Bio-Kartoffel (1,5 kg)	normale Kartoffel (2,5 kg)	vom Bauer gekauft (10 kg)
99, 87, 35, 59, 47, 26, 59, 40, 41, 53, 45, 139, 53, 113, 50, 58, 62, 55, 40, 107, 54, 48, 40, 139	116, 99, 88, 109, 114, 39, 50, 46, 99, 111, 89, 58, 141, 83, 92, 50, 113, 74, 80, 66, 72, 105, 116, 100, 79, 92, 142, 68, 47	182, 164, 134, 106, 99, 124, 124, 201, 62, 39, 58, 35, 47, 42, 132, 127, 123, 79, 131, 155, 109, 94, 118, 121, 62, 110, 60, 156, 196, 105, 70, 65, 122, 110, 123, 102, 117, 115, 265, 110, 83, 104, 80, 170, 165, 122, 95, 32, 92, 108, 52, 158, 119, 162, 85, 111, 193, 79, 61, 112, 15, 117, 129, 66, 144, 152, 147, 116, 66, 148, 132, 96, 69, 75, 62, 73, 44, 185, 75, 122, 67, 103, 183, 166, 108, 92, 108, 147, 100, 62, 66, 135, 175, 99, 45, 87, 110, 135, 86, 32, 20, 65, 63, 41, 51, 47, 173, 77, 26

Man kann bereits mit diesen Zahlen eine Reihe von Informationen gewinnen:

- Anzahl der Kartoffeln im Sack
- kleinster und größter Wert
- Spannweite

	Bio-Kartoffeln	normale Kartoffeln	10 kg-Packung
Anzahl Kartoffeln	24	29	109
kleinster Wert	26	39	15
größter Wert	139	141	265
Spannweite (größter - kleinster Wert)	113	102	250

Anmerkungen: die Spannweite (engl. Range, abgekürzt R) wird errechnet, indem man vom größten Wert den kleinsten Wert abzieht. Sie ist ein einfach zu errechnender Wert, der gelegentlich benutzt wird, wenn nur wenige Messwerte vorliegen. In letzter Zeit taucht dieser Wert bei Gehältern auf, wenn darüber diskutiert wird, um wieviel höher ein Managergehalt sein darf, verglichen mit dem Gehalt eines ‚einfachen' Arbeiters.
Die kleinste hier im Haus angetroffene Kartoffel hatte übrigens ein Gewicht von 6 g und einen Durchmesser von ca. 2 cm. So kleine Kartoffeln kamen in den drei eingekauften Packungen nicht vor!

Um die Daten zu beschreiben und um später Aussagen zu ermöglichen wie z.B.: ‚Biokartoffeln sind kleiner als herkömmlich erzeugte Kartoffeln' macht man folgendes:

- Man berechnet die Mittelwerte;
- man berechnet die Streuung der Einzelwerte um den arithmetischen Mittelwert;
- man rechnet einen Test, um herauszufinden, ob Unterschiede zwischen den Mittelwerten zufällig entstanden sind oder ob sie so deutlich sind, dass wir von einem signifikanten Unterschied sprechen können.

Mittelwerte

Typischerweise berechnet man immer drei Mittelwerte, um Daten zu beschreiben:

- arithmetischer Mittelwert (der wichtigste);
- Median;
- Modus.

Die Berechnung des arithmetischen Mittelwertes

Er wird berechnet, indem man die einzelnen Werte addiert und durch die Anzahl der Werte dividiert, man erhält hier dadurch eine Art ‚mittlere' Kartoffel, die man dann mit den anderen vergleichen kann.

In der folgenden Übersicht sind die Anzahl der Kartoffeln, die Summe der Einzelwerte und der daraus berechnete arithmetisch Mittelwert enthalten:

	Bio-Kartoffeln	normale Kartoffeln	10 kg-vom Bauer
Anzahl Kartoffeln	24	29	109
Summe der Einzelwerte	1549	2538	11379
arithmetischer Mittelwert	64,5	87,5	104,4

Halten wir fest: der Mittelwert der Bio-Kartoffeln ist der kleinste, dann folgen die herkömmlich erzeugten Kartoffel und am schwersten sind die aus der 10 kg-Packung, die direkt beim Bauer gekauft wurden.

Eine Frage, die sofort gestellt werden kann: Sind die Unterschiede zufällig oder sind sie so deutlich, dass wir von einem gesicherten Unterschied sprechen können? Mit den bisher berechneten Werten ist das noch nicht zu beantworten.

Obwohl die Berechnung mit nur wenigen Daten in einer Minute erledigt ist, kann es schon mühselig und ungenau sein, diese Berechnung mit dem Taschenrechner zu machen. Aber: lässt man eine andere Person die gleichen Zahlen addieren, kommt es leicht zu unteschiedlichen Ergebnissen. Deshalb ist es besser, Daten in einer separaten Datei zu speichern, so dass man jederzeit die Daten ausdrucken und kontrollieren kann. Auch das Speichern in Spreadsheets aus der Tabellenkalkulation (OpenOffice/Excel) ist eine gute Idee, große Datenmengen dagegen sollten in einer Datenbank-Tabelle gespeichert werden. Dies wird in einem späteren Kapitel erläutert.
Noch ein Hinweis zu Tabellenkalkulationen: obwohl auch bei denen Funktionen zur Statistik enthalten sind, sollte man sie nicht verwenden; schon vor über 20 Jahren war es so, dass in der pharmazeutischen Forschung Excel (und sein Vorgänger Multiplan) wegen fehlerhaften Berechnungen nicht zur Auswertung von Versuchsdaten verwendet werden durften und es scheint so, dass seriöse Auswertungen immer ohne die genannten Produkte gemacht werden.

Die allgemeine Formel zur Berechnung des arithmetischen Mittelwertes

Die Summe der Einzelwerte, dividiert durch die Anzahl der Einzelwerte ergibt den arithmetischen Mittelwert, hier beispielhaft für die erste Spalte angezeigt.

$$\frac{\text{Summe der Einzelwerte}}{\text{Anzahl der Einzelwerte}} = \frac{1549}{24} = 64{,}5$$

Hier die mathematische Formel für den arithmetischen Mittelwert, dabei steht $\bar{x}$ für den Mittelwert aller Werte und n für die Anzahl der Werte. Man kann die ein-

zelnen Werte auch mit x_1, x_2 bezeichnen und aufsummieren, aber am einfachsten geht es, wenn man dazu das Summenzeichen ∑ benutzt:

$$x = \frac{\text{Summe der Einzelwerte}}{\text{Anzahl der Einzelwerte}} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x}{n} = \frac{1549}{24} = 64{,}5$$

Sigma (∑) kommt aus dem Griechischen Alphabet, es ist dort das ‚große S' und wird auf der ganzen Welt für die Summe verwendet, auch in den üblichen Tabellenkalkulationsprogrammen.

Die vorstehene Formel ist im Grunde harmlos, sie wird in einer allgemeiner formulierten Variante auch nicht gefährlicher, dort steht nur zusätzlich dabei, dass beginnend vom ersten Wert ($i=1$) bis zum letzten Wert (n) alles summiert werden soll:

$$\overline{x} = \frac{\sum_{i=1}^{n} x}{n}$$

Die Berechnung des Median

Als Median bezeichnet man den Wert, der die vorhandene (sortierte) Menge aller Werte in zwei gleich große Hälften teilt. Ein Blick auf Ihre Hand zeigt Ihnen, dass es bei fünf Fingern genau einen Finger gibt, der in der Mitte steht, das ist bei einer ungeraden Anzahl von Dingen immer so.

Hier also die Daten der normalen Kartoffeln (29 Messwerte), nach dem Gewicht sortiert. Der 15. Messwert teilt die Daten in zwei gleichgroße Hälften, er ist der Median (hier unterstrichen)

39 46 47 50 50 58 66 68 72 74 79 80 83 88 89 92 92 99 99 100 105 109 111 113 114 116 116 141 142

Bei einer geraden Anzahl Messwerte (Bio-Kartoffel) gibt es keinen mittleren Wert, der zwei Hälften entstehen lässt. Deshalb sortiert man die Werte nach der Größe und nimmt die *beiden Werte aus der Mitte*, summiert sie und teilt sie anschließend durch 2. Die beiden mittleren Werte sind in der folgenden Datenreihe hervorgehoben:

26 35 40 40 40 41 45 47 48 50 53 54 55 58 59 59 62 87 99 107 113 139 139 153

Man summiert diese beiden Werte und teilt das Ergebnis durch 2 und erhält den Median:

(54+55) / 2 = 54,5

Die Berechnung des Modus

Wie zuvor schon erwähnt, haben typischerweise keine zwei Kartoffeln das gleiche Gewicht, deshalb muss man bei solchen Messwerten die Daten in Klassen sortieren, zum Beispiel die Klasse von 0-29 g, 30-59 g und so weiter. Dann ermittelt man, wieviele Kartoffeln in die jeweilige Klasse fallen. Der Modus ist dann die Klasse, in der die meisten Werte liegen. Etwas weiter unten folgt ein Beispiel mit Werten, die auf Klassen aufgeteilt werden.

Es gibt ja auch noch andere Daten, zum Beispiel die, die man mit einem Fragebogen gewinnt. Dort ist der Modus die am häufigsten genannte Antwort auf eine bestimmte Frage.

Die Streuung der Werte

Wenn wir Dinge miteinander vergleichen, spielen die Mittelwerte eine herausragende Rolle. Wollen wir z.B. über die Wirksamkeit eines Medikaments auf den Blutdruck eines Menschen oder über die Wirkung einer Pflanzendüngung auf das Wachstum von Getreide urteilen, dann machen wir als erstes einen Versuchsplan, behandeln entsprechend und messen am Ende die Werte.

Auf der Basis dieser Werte berechnen wir anschließend Mittelwerte (das ist auf den voranstehenden Seiten schon passiert) und die Streuung. Letztendlich hängt von ihr ab, ob man die Unterschiede nach einer Behandlung als *zufällig* oder als *signifikant* bezeichnen kann.

Hustensaft und seine Wirksamkeit könnte ein deutliches Beispiel dafür sein, dass es ‚Medikamente' gibt, die nicht wirksam sind (und die dann eigentlich gar nicht als Medizin bezeichnet werden dürften). Aber Menschen glauben oft an Dinge, und offensichtlich reicht das manchmal schon aus. Bei selbst hergestellten Hustensäften unter Verwendung von Kräutern ist es bestimmt so, dass diese wirken, aber eigentlich müsste man zum Beweis dieser Wirksamkeit einen Versuch machen, bei dem auch Menschen einbezogen werden, die skeptisch gegenüber diesen Dingen sind.

Die Streuungsmaße haben die folgenden Namen:

- Varianz (s^2);
- Standardabweichung (s);
- Variationskoeffizient (VK).

Die drei Streuungsmaße lassen sich leicht aus der Varianz berechnen. So ist die Standardabweichung die Quadratwurzel aus der Varianz und der Variationskoeffizient ist die Varianz dividiert durch den Mittelwert, wie das geht, wird nachfolgend erklärt.

Liegen die einzelnen Messwerte weit weg vom artithmetischen Mittelwert, so ist mit einer großen Standardabweichung zu rechnen, streuen sie nur wenig um den arithmetischen Mittelwert, ist die Standardabweichung gering, im Extremfall (bei immer gleichen Messwerten) wäre sie 0.

Ein paar Überlegungen dazu

Zwei kurze Datenreihen werden miteinander verglichen, sie bestehen aus drei Zahlen:

die erste Datenreihe

10, 20, 30

die zweite Datenreihe

20, 20, 20

Bei beiden Datenreihen ergibt sich ein arithmetischen Mittelwert von

20

Es ist offensichtlich, in welcher Datenreihe die Werte mehr streuen, nämlich in der ersten, bei der zweiten Datenreihe dagegen ist die Streuung 0.

Aber wie groß ist die Standardabweichung in der ersten Datenreihe? Sie ist 10 g. Wenn wir uns kurz vorstellen, dass um Gewichte von Kartoffeln geht, dann bedeutet das, dass die mittlere Abweichung vom Mittelwert 10 g ist. Die Einheit ist hier ‚g', wie beim Mittelwert.

Natürlich wird es nie wieder so einfach sein wie hier, aber das Prinzip ist immer das gleiche: man berechnet die Differenzen zwischen dem arithmetischen Mittelwert und den Einzelwerten. Da hierbei positive und negative Vorzeichen auftreten (+10, -10), quadriert man die Abweichungen (dabei verschwinden die negativen Vorzeichen) und schließlich berechnet man daraus die Summe. Das Ganze bezeichnet man als ‚Summe der Abweichungsquadrate'). Auf den folgenden Seiten folgt dazu ein realistisches Beispiel.

Wie macht man Streuung sichtbar? Und vergleichbar?

Streuung können wir mit der Taschenlampe sichtbar machen, die wir in die Nacht oder in den Nebel richten. Aber dort verteilt sich das Licht gleichmäßig um die Mitte, wohingegen das bei Daten nur manchmal so ist. Meist gibt es eine Häufung von Werten um das Zentrum, um den arithmetischen Mittelwert; weit davon weg liegende Werte (extreme Werte) kommen dagegen weniger häufig vor.

Wir wissen, dass ein Teil der Werte kleiner und ein anderer Teil größer als der arithemtische Mittelwert ist. Der erste Versuch könnte ein Zahlenstrahl sein (also im Grunde eine x-Achse), auf der man neben dem Mittelwert die Messwerte mit einem Kreuz markiert.

Das gibt eine gewisse Ahnung, aber ist doch recht unklar bei eng zusammenliegenden Werten.

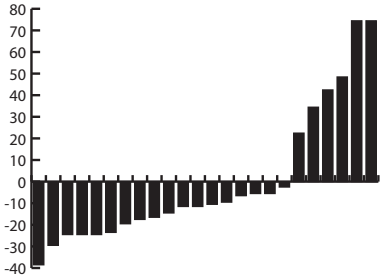
Man könnte es in anderer Form grafisch darstellen, indem man den Mittelwert als eine Gerade zeichnet und die einzelnen Messwerte rechts bzw. links davon (manche nutzen hierzu ein Balkendiagramm):

Aber selbst bei nur wenigen Daten wird dies schnell unübersichtlich, wie die vorstehende Zeichnung zeigt.

Eine andere Art der Darstelltung findet sich in älteren Büchern, dort ist anstelle von Kreuzen die Häufigkeit von Messwerten angegeben, so z.B. eine ‚1' für einen einzelnen Messwert und eine ‚2' für zwei eng beieinander liegende Messwerte (und entsprechend ‚3' für drei Messwerte und so weiter).

Hat man nur eine Datenreihe (und nur wenige Messwerte), dann ist das durchaus verwendbar, denn Mittelwert und Streuung lassen sich damit erklären. Aber bei mehreren Datenreihen oder vielen Messwerten wird diese Art der Darstellung unübersichtlich.

Die folgende Abbildung zeigt die negativen und positiven Abweichungen mit Balken an, sie wurde mit Illustrator erstellt, ähnliche Abbildungen gibt es auch in der Tabellenkalkulation.



Balkendiagramm
(verwendet sind die Daten: Differenz zum Mittelwert etwas weiter unten)

So griffig der arithmetische Mittelwert für uns ist, wünscht man sich doch eine zweite Kennzahl für die Streuung, die man einfach zusammen mit dem Mittelwert angeben kann. Diese Kennzahl ist die Standardabweichung. Sie wird im folgenden vorgestellt.

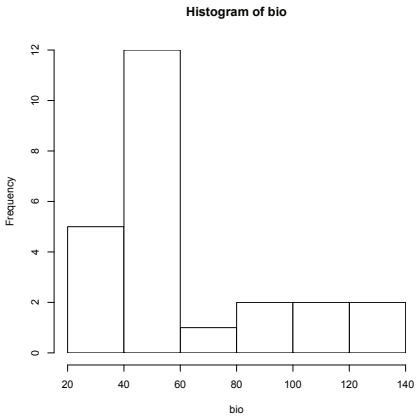
Histogramm zeichnen lassen

Zuerst werden die Daten in eine Variable mit dem Namen ‚bio' gelesen, dazu wird die Funktion ‚c()' benutzt

```
bio<-c(99, 87, 35, 59, 47, 26, 59, 40, 41, 53, 45, 139, 53, 113, 50, 58, 62, 55, 40, 107, 54, 48, 40, 139)
```

Die Funktion zum Zeichnen des Histogramms

```
hist(bio)
```



Ein Histogramm mit R, erzeugt mit der Funktion ‚hist()'

Anmerkungen:

- Die Einteilung in Klassen hat R selbst gemacht, man kann das aber auch selbst festlegen. So kann man mit der Anweisung ‚hist(bio, breaks=5)' dafür sorgen, dass 5 Klassen gebildet werden.
- Dass die Abbildung auf dem Bildschirm angezeigt wird, ist die Voreinstellung, dieses Bild kann mit der Tastenkombination ‚Strg+C' in die Zwischenablage kopiert und weiterverwendet werden (zum Beispiel, um es in einem anderen Programm mittels ‚Strg+V' einzufügen).
- Mit der Anweisung ‚?hist()' kann man in R die Hilfe zu dieser Funktion anfordern.

Histogramm PLUS Kurve der Normalverteilung:

Mit den folgenden Anweisungen kann man die Häufigkeitsverteilung (s. vorstehende Abbildung) zusammen mit einer Kurve anzeigen lassen:

Zuerst werden die Daten eingelesen in die Variable ‚bio'

```
bio<-c(99, 87, 35, 59, 47, 26, 59, 40, 41, 53, 45, 139, 53, 113, 50, 58, 62, 55, 40, 107, 54, 48, 40, 139)
```

Ein Histogramm wird gezeichnet

```
hist(bio, breaks=5, prob=TRUE)
```

Zur Erläuterung:

- hist() ist der Name der verwendeten Funktion, ihr werden Parameter beim Aufruf übergeben:
  - bio (der Name der Variablen, in der die Daten stehen);
  - breaks (legt die Klassengrenzen fest, hier: 5) (tatsächlich berechnet R die Anzahl der Klassen selbst, das komplexe Verfahren ist in einem Aufsatz beschrieben: Schumacher: http://planspace.org/20141225-how_does_r_calculate_histogram_break_points);
  - prob (falls der Wert TRUE ist, wird auf der *y*-Achse die Dichte aufgetrage, man kann daran die relative Häufigkeit ablesen, falls der Wert FALSE ist, werden die Häufigkeiten selbst gezeigt;

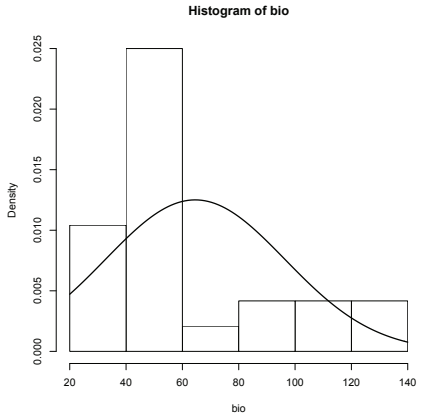
Zusätzlich zum Histogramm wird nun eine Kurve gemalt

```
curve(dnorm(x, mean(bio), sd(bio)), add=TRUE)
```

Zur Erläuterung der vier verwendeten Funktionen:

- ‚mean()' liefert den arithmetischen Mittelwert der Datenreihe ‚bio';
- ‚sd()' liefert die Standardabweichung der gleichen Datenreihe;
- ‚dnorm()' berechnet die Daten der Normalverteilung;
- ‚curve()' zeichnet die Kurve zu der Normalverteilung.

Dabei entsteht die folgende Grafik:



Histogramm mit Kurve

Rechenbeispiel mit dem TI 84 Plus

Wenn Sie die vorstehenden Zahlen nicht abtippen möchten, erfinden Sie einfach welche. Die Eingabe und Auswertung mit dem Taschenrechner ist so einfach, dass Sie bestimmt mehrfach Daten auswerten lassen. Und warum nicht das Gewicht von 20 Kirschen untersuchen?

1. Liste erzeugen
   STAT-Taste drücken
   erster Menüpunkt EDIT, erste Option EDIT wählen
   Enter-Taste drücken
   Zahlen untereinander eingeben, hier in die Spalte ‚L1'
   (zum Beispiel die Zahlen der Bio-Kartoffel)

2. Statistische Daten berechnen lassen
   STAT-Taste drücken
   Cursor-Taste rechts einmal drücken
   ‚1-Var Stats' mit Enter bestätigen (bereits vorgeschlagen)
   ‚List: L1' (dieser Name ist bereits auf der Taste mit der Ziffer 1 abgelegt)
   ‚Calculate' mit der Enter-Taste bestätigen.

3. Ergebnisse anzeigen lassen

$\overline{x}$	=	64,541	(arithmetischer Mittelwert)
$\sum x$	=	1549	(Summe aller aufsummierten x)
$\sum x^2$	=	123379	(quadrierte Summe)
Sx	=	31,899	(Standardabweichung)
σx	=	31,227	(Sigma)
n	=	24	(Anzahl der Messwerte)

Die Einteilung der Werte in Klassen zur Darstellung von Häufigkeiten

Liegen viele Werte vor, ist es umständlich von Hand zu rechnen. Um die Berechnung zu vereinfachen, bildet man Klassen und teilt jeden Messwert einer Klasse zu. Innerhalb einer Klasse zählt man dann die Anzahl der Messwerte. So liegen zum Beispiel innerhalb der Klasse von 0 bis 29 g insgesamt 15 Kartoffeln. Sind alle individuellen Messwerte auf die Klassen verteilt, nimmt man die Klassenmitte (hier zum Beispiel: 15 g, 45 g usw.) und rechnet mit dieser weiter.

Ein typischer Ansatz für die Einteilung in Klassen ist, dass man 10 (bis 20) Klassen herstellt. Es leuchtet ein, dass das nur dann Sinn macht, wenn man genügend Daten hat. Wir verwenden hier die Daten der beim Bauer gekauften Kartoffeln, die weiter oben gelistet sind, es handelt sich um 109 Messwerte:

1 182	20 155	39 265	58 79	77 44	96 87
2 164	21 109	40 110	59 61	78 185	97 110
3 134	22 94	41 83	60 112	79 75	98 135
4 106	23 118	42 104	61 15	80 122	99 86
5 99	24 121	43 80	62 117	81 67	100 32
6 124	25 62	44 170	63 129	82 103	101 20
7 124	26 110	45 165	64 66	83 183	102 65
8 201	27 60	46 122	65 144	84 166	103 63
9 62	28 156	47 95	66 152	85 108	104 41
10 39	29 196	48 32	67 147	86 92	105 51
11 58	30 105	49 92	68 116	87 108	106 47
12 35	31 70	50 108	69 66	88 147	107 173
13 47	32 65	51 52	70 148	89 100	108 77
14 42	33 122	52 158	71 132	90 62	109 26
15 132	34 110	53 119	72 96	91 66	
16 127	35 123	54 162	73 69	92 135	
17 123	36 102	55 85	74 75	93 175	
18 79	37 117	56 111	75 62	94 99	
19 131	38 115	57 193	76 73	95 45	

Einteilung in Klassen

Klassenbreite	Klassenmitte	Anzahl	prozentualer Anteil	summierte Prozent-Anteile
0 – 29	15	3	2,8	2,8
30 – 59	45	13	11,9	14,7
60 – 89	75	26	23,9	38,6
90 – 119	105	29	26,6	65,2
120 – 149	135	20	18,3	83,5
150 – 179	165	11	10,1	93,6
180 – 209	195	6	5,5	99,1
210 – 239	225	0	0	99,1
240 – 269	255	1	0,9	100
		109	100	

Erläuterungen zu den Spalten:

- Anzahl: die Anzahl der Messwerte in einer Klasse; die Summe (unten) wird benötigt, um den prozentualen Anteil an der Gesamtzahl (109) zu berechnen
- prozentualer Anteil: (3 / 109) * 100 = 2,8%
- summierte Anteile: zur Kontrolle, sie soll 100 ergeben (2,8 + 13 + ...) und dienen der Kontrolle

Berechnung der Standard-Abweichung

Standard-Abweichung mit Einzeldaten

Liegen nur wenige Daten vor, kann man die Abweichung der Einzelwerte vom arithmetischen Mittelwert leicht mit dem Taschenrechner berechen. Ausnahmsweise kann auch die Tabellenkalkulation dazu verwendet werden und nicht zuletzt R (oder irgendeine andere Statistik-Software). Verwendet werden hier die Daten der Bio-Kartoffeln.

Zuerst werden die Differenzen zwischen dem arithmetischen Mittelwert (der wurde schon davor berechnet: 64,54) und den Messwerten berechnet, danach werden diese quadriert (um die negativen Vorzeichen verschwinden zu lassen). Bei der Berechnung der Quadrate entstehen schnell große Zahlen, die in der letzten Spalte aufsummiert werden.

Die Summe wird danach durch 23 dividiert, das ist 24-1 oder allgemein formuliert: n-1. Schließlich wird aus dieser Zahl die Wurzel gezogen und – voilà – da steht die Standardabweichung.

Nr.	Messwert	arithmetischer Mittelwert	Differenz zum Mittelwert	Quadrat der Differenz
1	26	64,54	-38,54	1485,33
2	35	64,54	-29,54	872,61
3	40	64,54	-24,54	602,21
4	40	64,54	-24,54	602,21
5	40	64,54	-24,54	602,21
6	41	64,54	-23,54	554,13
7	45	64,54	-19,54	381,81
8	47	64,54	-17,54	307,65
9	48	64,54	-16,54	273,57
10	50	64,54	-14,54	211,41
11	53	64,54	-11,54	133,17
12	53	64,54	-11,54	133,17
13	54	64,54	-10,54	111,09
14	55	64,54	-9,54	91,01
15	58	64,54	-6,54	42,77
16	59	64,54	-5,54	30,69
17	59	64,54	-5,54	30,69
18	62	64,54	-2,54	6,45
19	87	64,54	22,46	504,45
20	99	64,54	34,46	1187,49
21	107	64,54	42,46	1802,85
22	113	64,54	48,46	2348,37
23	139	64,54	74,46	5544,29
24	139	64,54	74,46	5544,29
Summen	1549		0,04 *	23403,96

* ein sinnloses Ergebnis, da sich die Zahlen durch die verschiedenen Vorzeichen teilweise aufheben.

Die Summe der Abweichungsquadrate beträgt:

23403,96

dividiert durch (Anzahl der Beobachtungen -1) (n-1), hier also (24-1): 23

1017,56 (gerundet)

daraus die Quadrat-Wurzel ergibt die Standard-Abweichung

31,90

Das bedeutet, die mittlere Abweichung vom arithmetischen Mittelwert ist 31,90 g.

Standard-Abweichung mit Daten, die in Klassen eingeteilt sind

Liegen viele Daten vor, dann ist es etwas umständlich, mit den Einzelwerten zu rechnen. Alternativ kann man Klassen (von ... bis ...) bilden und die Einzelwerte zählen, die in einer Klasse liegen. Die Festlegung der Klassenbreiten ist nicht immer ganz einfach, denn keine der Klassen sollte leer bleiben und es sollten auch nicht zu wenige/zu viele Klassen sein (d.h., man muss manchmal etwas basteln). In der Praxis werden es meist 10 bis 20 Klassen sein.

Um die Klassen festzulegen, muss man den kleinsten und größten Wert kennen, die lassen sich hier (noch) von Hand ermitteln. Der kleinste Wert ist 15 (Gramm), der größte Wert ist 265 (Gramm), die Spannweite beträgt demnach 250. Wir haben uns hier für eine Klassenbreite von 30 g entschieden.

Sortiert sind die Gewichte der Kartoffeln, die beim Bauer gekauft wurden (109 Stück in einem Sack mit 10 kg), Damit lassen sich die Anzahl der Kartoffeln in einer Klasse leicht abgreifen (unterstrichen sind die letzten Werte innerhalb der einzelnen Klassen):

Sortierte Daten (109 Stück)

15	47	63	75	94	108	116	124	147	170
20	47	65	77	95	108	117	124	147	173
26	51	65	79	96	108	117	127	148	175
32	52	66	79	99	109	118	129	152	182
32	58	66	80	99	110	119	131	155	183
35	60	66	83	100	110	121	132	156	185
39	61	67	85	102	110	122	132	158	193
41	62	69	86	103	110	122	134	162	196
42	62	70	87	104	111	122	135	164	201
44	62	73	92	105	112	123	135	165	265
45	62	75	92	106	115	123	144	166	

Durch das Sortieren der Daten können bereits Informationen verloren gehen, wenn nämlich bei der Messung ein Muster auftritt (z.B. abwechselnd ein großer Wert und ein kleiner Wert). Deshalb sollte man die ursprünglichen Listen unbedingt aufbewahren.

Einteilung in Klassen

Wird mit Daten gerechnet, die in Klassen vorliegen, ist der Rechengang etwas anders. Es lässt sich mit Papier, Bleistift und Taschenrechner gerade noch so bewerkstelligen, aber es ist schon etwas mühsam.

Die Rechenschritte sind immer die gleichen: man verwendet die Klassenmitte und multipliziert mit der Anzahl (15 x 3 = 45). Danach werden die Werte der Klassenmitte quadriert und mit der Anzahl multipliziert, zuletzt werden die Summen gebildet. Es liegt auf der Hand, dass durch die Verwendung der Klassenmitten

kleine Abweichungen entstehen (gegenüber der Berechnung mit den Original-Daten), aber das ist vertretbar.

Klassenbreite	Klassenmitte	Anzahl	Anzahl x Klassenmitte	Klassenmitte²	Klassenmitte² x Anzahl
0-29	15	3	45	225	675
30-59	45	13	585	2025	26325
60-89	75	26	1950	5625	146250
90-119	105	29	3045	11025	319725
120-149	135	20	2700	18225	364500
150-179	165	11	1815	27225	299475
180-209	195	6	1170	38025	228150
210-239	225	0	0	50625	0
240-269	255	1	255	65025	65025
Summen		109	11565	218025	1450125

Berechnung vom arithmetischen Mittelwert:

11565/109=106,10

Berechnung der Standardabweichung:

$$\sqrt{\frac{1450125 - (106{,}10 * 11565)}{108}}$$

$$= \sqrt{2248{,}65}$$

= 45,45

Die Berechnung der Standardabweichung

$$\frac{1450125 - (11565^2 / 109)}{108}$$

= 2065,44

Die Quadratwurzel daraus

= 45,45

Die Standardabweichung ist hier größer als bei der zuvor gerechneten Datenreihe, das passt auch zu der größeren Spannbreite, die am Anfang festgestellt wurde.

Berechnung mit dem Taschenrechner

Anmerkung: die 109 Messwerte waren in fünf Minuten in den Taschenrechner eingetippt, dann nochmals drei Minuten zur Überprüfung der Daten, die Auswertung war in Sekunden gerechnet.

$\bar{x}$	=	104,39	(arithmetischer Mittelwert)
$\sum x$	=	11379	(Summe aller aufsummierten x)
$\sum x^2$	=	1415649	(quadrierte Summe)
Sx	=	45,92	(Standardabweichung)
σx	=	45,71	(Sigma)
n	=	109	(Anzahl der Messwerte)

Zusammenfassung

Hier sollen nun alle Informationen zusammengetragen werden, die wir auf den Seiten davor gesammelt und berechnet haben. Zur Erinnerung: am Anfang hatten wir drei Stichproben mit Werten. Aus den drei Datenreihen haben wir zunächst die Mittelwerte berechnet (arithmetischer Mittelwert, Median und Modus), nun dazu die Standardabweichungen.

Bio-Kartoffel

Daten einlesen

```
bio<-c(99, 87, 35, 59, 47, 26, 59, 40, 41, 53, 45, 139, 53, 113, 50, 58, 62, 55, 40, 107, 54, 48,
40, 139)
```

Mittelwert

```
mean(bio)
64.54167
```

Standardabweichung

```
 sd(bio)
31.89927
```

Normale Kartoffel

Daten einlesen

```
normal<-c(116, 99, 88, 109, 114, 39, 50, 46, 99, 111, 89, 58, 141, 83, 92, 50, 113, 74, 80, 66, 72,
105, 116, 100, 79, 92, 142, 68, 47)
```

Mittelwert

```
mean(normal)
87.51724
```

Standardabweichung

```
sd(normal)
27.6491
```

Vom Bauer gekaufte Kartoffeln

```
Bauer<-c(182, 164, 134, 106, 99, 124, 124, 201, 62, 39, 58, 35, 47, 42, 132, 127, 123, 79, 131,
155, 109, 94, 118, 121, 62, 110, 60, 156, 196, 105, 70, 65, 122, 110, 123, 102, 117, 115, 265,
110, 83, 104, 80, 170, 165, 122, 95, 32, 92, 108, 52, 158, 119, 162, 85, 111, 193, 79, 61, 112,
15, 117, 129, 66, 144, 152, 147, 116, 66, 148, 132, 96, 69, 75, 62, 73, 44, 185, 75, 122, 67, 103,
183, 166, 108, 92, 108, 147, 100, 62, 66, 135, 175, 99, 45, 87, 110, 135, 86, 32, 20, 65, 63, 41,
51, 47, 173, 77, 26)
```

Mittelwertberechnung

```
mean(Bauer)
[1] 104.3945
```

Standardabweichung

```
sd(Bauer)
45.92103
```

Damit sind die wichtigesten Berechnungen gemacht.

DER VARIATIONSKOEFFIZIENT

Schließlich soll zusätzlich noch der Variationskoeffizient berechnet werden. Der ist – im Gegensatz zu der Standardabweichung – besser geeignet, Datenmengen miteinander zu vergleichen, bei denen die Mittelwerte weit auseinanderliegen, denn bei denen liegen dann die Standardabweichungen auch weit auseinander und erschweren den Vergleich.

Um den Variationskoeffizienten zu berechnen teilt man die Standardabweichung durch den arithmetischen Mittelwert und multipliziert den Wert mit 100:

$$VK = \frac{\text{Standardabweichung}}{\text{arithmetischer Mittelwert}} \times 100$$

Bei den Bio-Kartoffeln ergibt das:

$$VK = \frac{31{,}89927}{64{,}54167} \times 100$$

$$VK = 49{,}43$$

Schaut man in der folgenden Tabelle auf die Standardabweichungen (hervorgehoben), dann kann man sehen, dass die Werte der ersten und dritten Datenspalte relativ weit auseinanderliegen.

Verwendet man dagegen den Variationskoeffizienten, dann sieht man klar, dass die Streuungen tatsächlich fast gleich sind.

Hier nun in einer Tabelle zusammengefasst, was wir nun über die Kartoffeln herausgefunden haben.

	Bio-Kartoffeln	normale Kartoffeln	10 kg-Packung
Anzahl Kartoffeln	24	29	109
kleinster Wert	26	39	15
größter Wert	139	141	265
Spannweite (größter - kleinster Wert)	113	102	250
Mittelwert	64,54	87,52	104,39
Standardabweichung	**31,90**	27,65	**45,92**
Variationskoeffizient	**49,43**	31,59	**43,99**

Sind die Daten normal verteilt?

Um die letzte Frage nach der Normalverteilung zu beantworten, werden die Daten zunächst grafisch dargestellt, danach wird ein Test bereechnet, der bei der Beantwortung dieser Frage hilft.

Die Anweisungen, die zu diesen drei Abbildungen führen, wurden zuvor schon erläutert, deshalb hier nur die Zusammenfassung der Anweisungen in R und danach die drei Abbildungen nebeneinander.

Für die erste Abbildung:

```
bio<-c(99, 87, 35, 59, 47, 26, 59, 40, 41, 53, 45, 139, 53, 113, 50, 58, 62, 55, 40, 107, 54, 48,
40, 139)
hist(bio, breaks=5, prob=TRUE)
curve(dnorm(x, mean(bio), sd(bio)), lwd=2, add=TRUE, yaxt="n")
```

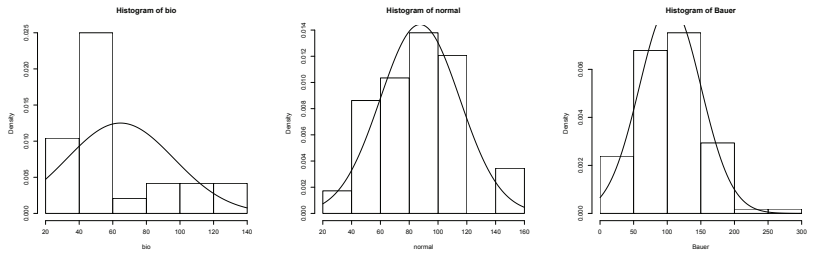
für die zweite Abbildung:

```
normal<-c(116, 99, 88, 109, 114, 39, 50, 46, 99, 111, 89, 58, 141, 83, 92, 50, 113, 74, 80, 66, 72,
105, 116, 100, 79, 92, 142, 68, 47)
hist(normal, breaks=5, prob=TRUE)
curve(dnorm(x, mean(normal), sd(normal)), lwd=2, add=TRUE, yaxt="n")
```

und für die dritte Abbildung

```
Bauer<-c(182, 164, 134, 106, 99, 124, 124, 201, 62, 39, 58, 35, 47, 42, 132, 127, 123, 79, 131,
155, 109, 94, 118, 121, 62, 110, 60, 156, 196, 105, 70, 65, 122, 110, 123, 102, 117, 115, 265,
110, 83, 104, 80, 170, 165, 122, 95, 32, 92, 108, 52, 158, 119, 162, 85, 111, 193, 79, 61, 112,
15, 117, 129, 66, 144, 152, 147, 116, 66, 148, 132, 96, 69, 75, 62, 73, 44, 185, 75, 122, 67, 103,
183, 166, 108, 92, 108, 147, 100, 62, 66, 135, 175, 99, 45, 87, 110, 135, 86, 32, 20, 65, 63, 41,
51, 47, 173, 77, 26)
hist(Bauer, breaks=5, prob=TRUE)
curve(dnorm(x, mean(Bauer), sd(Bauer)), lwd=2, add=TRUE, yaxt="n")
```

Bio-Kartoffeln	normale Kartoffeln	vom Bauer



Zur Erläuterung:

- Die Anweisungen unterscheiden sich voneinander nur durch den Namen der drei Variablen, die zuvor mit Daten gefüllt wurden;
- die erste Kurve wirkt nicht symmetrisch, sie ist nach links verzerrt, vermutlich sind die Daten nicht normalverteilt;
- die zweite und dritte Kurve scheinen symmetrisch, daraus können wir folgern, dass die Daten möglicherweise normalverteilt sind.

Test auf Normalverteilung

Endgültige Klarheit soll ein Test bringen, der prüft, ob die Daten normalverteilt sind, hier der Test mit dem Namen ‚Shapiro-Wilk'.

Es gibt eine ganze Reihe von Tests, die hierfür verwendet werden können, zum Beispiel der Kolmogorov-Smirnov-Test, der Lilliefors-Test oder der Chi-Quadrat-Test. Alle haben Stärken und Schwächen. Über die Jahre hinweg wurde mal der eine, mal der andere in der Literatur bevorzugt.

Liegt Normalverteilung vor? Eine Antwort mit R

Die zuvor angelegten Variablen (bio, normal, Bauer) können auch für diesen Test verwendet werden. Die folgende Tabelle fasst zusammen:

	Bio-Kartoffeln	normale Kartoffeln	vom Bauer
Variable	bio	normal	Bauer
Anweisung in R	shapiro.test(bio)	shapiro.test(normal)	shapiro.test(Bauer)
Ergebnis	Shapiro-Wilk normality test data: bio W = 0.81803, p-value = 0.0005894	Shapiro-Wilk normality test data: normal W = 0.96951, p-value = 0.5464	Shapiro-Wilk normality test data: Bauer W = 0.9798, p-value = 0.09667
liegt Normal-verteilung vor?	nein	ja	ja

Zur Erläuterung:

- Früher hat man den berechneten Wert (‚W') mit dem Wert in einer Tabelle verglichen, das ist mit R nicht mehr nötig: der zusätzlich angezeigte ‚p-Wert' zeigt uns, ob Normalverteilung vorliegt, ist er über 0,05 (unsere ‚übliche' Irrtumswahrscheinlichkeit von 5%), dann liegt Normalverteilung vor, liegt er unter 0,05, dann liegt keine Normalverteilung vor.
- Hier erkennen wir klar, dass bei den Bio-Kartoffeln keine Normalverteilung vorliegt, bei den anderen dagegen schon, etwas, was wir auch schon anhand der zuvor gezeigten Histogramme und Kurven vermutet hatten.

Fazit

An sich sind die Ergebnisse erstaunlich, aber ein Gespräch mit einem Bauer brachte Klarheit, denn Bauern spezialisieren sich

- entweder für den Speisemarkt, d.h. sie erzeugen Kartoffeln für den Endkunden
- oder für den Industriemarkt, die Kartoffeln werden für Industrielle Weiterverarbeitung angebaut, aus ihnen entstehen Pommes Frites, Chips, Stärke oder Alkohol.

Bei der Produktion für den Speisemarkt werden die Kartoffeln sortiert, im Sack landen die zwischen 35 – 65 mm (für die Sortierung werden Sortierwalzen oder Lochbleche verwendet). Die Kartoffeln über 65 mm werden aussortiert und werden als Ofenkartoffeln verkauft, das bringt dem Bauer mehr Gewinn.

Die kleinsten Kartoffeln (unter 35 mm) werden nicht mitgeerntet, sie bleiben oft im Acker liegen.