

Übersicht

2.1	Motivation	49
2.1.1	Grundbegriffe	50
2.2	Lage- und Streuungsmaße	51
2.2.1	Lagemaße	55
2.2.2	Streuungsmaße	59
2.3	Kenngrößen für den Zusammenhang von Merkmalen	63
2.3.1	Korrelation	63
2.3.2	Lineare Regression	67
2.4	Aufgaben	68

2.1 Motivation

Man kann keine wissenschaftliche Arbeit veröffentlichen, wenn man seine Messungen nicht korrekt darstellt oder durchgeführte Studien falsch auswertet. Als erfolgreicher Wissenschaftler sollte man mit beiden Beinen fest auf stati(sti)schem Boden stehen. Schließlich muss jedes Experiment mal ausgewertet und aussagekräftig erklärt werden. Dazu bedient man sich am besten der beschreibenden (deskriptiven) Statistik. Wir werden uns auf den folgenden Seiten genauer mit der Beschreibung und Darstellung von experimentellen Daten beschäftigen. Dabei sollten nicht voreilig Schlüsse aus diesen Darstellungen oder Parametern gezogen werden. Die deskriptive Statistik beschreibt die Daten nur, es werden aber keine Interpretationen gegeben. Möchte man die Verlässlichkeit der Daten überprüfen oder von einem Datensatz auf zukünftige Messungen schließen, müssen Methoden aus der induktiven Statistik genutzt werden (s. Abschn. 4.5, S. 121).

Außerdem sollte man sich genau mit den in Experimenten ermittelten Zahlen, deren Größenordnung und deren Skalierung auseinandersetzen. Schließlich können die

Skalierungen von biologischen Systemen sehr unterschiedlich sein. Während die Membranen, die alle Zellen unseres Körpers umspannen, nur eine Dicke von wenigen Nanometern (nm, 10^{-9} m) besitzen, können Pilzgeflechte eines einzelnen Organismus Gebiete mit einem Durchmesser von mehreren Kilometern (km, 10^3 m) umfassen. Zwischen diesen beiden Maßen liegen zwölf Größenordnungen. Ähnlich verhält es sich bei den zeitlichen Dimensionen biologischer Prozesse. Während die Moleküle aller biochemischen Reaktionen ihre Energiezustände innerhalb von Femtosekunden (fs, 10^{-15} s) ändern, benötigt die Evolution der Lebewesen, in denen ebendiese Reaktionen ablaufen, mitunter mehrere Millionen Jahre (zehn Billionen = 10^{13} s). Interessante Zahlen im biologischen Kontext liefert die Datenbank <http://www.BioNumbers.org>.

Wichtiges in Kürze

Die nachstehenden Gleichungen musst du dir einprägen, wenn du die Klausur bestehen willst.

- arithmetisches Mittel („Mittelwert“): $\mu = \frac{1}{n} \sum_{i=1}^n x_i$ (s. Abschn. 2.2.1, S. 55)
- Varianz: $\sigma(x)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$ (s. Abschn. 2.2.2, S. 60)
- Standardabweichung: $\sqrt{\sigma(x)^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$ (s. Abschn. 2.2.2, S. 60)
- linearer Korrelationskoeffizient: $r(x, y) = \frac{\sigma(x, y)}{\sigma(x) \cdot \sigma(y)}$ (s. Abschn. 2.3.1, S. 63)
- lineare Regression: $a = \mu_y - b\mu_x = \frac{\sum y_i}{n} - b \frac{\sum x_i}{n}$ (s. Abschn. 2.3.2, S. 67)

$$b = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^n (x_i - \mu_x)^2}$$

2.1.1 Grundbegriffe

Um sich in der Welt der deskriptiven Statistik zurecht zu finden, muss man die darin lebenden mathematischen Kreaturen und ihr Wesen kennen. Deshalb wollen wir uns zunächst mit den wichtigsten Begriffen vertraut machen. Die zentralen Spieler in der induktiven Statistik sind die sogenannten **statistischen Einheiten**. Grob gesagt, kann man sich unter einer statistischen Einheit auch einfach ein Objekt vorstellen. Für einen Biologen sind diese Objekte meist Zellen, Proteine, Versuchstiere oder Gene. Sie sind also die einzelnen untersuchten Objekte einer Datenerhebung, deren Eigenschaften von Interesse sind. Gebräuchlich ist deshalb auch der Begriff der **Merkmalsträger**. Vor einer Erhebung muss der Merkmalsträger genau definiert werden. Dies geschieht im idealen Fall anhand von einer sachlichen, räumlichen und zeitlichen Beschreibung. So können als Merkmalsträger für eine Datenerhebung aus einer Gewebeprobe alle Epithelzellen (sachliche Definition) definiert werden, welche drei Stunden nach Färbung (zeitliche Definition) mit einem speziellen Antikörper eine Färbung innerhalb des Zellkerns (räumliche Definition) aufweisen. Alle Zellen, auf die diese Definitionen zutreffen, werden weiterhin als Merkmalsträger bezeichnet. Die

Gesamtheit aller Merkmalsträger wird als **Grundgesamtheit** oder **Population** bezeichnet. Die Größe dieser Grundgesamtheit spielt vor allem in der induktiven Statistik eine wichtige Rolle. Allerdings müssen nicht alle Merkmalsträger auch tatsächlich materiellen Charakter haben. So können auch eine Bewegung (z. B. ein Migrationsverhalten von Zellen) oder ein Verhalten einen Merkmalsträger darstellen. Der Begriff Merkmalsträger wird auch verwendet, da meistens die Merkmale dieser Objekte von Interesse sind. Ein interessantes Merkmal der oben definierten Zellen könnte z. B. die Größe der Zellkerne sein. Typischerweise liegt diese im Bereich einiger Mikrometer. Die Größe wird dann als **Merkmalsausprägung** bezeichnet; ein Merkmal besitzt also eine Ausprägung. Die Ausprägungen eines bestimmten Merkmals eines Merkmalsträgers sind die in einem Experiment ermittelten Messwerte.

Grundbegriffe

Grundgesamtheit: die Gesamtheit aller Merkmalsträger mit übereinstimmenden (räumlichen, sachlichen und zeitlichen) Identifikationen

Merkmalsträger/ statistische Einheit: die Einzelobjekte einer statistischen Erhebung

Merkmal: untersuchte Eigenschaft eines Merkmalsträgers

Merkmalsausprägung: ermittelter Wert des Merkmals

2.2 Lage- und Streuungsmaße

Lage- und Streuungsmaße sind in der Biologie von immenser Bedeutung, da sie charakterisieren, wo sich ein Messwert auf einer Skala (Tab. 2.1) befindet und wie sehr er schwankt. Die Position des Messwertes bezeichnet man als Lage und seine Schwankung gemeinhin als Streuung. Abgesehen von Lage und Streuung eines Merkmals, muss auch stets dessen Skalierung im Hinterkopf behalten werden, denn „man kann Äpfel nicht mit Birnen vergleichen“. Naja, biologisch schon, aber in der Statistik nicht! Manche Eigenschaften von Merkmalen können nicht mit anderen Eigenschaften verglichen werden, da sie eine andere **Skala** besitzen. Eine Skalierung beschreibt in gewisser Weise die Natur von Daten. Es werden drei große Skalenklassen für Daten unterschieden: nominal, ordinal und kardinal skalierte Daten. In einer Versuchsgruppe von Mäusen kann zwischen Versuchsmäusen, also Tiere, die z. B. ein Medikament erhalten haben, und Kontrollmäusen, eine Gruppe, die unter gleichen Bedingungen gehalten wird, unterschieden werden. Das Merkmal „Gruppenzugehörigkeit“ ist **nominal** skaliert, da seine Ausprägungen Namen sind. Legen wir nun alle Mäuse der Größe nach nebeneinander, ist das Merkmal „Körpergröße“ **ordinal** skaliert, denn wir können die Mäuse nun mit „größer als“ oder „kleiner als“ beschreiben. Die Ausprägung ist allerdings immer noch ein Name. Messen wir hingegen die genaue Körpergröße, so wird dieses Merkmal als **kardinal** skaliert bezeichnet.

Tab. 2.1 Verschiedene Skalen für biologische Messwerte

qualitative Variablen kategorisch	quantitative Variablen numerisch
binär 0/1, ja/nein, ♂/♀, tot/lebendig	diskret Anzahl von Individuen
nominal Farben, Formen, Spezies	kontinuierlich Größe, Gewicht, Temperatur, Zeit
ordinal Schmerz, Lebensqualität, Tumorstadium	

Je nach Skalierung eines Merkmals ist dieses unterschiedlich informativ. So kann bei nominal skalierten Merkmalen nur eine Aussage über deren Gleichheit getroffen werden. Wir können keine in sich sinnvolle Reihenfolge oder Bewertung für diese Ausprägungen einführen. Anders bei ordinalen Skalen. Hier wird immerhin die Position einer Ausprägung auf einer Skala in Bezug zu einer anderen Ausprägung gegeben. Nominal und ordinal skalierte Merkmale werden als **qualitative Merkmale** bezeichnet, da sie keinerlei quantitative Informationen liefern. Kardinal skalierte Merkmale jedoch können nicht nur quantitative Informationen liefern, deshalb auch **quantitative Merkmale** genannt, sondern es können auch sinnvolle Rechenoperationen wie die Bildung von Differenzen mit deren Ausprägungen durchgeführt werden.

Kardinal, manchmal auch **numerisch** skalierte Merkmale genannt, können weiterhin noch in zwei Unterklassen unterteilt werden. Als **kontinuierlich** skalierte Merkmale werden solche bezeichnet, die alle möglichen Zwischenwerte auf einer Skala einnehmen können. So ist die gemessene Körpergröße eine kontinuierliche Ausprägung, da eine Maus sowohl 20 cm als auch 20,673564 cm lang sein kann. Anders bei **diskreten** Ausprägungen, denn die Anzahl an Beinen kann entweder 1, 2, 3 oder 4 betragen, nicht aber 2,56. Tabelle 2.1 liefert einen Überblick über die vorgestellten Skalen.

Da Lage- und Streuungsmaße biologische Daten (also beispielsweise unsere experimentellen Messreihen) charakterisieren, ist es wichtig, dass wir diese Maße gewissenhaft bestimmen. Es ist ratsam, einen aufgenommenen Datensatz unvoreingenommen auszuwerten. Die deskriptive Statistik ist hier hilfreich, weil sie unseren Datensatz zunächst nur darstellt und beschreibt, aber nicht interpretiert. Um Datensätze zu interpretieren, sollte man Methoden aus der induktiven Statistik heranziehen. Mit der deskriptive Statistik lassen sich Datensätze, die aufgrund ihrer Größe oder Komplexität einen großen Informationsgehalt besitzen, auf handliche Informationsstückchen reduzieren. Dies ist immer mit einem Informationsverlust verbunden, daher ist auch an dieser Stelle Vorsicht geboten. Die deskriptive Statistik versucht also, zwischen zwei Streitpartnern, der Übersichtlichkeit und dem Informationsverlust, zu vermitteln, ohne dass einer von beiden zu große Verluste bei diesem Deal macht.

Beispiel 2.1 Zellen in der Wasserrutsche

Neue Technologien, insbesondere in der Molekular- und Zellbiologie, erlauben es, im Hochdurchsatzverfahren in kurzer Zeit immense Datenmengen aufzunehmen. Am Beispiel der Durchflusszytometrie lässt sich dies gut verdeutlichen, weswegen diese Methode uns auch als Beispiel dienen soll, um experimentelle Daten zu charakterisieren. Bei der Durchflusszytometrie passieren einzelne Zellen in einem dünnen Flüssigkeitsstrom einen Lichtstrahl. Die Zellen werden dafür zuvor in einem Reagenzröhrchen in Flüssigkeit suspendiert und dann durch enge Kapillaren gezogen, sodass die Zellen, eine nach der anderen, wie auf einer Wasserrutsche am Laser vorbeiströmen. Das Licht, das auf die Zellen trifft, wird gebrochen und in alle Richtungen zurückgeworfen. Dieses sogenannte Streulicht kann dann von Detektoren registriert werden und gibt Aufschluss über die Eigenschaften der einzelnen Zellen. Das Vorderstreulicht (engl. *front scatter*, FSC) gilt dabei als Maß für die Zellgröße bzw. genau genommen deren Querschnittsfläche. Je größer die Zelle ist, desto mehr Licht wird auch wieder nach vorne zurückgeworfen. Das Seitenstreulicht (engl. *side scatter*, SSC) ist umso stärker, je mehr die Zelle mit Vesikeln, sogenannten Granula, gefüllt ist. Außerdem können mit Laserstrahlen bestimmter Wellenlängen auch Fluoreszenzfarbstoffe zum Leuchten angeregt werden. Im Fall unseres Experiments haben wir einen Rezeptor auf der Zelloberfläche mit einem Antikörper detektiert, an den ein Farbstoffmolekül gekoppelt ist. Die Fluoreszenzintensität steigt somit mit steigender Rezeptorzahl an. Der Datensatz, den wir aufgenommen haben, umfasst 1000 Zellen und sieht in etwa wie folgt aus (der komplette Datensatz kann auf der Onlineplattform eingesehen werden):

Größe	Granularität	Rezeptorzahl
57.280	49.792	85.824
95.808	72.000	102.656
...
2707,8	5699,8	2536,86

Die Einträge sind dabei Zahlenwerte mit der Einheit für Lichtintensitäten und bieten ein Maß für Größe, Granularität und Rezeptorzahl der einzelnen Zellen (eine Zelle pro Zeile). Zunächst wollen wir betrachten, wie unterschiedlich groß unsere Zellen sind. □

Wie sich die Größe in der gesamten Population der 1000 Zellen verteilt, sieht man am besten in einem Histogramm, das zeigt, wie viele Zellen mit einer bestimmten Querschnittsfläche vorkommen (s. Abb. 2.1). Solche Histogramme werden in unterschiedlichen Zusammenhängen verwendet. Die Erstellung ist sehr intuitiv. Wir sehen auf der x -Achse einzelne Merkmalsausprägungen; hier Werte für die Querschnittsfläche der einzelnen Zellen. Die Höhe des zugehörigen Balkens richtet sich nach der Häufigkeit der gemessenen Merkmalsausprägung. Natürlich handelt es sich bei der Zellgröße bzw. den ermittelten Lichtintensitäten, um kontinuierlich skalierte Werte. Es gibt allerdings so viele unterschiedliche Ausprägungen,

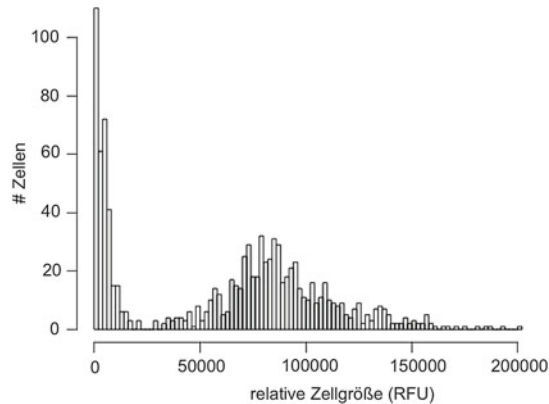
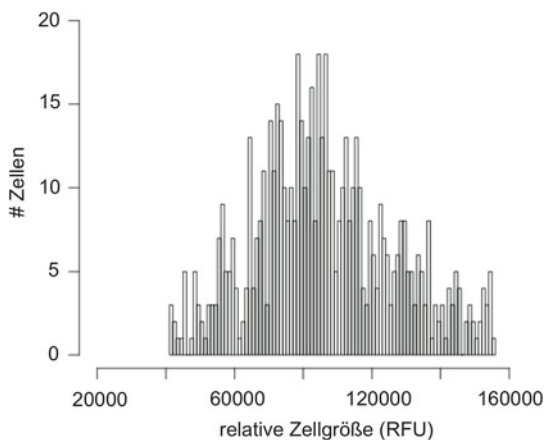


Abb. 2.1 Histogramm des via Durchflusszytometrie experimentell ermittelten Maßes für die Zellgröße. Je häufiger ein bestimmter Wert gemessen wurde, desto höher ist auch der zugehörige Balken. Die Anzahl der beobachteten Zellen (y -Achse) wurde gegen die gemessenen Werte für die Zellgröße (x -Achse) aufgetragen

dass es sich empfiehlt, Gruppen zu bilden und Werte zusammenzufassen. Hier wurden die Häufigkeiten von jeweils aufeinanderfolgenden Merkmalsausprägungen aufsummiert, also in einem Balken zusammengefasst, sodass insgesamt 100 Balken entstehen. Das gestaltet das Ganze übersichtlicher. Bei der Gruppierung sollten die Grenzen jedoch mit Bedacht gewählt werden, damit einzelne Ergebnisse nicht fälschlicherweise zusammengefasst werden. Je nachdem, wie die Grenzen für die Gruppierung gewählt werden, können einzelne Merkmalsausprägungen in der Darstellung verloren gehen. Diesen Kompromiss zwischen Übersichtlichkeit und Informationsverlust muss man auch hier eingehen. Das Mathematikerwort hierfür ist **Klassifizierung**, soll heißen, dass Merkmalsausprägungen in mindestens zwei Kategorien eingeteilt werden.

Wie wir dem Histogramm in Abb. 2.1 entnehmen können, gibt es einige sehr kleine Werte, zu erkennen an der Häufung am linken Ende der Verteilung. Da solch kleine Zellen nicht vorkommen, handelt es sich dabei wahrscheinlich um Schmutzpartikel, die bei der Analyse nicht berücksichtigt werden sollten. Solche Artefakte, später auch Ausreißer genannt, spielen eine wichtige Rolle in unseren Überlegungen. Mathematiker sprechen bei solchen Fällen gerne von Werten, die einer anderen Verteilung entstammen; soll heißen, dass sie eine andere Natur haben oder eben durch Fehler von uns aufgenommen wurden. Viele Diskussionen in der Biologie drehen sich um Ausreißer. Generell sollte man Daten aber nicht einfach aus der Auswertung entfernen, nur weil sie nicht den Erwartungen entsprechen. Hier wissen wir jedoch, dass kleine Messwerte auf Schmutzpartikel und nicht auf unsere Zellen zurückzuführen sind. Die höchsten Messwerte resultieren dagegen von Zelldubletten. Deshalb schließen wir besonders kleine und besonders große Messwerte aus, um keine falschen Schlüsse in der Analyse zu ziehen. Entfernt man die kleinsten 350 Zellen und

Abb. 2.2 Gecropptes Histogramm. Die Verteilung der relativen Zellgröße erhält man, wenn man kleine Messwerte, die nicht von Zellen, sondern von Schmutzpartikeln stammen, aus dem Histogramm entfernt



die größten 50, ein Vorgang, der als **Croppen** bezeichnet wird, erhält man das Histogramm mit einer tatsächlichen Verteilung der Zellgröße wie in Abb. 2.2.

Wie wir sehen können, enthält ein solches, noch recht simples Histogramm bereits eine große Menge an Information. Im Folgenden wollen wir einige Techniken kennenlernen, wie man bestimmte Charakteristika dieses Informationshaufens anhand von Kennzahlen herausstellt.

2.2.1 Lagemaße

Arithmetisches Mittel

Ein Kennzahl, die in den meisten Fällen von großem Interesse ist, ist das arithmetische Mittel: der Mittelwert μ . Er beschreibt den statistischen Durchschnitt einer Messreihe, also die Mitte aller Messpunkte. Seine Berechnung ist ganz einfach und lautet auf „Mathematisch“:

$$\mu = \frac{x_1 + x_2 + x_3 + \dots + x_i}{n} = \frac{1}{n} \sum_{i=1}^n x_{i-\text{ten Messung}}. \quad (2.1)$$

Im Alltag hat wohl schon jeder einmal den Mittelwert von etwas gebildet. Einfach gesagt, addieren wir alle Messwerte und teilen das Ergebnis dann durch die Anzahl der Messwerte. Mit unseren Durchflussszytometriedaten können wir also ein durchschnittliches Maß für die Zellgröße bestimmen. Für alle Messwerte gilt:

$$\begin{aligned}
 \mu &= \frac{1}{1000} \sum_{i=1}^{1000} \text{Zellgröße}_i \\
 &= \frac{1}{1000} (57.280 + 95.208 + \dots + 2707,8) \\
 &= 61.438
 \end{aligned}$$

Die beschnittenen Messwerte aus Abb. 2.2 haben hingegen einen Mittelwert von 86.140,17. Die zuvor entfernten (350 kleinsten und 50 größten) Werte, die einer anderen Verteilung angehören bzw. einfach Artefakte sind, werden Ausreißer genannt. Wie wir an unserer Berechnung sehen können, ist das arithmetische Mittel nicht robust gegen Ausreißer. Das soll heißen, dass das μ aller gemessenen Zellgrößen uns eine falsche Mitte unserer Verteilung vorgaukelt, wenn wir in der Berechnung Ausreißer berücksichtigen. Dies sollte man zunächst immer tun, weil man nie weiß, ob diese Werte biologisch relevant sind. Wir wollen im Folgenden eine Methode kennenlernen, um die durchschnittliche Lage einer Verteilung zu ermitteln, ohne uns Sorgen um Ausreißer machen zu müssen.

Median

Eine robuste, sprich ausreißerinsensitive Methode zur Ermittlung der durchschnittlichen Lage einer Verteilung ist der Median. Man sortiert dazu die Daten einfach nach ihrer Größe und ermittelt, welcher Wert in der Mitte der Reihe liegt. Von sieben geordneten Messpunkten wäre dies der vierte Wert (links drei, rechts drei). Bei acht Werten handelt es sich beim Median um den Durchschnitt aus dem vierten und dem fünften Wert. Der Median beziffert damit einen Wert, der die Verteilung in zwei Hälften unterteilt, wobei 50 % der Werte links davon und 50 % der Werte rechts davon liegen. Kommt ein Wert mehrfach in einer Datenerhebung vor, so muss er bei der Ermittlung des Medians auch mehrfach beachtet werden. Die allgemeine Formel für den Median lautet:

$$\tilde{x}_{Med} = \begin{cases} x_{\frac{n+1}{2}} & n \text{ ungerade} \\ \frac{1}{2} (x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n \text{ gerade.} \end{cases} \quad (2.2)$$

Für ein kleines Rechenbeispiel mit überschaubarer Anzahl an Messpunkten soll uns folgender Datensatz dienen:

5 8 3 4 3 6 8 7 4 8 25.

Zunächst müssen wir die Daten der Größe nach ordnen. Hierbei ist es egal, ob man aufsteigend oder absteigend ordnet:

3 3 4 4 5 6 7 8 8 8 25.

Da es sich hierbei um 11 Messpunkte handelt ($n = 11$), müssen wir die Formel für ungerade Datensätze heranziehen:

$$\tilde{x}_{Med} = x_{\frac{n+1}{2}}.$$

Setzen wir nun 11 für n ein, sehen wir, dass der sechste Datenpunkt unserem Median entspricht:

$$\tilde{x}_{Med} = x_{\frac{n+1}{2}} = x_{\frac{11+1}{2}} = x_6 = 6.$$

Zieht man den Datensatz der Durchflusszytometriedaten heran, merkt man, dass der Median für alle Messwerte mit 72.256 deutlich näher am Mittelwert der Population liegt, bei der wir die kleinen Messwerte ausgeschlossen haben. Im Median fallen die kleinen Ausreißer folglich nicht so sehr ins Gewicht. Bei der beschnittenen, in etwa normalverteilten Population liegt der Median mit 84.672 auch deutlich näher am Mittelwert von 86.140,7.

Modus

Außer dem Median und dem Mittelwert kann der Modus (oder Modalwert) als Lagemaß einer Messreihe dienen. Der Modus gibt den am häufigsten vorkommenden Wert an. Die Bildung des Modus macht vor allem bei Datensätzen mit **geringer Streuung** und **diskreten** Messwerten Sinn. Der Modus des Datensatzes 3 2 4 2 1 3 2 2 4 1 2 beträgt 2.

Bei kontinuierlichen Messwerten wie im Durchflusszytometriebeispiel kommen selten Werte exakt zweimal oder gar häufiger vor. Die Bildung des Modus würde sich nur für die diskreten Gruppen empfehlen, die jeweils in einem Balken zusammengefasst sind. Der Modus läge dann an der Spitze des Histogramms. Der höchste Balken in Abb. 2.1 befindet sich allerdings ganz links und ist ein Ausreißer. Wollten wir diese Ergebnisse publizieren, wären wir wohl die Lachnummer unserer Fachbereichs, schließlich beruhen unsere Ergebnisse im wahrsten Sinne des Wortes auf Dreck. Der Modus kann also sehr empfindlich für Ausreißer sein.

Quantile

Wer sich bereits mit dem Median beschäftigt hat, hat sich – bewusst oder unbewusst – auch schon mit Quantilen auseinandergesetzt. Denn der Median ist nur ein Trivialname für das 0,5-Quantil. Quantile heißen mit vollem Namen p -Quantile, wobei das p einen Wert zwischen 0 und 1 annehmen kann. Das p sagt aus, wie viel Prozent der Werte einer Messreihe sich links von diesem Wert befinden sollen (also kleiner sind). Links vom 0,25-Quantil liegt folglich ein Viertel aller Messpunkte und rechts die restlichen drei Viertel. Das 0,4-Quantil der Messwerte aus Abb. 2.2 liegt bei 80.000 und damit etwas links von der Mitte der Verteilung. In der Mitte, bei 84.672, liegt der Median, von dem links 50 % aller Werte platziert sind.

Die Berechnung eines Quantils ist sehr einfach, man muss nur zwei kleine Formeln im Kopf haben:

$$\tilde{Q}_p = \begin{cases} \frac{1}{2}(x_{n \cdot p} + x_{n \cdot p + 1}), & \text{wenn } n \cdot p \text{ ganzzahlig,} \\ x_{\lceil n \cdot p \rceil}, & \text{wenn } n \cdot p \text{ nicht ganzzahlig.} \end{cases}$$

Wobei n die Anzahl der Messwerte und p das zu errechnende Quantil ist. Betrachten wir als Beispiel kurz das oben schon erwähnte 0,4-Quantil. In unserem Fall ist $n = 601$ und $p = 0,4$. Multiplizieren wir diese Werte, erhalten wir den nicht ganzzahligen Wert 240,4. Wir müssen also die Gleichung für nicht ganzzahligen Werte nutzen und aufrunden – genau das wird durch die Klammern $\lceil \cdot \rceil$ symbolisiert. Der 241. Wert in der geordneten, beschnittenen Messreihe der 601 Messwerte aus Abb. 2.2 lautet 80.000.

Es gibt einige p -Quantile, die so oft benutzt werden, dass sie eigene Bezeichnungen erhalten haben. Den Median haben wir bereits kennengelernt (S. 56). Er bezeichnet das 0,5-Quantil.

Das 0,25-Quantil und das 0,75-Quantil werden als unteres und oberes Quartil bezeichnet. Quartil deshalb, weil sie die Werte angeben, bei denen jeweils ein Viertel der Messwerte unter- beziehungsweise oberhalb liegen. Gebräuchlich sind auch die Begriffe Tertil (0,3-Quantil), Quintil (0,2-Quantil), Dezil (0,1-Quantil) oder Perzentil (0,01-Quantil). Verwendung finden solche Quantile oft nicht nur bei der Ermittlung der Lage einer Verteilung, sondern auch bei Normalisierungsoperationen. Hierbei geht es darum, unterschiedliche Datensätze miteinander vergleichbar zu machen.

Beispiel 2.2 Kleinvieh macht auch Mist

In vielen molekularbiologisch ausgerichteten Labors findet man immer öfter seltsam aussehende Chips mit kleinen schwarzen Arbeitsflächen. Kaum größer als eine Briefmarke sind diese allerdings nicht als Dopingtest für den ermüdeten Doktoranden gedacht. Nein, diese kleinen Dinger sind eine Revolution in der Biologie. *Lab on a chip*, oder DNA-Microarray heißt das Konzept. Auf diesen Chips befinden sich beispielsweise Tausende von kurzen DNA-Molekülen, die fest mit der Chipoberfläche verbunden sind. Jeder dieser DNA-Schnipsel passt dann z. B. genau zu einem ganz bestimmten Gen des Menschen. Gibt man nun die mit Fluoreszenzfarbstoffen versehenen mRNA-Moleküle einer Probe auf diese Chips, suchen diese RNA-Moleküle ihren passenden Bindungspartner auf der Oberfläche und bleiben auch nach dem Waschen des Chips fest mit ihm verbunden. Unter dem Mikroskop kann man nun ein fein gerastertes Gitter mit Fluoreszenzsignalen sehen und mit einer hochempfindlichen CCD-Kamera aufnehmen. Die unterschiedlichen Farbpunkte liefern dann eine unterschiedliche Fluoreszenzintensität oder Farbe. Die Intensität der einzelnen Felder hängt wiederum von der Menge an gebundener RNA ab. So kann die Expressionsrate mehrerer Tausend Gene auf einen Schlag getestet werden. \square

Leider machen diese kleinen Viecher auch mal etwas Mist. Bei der Durchführung von Microarrayexperimenten gibt es, wie bei allen anderen Experimenten auch, einige Fehlerquellen, die man bei der Auswertung beachten sollte. Zwei wichtige sind z. B. die Abhängigkeit der Fluoreszenz von der RNA-Menge in der Probe oder die Eigenreflexion und Quencheigenschaften der Chipoberfläche. Um Ergebnisse unterschiedlicher Microarrays vergleichbar zu machen, müssen diese normalisiert werden. Eine effektive und

einfache Methode hierfür ist die am Deutschen Krebsforschungszentrum entwickelte Methode von Tim Beissbarth [20]. Hierbei wird das 0,05-Quantil der Datenerhebung von allen Messwerten abgezogen. So kann der Effekt der Lichtreflexion der Oberfläche auf die Daten kompensiert werden. Man nimmt an, dass es einen Grundwert gibt, der genau so groß ist wie das 0,05-Quantil und der nur auf Reflexionen zurückzuführen ist. Dies ist eine Annahme, die nicht auf jeden einzelnen Wert genau zutreffen muss, im Schnitt den Datensatz aber etwas näher an die Wahrheit rückt.

2.2.2 Streuungsmaße

In Abschn. 2.2.1 haben wir Kennzahlen für Messreihen kennengelernt. Oft ist eine bloße Kennzahl der Lage einer Verteilung aber nicht sehr aussagekräftig. So werden Mittelwerte oder Mediane oft mit der dazugehörigen Streuung publiziert. Mit Streuung ist der Schwankungsbereich um den Mittelwert von Messwerten bzw. den Lagemaßen gemeint. Hat ein Mittelwert eine sehr kleine Streuung, so kann man davon ausgehen, dass er die Mitte einer Verteilung zuverlässig kennzeichnet. Denn auch wenn Datensätze identische Lagemaße haben, müssen sich diese nicht genau gleichen. Im Folgenden werden wir deshalb einige Methoden zur Ermittlung der Streuung von Messdaten vorstellen.

Spannweite

Das Streuungsmaß, das am einfachsten zu berechnen ist, ist die Spannweite. Hierbei ziehen wir einfach den kleinsten gemessenen Wert vom größten gemessenen Wert ab:

$$\text{Spannweite} = x_{\max} - x_{\min}. \quad (2.3)$$

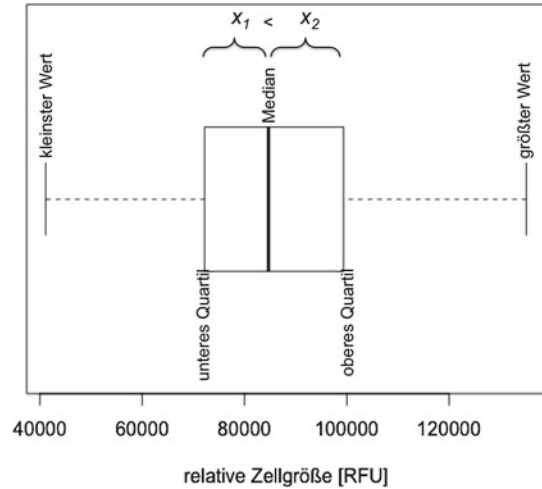
Eine Spannweite von 2 μm , sprich die kleinste und die größte Zelle haben nur einen Größenunterschied von 2 μm , würde bedeuten, dass unsere Zellen eine sehr homogene Größe haben und damit vielleicht ähnliche morphologische Eigenschaften. Es könnte aber auch bedeuten, dass sich alle Zellen in der gleichen Phase des Zellzyklus befinden (der Interpretationsspielraum ist groß). In unserem Datensatz beträgt die Spannweite $41.125 - 135.168 = 94.016$.

Natürlich ist die Einheit dieser Werte nicht Mikrometer sondern relative Lichtintensität (*relative fluorescence unit*, RFU). Auf den ersten Blick ist dies eine unheimlich große Spannweite, allerdings ist sie auch nicht weiter verwunderlich, schließlich haben wir die Zellgröße nur indirekt gemessen und müssten noch eine Kalibrierung durchführen, um zu wissen, wie man RFU in Mikrometer umrechnet.

Interquartilsabstand

In Abschn. 2.2.1 haben wir uns bereits mit Quantilen beschäftigt. Diese, vor allem die Quartile, können auch für die Ermittlung der Streuung wichtig sein. Bildet man die Differenz

Abb. 2.3 Ein Boxplot für die relative Zellgröße



zwischen dem oberen und unteren Quartil, erhält man den (Inter-)Quartilsabstand. Er enthält die zentralen 50 % aller Messwerte und ist deshalb ein wichtiges Streuungsmaß:

$$\text{Interquartilsabstand} = Q_{0,75} - Q_{0,25}. \quad (2.4)$$

In vielen Publikationen im Bereich der Naturwissenschaften trifft man auf Plots, in denen Quantile eingezeichnet sind. Einer dieser sogenannten Boxplots ist in Abb. 2.3 gezeigt.

Solch ein Boxplot verschafft dem Betrachter einen schnellen Überblick über eine Messreihe beziehungsweise über verschiedene Lageparameter dieser Reihe. Hierbei ist der Kasten in der Mitte von zentraler Bedeutung (und auch namensgebend). Er stellt den Interquartilsabstand dar. Der meist dickere Strich in der Mitte steht, je nach Publikation, für den Median (das 0,5-Quantil) oder den Mittelwert der Verteilung. Für die äußersten zwei Striche, auch Whisker genannt, gibt es leider keine wirklich klaren Konventionen. Manchmal stellen sie die äußersten Werte, also die Spannweite, dar, in anderen Darstellungen das 2,5-Quantil und das 97,5-Quantil.

Außer dem Interquartilsabstand können natürlich auch beliebige andere Quantilsabstände gebildet werden. Der Bereich zwischen dem 0,2-Quantil und dem 0,8-Quantil z. B., enthält die zentralen 60 % aller Messpunkte. Je größer man den Quantilsabstand wählt, umso größer ist auch das Risiko, Ausreißer in der Streumaßberechnung zu berücksichtigen.

Varianz und Standardabweichung

Sowohl bei der Ermittlung der Spannweite als auch bei der Ermittlung des Interquartilsabstands wurde immer nur Bezug auf einige genau definierte Messwerte einer geordneten Datenreihe genommen. Es leuchtet wohl schnell ein, dass eine Einbeziehung aller ermittelten Werte zur Beschreibung der durchschnittlichen Schwankung Sinn macht. Ein häufig

verwendetes Streuungsmaß ist daher die Standardabweichung bzw. die Varianz. Diese ermittelt die Streuung unter Einbeziehung aller Messpunkte. Um diese Streuungsmaße genau zu verstehen, wollen wir ihre Berechnung in einzeln abgegrenzte Schritte unterteilen:

1. Da wir alle ermittelten Werte mit einbeziehen wollen, bietet es sich an, zunächst den gesamten Abstand einzelner Messwerte zum Mittelwert (μ) zu ermitteln. Hierfür zieht man von jedem Wert den Mittelwert ab, erhält also deren Abstand, und summiert alle Abstände auf:

$$\sum_{i=1}^n (x_i - \mu).$$

Hierbei ergeben sich allerdings zwei grundlegende Probleme. Da manche Werte größer, andere wiederum kleiner als der Mittelwert sind, erhält man auch Abstände, die ein negatives Vorzeichen haben. Diese heben wiederum positive Werte auf. Um dieses Problem zu umgehen, kann man die erhaltenen Abstände jeweils quadrieren:

$$\sum_{i=1}^n (x_i - \mu)^2.$$

Somit werden auch negative Werte „positiviert“. Diese Summe wird **Summe der Abweichungsquadrate** genannt.

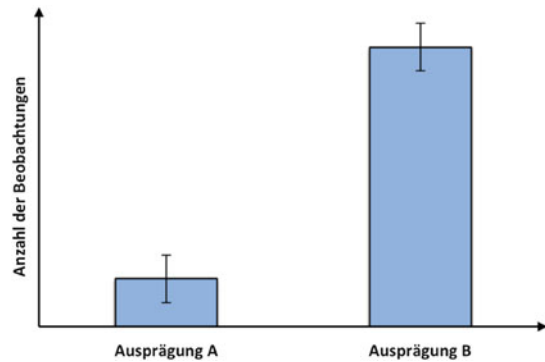
2. Das zweite Problem ergibt sich dadurch, dass diese Summe, also die so ermittelte Streuung, mit zunehmender Größe des Datensatzes automatisch größer wird. Das macht natürlich keinen Sinn, denn nur weil ein Datensatz groß ist, muss er keine größere Streuung aufweisen als kleinere Datensätze. Dieses Problem lässt sich einfach damit umgehen, dass man die erhaltene Summe der Abweichungsquadrate durch die Anzahl der Messwerte n teilt. Somit erzeugt man eine relative Streuung um den Mittelwert in Abhängigkeit von der Größe des Datensatzes. Diesen Wert nennt man Varianz S^2 :

$$S(x)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \quad (2.5)$$

Ein verbleibendes Problem ist nun, dass wir bei der Quadrierung der Abstände von x_i zu μ auch die Dimensionen, also die Einheiten, der Messwerte quadriert haben. In unserem Fall würden wir also die Varianz in der Einheit RFU^2 bzw. μm^2 angeben. Dies macht biologisch und vor allem physikalisch in diesem Kontext keinen Sinn. Deshalb wird die Varianz in solchen Fällen radiziert; es wird ihre Wurzel gezogen. Durch diesen Vorgang erhält man die Standardabweichung $\sqrt{S(x)^2} = S(x)$:

$$S(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}. \quad (2.6)$$

Abb. 2.4 Darstellung von zwei Merkmalsausprägungen in einem Balkendiagramm mit dazugehörigen Fehlerindikatoren



In vielen wissenschaftlichen Publikationen werden Messwerte mit dazugehörigen Standardabweichungen angegeben. In vielen Fällen geschieht dies durch Fehlerbalken in Diagrammen. Diese müssen nicht immer der Standardabweichung beziehungsweise der Varianz entsprechen, es ist allerdings in biologisch ausgerichteten Publikationen die Regel. Ein Balkendiagramm mit dazugehörigen Fehlerbalken ist in Abb. 2.4 gezeigt. Solche Balkendiagramme werden oft verwendet, wenn die Ausprägung eines Merkmals unter verschiedenen Bedingungen gezeigt werden soll. So könnte in dieser Darstellung das untersuchte Merkmal einem bestimmten Hormonrezeptor entsprechen, dessen Ausprägung im Fall A bei weiblichen Personen und im Fall B bei männlichen Personen ermittelt wurde. Die Höhe der Balken entspricht hierbei dem Mittelwert an gemessenen Rezeptorzahlen eines bestimmten Zelltyps, die kleinen Striche an der Spitze der Balken geben die Varianz der Ausprägungen an.

Diese Darstellung verleitet schnell dazu, konkrete Schlüsse bezüglich der Relevanz des Einflussfaktors, in unserem Beispiel das Geschlecht, zu ziehen. Im Extremfall ist die Varianz, also die Länge der Fehlerindikatoren, um ein Vielfaches größer als der eigentliche Balken. Überschneiden sich die Fehlerindikatoren über einen großen Bereich, ist das ein Hinweis dafür, dass der Einflussfaktor keinen Effekt auf das Merkmal hat. Allerdings sollte hier nicht zu vorschnell interpretiert werden. Um solche Schlüsse zu ziehen, sollte man stets auf Relevanztests der induktiven Statistik zurückgreifen (s. Abschn. 4.5, S. 121).

Standardabweichung und Varianz

Standardabweichung und Varianz sind Maße für die Streuung der Ausprägungen eines Merkmals um den Mittelwert.

2.3 Kenngrößen für den Zusammenhang von Merkmalen

2.3.1 Korrelation

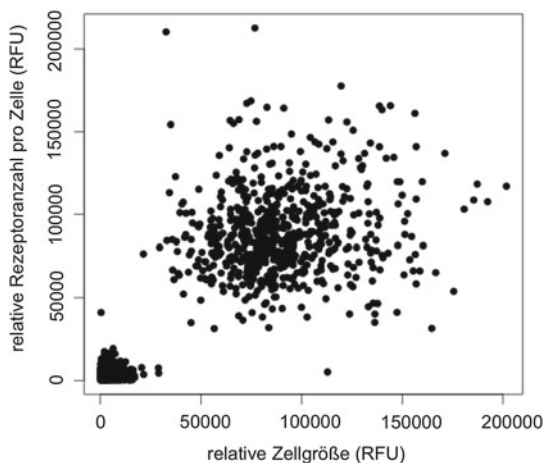
„Dr. Arnold, kommen Sie schnell mal ans Mikroskop! Schauen Sie, je mehr Luciferin ich zu den luciferasetransformierten Zellen dazugebe, umso stärker leuchten sie“. Zunächst wohl keine sehr überraschende Erkenntnis, wenn man weiß, dass Luciferase ein Enzym ist, welches eine Reaktion katalysiert, bei der unter Verwendung von Luciferin ein Leuchtsignal emittiert wird.

Aber genau um solche Fragestellungen drehen sich viele Experimente im Labor. Denn in der Biologie sind in vielen Fällen nicht die absolut gemessenen Werte von zentraler Bedeutung, sondern vielmehr der Zusammenhang zwischen zwei Merkmalen und deren Ausprägungen. Oft möchte man wissen, ob die Ausprägung eines Merkmals durch Veränderung der Ausprägung eines zweiten Merkmals beeinflusst wird und am besten noch in welcher Form. Man fragt daher nach einer **Korrelation** dieser zwei Merkmale, also nach der zwischen ihnen bestehenden **Beziehung**. Diese kann von unterschiedlicher Natur sein. Merkmale können **linear korrelieren**. Hierbei vermindert sich die Ausprägung von Merkmal A in gleicher Weise wie die von Merkmal B. Aber auch eine **nichtlineare Korrelation** kann bestehen. Dies ist z. B. der Fall, wenn sich die Ausprägung von Merkmal A zunächst stark mit der Änderung von Merkmal B verändert, dann ein Plateau erreicht und anschließend wieder abfällt. Eine **einseitige Abhängigkeit** bedeutet, dass Merkmal 1 (fortan M1) die Ursache für Merkmal 2 (M2) ist. So ist die Streuung des Lichts im Durchflusszytometer direkt abhängig von der Granularität der vorbeifließenden Zellen. Eine **wechselseitige Abhängigkeit** liegt vor, wenn M1 die Ursache für M2, allerdings M2 auch die Ursache für M1 darstellt. So beeinflusst die Glucosekonzentration in einem Nährmedium die Anzahl der wachsenden Krebszellen. Die Anzahl der Krebszellen ändert jedoch auch die Glucosekonzentration, weil der Zucker von den Krebszellen verbraucht wird. Eine Korrelation kann auch errechnet werden, wenn zwei Merkmale nur eine gemeinsame Ursache haben. So ist es gut möglich, dass die Glucosekonzentration in einem Wachstumsmedium mit der Anzahl an apoptotischen (also absterbenden) Krebszellen korreliert. Dies bedeutet allerdings nicht, dass der Zucker fortan als Krebsmedikament eingesetzt werden kann. Die Zellen vermehren sich jedoch mit einer erhöhten Zuckerdosis schneller und bei mehr Zellen ist auch die absolute Zahl an sterbenden Zellen größer. Hier wäre es wichtig, den relativen Anteil apoptotischer Zellen zu bestimmen, also durch die Gesamtzahl der Zellen zu teilen und den Wert somit entsprechend zu skalieren.

Nur weil eine mathematische Korrelation vorliegt, müssen die untersuchten Merkmale noch lange nicht in einem kausalen (Ursache-Wirkungs-)Zusammenhang zueinander stehen. Korrelationsauswertungen sollten daher immer genauestens überprüft werden, denn eine Korrelation gibt keine Aussage über **Kausalität**.

Bei Datensätzen mit nur wenigen Messpunkten können oft bereits mit bloßem Auge korrelierende Merkmale ausfindig gemacht werden. Zieht man allerdings den Durchfluss-

Abb. 2.5 Eine Punktwolke für die relativen Zellgrößen und Rezeptorzahlen. Die Häufung in der linken unteren Ecke geht auf die schon vorher beschriebenen Schmutzpartikel zurück



zytometriedatensatz mit solch einer großen Anzahl an Messpunkten heran, ist es leider sehr schwer, auf Anhieb die Beziehungen zwischen den zwei Merkmalen zu erkennen. Blickt man auf die Punktwolke, wenn die Zahl an Rezeptoren gegen die Zellgröße aufgetragen ist, so lässt sich der Zusammenhang nur schwerlich beurteilen (s. Abb. 2.5).

Abhilfe bringen hier verschiedene Korrelationsverfahren, mit denen man versucht herauszufinden, in welcher Abhängigkeit Merkmale zueinander stehen. Um die Korrelation zweier Merkmale zu beziffern, wird ein Korrelationskoeffizient gebildet. Dieser kann einen Wert zwischen -1 und $+1$ annehmen. Der Wert dieses Koeffizienten quantifiziert hierbei die Linearität der Korrelation. Das bedeutet, je näher der Wert an -1 oder $+1$ liegt, umso perfekter korrelieren die Merkmale linear. Beträgt der Wert hingegen 0 , so besteht keine lineare Korrelation. Das Vorzeichen des Korrelationskoeffizienten gibt hierbei die Art der Abhängigkeit an, also ob M_2 bei Zunahme von M_1 ebenfalls zunimmt (positives Vorzeichen) oder ob M_2 bei Zunahme von M_1 abnimmt (negatives Vorzeichen). In Abb. 2.6 sind einige Punktdiagramme mit dazugehörigen Korrelationskoeffizienten (r) gezeigt.

Die Grundlage zur Berechnung des Korrelationskoeffizienten sind Messwertpaare. An einem Merkmalsträger wird sowohl die Ausprägung x (z. B. Zellgröße) sowie die zugehörige Ausprägung y (z. B. Rezeptorzahl pro Zelle) desselben Merkmalsträgers i gemessen.

Zur Ermittlung des Korrelationskoeffizienten werden zunächst die standardisierten Daten betrachtet:

$$\tilde{x}_i = \frac{x_i - \mu_x}{S_x} \quad \tilde{y}_i = \frac{y_i - \mu_y}{S_y} \quad \text{mit} \quad \mu_x, \mu_y: \text{Mittelwert und } S_x, S_y: \text{Standardabweichungen.}$$

Hat ein Wertepaar hier das gleiche Vorzeichen, spricht dies für eine positive Beziehung zwischen den beiden Merkmalen. Haben sie hingegen unterschiedliche Vorzeichen, könnte dies auf eine negative Beziehung hinweisen. Um diese Beziehung mathematisch auch sinnvoll auszudrücken, kann das Produkt dieser Wertepaare gebildet werden. Möchte man

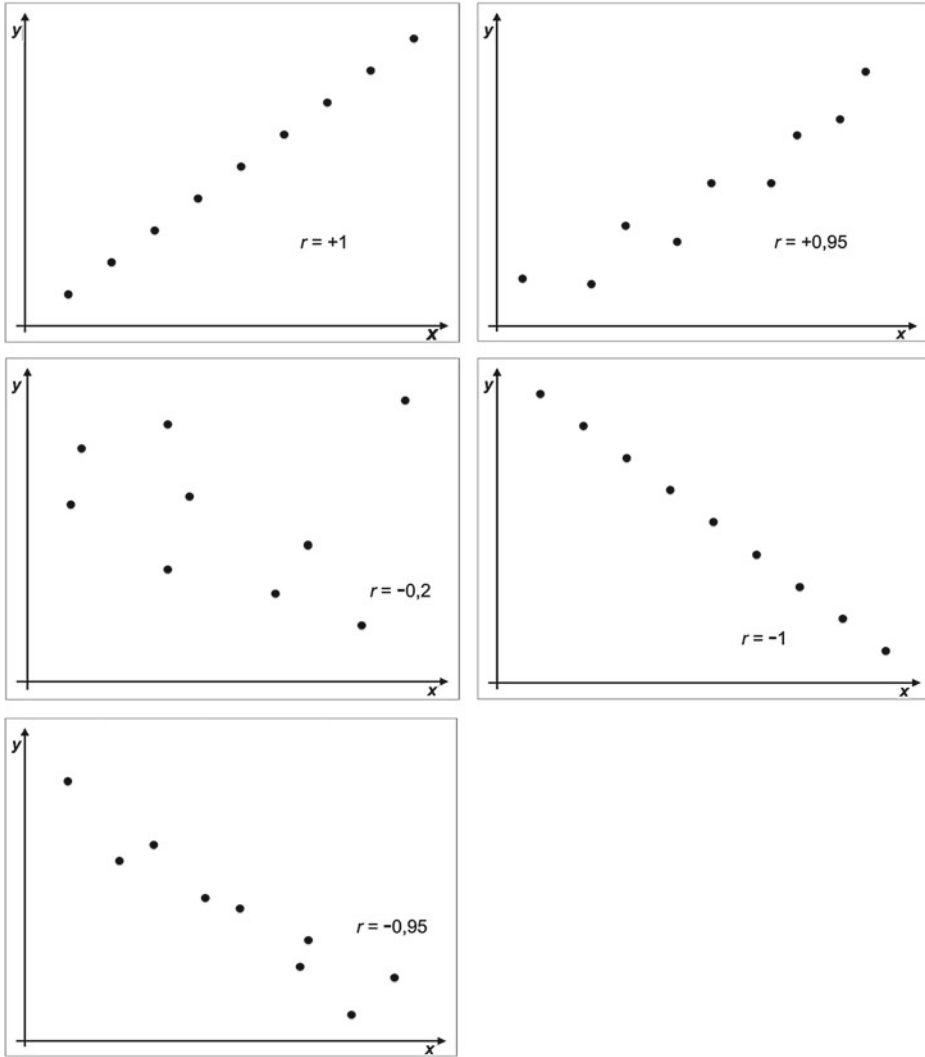


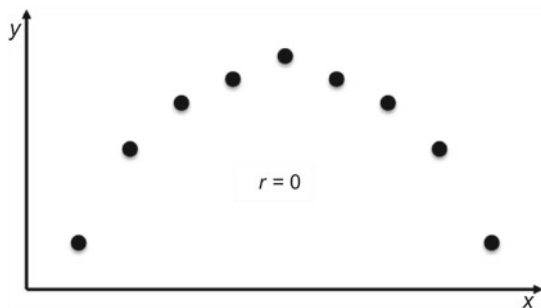
Abb. 2.6 Datenwolken mit den dazugehörigen r -Werten. Dem Vorzeichen ist zu entnehmen, welches Merkmal das andere negativ oder positiv beeinflusst

diese Beziehung nicht nur für einzelne Wertepaare ermitteln, werden alle Datenpunkte ausgemittelt:

$$r(x, y) = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i \cdot \tilde{y}_i.$$

Damit ergibt sich die Formel für den Korrelationskoeffizienten:

Abb. 2.7 Zwei nicht linear korrelierende Merkmale mit linearem Korrelationskoeffizienten. Obwohl klar ersichtlich ein Zusammenhang besteht, beträgt der lineare Korrelationskoeffizient 0



$$r(x, y) = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x) \cdot (y_i - \mu_y)}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \mu_y)^2}} = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \mu_y)^2}}$$

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_x)^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \mu_y)^2}}. \quad (2.7)$$

Die Gleichung sieht auf den ersten Blick sehr kompliziert aus, im Endeffekt müssen allerdings nur viele Summen gebildet werden. Im Nenner steht die Wurzel aus dem Produkt der Standardabweichungen für x und y , wohingegen im Zähler die Kovarianz der beiden Merkmale aufgeführt ist. Der Koeffizient r beschreibt, wie viel Streuung in y durch die Streuung in x hervorgerufen wird. Glatte r -Werte ($-1, 0, +1$) ergeben sich in der Praxis so gut wie nie, da Messdaten, auch wenn eine strenge lineare Korrelation vorliegt, aufgrund von Messfehlern immer etwas von dem perfekten linearen Zusammenhang abweichen. Der lineare Korrelationskoeffizient für die Datenwolke aus Abb. 2.5 beträgt übrigens $\approx 0,79$.

Wie erwähnt, gibt es neben einer linearen Korrelation aber auch viele andere Korrelationstypen. Diese lassen sich mit der oben vorgestellten Methode allerdings nicht quantifizieren. Ein r -Wert, der nahe bei 0 liegt, ist daher kein Todesurteil für eine Korrelation. Aus Symmetriegründen kann solch eine Korrelation oft übersehen werden. In Abb. 2.7 ist ein Punktediagramm gezeigt, in dem die gegeneinander aufgetragenen Merkmale eindeutig korrelieren, allerdings eben nicht linear. Würden wir nur den Korrelationskoeffizienten für eine lineare Korrelation ($r = 0$) berücksichtigen, würden wir unsere Ergebnisse vielleicht fälschlicherweise in den Müll werfen.

Korrelationskoeffizient

Der Korrelationskoeffizient gibt eine quantifizierbare Aussage über eine bestehende oder nicht bestehende lineare Beziehung zwischen den Ausprägungen von zwei Merkmalen. Liegt der sich hierfür ergebende r -Wert nahe an $+1$ oder -1 , liegt eine gute lineare Korrelation vor. Liegt er nahe an 0, liegt eine äußerst schlechte lineare Korrelation vor.

2.3.2 Lineare Regression

Nicht nur der gegenseitige Einfluss von zwei Merkmalen kann von Bedeutung sein, vielmehr ist in biologischen Experimenten der Einfluss einer Einflussgröße auf ein Merkmal von Interesse. So sind Krebszellen oftmals deutlich größer als nicht entartete Zellen. Wenn man nun selektiv und abgestuft ein Onkogen überexprimiert und anschließend die relative Zellgröße mittels Durchflusszytometrie misst, kann der Korrelationskoeffizient nicht zur Beurteilung der Abhängigkeit herangezogen werden. Zwar handelt es sich um zwei unterschiedliche Merkmale, die gleichzeitig an einem Merkmalsträger gemessen wurden, jedoch wurde eines dieser Merkmale systematisch variiert (Expression des Onkogens) und stellt damit keine Stichprobe mehr dar. Dies ist allerdings eine essenzielle Bedingung zur legitimen Ermittlung des Korrelationskoeffizienten. Das variierte Merkmal wird als unabhängige Variable beziehungsweise Einflussgröße bezeichnet, das gemessene Merkmal als abhängige Variable oder Zielgröße. Die Regression bietet eine Möglichkeit, den Zusammenhang zwischen Einflussgröße und Zielgröße zu quantifizieren. Es besteht die Möglichkeit, durch Regressionsanalysen auch andere Zusammenhänge zu bestimmen, z. B. solche, bei denen keine lineare Abhängigkeit vorliegt. Allerdings sind diese Zusammenhänge höheren Grades oft sehr komplex, sodass wir hier nur auf die lineare Regression eingehen werden.

Eine möglichst treffende Beschreibung des Zusammenhangs zwischen den Messwerten wäre eine Gerade, die einen minimalen Abstand zu allen Punkten hat. Man könnte einfach per Hand eine Gerade durch die Punktwolke ziehen. Die optimale Gerade dabei zu finden, ist unmöglich. Die mathematische Lösung dieses Problems stellt ein Optimierungsproblem dar. Welche Gerade hat den geringsten Abstand in y -Richtung zu allen Punkten, wenn man die quadrierten Abstände der Punkte zur Gerade aufsummiert? Die Grundlage der Optimierung ist hierbei die Geradengleichung $y = ax + b$, wobei a die Steigung der Gerade und b den y -Achsenabschnitt darstellt (s. Abschn. 1.2, S. 4 zur Erläuterung). Mit diesen Formeln lassen sich die beiden Faktoren für einen gegebenen Datensatz ermitteln:

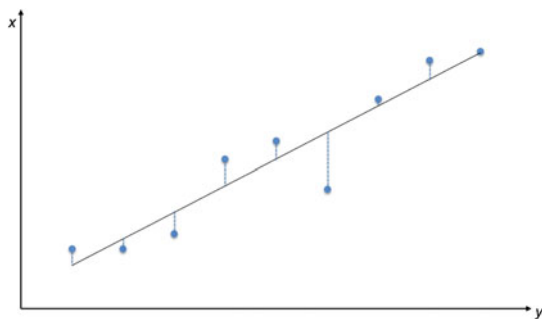
$$a = \mu_y - b\mu_x = \frac{\sum y_i}{n} - b \frac{\sum x_i}{n}$$

$$b = \frac{\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=1}^n (x_i - \mu_x)^2}. \quad (2.8)$$

Ein Beispiel für solch eine Regressionsgerade ist in Abb. 2.8 gezeigt. Die Gerade stellt den optimalen linearen Zusammenhang der beiden Merkmale dar. Die gepunkteten Linien stellen die Abstände der Datenpunkte zur Gerade dar. Es gibt keine andere Gerade, bei der die Summe der quadrierten Längen der gepunkteten Linien kleiner ist.

Für die in Abb. 2.8 gezeigten Werte ist die Regressionsgerade eingezeichnet. Wie zu sehen ist, liegen einige der Punkte unterhalb und andere oberhalb der Gerade. Das ist nicht verwunderlich, schließlich war das Ziel die Minimierung der quadrierten Abstände. Auf welcher Seite der Gerade die Punkte liegen, ist hierbei nicht relevant. Da die Abstände

Abb. 2.8 Datenpunkte zweier Merkmale mit dazugehöriger Regressionsgerade. Die gepunkteten Linien zeigt die Linie an, bei der der quadrierte Abstand minimal wird



quadriert werden, heben die Abstände von Punkten unter der Linie die Abstände von Punkten über der Linie nicht auf.

Ermittelt man die Steigung der Gerade (a oder auch Regressionskoeffizient genannt), ergibt sich eine quantifizierbare Aussage über den Einfluss der variablen Größe auf die Beobachtungsgröße. a sagt dabei aus, dass die Beobachtungsgröße pro Steigerung der variablen Größe um eine Einheit, um a ansteigt. Die Quantifizierung einer solchen Beziehung nach dieser Methode wird auch **Methode der kleinsten Quadrate** genannt. Dem aufmerksamen Leser wird aufgefallen sein, dass es durchaus kürzere Linien von den Punkten zu der Gerade gibt. Wir haben uns hier auf vertikale Linien beschränkt. Ohne diese Beschränkung können genauere Aussagen getroffen werden, allerdings ist in diesem Fall die Berechnung auch deutlich schwerer.

Lineare Regression

Die lineare Regression ermittelt die Geradengleichung $y = m \cdot x + b$, bei der die Summe der quadrierten Abstände zu den Datenpunkten minimal ist.

2.4 Aufgaben

- A1** Bilde für die folgenden kleinen Datensätze Mittelwert, Modalwert, Median, 0,25-Quantil und 0,75-Quantil.

Datensatz 1: 3 5 7 2 5 2 5 7 9 3 9 5 7 3 0 5 7 3

Datensatz 2: 4 7 8 2 6 5 9 0 2 7 4 8 6 2 7 4 8

- A2** Zeichne ein Histogramm für folgenden Datensatz und ermittle Mittelwert sowie den Median.

25 5 7 2 25 5 2 25 5 7 9 3 23 9 5 3 5 7 3

Croppe nun den Datensatz und ermittle beide Werte nochmals. Warum und in welcher Weise (wie stark beziehungsweise in welche Richtung) ändern sich die Werte?

- A3** Zeichne die jeweiligen Boxplots für folgende Datensätze (Spannweite des Datensatzes für die Whisker) und vergleiche sie hinsichtlich der Auswirkungen von Ausreißern.

Datensatz 1: 3 6 7 2 4 6 3 9 4 8 7 2 5 4 8 1 7 3 6 4 8 2 7 4 4 2

Datensatz 2: 3 6 23 4 8 3 6 4 2 5 5 1 3 7 4 9 3 18 6 2 9 25 3 6 5

Welche Datenpunkte würdest du als Ausreißer betrachten? Welche Parameter ändern sich, wenn diese in den Berechnungen betrachtet werden?

- A4** Bilde für folgende Datensätze Mittelwert, Varianz und Standardabweichung. Zeichne anschließend ein Balkendiagramm, in dem alle drei Datensätze zusammengefasst werden, mit dazugehörigen Fehlerbalken.

Datensatz A: 3 2 5 7 4 6 3 5

Datensatz B: 2 4 2 3 1 4 3 2

Datensatz C: 2 5 3 4 3 5 25 4

- A5** Zeichne für folgende Datensätze die dazugehörigen Punktwolken (Punktdiagramme) und schätze einen Korrelationskoeffizienten. Berechne anschließend den Korrelationskoeffizienten (runde wenn nötig auf drei Stellen hinter dem Komma).

Ausprägungen von Merkmal A: 1 2 3 4 5 6

Ausprägungen von Merkmal B: 1 2 3 4 5 6

Ausprägungen von Merkmal C: 1 2 3 4 5 6

Ausprägungen von Merkmal D: 0,5 1 2,5 3 3,5 4

Ausprägungen von Merkmal E: 1 2 3 4 5 6

Ausprägungen von Merkmal F: 4,5 4,5 3,5 3 3 2

Um welche Form von Abhängigkeiten handelt es sich in den einzelnen Fällen?

- A6** Zeichne für folgende Datensätze die dazugehörigen Punktdiagramme, in denen jeweils die beiden Ausprägungen gegeneinander aufgetragen werden. Zeichne nun per Hand eine Linie ein, welche nach deiner Einschätzung die beste Regressionsgerade darstellt (kleinste Summe der quadrierten Abstände zur Kurve). Berechne anschließend die Regressionsgerade und zeichne diese in dieselbe Abbildung ein (runde auf eine Stelle hinter dem Komma).

Ausprägungen A: 1 2 3 4 5 6 7 8 9

Ausprägungen B: 2 2 2,5 5 5,6 4 7 8,3 8,6

Tutorium Mathe für Biologen

Von Studenten für Studenten

Adlung, L.; Hopp, C.; Köthe, A.; Schnellbacher, N.;

Staufer, O.

2014, X, 287 S. 66 Abb., 53 Abb. in Farbe., Softcover

ISBN: 978-3-642-37785-3