

2 Grundlagen der Statistik

Wir haben nun das theoretische Handwerkszeug beisammen, um uns praktischen Fragen der Art zu widmen, wie wir sie zu Anfang des ersten Kapitels gestellt haben: Wenn Sie beispielsweise einen Werkstoff hergestellt haben, wollen Sie wissen, ob er die Anforderungen erfüllt, das heißt, ob er etwa die richtige Biegesteifigkeit hat oder ob diese vom Sollwert abweicht. Wie viele Proben sollen Sie vermessen, um eine verlässliche Aussage treffen zu können? Gibt es einen signifikanten Unterschied in der Scherfestigkeit zwischen einem Werkstoff und einem anderen, oder sind die Unterschiede bloß Messungenauigkeiten? In diesem Kapitel lernen wir, wie wir *anhand von Messdaten Aussagen treffen*, wie wir sie mit Wahrscheinlichkeiten verbinden und was wir überhaupt mit Ausdrücken wie *signifikant* meinen. Das ist Statistik.

Die Aufgabe der Statistik besteht darin, aus Daten Informationen zu destillieren. Dies kann auf verschiedene Arten geschehen. Wir unterscheiden zwischen *deskriptiver* (beschreibender) und *inferentieller* (schließender) Statistik. In der deskriptiven Statistik geht es darum, *geeignete Zusammenfassungen* der Daten zu finden, die die wesentlichen Aspekte wiedergeben; das können sowohl numerische als auch grafische Zusammenfassungen sein. In der inferentiellen Statistik geht es darum, aus den Daten *Rückschlüsse auf die Mechanismen* zu ziehen, die die Daten erzeugt haben, und zum Beispiel *Vorhersagen* für die Zukunft zu treffen.

Betrachten wir als Beispiel die Abschlussklausur der Vorlesung Mathematik I des letzten Semesters. Zunächst haben wir die erreichten Punktzahlen $x_1, \dots, x_n$, wobei $n = 647$ die Anzahl der Klausurteilnehmer ist. Das ist ein immenses Datenmaterial, in dem die relevanten Informationen, die uns interessieren, verborgen liegen, etwa: Wie viele Studenten haben eine gute Klausur geschrieben, wie viele sind durchgefallen, wie ist die Klausur ausgefallen? Wir können zunächst versuchen, diese Informationen durch geeignete numerische und grafische Zusammenfassungen sichtbar zu machen. Das ist deskriptive Statistik. Darüberhinaus können wir uns fragen, welche Zufallsmechanismen die Klausurergebnisse hervorgebracht haben. Wir fassen dazu die erreichten Punkte $x_1, \dots, x_n$ als Realisierungen von Zufallsvariablen $X_1, \dots, X_n$ auf und fragen uns: Was können wir auf Grundlage der Daten $x_1, \dots, x_n$ über die Verteilung dieser Zufallsvariablen aussagen? Können wir Notendurchschnitt und Durchfallquote für die nächste Klausur vorhersagen? Zumindest in einem gewissen Rahmen? Das ist inferentielle Statistik.

In diesem Kapitel werden wir kurz die wichtigsten Werkzeuge vorstellen, um Daten übersichtlich zusammenzufassen, etwa *Mittelwert* und *empirische Varianz*,

die die Lage und die Streuung der Daten beschreiben, oder den *Boxplot*, der sie graphisch veranschaulicht. Wir lernen *Schätzer* kennen, mathematische Funktionen, in die wir Messdaten einsetzen, um etwas über die Natur der Daten, das heißt über die hinter ihnen stehenden Zufallsvariablen, zu erfahren (zum Beispiel ihren Erwartungswert oder bestimmte Wahrscheinlichkeiten). Es gibt gute und schlechte Schätzer, und wir werden Methoden behandeln, um *Schätzer zu vergleichen*. Mithilfe von *Konfidenzintervallen* werden wir angeben können, wie gut die Schätzung eines unbekannten Parameters den wahren Parameter trifft, und mit *Signifikanztests* können wir schließlich *Vermutungen über Daten bewerten*: Unterscheiden sich zwei Datensätze in ihrer Art? Sind bei zweidimensionalen Daten beide Merkmale unabhängig? Wie stark streuen die Messwerte? Wir lernen, wie solche Tests funktionieren und wie man verschiedene Tests miteinander vergleicht, und wir behandeln viele verschiedene Testverfahren anhand konkreter Beispiele.

2.1 Deskriptive Statistik

Daten übersichtlich zusammenfassen

In der deskriptiven Statistik, auch beschreibende Statistik genannt, sucht man nach geeigneten Zusammenfassungen, die die wesentlichen Aspekte der Daten wiedergeben. Zum einen gibt es *numerische Zusammenfassungen* wie den *Mittelwert* und den *Median*, die die Lage der Daten angeben, die *empirische Varianz* oder den *Interquartilsabstand*, die beschreiben, wie die Daten streuen; zum anderen können wir mit *graphischen Zusammenfassungen* wie einem *Boxplot* und einem *Histogramm* den Datensatz bildlich veranschaulichen. Wir lernen die *empirische Verteilungsfunktion* kennen, mit der wir die wahre Verteilungsfunktion, die die Daten generiert hat, schätzen können. Mithilfe des *Q-Q-Plots* können wir außerdem mit dem Auge die Verteilung eines Datensatzes mit einer vorgegebenen Verteilung vergleichen.

Im Rahmen dieses Buches werden wir uns bloß am Rande mit deskriptiver Statistik befassen und nur die wichtigsten Verfahren behandeln. Für ein gutes Verständnis ist es dennoch wichtig, sie mithilfe eines Rechners selbst anzuwenden – versuchen Sie es einfach, zum Beispiel in Tabellenkalkulationsprogrammen wie **Excel** oder in Statistiksoftware wie **R**.

2.1.1 Numerische Zusammenfassungen eindimensionaler Daten

Einen Datensatz durch Kennzahlen beschreiben

Wir wollen n Daten $x_1, \dots, x_n \in \mathbb{R}$ analysieren. Es ist oft hilfreich, die Daten dazu der Größe nach zu sortieren, denn durch die sortierten Daten lassen sich wichtige Kenngrößen ausdrücken.

Definition 2.1 (Ordnungsstatistik und Rang). Für Daten $x_1, \dots, x_n \in \mathbb{R}$ bezeichnen wir mit $x_{(1)}$ die kleinste Beobachtung, mit $x_{(2)}$ die zweitkleinste Beobachtung und so weiter und mit $x_{(n)}$ die größte Beobachtung. Der neue sortierte Datensatz

$$x_{(1)} \leq \dots \leq x_{(n)}$$

heißt *Ordnungsstatistik*. Tritt ein Wert in der Stichprobe mehrfach auf, so erscheint dieser in der Ordnungsstatistik mit derselben Vielfachheit.

Die Beobachtung mit Wert $x_{(i)}$ hat somit den *Rang* i erhalten. Der Rang der Beobachtung x_j wird mit r_j oder $r(x_j)$ bezeichnet. Treten mehrere gleiche Werte in der Stichprobe auf (sogenannte *Bindungen*), ist der Rang nicht eindeutig; häufig wird gleichgroßen Beobachtungen der Mittelwert der Ränge, die sie belegen, zugeordnet.

Beispiel 2.2. Sie haben nacheinander die Daten

7, 2, 24, 17, 7, 6, 7, 10, 16

gemessen, das heißt

$$\begin{aligned} x_1 = 7, \quad x_2 = 2, \quad x_3 = 24, \quad x_4 = 17, \quad x_5 = 7, \\ x_6 = 6, \quad x_7 = 7, \quad x_8 = 10, \quad x_9 = 16. \end{aligned}$$

Die zugehörige Ordnungsstatistik ist

$$\begin{aligned} x_{(1)} = 2, \quad x_{(2)} = 6, \quad x_{(3)} = x_{(4)} = x_{(5)} = 7, \\ x_{(6)} = 10, \quad x_{(7)} = 16, \quad x_{(8)} = 17, \quad x_{(9)} = 24, \end{aligned}$$

und die Ränge sind

$$\begin{aligned} r_1 = r(x_1) = 4, \quad r_2 = r(x_2) = 1, \quad r_3 = r(x_3) = 9, \\ r_4 = r(x_4) = 8, \quad r_5 = r(x_5) = 4, \quad r_6 = r(x_6) = 2, \\ r_7 = r(x_7) = 4, \quad r_8 = r(x_8) = 6, \quad r_9 = r(x_9) = 7. \end{aligned}$$

Beschreibung der Lage

Wir fassen einige wichtige Kennzahlen zusammen, die die Lage der Daten beschreiben.

Mittelwert

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n x_{(i)}$$

Der Mittelwert ist der Durchschnittswert der Daten. Schon ein einziger Messfehler (zum Beispiel ein verrutschtes Komma) kann den Mittelwert der Datenreihe so erheblich beeinflussen, dass er ein völlig falsches Bild liefert.

Median

$$\text{med}(x_1, \dots, x_n) = \begin{cases} x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2} \left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right) & \text{falls } n \text{ gerade} \end{cases}$$

Der Median teilt die Daten in eine obere und eine untere Hälfte und ignoriert dabei, wie groß die Daten sind. Das macht ihn unempfindlich gegen besonders große und kleine Messwerte (sogenannte *Ausreißer*), die vielleicht nur Messfehler sind: Ob der größte Messwert nun 2,5 oder 25 beträgt, er bleibt der größte, und damit bleibt die Trenngrenze, die die Daten in zwei Hälften teilt, auch die gleiche. Diese Eigenschaft nennen wir *robust (gegen Ausreißer)*.

Quartile

$$Q_1 = x_{(\frac{n+1}{4})}, \quad Q_3 = x_{(3\frac{n+1}{4})}$$

Ist $\frac{n+1}{4}$ keine ganze Zahl (was soll zum Beispiel die 3,75. Beobachtung sein?), so nehmen wir den entsprechenden gewichteten Mittelwert der beiden benachbarten Ordnungsstatistiken, etwa

$$x_{(3,75)} = \frac{1}{4}x_{(3)} + \frac{3}{4}x_{(4)},$$

da $3,75 = \frac{1}{4}3 + \frac{3}{4}4$.

Die Quartile Q_1 und Q_3 geben die Werte an, die von 25 % der Daten unter- beziehungsweise überschritten werden. Das Quartil Q_2 gibt den Wert an, der von der Hälfte der Daten überschritten wird – es ist also nichts anderes als der Median.

Bei der Darstellung als gewichteter Mittelwert aus der nächst kleineren und der nächst größeren Beobachtung müssen die Gewichte zusammen immer 1 ergeben.

Quartile sind Sonderfälle sogenannter *p-Quantile*. Ein *p-Quantil* teilt die (nach der Größe sortierten) Daten in zwei Teile, und zwar so, dass $100 \cdot p\%$ der Daten links davon und der Rest $(100 \cdot (1 - p)\%)$ rechts davon liegt. p ist eine reelle Zahl zwischen 0 und 1. Häufige Sonderfälle sind der Median (Halbwert, $p = 0,5$), die Quartile (Viertelwerte, $p = 0,25$, $p = 0,5$ und $p = 0,75$), die Quintile (Fünftelwerte, $p = 0,2$, $p = 0,4$, $p = 0,6$, $p = 0,8$), die Dezile (Zehntelwerte, $p = 0,1, 0,2, \dots, 0,9$) und die Perzentile (Hundertstelwerte, $p = 0,01, 0,02, \dots, 0,98, 0,99$).

Getrimmter Mittelwert

$$\bar{x}_\alpha = \frac{1}{[n(1 - 2\alpha)]} \sum_{i=[n\alpha+1]}^{[n(1-\alpha)]} x_{(i)},$$

wobei $[x]$ die ganze Zahl ist, die man durch Abrunden von x erhält, und $\lceil x \rceil$ die Zahl, die man durch Aufrunden von x erhält.

Das getrimmte Mittel ist der gewöhnliche Mittelwert, allerdings werden vor der Berechnung die kleinsten und größten $\alpha \cdot 100\%$ der Daten abgeschnitten und weggeworfen. Damit ist das getrimmte Mittel nicht mehr so anfällig für Ausreißer wie der Mittelwert, da extrem große und kleine Werte nicht miteinbezogen werden; es ist robust.

Beschreibung der Streuung

Als nächstes stellen wir Kennzahlen vor, die die Streuung der Daten beschreiben.

Empirische Varianz

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

beziehungsweise die empirische Standardabweichung $\sqrt{s_x^2}$.

Die empirische Varianz misst, wie stark die Daten x_i vom Mittelwert $\bar{x}$ abweichen. Die Abweichungen werden quadriert (damit sich positive und negative Abweichungen nicht zufällig gegeneinander aufheben) und aufsummiert. Durch das Quadrieren können Ausreißer besonders stark ins Gewicht fallen und die Varianz beeinflussen.

Es mag verwirren, dass wir die Summe am Schluss durch $n - 1$ teilen, und nicht durch n . Bei sehr vielen Daten, wenn n also sehr groß ist, macht es keinen Unterschied, aber es lässt sich mathematisch nachweisen, dass man mit dem Vorfaktor

$\frac{1}{n-1}$ im Schnitt die Varianz der Datenart realistischer beschreiben kann: s_x^2 ist ein sogenannter *erwartungstreu*er Schätzer der Varianz. Wir werden später noch sehen, dass die empirische Varianz der beste Schätzer der Varianz ist, wenn die Daten normalverteilt sind (siehe Definition 2.20).

Interquartilsabstand (IQR = interquartile range)

$$\text{IQR} = Q_3 - Q_1$$

Der IQR ist die Differenz aus dem oberen und dem unteren Quartil, also ist es die Länge des Bereichs, in dem die mittleren 50 % der Daten liegen. Es ist eine robuste Statistik.

Median absolute deviation (MAD)

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \text{med}(x_1, \dots, x_n)|$$

Die MAD, die mittlere absolute Abweichung vom Median, ist robuster gegen Ausreißer als etwa die empirische Varianz. Außerdem gibt es Verteilungen, für die Erwartungswert und Varianz überhaupt nicht existieren, etwa die *Cauchy-Verteilung*, und wenn wir Daten analysieren, die einer solchen Verteilung folgen, können wir die Streuung durch die MAD beschreiben.

Gelegentlich wird auch die *Mean Absolute Deviation* betrachtet, die absolute Abweichung vom Mittelwert; dabei wird in obiger Formel der Median durch den Mittelwert $\bar{x}$ ersetzt.

Beispiel 2.3. Es liegen Ihnen 19 Luftproben vor, die Sie auf Verschmutzung untersuchen. Sie messen, wie viel Kohlenstoffmonoxid (in ppm, *parts per million*) enthalten ist und erhalten diese Messreihe:

45, 30, 38, 42, 63, 43, 102, 86, 99, 63, 58, 34, 37, 55, 58, 153, 75, 58, 36

Diese Daten werden der Größe nach sortiert:

30, 34, 36, 37, 38, 42, 43, 45, 55, 58, 58, 58, 63, 63, 75, 86, 99, 102, 153

Jetzt können wir die Daten wie folgt zusammenfassen:

- Mittelwert $\bar{x} = 61,84$
- Median $\tilde{x} = x_{(10)} = 58,0$
- Quartile $Q_1 = x_{(20/4)} = x_{(5)} = 38$ und $Q_3 = x_{(3 \cdot 20/4)} = x_{(15)} = 75$
- Getrimmtes Mittel mit $\alpha = 10\%$

$$\bar{x}_\alpha = \frac{1}{[15,2]} \sum_{i=[2,9]}^{[17,1]} x_{(i)} = \frac{1}{15} \sum_{i=3}^{17} x_{(i)} = 57,07$$

- Empirische Varianz $s_x^2 = 934,92$
- Interquartilsabstand IQR = $75 - 38 = 37$
- Median absolute deviation MAD = $20,89$

2.1.2 Grafische Zusammenfassungen eindimensionaler Daten

Einen Datensatz auf einen Blick überschauen

Boxplot

In einem *Boxplot* werden die Kennzahlen Median, Quartile Q_1 , Q_3 und der Interquartilsabstand grafisch dargestellt. Außerdem werden *Ausreißer* erfasst, das sind Daten, die extrem von der Mehrheit abweichen und besonders betrachtet werden sollten (eventuell liegen hier grobe Messfehler vor).

Ein Boxplot wird so konstruiert: Die Box, das zentrale Rechteck des Boxplots, verläuft vom oberen (75 %) Quartil zum unteren (25 %) Quartil, das mittlere (50 %) Quartil, der Median, wird deutlich erkennbar in die Box eingezeichnet (der Median liegt keineswegs immer in der Mitte der Box; seine Lage hängt von der Form der Verteilung ab, die somit ebenfalls direkt aus dem Boxplot abgelesen werden kann). Da die Box zwischen dem oberen und dem unteren Quartil verläuft, entspricht ihre Länge genau dem Interquartilsabstand IQR.

Nun wird ein Abstand von jeweils 1,5 IQR auf die obere und die untere Kante der Box addiert, sodass sich ein Feld mit einer Gesamtlänge von 4 IQR ergibt. Zwei Werte, der größte und der kleinste beobachtete Wert, die noch in diesem Bereich von 4 IQR liegen, bilden nun die Grenzpunkte für den oberen und den unteren Zaun (whisker) des Boxplots, die jeweils durch eine Linie mit der Box verbunden werden. Zu beachten ist, dass die Zäune nicht an der Grenze von $\pm 1,5$ IQR um

die beiden Enden der Box liegen, sondern dort, wo der größte beziehungsweise der kleinste Wert der Verteilung innerhalb dieser beiden Abstände liegt. Deshalb sind die Zäune in der Regel nicht gleich lang.

Alle Werte, die außerhalb der Zäune liegen, sind Ausreißer und werden mit einem Kreis gekennzeichnet.

Während die Box eines Boxplots immer das erste und dritte Quartil umfasst, gibt es bei den Zäunen und Ausreißern verschiedene Konventionen: Manchmal werden die Ausreißer mit einem Punkt oder einem Stern markiert, selten auch gar nicht; die Zäune zeigen beispielsweise manchmal das Minimum/Maximum der Daten oder verschiedene Quantile an. Es ist daher ratsam, bei einem Boxplot anzugeben, nach welcher Konvention die Zäune eingezeichnet wurden.



Abb. 2.1: Boxplot der Daten aus Beispiel 2.3

Empirische Verteilungsfunktion

Wenn wir die Daten $x_1, \dots, x_n$ als Realisierungen einer Zufallsvariablen X betrachten, hat diese Zufallsvariable eine Verteilungsfunktion F_X , die wir aus den Daten heraus schätzen können.

Definition 2.4 (Empirische Verteilungsfunktion). Die *empirische Verteilungsfunktion* der Daten $x_1, \dots, x_n$ ist

$$F_n(x) = \frac{1}{n} \# \{1 \leq i \leq n : x_i \leq x\}.$$

$F_n(x)$ gibt an, welcher Anteil der Beobachtungen kleiner oder gleich x ist.

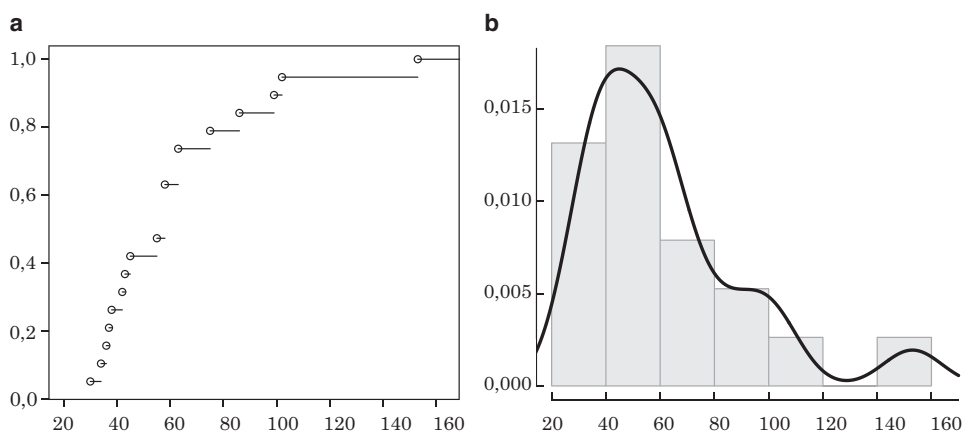


Abb. 2.2: Empirische Verteilungsfunktion (a) und Histogramm (b) der Daten aus Beispiel 2.3. Beim Histogramm ist eine Dichteschätzung der Daten eingezeichnet, um zu zeigen, dass das Histogramm, wenn es hinreichend feine Klassen besitzt, die Dichte schätzt

Wenn man die Daten als Realisierungen unabhängiger Zufallsvariablen auffassen kann, ist die empirische Verteilungsfunktion der beste Schätzer der Verteilungsfunktion der Zufallsvariablen. Mithilfe der Ordnungsstatistik kann man die empirische Verteilungsfunktion übrigens einfach bestimmen; es gilt

$$F_n(x_{(k)}) = \frac{k}{n},$$

und bis zum nächsten Wert $x_{(k+1)}$ bleibt die Funktion konstant (fallen mehrere Ordnungsstatistiken zusammen, ist der größte der so bestimmten Werte zu nehmen; das heißt, ist zum Beispiel $x_{(4)} = x_{(5)} = x_{(6)}$, so ist $F_n(x_{(3)}) = \frac{3}{n}$ und $F_n(x_{(4)}) = F_n(x_{(5)}) = F_n(x_{(6)}) = \frac{6}{n}$).

Histogramm

In einem *Histogramm* werden die Daten in Klassen eingeteilt. Jede Klasse wird durch einen Balken visualisiert, der so hoch ist, dass die Fläche des Balkens proportional zur Anzahl der Daten ist, die in der Klasse liegen.

Wir wollen unsere Daten $x_1, \dots, x_n$ in Klassen einteilen. Dazu benötigen wir zuerst Trenngrenzen $a_0, \dots, a_K$:

$$-\infty < a_0 < a_1 < \dots < a_K < \infty,$$

wobei die äußeren Grenzen so gewählt sind, dass alle Beobachtungen in das Intervall $(a_0, a_K]$ fallen. Das *Histogramm* ist dann die Funktion $h : \mathbb{R} \rightarrow \mathbb{R}$, die auf dem Intervall $(a_{k-1}, a_k]$ den Wert

$$\begin{aligned} h(x) &= \frac{1}{n(a_k - a_{k-1})} \#\{1 \leq i \leq n : a_{k-1} < x_i \leq a_k\} \\ &= \frac{F_n(a_k) - F_n(a_{k-1})}{a_k - a_{k-1}} \end{aligned}$$

hat. Dieser Wert ist der Anteil der Daten, die größer als a_{k-1} und kleiner oder gleich a_k sind, geteilt durch die Breite der Klasse $a_k - a_{k-1}$. Die Höhe des Rechtecks über der Klasse $(a_{k-1}, a_k]$ ist also die relative Häufigkeit der Klasse in der Datenmenge, geteilt durch die Klassenbreite. Man kann das Histogramm als Schätzer der Dichtefunktion auffassen.

Beachten Sie, dass in einem ordnungsgemäßen Histogramm die *Fläche* eines Klassenrechtecks proportional zur relativen Häufigkeit der Klasse ist, und nicht die Höhe, denn die Klassen können verschieden breit sein. In der obigen Definition summieren sich die Flächeninhalte zu 1, man spricht von einem *normierten* Histogramm.

In der Praxis tauchen viele verschiedene Arten von Histogrammen auf, und auch viele Grafiken, die als Histogramm bezeichnet werden, aber keines sind, sondern etwa nur ein Balkendiagramm. Beispielsweise wird manchmal der Vorfaktor $1/(n(a_k - a_{k-1}))$ weggelassen und auf der y -Achse nur die absolute Anzahl eingetragen. Dann jedoch kann das Histogramm missverständlich sein und einen falschen Eindruck der Daten vermitteln.

2.1.3 Numerische Zusammenfassungen zweidimensionaler Daten

Durch Kennzahlen beschreiben, wie zwei Datensätze zusammenhängen

Wir betrachten jetzt Situationen, in der die Daten Zahlenpaare

$$(x_1, y_1), \dots, (x_n, y_n)$$

sind (zum Beispiel Elastizitätsmodul und Zugfestigkeit von verschiedenen Werkstoffen). Wir interessieren uns für Zusammenhänge zwischen den beiden Größen x und y (die sich in den Realisierungen $x_1, \dots, x_n$ und $y_1, \dots, y_n$ zeigen).

Wir können erst die beiden Beobachtungen getrennt betrachten, also $x_1, \dots, x_n$ und $y_1, \dots, y_n$, und dabei die Mittelwerte $\bar{x}$, $\bar{y}$ und die Varianzen σ_x^2 und σ_y^2 berechnen.

Statistik für Ingenieure

Wahrscheinlichkeitsrechnung und Datenauswertung

endlich verständlich

Rooch, A.

2014, XI, 229 S. 25 Abb., Softcover

ISBN: 978-3-642-54856-7