

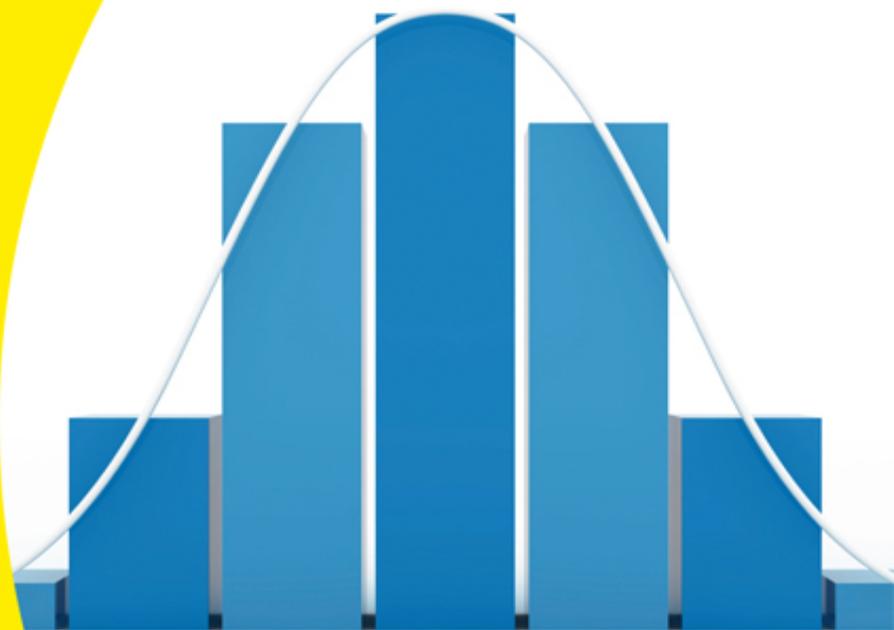
Von Mittelwerten, Verteilungen und Hypothesentests

# Statistik-Formeln

## FÜR DUMMIES<sup>®</sup>

### *Auf einen Blick:*

- Die wichtigsten Statistik-Formeln verstehen und anwenden
- Arithmetisches Mittel, Chi-Quadrat-Test, Histogramm & Co.
- Wahrscheinlichkeitsrechnung, beschreibende und schließende Statistik
- Zum schnellen Nachschlagen und Lernen



**Timm Sigg**

## Statistik-Formeln für Dummies – Schummelseite

### Formeln aus der beschreibenden Statistik

1. Arithmetisches Mittel:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
2. Empirische Varianz:  $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \cdot \left( \sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2 \right)$
3. Empirische Standardabweichung:  $s = \sqrt{s^2}$
4. Empirische Kovarianz:  

$$s_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \cdot \left( \sum_{i=1}^n x_i y_i - n \bar{x} \cdot \bar{y} \right)$$
5. Empirischer Korrelationskoeffizient:  $r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$
6. Die Regressionsgerade  $y = mx + b$  hat die Steigung:  $m = \frac{s_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}$ .
7. Die Regressionsgerade  $y = mx + b$  hat den  $y$ -Achsenabschnitt:  $b = \bar{y} - m \cdot \bar{x}$ .

### Formeln aus der Wahrscheinlichkeitstheorie

	Diskrete Zufallsvariable	Stetige Zufallsvariable
Erwartungswert	$\mu = \sum_i x_i \cdot f(x_i)$	$\mu = \int_{-\infty}^{\infty} x \cdot f(x) dx$
Varianz	$\begin{aligned} \sigma^2 &= \sum_i (x_i - \mu)^2 \cdot f(x_i) \\ &= \sum_i x_i^2 \cdot f(x_i) - \mu^2 \end{aligned}$	$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - \mu^2 \end{aligned}$
Standardabweichung	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$
Kovarianz	$\begin{aligned} \sigma_{XY} &= E((X - \mu_X) \cdot (Y - \mu_Y)) \\ &= E(X \cdot Y) - \mu_X \cdot \mu_Y \end{aligned}$	$\begin{aligned} \sigma_{XY} &= E((X - \mu_X) \cdot (Y - \mu_Y)) \\ &= E(X \cdot Y) - \mu_X \cdot \mu_Y \end{aligned}$
Korrelationskoeffizient	$\rho(X, Y) = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$	$\rho(X, Y) = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$



### Verteilungen diskreter Zufallsvariablen

Name der Verteilung	Diskrete Dichte $P(X = k)$	Erwartungswert $E(X)$	Varianz $Var(X)$
Diskrete Gleichverteilung	$\frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Binomialverteilung	$\binom{n}{k} p^k (1-p)^{n-k}$	$np$	$np(1-p)$
Hypergeometrische Verteilung	$\frac{\binom{M}{k} \cdot \binom{N-M}{n-k}}{\binom{N}{n}}$	$np$	$np(1-p) \cdot \frac{N-n}{N-1}$
Geometrische Verteilung	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poissonverteilung	$\frac{\lambda^k}{k!} \cdot e^{-\lambda}$	$\lambda$	$\lambda$

### Verteilungen stetiger Zufallsvariablen

Name der Verteilung	Dichtefunktion $f(x)$	Verteilungsfunktion $F(x)$	Erwartungswert $E(X)$	Varianz $Var(X)$
Stetige Gleichverteilung	$\begin{cases} 0, & \text{für } x < a \\ \frac{1}{b-a}, & \text{für } a \leq x \leq b \\ 0, & \text{für } x > b \end{cases}$	$\begin{cases} 0, & \text{für } x < a \\ \frac{x-a}{b-a}, & \text{für } a \leq x \leq b \\ 1, & \text{für } x > b \end{cases}$	$\frac{a+b}{2}$	$\frac{(a+b)^2}{12}$
Exponentialverteilung	$\begin{cases} 0, & \text{für } x < 0 \\ \lambda e^{-\lambda x}, & \text{für } x \geq 0 \end{cases}$	$\begin{cases} 0, & \text{für } x < 0 \\ 1 - e^{-\lambda x}, & \text{für } x \geq 0 \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normalverteilung	$\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$F(x) = \int_{-\infty}^x f(u) du$	$\mu$	$\sigma^2$
Standardnormalverteilung	$\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$	$\Phi(z) = \int_{-\infty}^z \varphi(u) du$	0	1



*Timm Sigg*

***Statistik-Formeln für Dummies***

*Fachkorrektur von  
Christoph Maas und Joachim Gaukel*

**WILEY**

WILEY-VCH Verlag GmbH & Co. KGaA

## **Bibliografische Information der Deutschen Nationalbibliothek**

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

1. Auflage 2014

© 2014 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

All rights reserved including the right of reproduction in whole or in part in any form.

Alle Rechte vorbehalten inklusive des Rechtes auf Reproduktion im Ganzen oder in Teilen und in jeglicher Form.

Wiley, the Wiley logo, Für Dummies, the Dummies Man logo, and related trademarks and trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries. Used by permission.

Wiley, die Bezeichnung »Für Dummies«, das Dummies-Mann-Logo und darauf bezogene Gestaltungen sind Marken oder eingetragene Marken von John Wiley & Sons, Inc., USA, Deutschland und in anderen Ländern.

Das vorliegende Werk wurde sorgfältig erarbeitet. Dennoch übernehmen Autoren und Verlag für die Richtigkeit von Angaben, Hinweisen und Ratschlägen sowie eventuelle Druckfehler keine Haftung.

Coverfoto: © Fotolia. de/erspek

Korrektur: Petra Heubach-Erdmann und Jürgen Erdmann, Düsseldorf

E-Book: Beltz Bad Langensalza GmbH, Bad Langensalza

Print ISBN: 978-3-527-71014-0

ePub ISBN: 978-3-527-69057-2

mobi ISBN: 978-3-527-69058-9

# Inhaltsverzeichnis

## Einleitung

### Über dieses Buch

Konventionen in diesem Buch

Was Sie nicht lesen müssen

Törichte Annahmen über die Leser

Wie dieses Buch aufgebaut ist

Symbole, die in diesem Buch verwendet werden

Wie es weitergeht

## Teil I: Formeln aus der beschreibenden Statistik

### Kapitel 1: Was genau beschreibt die beschreibende Statistik?

Erste Daten werden erhoben, erste Stichproben genommen

Einteilung der Merkmale

### Kapitel 2: Ein Bild sagt mehr als 1000 Worte: Diagramme

Nicht schön, aber nützlich: Häufigkeitstabellen

Himbeer- oder Käsesahne? Kreis- oder Tortendiagramme?

Diagramme mit Säulen, Balken und Stäben

Punktewolken für zweidimensionale Darstellungen

### Kapitel 3: Formeln zu eindimensionalen Stichproben

Häufigkeiten und empirische Verteilungsfunktion

Die Lage peilen mit den Lagemaßen

Weitere Maße, die Streumaße

### Kapitel 4: Formeln zu zweidimensionalen Stichproben

Korreliert Ihre Lesedauer mit Ihrem Spaß an der Statistik?

Regressionen aller Arten

## Teil II: Formeln aus der Wahrscheinlichkeitsrechnung

### Kapitel 5: Elementare Wahrscheinlichkeitsrechnung

Ein Klassiker: Die klassische Wahrscheinlichkeit

Der Abschnitt für die Erbsenzähler: Kombinatorik

Bedingte Wahrscheinlichkeit und Unabhängigkeit

### Kapitel 6: Diskrete Zufallsvariable

Der Begriff der Zufallsvariablen

Ein Hauptdarsteller: Die diskrete Zufallsvariable

Erwartungswert, Varianz und Standardabweichung

Bekanntere Verteilungen diskreter Zufallsvariablen

## **Kapitel 7: Stetige Zufallsvariable**

Dichte und Verteilungsfunktion stetiger Zufallsvariablen

Erwartungswert, Varianz und Standardabweichung

Bekannte Verteilungen stetiger Zufallsvariablen

## **Kapitel 8: Gesetze der großen Zahlen**

Zentraler Grenzwertsatz

Schwaches Gesetz der großen Zahlen

## **Teil III: Formeln aus der schließenden Statistik**

### **Kapitel 9: Punktschätzer**

Die Brücke zwischen Teil I und Teil II

Punktschätzer schätzen punktgenau

### **Kapitel 10: Zufallsstrebereiche und Konfidenzintervalle**

Zufallsstrebereiche und die Signifikanz

Konfidenzintervalle schaffen Vertrauen

Wann nimmt man Zufallsstrebereiche, wann Konfidenzintervalle?

### **Kapitel 11: Parametertests**

So gehen Sie bei einem Parametertest vor

Parametertests für Erwartungswerte

Parametertests von Varianzen

Parametertest einer Wahrscheinlichkeit

### **Kapitel 12: Chi-Quadrat-Tests**

Anpassungstests: alles eine Frage der Anpassung

Der Chi-Quadrat-Unabhängigkeitstest

## **Teil IV: Der Top-Ten-Teil**

### **Kapitel 13: Zehn typische Fehler, die Sie vermeiden sollten**

Fehlerquelle 1: Im Allgemeinen gilt nicht  $P(A \cap B) = P(A) \cdot P(B)$ .

Fehlerquelle 2: Im Allgemeinen gilt nicht  $P(A \cup B) = P(A) + P(B)$ .

Fehlerquelle 3: Intervalle sind nicht gleich groß

Fehlerquelle 4: Ausreißer haben zu großes Gewicht

Fehlerquelle 5: Rundungsfehler verfälschen die Sache

Fehlerquelle 6: Korrelation bedeutet nicht zwingend Kausalität

Fehlerquelle 7: Geeignete Wahl der Grundmenge  $\Omega$

Fehlerquelle 8: Falsche Schlüsse aus den Testergebnissen

Fehlerquelle 9: Die Voraussetzungen stimmen nicht

Fehlerquelle 10: »Traue keiner Statistik, ...

## **Anhang - Tabellen**

### **Anhang**

[Werte der Standardnormalverteilung  \$\Phi\(z\)\$  für verschiedene Werte von  \$z\$](#)

[Quantile der Standardnormalverteilung](#)

[Quantile der  \$t\$ -Verteilung](#)

[Quantile der  \$\chi^2\$ -Verteilung](#)

# *Danksagung*

Ich möchte mich ausdrücklich bei meinem Kollegen Peter Plappert bedanken! Viele Ideen der didaktischen Aufbereitung des Themas stammen von ihm.

Außerdem danke ich meinem Kollegen Joachim Gaukel sehr herzlich fürs aufwendige Nachrechnen, Nachvollziehen und Korrekturlesen!

# *Über den Autor*

**Timm Sigg** studierte an der Universität Stuttgart Mathematik und Physik und promovierte anschließend in theoretischer Physik. Erfahrung in angewandter Statistik konnte er in seiner Tätigkeit als Statistiker und Risikomanager bei der EnBW Vertriebsgesellschaft mbH sammeln. Seit 2004 ist er Professor für Mathematik an der Hochschule Esslingen, wo er begeistert Kurs- und Statistikvorlesungen hält. Darüber hinaus hat er einige Arbeiten zu den Themen stochastische Physik und zur Quantenkosmologie veröffentlicht. Sein Buch *Grundlagen der Differenzialgleichungen für Dummies* ist seit Oktober 2010 auf dem Markt.

# *Über die Fachkorrektoren*

**Prof. Dr. Christoph Maas** lehrt Mathematik in Ingenieurstudiengängen an der Hochschule für Angewandte Wissenschaften Hamburg und ist Autor von *Stochastik für Dummies*. **Prof. Dr. Joachim Gaukel** lehrt an der Hochschule Esslingen seit 2009 die Fächer Mathematik, Statistik und Numerik.

# Einleitung

---

Was fällt Ihnen zu dem Begriff *Statistik* ein?

Ist es

1. der Ausspruch: »Traue keiner Statistik, die du nicht selbst gefälscht hast!«
2. die Hochrechnungen der letzten Bundestagswahlen
3. Ihre aktuelle Statistik-Vorlesung mit der noch ausstehenden Prüfung
4. Ihr letzter Besuch im Spielkasino

oder ist es etwas ganz anderes? Die Statistik taucht an so vielen verschiedenen Stellen im Leben und in so vielfältiger Form auf, dass man diese Liste fast beliebig fortsetzen könnte. Welcher Aspekt davon auch immer für Sie von Bedeutung sein mag, die *Statistik-Formeln für Dummies* sollen Sie dabei in doppelter Hinsicht unterstützen.

Zum einen ist das Buch so angelegt, dass Sie alle zentralen statistischen Begriffe leicht nachschlagen können und auch griffige Kurzerklärungen dazu finden.

Zum anderen sind alle wichtigen statistischen Formeln ausführlich erklärt und mit Beispielen versehen. Sie werden feststellen, dass Sie die Formeln viel leichter verinnerlichen, wenn Sie sie an einem Beispiel nachrechnen.

## *Über dieses Buch*

Dieses Buch ist als Nachschlagewerk konzipiert. Wann immer Ihnen also ein Begriff, eine Fragestellung oder eine spezielle Übungsaufgabe aus der Statistik begegnet,

können Sie hier nachschlagen. Die *Statistik-Formeln für Dummies* bieten Ihnen darüber hinaus zu jedem Begriff, jeder Formel und jedem mathematischen Zusammenhang Erläuterungen und Beispiele. Darin unterscheidet sich das Buch von einer reinen Formelsammlung oder einem Lexikon.

Formeln und Begriffsdefinitionen findet man im Internet heutzutage in Hülle und Fülle. Mir war es aber wichtig, die Formeln und Begriffe mit Leben zu füllen, sie verständlich zu erläutern und mit Beispielen zu versehen. Dieses Buch soll also neben seinem Charakter als Nachschlagewerk durchaus auch zur Lektüre anregen.

## ***Konventionen in diesem Buch***

Alle statistischen und sonstigen mathematischen Begriffe und Symbole werden an Ort und Stelle erklärt.

Typischerweise sind sie *kursiv* gesetzt, wenn sie eine zentrale Bedeutung haben. Sie finden diese auch im Index am Ende des Buches.

Dezimalzahlen schreibe ich mit einem Punkt und nicht mit Komma, also beispielsweise  $\frac{1}{4} = 0.25$ .

## ***Was Sie nicht lesen müssen***

Alle Abschnitte sind in diesem Buch in sich abgeschlossen und somit für sich alleine genommen verständlich. Daher schlage ich vor, Sie lesen stets genau das, was für Sie aktuell von Bedeutung ist.

Abgesehen davon gibt es viele Querverweise innerhalb des Buches, die insgesamt zu einem besseren Verständnis beitragen. Und schließlich habe ich das Buch so geschrieben, dass man es auch komplett von vorne bis hinten durchlesen kann, ohne ständig das Gefühl zu haben, mit etwas völlig Neuem konfrontiert zu werden, wie beispielsweise bei einem Lexikon oder einer reinen Formelsammlung.

## ***Törichte Annahmen über die Leser***

Zugegeben, ich kenne Sie nicht. Aber dennoch habe ich eine vage Vorstellung von Ihnen. Ich stelle Sie mir so vor:

- ✓ Sie sind Schüler beziehungsweise Student und kommen an Ihrer Schule beziehungsweise Hochschule mit Statistik oder Wahrscheinlichkeitsrechnung in Kontakt oder
- ✓ Sie sind berufstätig und immer wieder begegnen Ihnen bei der Arbeit Begriffe wie Mittelwert, signifikant, Streuung oder Normalverteilung und Sie wollen endlich mal genau wissen, was das alles ist, oder
- ✓ Sie sollen eine statistische Auswertung verstehen oder selbst eine erstellen oder
- ✓ Sie sind einfach so an Statistik interessiert, weil man in so vielen Situationen von Statistik hört, beim Lesen von Bilanzen, im Spielkasino, vor und nach politischen Wahlen, in der Medizin, bei Umfragen.

Aus welcher Motivation heraus Sie sich auch immer mit Statistik beschäftigen, um die Formeln zu verstehen und nachvollziehen zu können, benötigen Sie etwas Mathematik, allerdings keine allzu anspruchsvolle. Dazu

gehören zunächst Funktionen und Summen. Vermutlich sind Ihnen diese schon an den unterschiedlichsten Stellen im Leben begegnet. Und zu weiten Teilen kommt das Buch auch ohne Ableitungen und Integrale aus; sie lassen sich aber nicht ganz vermeiden, insbesondere wenn es um die Begriffe Dichtefunktion und Verteilungsfunktion geht. Hier ist nämlich die eine Funktion die Ableitung der anderen. Wenn Sie sich im Ableiten und Integrieren noch unsicher fühlen, es aber unbedingt können wollen, so kann ich Ihnen das Buch *Analysis für Dummies* ans Herz legen.

## ***Wie dieses Buch aufgebaut ist***

Dieses Buch besteht aus vier Teilen und einem Anhang. Die ersten drei Teile behandeln die beschreibende Statistik, die Wahrscheinlichkeitsrechnung und zu guter Letzt die schließende Statistik. Der vierte Teil des Buches ist der Top-Ten-Teil. Danach folgt noch der Anhang mit vier Tabellen. Und jetzt das Ganze nochmals etwas ausführlicher:

### ***Teil I: Formeln zur beschreibenden Statistik***

Dieser Teil widmet sich der Datenerhebung, der Klassifizierung von Merkmalen und den Darstellungsmöglichkeiten von größeren Datenmengen, vor allem aber den Formeln der *beschreibenden* Statistik (man sagt auch *deskriptive* Statistik).

Wenn Sie also wissen wollen, wie man das Durchschnittsalter einer Fußballmannschaft berechnet, wie groß die Streuung des Gewichts Ihrer Schokoladentafelsammlung ist, was es bedeutet, wenn das Gewicht Ihres Kindes auf dem 3%-Perzentil liegt, oder ob

die Größe eines Menschen mit der Zahl seiner Sommersprossen korreliert, dann sind Sie hier richtig.

## ***Teil II: Formeln zur Wahrscheinlichkeitsrechnung***

Streng genommen müsste man die Wahrscheinlichkeitsrechnung nicht zur Statistik zählen, sie ist aber einerseits eng mit ihr verbunden und wird andererseits vor allem für die schließende Statistik (Teil III) benötigt. Daher sind die Formeln zur Wahrscheinlichkeitsrechnung unumgänglich, wenn man Statistik betreiben möchte.

Mit diesen Formeln sind Sie in der Lage, auszurechnen, wie groß die Wahrscheinlichkeit ist, einen Sechserpasch zu würfeln, Sie erfahren, was man unter bedingter Wahrscheinlichkeit versteht, was unter unabhängigen Ereignissen, wofür man eine Zufallsvariable braucht, welche Verteilungsfunktionen es gibt und was das überhaupt ist.

## ***Teil III: Formeln zur schließenden Statistik***

Die sogenannte *schließende* Statistik (man sagt auch *induktive* Statistik) ist die Krönung der Statistik. Die Formeln werden etwas länger, die Verfahren aufwendiger und die Interpretationen umfassender. Sie lernen hier unter anderem statistische Tests und Konfidenzintervalle kennen.

Mit den Formeln der schließenden Statistik lassen sich richtig tolle Sachen machen. Ist ein Würfel, den Sie 30-mal werfen und der kein einziges Mal eine 6 zeigt, gezinkt? Ja oder nein? Und wenn Ja, mit welcher Wahrscheinlichkeit stimmt dieses Ja?

## ***Teil IV: Der Top-Ten-Teil***

Der *Dummies*-Buch-typische Top-Ten-Teil schließt das Buch ab. Er enthält zehn typische Fehlerquellen, in die man in der Statistik und der Wahrscheinlichkeitstheorie gerne tappt.

## ***Anhang: Tabellen***

Im Anhang finden Sie vier Tabellen A.1 bis A.4, um die Werte, die Sie für einige statistische Methoden benötigen, nachzuschlagen.

# ***Symbole, die in diesem Buch verwendet werden***

Die folgenden Symbole werden Sie durch das ganze Buch begleiten:



Dieses Symbol weist auf einen Tipp hin, der Ihnen das Leben vereinfachen soll. Meine Empfehlung: Lesen Sie sich den Tipp auf alle Fälle durch! Dadurch sparen Sie in Zukunft Zeit und Nerven.



Die Statistik ist bekannt für ihre vielen Definitionen. Taucht ein zentraler Begriff erstmals im Buch auf, so wird er unter diesem Symbol definiert beziehungsweise erklärt. Meine Empfehlung: Lesen Sie sich die Definitionen durch und suchen Sie auch gegebenenfalls im Index danach, wenn Ihnen ein zentraler Begriff nicht klar ist.



Wenn Sie die Bombe sehen, seien Sie gewarnt. Etwas ist im Busch. Womöglich eine Falle, in die man leicht tappt, oder eine Kleinigkeit, die man leicht übersieht. Meine Empfehlung: Lesen Sie sich auch diesen Text auf alle Fälle durch. Wenn Sie erkennen, dass Sie nie und nimmer in die Falle tappen würden, umso besser!



Hier wird's technisch und gerne auch mal ein bisschen komplizierter. Für das weitere Verständnis des Buches sind die hier stehenden Informationen nicht von Belang. Meine Empfehlung: Entscheiden Sie ganz frei von Fall zu Fall, wie Sie mit diesen Insiderinformationen umgehen wollen!

## ***Wie es weitergeht***

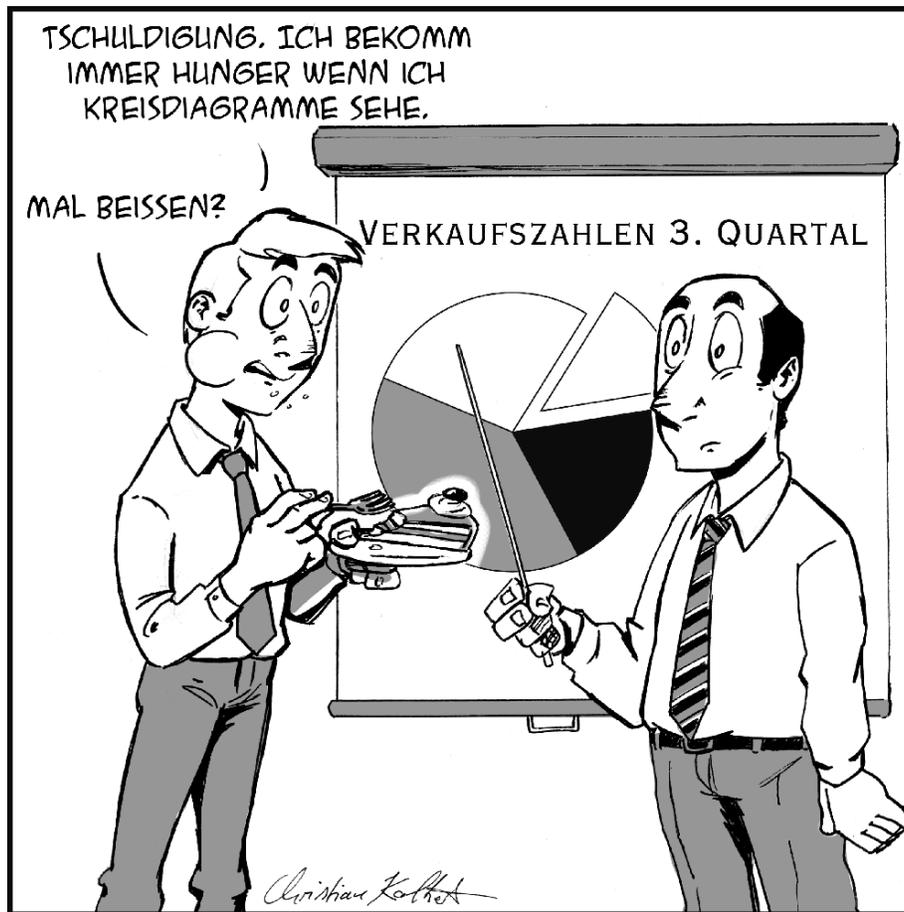
Ich sehe nun zwei Möglichkeiten, wie es weitergehen kann:

1. Sie wollen dieses Buch der Reihe nach durchlesen. Dann beginnt für Sie nun der Teil I, in dem die beschreibende Statistik vorgestellt wird.
2. Sie haben eine konkrete Frage zu einer Formel oder zu einem Sachverhalt. Dann blättern Sie doch einfach direkt dorthin. Im Index finden Sie die einschlägigen Stichwörter.

In allen Fällen wünsche ich Ihnen viel Spaß mit dem Buch!

# Teil I

## Formeln aus der beschreibenden Statistik



## ***In diesem Teil ...***

geht es um die *beschreibende* Statistik. Sie steht nicht nur in diesem Buch ganz vorne. Sie steht auch immer zeitlich am Anfang, wenn Sie Statistik betreiben. Und darum habe ich diesen Teil auch so aufgebaut, wie man in der Realität typischerweise vorgeht: Er beginnt mit der Datengewinnung. Und selbst wenn Sie nicht für die Erhebung an sich zuständig sind, ist es von Vorteil, wenn Sie sich hier etwas auskennen; insbesondere wenn es darum geht, welche Möglichkeiten Sie haben, die Daten einzuteilen. Weiter geht es mit den zahlreichen Möglichkeiten der grafischen Darstellung dieser Daten. Und schließlich lernen Sie die wesentlichen Formeln der beschreibenden Statistik kennen.

# Kapitel 1

## Was genau beschreibt die beschreibende Statistik?

---

### *In diesem Kapitel*

- ▶ Erfahren, worum es überhaupt geht in der beschreibenden Statistik
  - ▶ Die guten Daten ins Töpfchen, die schlechten ins Kröpfchen
  - ▶ Quantitativ von qualitativ unterscheiden
- 

Das Arbeitsgebiet der beschreibenden Statistik lässt sich so zusammenfassen: Mit der beschreibenden Statistik lassen sich Daten übersichtlich darstellen. Dazu werden nicht nur Tabellen oder grafische Methoden verwendet, sondern darüber hinaus Kennzahlen definiert und berechnet, die dieser Übersichtlichkeit dienen.

## *Erste Daten werden erhoben, erste Stichproben genommen*

Das Wichtigste gleich vorweg: Eine statistische Auswertung ist nur sinnvoll, wenn genügend Daten vorliegen und wenn Sie den Daten trauen können. Taugen die Daten nichts oder sind es (viel) zu wenige, so ist jede weitere Rechnung für die Mülltonne. So toll Ihr

Statistikprogramm auch sein mag, es kann von sich aus nicht beurteilen, ob die Daten etwas taugen.

Als Mathematiker wird man häufig gefragt, ob man bei dieser oder jener statistischen Auswertung nicht helfen könne. Ich mache das immer gerne und habe dabei festgestellt, dass es meistens nicht die statistischen Verfahren sind, die Probleme machen, sondern die Daten und dabei vor allem die folgenden Tücken:

- ✓ Die Datenmenge ist so gering, dass keine vernünftigen Schlüsse gezogen werden können. Beispiel: Sie untersuchen die Wirksamkeit einer Therapiemethode in einem Krankenhaus. Sie haben aber nur vier Patienten.
- ✓ Es gibt zu viele Daten, denen Sie nicht trauen können, weil nicht klar ist, woher sie kommen oder ob in der Übermittlung der Daten etwas schiefgelaufen ist. Beispiel: Sie erhalten Daten von verschiedenen Institutionen.
- ✓ Es gibt Ausreißer, die das Ergebnis verfälschen. Beispiel: Ihnen liegen 100 Messdaten aus einer Messung vor. 98 davon liegen zwischen 11 und 12. Die übrigen beiden aber lauten 1000 und 1200. Was tun? Jede Mittelwertberechnung und sonstige statistische Auswertung wird durch die beiden Ausreißer extrem beeinflusst, aber können Sie sie einfach weglassen?



Da es meistens nicht möglich ist, eine Totalerhebung (die komplette Grundgesamtheit liegt als Datenmaterial vor) zu machen, beschränkt man sich auf eine Teilerhebung. Diese wird *Stichprobe* genannt.

# *Einteilung der Merkmale*

Grundsätzlich lassen sich Merkmale in *quantitative* und *qualitative* Merkmale aufteilen. Die quantitativen Merkmale haben in der Statistik die bei Weitem größere Bedeutung.

## *Quantitative Merkmale - zählen und messen*

Die Ausprägungen quantitativer Merkmale sind Zahlen aus Messungen oder aus Zählungen. Man unterscheidet dabei

- ✓ **Quantitativ-diskrete Merkmale**, bei denen die Ausprägungen einzelne Punkte sind und die Daten vorzugsweise aus Zählungen stammen. Beispiele hierfür sind Ergebnisse von Verkehrszählungen, Anzahl von Defektstücken, Alter (in Jahren) und so weiter.
- ✓ **Quantitativ-stetige Merkmale**, bei denen die Ausprägungen typischerweise aus einem bestimmten Intervall sind und die Daten vorzugsweise aus Messungen stammen. Beispiele hierfür sind Längen, Größen, Gewicht, Temperaturen und so weiter.

Diese Unterscheidung zwischen *stetig* und *diskret* zieht sich wie ein roter Faden durch die Statistik und somit auch durch dieses Buch.



Die Ausprägungen quantitativ-diskreter Merkmale sind beispielsweise die natürlichen Zahlen inklusive der Null, also 0, 1, 2, 3 ...

Die Ausprägungen quantitativ-stetiger Merkmale hingegen sind prinzipiell die reellen Zahlen  $\mathbb{R}$ . Da

aber keine Messung so exakt durchgeführt werden kann, werden die Ergebnisse in Intervallen angegeben. Dazu ein Beispiel:

Sie messen von mehreren Personen die Körpergröße. Bei der ersten Messung notieren Sie 173.8 cm. Klar ist, dass diese Person nicht exakt 173.8 cm groß ist, Sie messen es eben nur nicht genauer. Dennoch handelt es sich um eine quantitativ-stetige Größe. Mit der Angabe 173.8 cm meinen Sie letztlich das Intervall zwischen 173.75 cm und 173.84 cm.

## ***Qualitative Merkmale - beschreiben und bestaunen***

Qualitative Merkmale sind beschreibende Eigenschaften.



Achtung: Beschreibende Eigenschaften können mit Zahlen codiert sein. Vor allem in Datenbanken taucht das häufig auf, um Speicherplatz zu sparen. Dann steht beispielsweise die Zahl 1 für die Farbe Grün, die Zahl 2 für die Farbe Rot und so weiter. Dennoch handelt es sich aber natürlich um ein qualitatives Merkmal, denn die Zahlen stellen ja nur Platzhalter dar und haben nicht die gleiche Bedeutung wie bei den quantitativ-diskreten Merkmalen.

Man unterscheidet

✓ **Ordinale Merkmale**, denen eine natürliche Reihenfolge zugrunde liegt. Ein Beispiel: Ich stelle Ihnen die Frage »Wie gerne lesen Sie dieses Buch?« und biete Ihnen die Antwortmöglichkeiten: »sehr gerne«, »gerne«, »mäßig gerne«, »nicht so gerne«, »ich hasse es!«. Die Antworten haben eine natürliche

Reihenfolge, aber sie repräsentieren nicht ein Ergebnis wie aus einer Messung.

- ✓ **Nominale Merkmale**, bei denen keine natürliche Reihenfolge zugrunde liegt. Beispiele hierfür sind Farbe, Wohnort, Geschlecht und so weiter.



Ein recht nützliches Vorgehen, quantitative von qualitativen Merkmalen zu unterscheiden, ist auch das Folgende:

Sie überlegen sich, ob die Differenz zwischen zwei Ausprägungen eine Bedeutung hat oder nicht. Hat sie eine, so ist das Merkmal quantitativ, ansonsten qualitativ.

Zum Abschluss des Kapitels noch zu jeder Kategorie ein Beispiel:

- ✓ **Quantitativ-diskret**: Gestern fuhren 150 Autos über die Kreuzung. Heute waren es 162. Also fuhren heute 12 Autos mehr über die Kreuzung. Die Differenz hat eine Bedeutung, also ist das Merkmal quantitativ. Genauer gesagt ist es *quantitativ-diskret*.
- ✓ **Quantitativ-stetig**: In Eisdorf war die durchschnittliche Temperatur im letzten Jahr 6.4 °C. In Ofenhausen war sie 12.2 °C. Also war es in Ofenhausen im Schnitt um 5.8 °C wärmer. Die Differenz hat eine Bedeutung, also ist das Merkmal quantitativ. Genauer gesagt ist es *quantitativ-stetig*.
- ✓ **Ordinal**: Franz kreuzte im Hotel bei der Frage, wie sehr er mit der Sauberkeit des Hotels zufrieden war, die Antwort »sehr« an. Heike, seine Frau, hingegen

»mittelmäßig«. Es kann keine Differenz ermittelt werden. Das Merkmal ist qualitativ, genauer gesagt *ordinal*.

✓ **Nominal:** Studenten werden nach ihren Studiengängen befragt. Diese werden mit Kennziffern codiert. Pascal studiert Maschinenbau (Kennziffer 3), Nadine studiert Mathematik (Kennziffer 5). Die Differenz zwischen den beiden Kennziffern beträgt 2 ( $5 - 3 = 2$ ). Die Kennziffer 2 stünde für Jura, aber das hat gar keine Bedeutung für die Auswertung. Das Merkmal ist *nominal*.

Das ist nicht so schwierig, nicht wahr? Dann kommt jetzt was fürs Auge, und zwar die verschiedenen Darstellungsmöglichkeiten dieser Merkmale.

## Kapitel 2

# Ein Bild sagt mehr als 1000 Worte: Diagramme

---

### *In diesem Kapitel*

- ▶ Den Farbkasten auspacken
  - ▶ Torten malen statt essen
  - ▶ In Punktwolken Figuren erkennen
- 

Kein Mensch will sich ewig lange Zahlenkolonnen durchlesen, die womöglich noch ungeordnet angeboten werden. In diesem Kapitel geht es darum, in welchem Fall welches Diagramm geeignet ist, einen Sachverhalt übersichtlich und unverfälscht darzustellen.

## *Nicht schön, aber nützlich: Häufigkeitstabellen*

Am Anfang der Datenbearbeitung steht immer die Tabelle. Nach der Erfassung der Daten werden sie typischerweise in Tabellen aufgelistet und stehen zur Weiterverwertung bereit. Und natürlich können Sie Tabellen auch dazu verwenden, Messreihen darzustellen, unabhängig davon, ob sie qualitative oder quantitative Merkmale enthalten. Dies geschieht in sogenannten *Häufigkeitstabellen*, mehr über Häufigkeitstabellen finden Sie in [Kapitel 3](#).

Dazu ein Beispiel mit einem qualitativen Merkmal: Ich habe 50 Studenten befragt, wie groß ihr Interesse an Statistik ist, und ihnen 5 Antwortmöglichkeiten gegeben: »sehr groß«,

»groß«, »mittel«, »gering« und »sehr gering«. Das Ergebnis können Sie in [Tabelle 2.1](#) bestaunen.

Die Zahlen in der rechten Spalte der Tabelle geben die Anzahl der Studenten wieder, die für diese Kategorie gestimmt haben. Man nennt sie die *absoluten Häufigkeiten*  $n_t$ . Die Summe der absoluten Häufigkeiten entspricht der Größe der Grundmenge  $N$ , also hier der Zahl 50. Es gilt also:

Interesse	Anzahl der Studenten
sehr groß	4
groß	14
mittel	21
gering	8
sehr gering	3

**Tabelle 2.1: Beispiel einer Häufigkeitstabelle eines qualitativen Merkmals (Interesse)**

$$\sum_i n_i = N.$$

Die verschiedenen Arten von Häufigkeiten werden noch in [Kapitel 5](#) eine Rolle spielen. Zugegeben, ich habe diese Umfrage nicht wirklich durchgeführt. Vermutlich würde die Tabelle sonst etwas anders aussehen.

## ***Himbeer- oder Käsesahne? Kreis- oder Tortendiagramme?***

Wer kennt sie nicht? Die Torten-, Kuchen- oder Kreisdiagramme? Sie haben mehrere Bezeichnungen und ich

werde mich bewusst auf keine festlegen. Mit ihnen werden meistens prozentuale Verhältnisse dargestellt. Beispiele dafür können sein:

- ✓ Sitzaufteilung in einem Parlament
- ✓ Rohstoffvorkommen auf der Erde
- ✓ Zusammensetzung eines Gases

Es ist Ihnen sicher klar, dass man diese Aufzählung beliebig fortführen könnte.

Wenn Sie damit das fiktive Umfrageergebnis (siehe [Tabelle 2.1](#)) darstellen wollen, müssen Sie die vorhandenen Zahlen, also die absoluten Häufigkeiten, erst in dazu passende Winkel umrechnen. Dies führt auf die erste, wenn auch harmlose, Formel des Buches.

## ***Umrechnungsformel zwischen absoluter Häufigkeit und Winkel im Tortendiagramm***

Diese Formel brauchen Sie, um die Größe des Tortenstücks zu bestimmen:

$n_i$  ... absolute Häufigkeiten

$N$  ... Summe der absoluten Häufigkeiten, also  $\sum n_i = N$

$\alpha_i$  ... Winkel der  $i$ -ten absoluten Häufigkeit in Grad

Dann gilt

$$\frac{\alpha_i}{360^\circ} = \frac{n_i}{N},$$

beziehungsweise nach  $\alpha_i$  aufgelöst:

$$\alpha_i = \frac{n_i}{N} \cdot 360^\circ.$$

Die Formel besagt, dass das Verhältnis zwischen einer absoluten Häufigkeit  $n_i$  und der Summe der Häufigkeiten  $N$  genau dem Verhältnis zwischen dem Winkel  $\alpha_i$  und dem Vollwinkel  $360^\circ$  entspricht.

Mit dieser Formel ergeben sich zu den schon bekannten Häufigkeiten  $n_i$  und zur Summe  $N = 50$  die entsprechenden Winkel  $\alpha_i$  in [Tabelle 2.2](#).

Interesse	Anzahl der Studenten $n_i$	Winkel $\alpha_i$
sehr groß	4	$\frac{4}{50} \cdot 360^\circ = 28.8^\circ$
groß	14	$\frac{14}{50} \cdot 360^\circ = 100.8^\circ$
mittel	21	$\frac{21}{50} \cdot 360^\circ = 151.2^\circ$
gering	8	$\frac{8}{50} \cdot 360^\circ = 57.6^\circ$
sehr gering	3	$\frac{3}{50} \cdot 360^\circ = 21.6^\circ$

**Tabelle 2.2: Häufigkeitstabelle mit zugehörigen Winkeln für ein Tortendiagramm**

Jetzt, wo die Winkel bekannt sind, können Sie ohne Weiteres das Kreisdiagramm zeichnen. Sie sehen es in [Abbildung 2.1](#).