

# Kapitel 2

## Einfache Stichprobenverfahren

### 2.1 Grundbegriffe

Bei der Durchführung einer statistischen Erhebung besteht die Absicht, Informationen über eine (üblicherweise große) Menge von Individuen zu erhalten. So kann ein Unternehmen Interesse daran haben, sich einen Überblick über die Kundenzufriedenheit zu verschaffen oder ein Meinungsforschungsinstitut möchte im Auftrag einer Fernsehanstalt Informationen über die politische Stimmung in einem Land erhalten. Bei der Durchführung der Erhebung muss zuerst die Menge der Individuen, über die eine Aussage getroffen werden soll, bestimmt und abgegrenzt werden. Diese Menge besteht bei Umfragen in der empirischen Sozialforschung typischerweise aus der Bevölkerung eines Landes oder einer Untergruppe daraus, wie z.B. aus den wahlberechtigten Bürgern. Daher wird in der Stichprobentheorie und nachfolgend in diesem Buch der Begriff *Population* für diese abgegrenzte Menge von Individuen verwendet. Im deutschsprachigen Raum wird diese auch als *Grundgesamtheit* bezeichnet.

- Die **Population** oder **Grundgesamtheit** ist die Menge aller Individuen oder Objekte, über die eine Aussage getroffen werden soll.

Die Grundgesamtheit muss nicht zwingend aus Personen bestehen. Bei einer ökologischen Fragestellung kann sie z.B. aus Planquadraten einer Fläche oder aus Seen eines Landes bestehen. Wir definieren daher allgemein die Elemente der Population wie folgt.

- **Merkmalsträger** oder **statistische Einheiten** sind die Einheiten oder Objekte, an denen Untersuchungen, Messungen oder Beobachtungen vorgenommen werden. Gelegentlich werden die Merkmalsträger auch **Individuen** genannt.

Der Bezug zur inhaltlichen Fragestellung wird durch den Begriff des Merkmals hergestellt.

- **Merkmale** sind die Eigenschaften der statistischen Einheiten, die untersucht, beobachtet oder gemessen werden sollen.

Merkmale sind dabei zum Beispiel Einkommen, Wahlverhalten oder Meinungen zu bestimmten Themen, die in einem Interview erfragt werden können. Merkmale können auch konkrete Messungen oder Beobachtungen sein. Beispielsweise kann eine statistische Erhebung Aufschluss über die Schädigung des deutschen Waldes geben. In diesem Fall ist das Merkmal der konkrete Schädigungsgrad eines einzelnen Baumes oder Waldabschnittes.

Der Ausgangspunkt einer empirischen Untersuchung ist damit die Definition von Grundgesamtheit, statistischen Einheiten und zu erhebenden Merkmalen. Das bedeutet im Grunde genommen nichts anderes als die folgenden Fragen zu beantworten:

- Worauf bezieht sich die Untersuchung?
- Welche Inhalte sollen betrachtet werden?

In den meisten Anwendungen ist die Grundgesamtheit relativ groß, so dass man sich darauf beschränkt, anstelle der gesamten Population eine (kleine) Teilmenge zu untersuchen oder zu befragen. Diese Teilmenge nennen wir Stichprobe.

- Eine **Stichprobe** ist die Teilmenge der Population, an der die Merkmale erhoben werden.

Wenn man sich dafür entscheidet, die gesamte Population zu untersuchen, spricht man von einer **Vollerhebung** (oder auch Zensus genannt). In diesem Fall sind Stichprobe und Population identisch. Ansonsten spricht man von einer **Teilerhebung**. Typische Beispiele für Vollerhebungen sind Volkszählungen und Wahlen. Bevor wir uns im Folgenden ausschließlich mit Teilerhebungen beschäftigen, wollen wir kurz die Vor- und Nachteile von Teilerhebungen im Vergleich zu einer Vollerhebung diskutieren. In vielen Fällen ist eine Vollerhebung weder praktikabel noch finanzierbar. Dies gilt insbesondere für Untersuchungen, die sich auf die Gesamtbevölkerung eines Landes beziehen. In anderen Fällen ist eine Vollerhebung rein technisch nicht ratsam, wie das folgende Beispiel zeigt.

*Beispiel 2.1:* Bei der Lieferung einer Charge von Äpfeln soll der Schadstoffgehalt bestimmt werden. Hier besteht die Population also aus allen Äpfeln der Charge, die Merkmalsträger sind die einzelnen Äpfel und das interessierende Merkmal ist z.B. der Schadstoffgehalt bezogen auf das Gewicht. Um diesen zu bestimmen, muss der Apfel chemisch untersucht werden und kann anschließend weder verzehrt noch verkauft werden. Daher hätte man nach einer Vollerhebung zwar genaue Angaben zur Schadstoffbelastung, aber keine Äpfel mehr. Somit ist es hier sinnlos, eine Vollerhebung durchzuführen. ◀

Des Weiteren haben Vollerhebungen den Nachteil, dass bedingt durch den hohen Aufwand häufig Mess- und Erhebungsfehler in stärkerem Maße auftreten. Zusätzlich kann es zu einer höheren Ausfallrate als bei einer Teilerhebung kommen. Das heißt, Individuen gelangen in die Stichprobe, ihre Antwort oder Messung bleibt jedoch aus, weil sie beispielsweise die Antwort verweigern. Unter Berücksichtigung derartiger Probleme bietet sich eine Stichprobe in Form einer Teilerhebung an. Entschließt man sich somit, eine Teilerhebung anstelle einer Vollerhebung durchzuführen, so stellen sich in der Anwendung drei konkrete Fragen:

1. Welche Merkmalsträger der Population sollen beobachtet oder gemessen werden?
2. Wie viele Merkmalsträger sollen erhoben werden, das heißt, wie groß soll die Stichprobe sein?
3. Wie groß ist der Informationsverlust im Vergleich zu einer Vollerhebung?

Wir werden diesen Fragen in den folgenden Kapiteln nachgehen. Dabei wird sich zeigen, dass eine Teilerhebung in vielen Fällen völlig ausreicht, um eine geforderte Genauigkeit einer Erhebung zu garantieren.

*Beispiel 2.2:* Eine öffentliche Verwaltung möchte den Einsatz der Arbeitszeit ihres Personals genauer erfassen. Sie interessiert sich dafür, wie lange ein Arbeitnehmer mit Vorgängen wie „Bearbeitung von externen Anfragen“, „Bearbeitung von internen Anfragen“, „Kundenverkehr“, „Schriftverkehr“, etc. beschäftigt ist. Die Verwaltungsleitung zieht zwei mögliche Methoden in Betracht, die Daten zu sammeln. Einerseits durch eine Vollerhebung, bei der jeder Arbeitnehmer detailliert die aufgewendete Arbeitszeit in den einzelnen Bereichen auflistet. Andererseits kann eine Teilerhebung durchgeführt werden, bei der ausgewählte Arbeitnehmer die entsprechende Information liefern.

Bei einer Vollerhebung ist sicher mit Problemen zu rechnen. Arbeitnehmer mögen sich kontrolliert fühlen und/oder ungenaue Angaben machen. Bei einer Teilerhebung können sich die ausgewählten Arbeitnehmer hingegen benachteiligt fühlen und ihre Mitarbeit verweigern. Da es sich jedoch dabei im Gegensatz zur Vollerhebung um eine kleinere Gruppe von Arbeitnehmern handelt, kann von Seiten der Verwaltungsleitung Motivation zur Mitarbeit gegeben werden. Somit kann durch eine Teilerhebung durchaus ein genaueres Ergebnis als durch eine Vollerhebung erzielt werden. ◀

## 2.2 Nicht-zufällige Auswahlverfahren

Generell unterscheidet man zwei Arten von Stichproben, **zufällige** und **nicht-zufällige** Verfahren. In diesem Buch werden wir uns fast ausschließlich mit zufälligen Verfahren beschäftigen. Bei den nicht-zufälligen Stichprobenverfahren ist der Prozess der Auswahl weder kontrollierbar noch kann er mit statistischen Modellen beschrieben werden. Daher gibt es auch keine theoretische Grundlage für diese Verfahren. Wir diskutieren die Problematik kurz anhand der wichtigsten in der Praxis verwendeten nicht-zufälligen Stichprobenverfahren.

### 2.2.1 „Auswahl auf's Geratewohl“

Hierbei wird der „Mann auf der Straße“ befragt. Diese einfache Methode kann durchaus zu interessanten Ergebnissen führen, lässt aber in der Regel keine Schlüsse auf die Population zu. Insbesondere kann nicht kontrolliert werden, welche Individuen keine Möglichkeit haben, in die Stichprobe zu gelangen. Ein Beispiel soll die Problematik offenlegen.

*Beispiel 2.3:* Ein Supermarkt möchte Informationen über die Kundenzufriedenheit sammeln und entschließt sich zu einer Umfrage. Dabei wird eine Auswahl auf's Geratewohl getroffen, indem an einem Vormittag 100 Kunden nach Bezahlen an der Kasse nach ihrer Zufriedenheit mit Service und Sortiment befragt werden. Die so erhaltenen Ergebnisse mögen für das Unternehmen von Interesse sein, sofern die Population, über die eine Aussage getroffen werden soll, die derzeitigen Kunden des Supermarktes sind, die zu der entsprechenden Zeit, in der die Befragung durchgeführt wird, üblicherweise einkaufen. Kunden, die zu anderen Zeiten einkaufen oder enttäuschte Ex-Kunden, die den Supermarkt inzwischen meiden, erreicht man mit so einer Auswahl auf's Geratewohl nicht. Der Informationsgewinn der Stichprobe ist somit recht gering und fragwürdig. Bei einer Auswahl aufs Geratewohl ist somit weder garantiert, dass alle Individuen der Population eine positive Wahrscheinlichkeit haben in die Stichprobe gezogen zu werden, noch ist in irgendeiner Form bewertbar, wie groß die Wahrscheinlichkeit für ein Individuum ist gezogen zu werden. Kurzum, die Auswahl ist nicht zufällig. ◀

*Beispiel 2.4:* Sogenannte TED-Umfragen, bei denen Fernsehzuschauer um eine Meinung per Telefon gebeten werden, gehören ebenfalls zur Auswahl auf's Geratewohl. Hier ist davon auszugehen, dass die Entscheidung, seine Meinung abzugeben, von der Meinung selbst stark abhängt. Die Gruppe der nicht interessierten Zuschauer wird in der Regel keine Meinung abgeben oder gar ein anderes Programm schauen. ◀

## 2.2.2 Typische Stichprobe

Hierbei befragt man eine „typische“ Person oder wählt ein typisches Element der Population. Dieses Verfahren ist so gut wie die Experten, die festlegen, was „typisch“ ist. Ein Nachteil des Verfahrens besteht darin, dass eine Genauigkeits-schätzung praktisch nicht möglich ist. Dennoch findet das Verfahren Anwendung wie Beispiele 2.5 und 2.6 zeigen.

*Beispiel 2.5:* Zur Ermittlung der Inflationsrate wird zur Erstellung einer Preisstatistik ein bestimmter Warenkorb ausgewählt und die Preise in typischen Geschäften festgestellt. Aufgrund der Vielzahl der Waren und Verkaufseinheiten ist eine Ziehung mittels einer Zufallsstichprobe nicht möglich. Aufgrund der Erhebung gleicher Produkte in gleichen Verkaufseinheiten zu verschiedenen Zeitpunkten lässt sich damit die Preisentwicklung durchaus zuverlässig erheben. ◀

*Beispiel 2.6:* Die Stadt Haßloch in Rheinland-Pfalz dient der Gesellschaft für Konsumforschung (GfK) als Testmarkt für neue Produkte. Hier gelangen Innovationen in die Regale von Geschäften und Supermärkten, bevor sie im Bundesgebiet auf den Markt kommen. Ist das Produkt in Haßloch erfolgreich, so lohnt die bundesweite Einführung, ansonsten wird das Produkt nicht auf den

Markt kommen. Die Konsumenten in Haßloch dienen somit als typische Stichprobe für die bundesdeutsche Bevölkerung im Hinblick auf den Konsum von Lebensmitteln. ◀

### 2.2.3 Quotenstichprobe

Dieses Verfahren wird hauptsächlich bei Umfragen verwendet. Die Idee besteht darin, ein möglichst gutes Abbild der Population (Bevölkerung) in der Stichprobe zu bekommen. Dies wird dadurch erreicht, dass zunächst gewisse Quotenmerkmale (z.B. Geschlecht, Altersgruppe, Berufstätigkeit) festgelegt werden, deren Verteilungen in der Population bekannt sind. Anschließend wird die Stichprobe so gezogen, dass die Anteile dieser Merkmale in der Stichprobe genau denen in der Population entsprechen. Eine Befragung von 1 000 Personen ist dann zum Beispiel so zu organisieren, dass 500 Personen weiblich sind, dass 200 Personen zwischen 21 und 30 Jahren alt sind usw. In der Praxis wird das so erreicht, dass jeder beteiligte Interviewer genaue Vorgaben erhält, sich Personen mit bestimmten Eigenschaften bezüglich der Quotenmerkmale zu suchen.

Die Diskussion über Vor- und Nachteile einer Quotenauswahl war für die Entwicklung der Statistik sehr nützlich, siehe dazu z.B. Noelle-Neumann (2000) und Quatember (1996). Wesentliches Argument für die Quotenstichprobe ist die Kontrolle relevanter Störgrößen. Hier gibt es Ähnlichkeiten zur Strategie der geschichteten Stichprobe, siehe dazu Abschn. 5.1. Allerdings handelt es sich bei der Auswahl innerhalb der Quoten wieder um eine Auswahl auf's Geratewohl. Daher ist auch für eine Quotenstichprobe eine zuverlässige Abschätzung der Genauigkeit problematisch. Verfahren zur Genauigkeitsabschätzung basieren in der Regel auf der Annahme, dass eine Quotenstichprobe ähnliche Eigenschaften wie eine Zufallsstichprobe aufweist.

## 2.3 Repräsentativität und Verzerrung

In der empirischen Forschung wird der Begriff „repräsentative Stichprobe“ in unterschiedlichen Bedeutungen verwendet. In der Marktforschung und in der empirischen Sozialforschung wird manchmal eine repräsentative Stichprobe als ein verkleinertes Abbild der Grundgesamtheit definiert. Typischerweise wird verlangt, dass personenbezogene Merkmale wie z.B. Alter, Geschlecht, Bildung und Berufstätigkeit in der Stichprobe eine möglichst ähnliche Verteilung haben wie in der Grundgesamtheit. Diese Forderung hat den Vorteil, dass sie in der Praxis einfach zu überprüfen ist, wenn die entsprechenden Anteile in der Grundgesamtheit bekannt sind. Allerdings ist dadurch noch nicht gesichert, dass bezüglich der interessierenden Variablen die Ergebnisse der Stichprobe auf die Grundgesamtheit übertragbar sind.

*Beispiel 2.7:* Parteipräferenz bei der Kommunalwahl

Ein Landkreis besteht aus 10 Gemeinden, von welchen angenommen wird, dass diese bezüglich der Bevölkerungsmerkmale eine sehr ähnliche Struktur

aufweisen. Nach obiger Definition wären die Bürger der Gemeinde *A* also eine repräsentative Stichprobe für den gesamten Landkreis. Befragt man diese nach ihren Konsumgewohnheiten, ist das Ergebnis von Gemeinde *A* vermutlich auf den Landkreis übertragbar.

Das Ergebnis der Frage nach der Parteipräferenz für die Partei *S* bei der nächsten Kommunalwahl könnte sich aber in der Gemeinde *A* von dem Ergebnis im Landkreis deutlich unterscheiden, wenn die Gemeinde *A* beispielsweise einen besonders beliebten Bürgermeister aus der Partei *S* hat. Insofern sind Schlüsse auf die Grundgesamtheit bezüglich des Konsumverhaltens möglich, aber nicht bezüglich der Parteipräferenz. ◀

Das Beispiel zeigt die Problematik des Begriffs der repräsentativen Stichprobe. Die grundsätzliche Frage ist, ob Schlüsse von der Stichprobe auf die Grundgesamtheit zulässig sind. Das lässt sich bei nicht-zufälligen Stichproben kaum allgemein beantworten. Wir ziehen daher vor, den Begriff der **Repräsentativität** eher als die Zulässigkeit von Schlüssen auf die Grundgesamtheit zu definieren. Dabei gehört zu dem Begriff der Bezug zu den Merkmalen. Im obigen Beispiel ist die Gemeinde *A* eine repräsentative Stichprobe bezüglich des Konsumverhaltens, aber nicht bezüglich der Parteipräferenz. Man spricht im letzteren Fall von einer **verzerrten Stichprobe** oder von einer **Stichprobe mit systematischem Fehler**, auch Bias genannt. Dieser Begriff wird später exakt definiert und diskutiert. Wir wollen hier noch analysieren, wie es zu einer Verzerrung kommt.

### 2.3.1 Gründe für Verzerrung

Bei nicht-zufälligen Auswahlverfahren kommt es besonders dann zu einer Verzerrung, wenn das Verfahren der Auswahl mit dem Zielmerkmal in Zusammenhang steht. Im obigen Beispiel wird als Auswahlkriterium der Wohnort gewählt (alle Bewohner der Gemeinde *A*). Wenn nun – dank des Bürgermeisters – der Wohnort mit der Parteipräferenz in Zusammenhang steht, kommt es zu einer Verzerrung.

Ebenso ergibt sich eine Verzerrung, wenn man versucht, die Verteilung der Berufe in einer Stadt durch eine Befragung mittags vor einem Kaufhaus zu erheben, da bestimmte Berufsgruppen zu diesem Zeitpunkt nicht die Gelegenheit haben, einzukaufen. Hier ist also durch die Auswahlstrategie die Unbrauchbarkeit der Ergebnisse vorprogrammiert. Da nutzt es auch nichts zu versuchen, die Repräsentativität dadurch herzustellen, dass die Anteile der Geschlechter und die Altersverteilung der der Gesamtbevölkerung entsprechen. Man beachte auch, dass die Befragung von vielen Personen die Verzerrung nicht beseitigt.

Ein wesentlicher Vorteil einer Zufallsstichprobe liegt in der Vermeidung solcher Verzerrungen. Der Mechanismus des Ziehens ist dabei unabhängig von dem zu betrachtenden Merkmal. Zusammenhänge sind somit zufällig und nicht systematisch. Die gerade angesprochenen Probleme treten aber bei Zufallsstichproben durch Antwortverweigerung bzw. durch nicht erreichbare Individuen auf und werden daher in Kap. 7 diskutiert.

## 2.4 Design einer Zufallsstichprobe

Bei den nicht-zufälligen Auswahlverfahren ist der Auswahlmechanismus immer von dem Verhalten der Personen, die die Auswahl durchführen, abhängig. Sie ist im Extremfall der Auswahl auf's Geratewohl völlig der Stimmung der Interviewer oder der Beteiligten überlassen. Das Ergebnis wird zwar umgangssprachlich als „zufällig“ bezeichnet, aber es ist genau genommen vom subjektiven, nicht kontrollierbaren Verhalten beeinflusst. Im Gegensatz dazu wird bei Zufallsstichproben der Prozess der Ziehung genau definiert. Die Zufälligkeit einer Ziehung setzt damit einen echten Zufallsprozess voraus, was in der Praxis meist durch einen Zufallszahlengenerator realisiert wird. Salopp gesprochen unterscheidet sich eine zufällige Stichprobe von einer nicht zufälligen dadurch, dass wir quantifizierbare Wahrscheinlichkeiten dafür angeben können, dass ein Merkmalsträger in die Stichprobe gezogen wird. Diese Wahrscheinlichkeiten müssen nicht für alle Merkmalsträger gleich sein, sie müssen aber **vor** der Stichprobenziehung bekannt sein. Die explizite Angabe von derartigen Wahrscheinlichkeiten und entsprechenden Wahrscheinlichkeitsverteilungen in den Stichproben bezeichnen wir im Folgenden auch als **Design** oder **Stichprobendesign**.

*Beispiel 2.8:* Gegeben sei eine Population von 5 Merkmalsträgern (A,B,C,D,E). Es sollen 2 Einheiten in Form einer Stichprobe gezogen werden. Als Ergebnis der Stichprobe ergeben sich damit die folgenden Möglichkeiten:

$$\begin{aligned} S_1 &= (A, B), S_2 = (A, C), S_3 = (A, D), S_4 = (A, E), \\ S_5 &= (B, C), S_6 = (B, D), S_7 = (B, E), S_8 = (C, D), \\ S_9 &= (C, E), S_{10} = (D, E). \end{aligned}$$

Eine naheliegende Möglichkeit ist es, allen 10 Stichproben die gleiche Wahrscheinlichkeit zuzuordnen. Jede Stichprobe hat somit die Wahrscheinlichkeit  $1/10$ . Dieses Design wird als **einfache Zufallsstichprobe** bezeichnet. Es können aber auch andere Strategien verfolgt werden. Beispielsweise könnten wir verlangen, dass in der Stichprobe ein Konsonant und ein Vokal vorkommen, womit nur die folgenden Stichproben

$$\begin{aligned} S_1 &= (A, B), S_2 = (A, C), S_3 = (A, D), \\ S_7 &= (B, E), S_9 = (C, E), S_{10} = (D, E) \end{aligned}$$

zulässig wären. Diesen ordnet man dann jeweils die Wahrscheinlichkeit  $1/6$  zu. Ein derartiges Design werden wir als geschichtete Stichprobe kennen lernen. Weiter nehmen wir an, dass das Element A besonders wichtig sei und man deswegen eine Stichprobe ziehen möchte, in der A ein höheres Gewicht bekommt, d.h. dass alle Stichproben, die A enthalten, eine größere Wahrscheinlichkeit erhalten. Die Wahrscheinlichkeiten für die einzelnen Stichproben könnten wie folgt gesetzt werden:

$$P(S_1) = \dots = P(S_4) = \frac{2}{14}, P(S_5) = \dots = P(S_{10}) = \frac{1}{14}.$$

Auch solche Designs werden wir in diesem Buch betrachten. Wir werden sie als Ziehen proportional zur Größe (oder englisch „probabilities proportional to size“, kurz PPS) bezeichnen. Wir fassen die angesprochenen Designs in Tabelle 2.1 zusammen. Diese gibt die Wahrscheinlichkeiten für die einzelnen Stichproben wieder:

**Tabelle 2.1** Wahrscheinlichkeiten bei verschiedenen Designs

Stichprobe	Wahrscheinlichkeit bei		
	Design 1	Design 2	Design 3
(A,B)	1/10	1/6	1/7
(A,C)	1/10	1/6	1/7
(A,D)	1/10	1/6	1/7
(A,E)	1/10	0	1/7
(B,C)	1/10	0	1/14
(B,D)	1/10	0	1/14
(B,E)	1/10	1/6	1/14
(C,D)	1/10	0	1/14
(C,E)	1/10	1/6	1/14
(D,E)	1/10	1/6	1/14



Der wesentliche Vorteil von Zufallsstichproben besteht darin, dass mit Hilfe der Wahrscheinlichkeitsrechnung unter Berücksichtigung des Designs statistische Schlüsse auf die Population gezogen werden können. Insbesondere ist es möglich, neben Schätzungen für die interessierenden Größen der Grundgesamtheit, Angaben zur Genauigkeit der Schätzung zu machen. Die Genauigkeit hängt dabei von dem gewählten Design, vom Stichprobenumfang und von den Verhältnissen in der Population ab. Wir beginnen in diesem Kapitel mit dem einfachsten und am häufigsten verwendeten Design der einfachen Zufallsstichprobe. In den nachfolgenden Kapiteln diskutieren wir dann komplexere Designs.

## 2.5 Einfache Zufallsstichprobe

Das Design der einfachen Zufallsstichprobe zeichnet sich dadurch aus, dass jede Stichprobe vom Umfang  $n$  mit gleicher Wahrscheinlichkeit gezogen wird. Betrachten wir eine Population vom Umfang  $N$ , aus der wir eine Stichprobe vom Umfang  $n \leq N$  ziehen. Wir ziehen dabei **ohne Zurücklegen**, das heißt alle  $n$  gezogenen Individuen in der Stichprobe sind unterschiedlich. Mit Regeln der Kombinatorik erhalten wir

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

mögliche Stichproben vom Umfang  $n$ , wobei  $n! = n \cdot (n-1) \cdot \dots \cdot 1$  ist und  $0! = 1$  per Definition. Wenn  $n = N$  ist, so ergibt sich exakt eine mögliche Stichprobe, was einer Vollerhebung entspricht. Wenn  $n = 1$  ist, so erhält man  $N$  mögliche Stichproben. Eine einfache Zufallsstichprobe liegt nun vor, wenn jede mögliche Stichprobe mit gleicher Wahrscheinlichkeit gezogen wird.

### Das Design der einfachen Zufallsstichprobe

Gegeben sei eine Population  $G$  von  $N$  Elementen. Wir ziehen  $n$  verschiedene Elemente und erhalten die Stichprobe  $s$ . Dabei haben alle möglichen Stichproben vom Umfang  $n$  die gleiche Wahrscheinlichkeit, gezogen zu werden. Es gilt:

$$P(s) = \frac{1}{\binom{N}{n}},$$

für alle Stichproben (Teilmengen von  $G$ ) vom Umfang  $n$ .

Da bei der Ziehung kein Element der Population bevorzugt wird, hat jedes Element der Grundgesamtheit die gleiche Wahrscheinlichkeit, in die Stichprobe gezogen zu werden. Diese Wahrscheinlichkeit beträgt  $\pi = n/N$  und wird als **Auswahlwahrscheinlichkeit** bezeichnet. Intuitiv lässt es sich damit begründen, dass wir  $n$  Elemente aus  $N$  verfügbaren Elementen ziehen. Greifen wir dazu das Beispiel 2.8 mit Design 1 nochmals auf. In Tabelle 2.1 sind die Wahrscheinlichkeiten für jede mögliche Zufallsstichprobe vom Umfang  $n = 2$  aus der Grundgesamtheit vom Umfang  $N = 5$  gegeben. Wir betrachten Design 1. Es ist ersichtlich, dass z.B. der Buchstabe A in 4 Stichproben vorkommt. Also ist die Auswahlwahrscheinlichkeit für A gerade  $4/10 = 2/5$ .

*Herleitung:* Allgemein lässt sich die Auswahlwahrscheinlichkeit  $\pi$  für einfache Zufallsstichproben wie folgt herleiten. Bei der einfachen Zufallsstichprobe tritt jede Stichprobe vom Umfang  $n$  mit der gleichen Wahrscheinlichkeit  $1/\binom{N}{n}$  auf. Um nun die Auswahlwahrscheinlichkeit  $\pi$  für ein Individuum zu berechnen, müssen wir die Anzahl der Stichproben bestimmen, die jenes Element enthalten. Da das entsprechende Element in der Stichprobe sein muss, können wir nur noch  $n-1$  Elemente aus den verbleibenden  $N-1$  Elementen ziehen, um die Stichprobe aufzufüllen. Wir erhalten also  $\binom{N-1}{n-1}$  Stichproben, die das entsprechende Element enthalten. Die Wahrscheinlichkeit  $\pi$ , dass wir das entsprechende Element ziehen, ergibt sich daher zu

$$\pi = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

Die Anzahl der „günstigen“ Stichproben geteilt durch die Anzahl aller möglichen Stichproben, liefert die Wahrscheinlichkeit für ein Individuum, in die Stichprobe vom Umfang  $n$  zu gelangen. ■

Wir wollen uns an dieser Stelle Gedanken darüber machen, wie eine einfache Zufallsstichprobe praktisch vollzogen werden kann. Um die Zufallsauswahl durchzuführen, könnten aus einer Urne mit  $N$  Losen  $n$  Lose gezogen werden, wie dies z.B. bei der Ziehung der Lottozahlen durchgeführt wird. In der Praxis werden dazu heutzutage Computerprogramme benutzt. Man bezeichnet sie als Zufallsgeneratoren. Auf technische Aspekte und die Realisierung in Programmpaketen gehen wir in Abschn. 2.12 ein. An dieser Stelle wollen wir ein anderes Problem bei der Umsetzung von Stichproben ansprechen. Die Frage ist, wie eine numerische Zufallszahl mit den Individuen der Population in Verbindung zu bringen ist. Dazu nehmen wir an, dass die Elemente der Grundgesamtheit durchnummeriert sind. Wir haben also eine Liste der Zahlen 1 bis  $N$  vorliegen, von denen jeder Eintrag exakt einem Merkmalsträger der Grundgesamtheit zugeordnet wird. Exemplarisch ist dies in Abb. 2.1 dargestellt. Wir bezeichnen die Liste im Folgenden auch als **Populationsliste**. Für eine einfache Zufallsstichprobe ziehen wir nun  $n$  Zufallszahlen aus der Populationsliste. Da jede Zahl in der Liste exakt einem Merkmalsträger in der Grundgesamtheit entspricht, haben wir somit eine einfache Zufallsstichprobe gezogen.

Auch wenn sich das Verfahren im Prinzip einfach anhört, so sind mit der Realisation durchaus große Schwierigkeiten verbunden. Wie kann man zum Beispiel

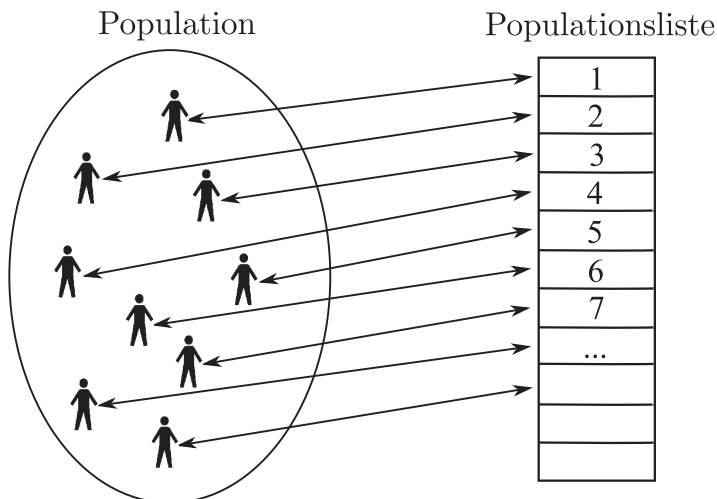
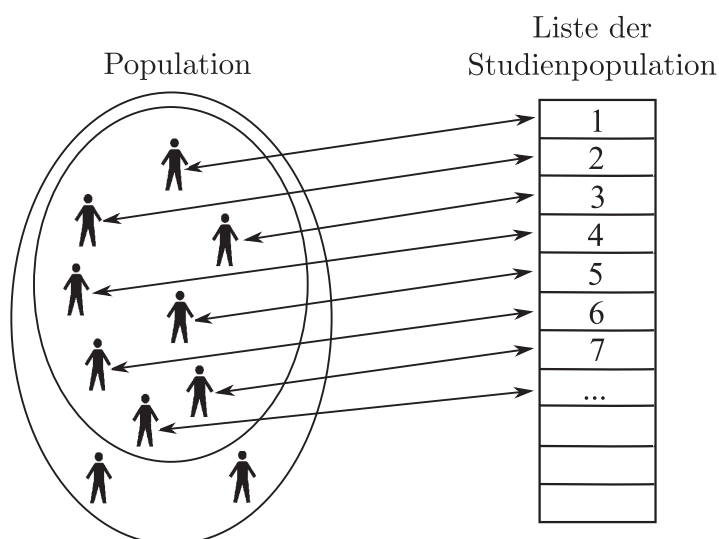


Abb. 2.1 Population und Populationsliste

auf einfache Weise eine derartige numerische Populationsliste erstellen oder wie kann man auf vorhandene Listen zurückgreifen? Betrachten wir dazu das folgende Beispiel.

*Beispiel 2.9:* Für eine Meinungsumfrage sollen 1 000 Haushalte in Deutschland kontaktiert werden. Diese Haushalte sollen nach dem Design der einfachen Zufallsstichprobe ausgewählt werden. Um diese Stichprobe zu ziehen, könnten z.B. zufällig Telefonnummern gewählt werden. Hierzu werden üblicherweise per Computer Telefonnummern zufällig gewählt. Der Interviewer, der die Befragung durchführt, wird dann mit dem ausgewählten Telefonanschluss verbunden. Mögliche Probleme bei diesem Verfahren sind unter anderem, dass Haushalte ohne Telefon eine Wahrscheinlichkeit von 0 haben, in die Stichprobe gezogen zu werden. Somit kann als Population nicht die Menge der Haushalte in Deutschland dienen, sondern nur die Menge der Haushalte mit Telefonanschluss. Diese Einschränkung kann zu einer systematischen Verzerrung führen, was wir im späteren Verlauf nochmals aufgreifen und weiter thematisieren wollen. ◀

Das Beispiel zeigt, dass in vielen Fällen, in denen eine einfache Zufallsstichprobe gezogen werden soll, Elemente in der Population existieren können, die eine Wahrscheinlichkeit von 0 besitzen, in die Stichprobe gezogen zu werden. Wir unterscheiden daher zwischen Population und Studienpopulation. Graphisch ist dies in Abb. 2.2 dargestellt.



**Abb. 2.2** Population, Studienpopulation und Liste der Studienpopulation

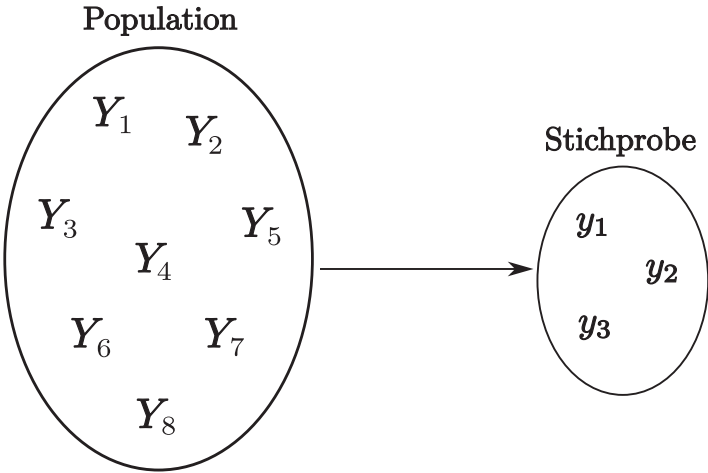
- Die **Studienpopulation** ist die Teilmenge der Population, die eine echt positive Wahrscheinlichkeit hat, in die Stichprobe gezogen zu werden. Allgemein gilt, dass wir generell nur eine Aussage über die Studienpopulation treffen können. Bestehen also zwischen Population und Studienpopulation relevante Unterschiede bezüglich des oder der interessierenden Merkmale, so ist die Stichprobenziehung basierend auf der gewählten Studienpopulation als kritisch zu betrachten.

*Beispiel 2.10:* Ein Internetversandhaus möchte eine Untersuchung zum Zahlungsverhalten der Kunden durchführen, bei der die Zeit zwischen Versand und Bezahlung als interessierende Variable erhoben werden soll. Als Population, sprich Menge der interessierenden Objekte, definiert man daher die Bestellungen beim Versandhaus. Nun will das Unternehmen natürlich nicht nur rückblickend, sondern auch vorausschauend die Ergebnisse der Untersuchung nutzen. Das bedeutet inhaltlich, dass die Population aus bisherigen **und** zukünftigen Bestellungen bestehen soll. Die Stichprobenziehung kann sich aber nur auf abgeschlossene Bestellungen beziehen. Das heißt, die Studienpopulation (bestehend aus den abgeschlossenen Bestellungen) ist nur eine Teilmenge der Bestellungen bei dem Unternehmen, über die eine Aussage getroffen werden soll. ◀

## 2.6 Statistische Inferenz

### 2.6.1 Notation

Wir wollen nun den Informationsgehalt einer einfachen Zufallsstichprobe mit statistischem Instrumentarium bewerten. Hierzu führen wir im Folgenden eine Notationskonvention ein, um die Population zu beschreiben. Wir gehen zunächst von einem Merkmal  $Y$  aus. Die Größen  $Y_1, \dots, Y_N$  sind die Merkmalsausprägungen in der Grundgesamtheit, das heißt  $Y_i$  ist beispielsweise das Alter oder das monatliche Einkommen der  $i$ -ten Person in der Population. Wir interessieren uns in der Regel für die Werte, die aus den  $Y_i$  abgeleitet werden, wie z.B. den Mittelwert oder die Varianz in der Population, also das mittlere Alter oder das mittlere Einkommen als Beispiele für Mittelwerte. Solche abgeleiteten Größen bezeichnen wir als **Parameter**. Ziel einer statistischen Erhebung ist es, diese Parameter zu schätzen. Dazu nutzen wir die Merkmalsausprägungen in der Stichprobe. Diese bezeichnen wir mit kleinen Buchstaben, also mit  $y_1, \dots, y_n$ , und nennen sie **Beobachtungen**. Damit ist  $y_k$  beispielsweise das Alter oder das Monatseinkommen der  $k$ -ten befragten und in die Stichprobe aufgenommenen Person. Aus der Stichprobe leiten wir sogenannte **Statistiken** oder **Schätzer** her, wie zum Beispiel den Mittelwert oder die Varianz in der Stichprobe. Somit beziehen sich große Buchstaben auf die Population, kleine Buchstaben sind Größen der Stichprobe. Schematisch ist dies in Abb. 2.3 dargestellt. Schätzer von Parametern einer Population notieren wir nachfolgend auch mit einem Dach  $\hat{\phantom{x}}$ .



**Abb. 2.3** Schematische Darstellung einer Stichprobenziehung

Für unsere weiteren Betrachtungen werden wir die folgende Notation verwenden. Auf die jeweiligen Größen wird in den nachfolgenden Abschnitten näher eingegangen.

Größe	Bedeutung
<b>In der Population:</b>	
$Y_i, i = 1, \dots, N$	Variable oder Merkmal des $i$ -ten Merkmalsträgers in der Population
$N$	Populationsumfang
$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$	Mittelwert des Merkmals in der Population
$S^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2$	Varianz des Merkmals in der Population
<b>In der Stichprobe:</b>	
$y_k, k = 1, \dots, n$	Variable oder Merkmal des $k$ -ten Merkmalsträgers in der Stichprobe
$n$	Stichprobenumfang
$\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k$	Mittelwert des Merkmals in der Stichprobe

$s^2 = \frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2$   
 $\hat{\bar{Y}}$

Varianz des Merkmals in der Stichprobe  
Schätzer für den Mittelwert in der Population

Wir verwenden im Folgenden in der Regel die Indizes  $i$  und  $j$  für Größen, die sich auf die Population beziehen und  $k$  und  $l$  als Indizes für die Variablen der Stichprobe.

2.6.2 Mittelwertschätzung

Um mit den eingeführten Begriffen vertraut zu werden und um verschiedene Eigenschaften zu veranschaulichen, betrachten wir zunächst ein kleines Beispiel.

*Beispiel 2.11:* Wir stellen uns eine kleine Population vom Umfang  $N = 5$  vor. Das interessierende Merkmal hat die Ausprägungen

$Y_1 = 9, \quad Y_2 = 10, \quad Y_3 = 11, \quad Y_4 = 18, \quad Y_5 = 22.$

Der interessierende Parameter ist der Mittelwert der  $Y$ -Werte. Hier ist  $\bar{Y} = 14$ . Um diesen Wert zu schätzen, ziehen wir eine einfache Zufallsstichprobe vom Umfang  $n = 3$  ohne Zurücklegen. Die Definition der einfachen Zufallsstichprobe besagt, dass jede mögliche Stichprobe mit gleicher Wahrscheinlichkeit auftritt. Es ergeben sich  $\binom{N}{n} = \binom{5}{3} = 5!/(2! 3!) = 10$  mögliche Stichproben, von denen jede mit gleicher Wahrscheinlichkeit, nämlich  $1/10$  auftritt. Damit erhalten wir folgende Schätzer und die Wahrscheinlichkeitsverteilung für den Mittelwert  $\bar{y}$  der Stichprobe.

Gezogene Individuen			Mittelwert der Stichprobe $\bar{y}$	Wahrscheinlichkeit
1	2	3	10,00	1/10
1	2	4	12,33	1/10
1	2	5	13,67	1/10
1	3	4	12,67	1/10
1	3	5	14,00	1/10
1	4	5	16,33	1/10
2	3	4	13,00	1/10
2	3	5	14,33	1/10
2	4	5	16,67	1/10
3	4	5	17,00	1/10

Wir können nun den Erwartungswert, d.h. den mittleren Wert über alle möglichen Stichproben, und die Streuung des Schätzers in Form der Varianz berechnen. Es ergibt sich als **Erwartungswert**

$$E(\bar{y}) = \frac{1}{10} 10,00 + \frac{1}{10} 12,33 + \dots + \frac{1}{10} 17,00 = 14,00$$

und als **Varianz**, d.h. als mittlere quadratische Abweichung vom Mittelwert

$$\begin{aligned}\text{Var}(\bar{y}) &= \frac{1}{10} (10,00 - 14,00)^2 + \frac{1}{10} (12,33 - 14,00)^2 + \dots \\ &\quad + \frac{1}{10} (17,00 - 14,00)^2 \\ &= 4,33.\end{aligned}$$

Beide Größen werden nachfolgend noch genauer definiert. Die **Standardabweichung** (sie entspricht der Quadratwurzel aus der Varianz) kann als ein Maß für die mittlere Abweichung vom Erwartungswert interpretiert werden. Sie hat hier den Wert  $\sqrt{4,33} = 2,08$ . ◀

Allgemein nehmen wir zunächst an, dass das interessierende Merkmal  $Y$  metrisch ist, also beispielsweise das Alter oder das Einkommen einer Person. Wie in obigem Beispiel ist man hierbei am Mittelwert der Merkmale  $Y_i, i = 1, \dots, N$  interessiert. Wir unterscheiden dabei Größen der Population und Größen der Stichprobe. Den Mittelwert der Population erhalten wir durch

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i,$$

wohingegen die einfache Zufallsstichprobe den Mittelwert

$$\hat{\bar{Y}}_{ES} = \bar{y} = \frac{1}{n} \sum_{k=1}^n y_k$$

liefert. Dabei ist die Größe  $\bar{Y}$  der gesuchte und unbekannte Mittelwert in der Population, der durch  $\hat{\bar{Y}}_{ES}$  als Stichprobenmittelwert geschätzt werden kann. Der Index  $ES$  steht für **Einfache Stichprobe**. Man beachte, dass  $\bar{Y}$  ein (unbekannter) fester Wert ist, während  $\hat{\bar{Y}}_{ES}$  eine Zufallsgröße ist, da diese von der Stichprobe abhängt.

Wir stellen uns nun die Frage welche Eigenschaften der Schätzer  $\hat{\bar{Y}}_{ES}$  hat. Ganz allgemein bewertet man einen Schätzer nach dem Schätzfehler. Dabei unterscheidet man zwischen dem systematischen und dem zufälligen Schätzfehler. Den systematischen Schätzfehler bezeichnen wir im Folgenden als **Bias**, den zufälligen Fehler messen wir in Form der Varianz. Diese Gütekriterien werden im Folgenden definiert.

Betrachten wir zuerst die Definition des Bias. Hierzu berechnen wir den **Erwartungswert** des Schätzers. Wir notieren den Erwartungswert mit  $E(\cdot)$ . Für die einfache Zufallsstichprobe gilt (Herleitung folgt später)

$$E(\hat{\bar{Y}}_{ES}) = \bar{Y},$$

und somit  $\text{Bias}(\widehat{\bar{Y}}_{ES}) = 0$ . Also liefert im Mittel das arithmetische Mittel der Stichprobe den gesuchten Parameter  $\bar{Y}$  der Population. Wir haben diesen Sachverhalt schon in dem kleinen Beispiel oben überprüft. Salopp gesprochen können wir sagen, dass wir im Mittel mit unserer Stichprobe richtig liegen.

Allerdings kann der Wert von  $\widehat{\bar{Y}}_{ES}$  von dem wahren Wert  $\bar{Y}$  je nach gezogener Stichprobe abweichen. Diese zufällige Abweichung wird durch die Varianz des Schätzers quantifiziert. Sie hängt von der Varianz der Variablen in der Population und dem Stichprobenumfang ab. Die **Varianz in der Population** ist ein Maß für die Streuung der einzelnen  $Y_i$ -Werte,  $i \in \{1, \dots, N\}$ , und ist definiert durch

$$S^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2.$$

Analog ist die **Varianz in der Stichprobe** definiert durch

$$s^2 = \frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2.$$

Die Größe  $S^2$  ist die Varianz von  $Y$  in der Population, die wir gelegentlich auch mit  $S_Y^2$  notieren. Diese ist, genau wie  $\bar{Y}$ , unbekannt. Basierend auf einer Stichprobe kann  $s^2$  als Schätzer für  $S^2$  herangezogen werden. Wie in den Herleitungen später gezeigt wird, führt die Division durch  $n-1$  (statt  $n$ ) zu einem annähernd unverzerrten Schätzer. Man achte an dieser Stelle auch auf die gewählte Notationskonvention, bei der Größen der Population mit großen Buchstaben notiert werden, wohingegen kleine Buchstaben für Größen der Stichprobe stehen.

Sofern der Stichprobenumfang kleiner ist als die Populationsgröße ( $n < N$ ), das heißt, sofern keine Vollerhebung (Zensus) durchgeführt wird, liefert eine Stichprobe **nicht** das exakte Ergebnis. Wir berechnen daher die Varianz des Schätzers als Maß für die Genauigkeit. Wie die weiter unten folgende Herleitung zeigt, gilt

$$\text{Var}(\widehat{\bar{Y}}_{ES}) = \frac{S^2}{n} \frac{N-n}{N-1}.$$

Damit ist die Standardabweichung

$$\text{STD}(\widehat{\bar{Y}}_{ES}) = \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

Nun ist  $S^2$  nicht bekannt, kann aber durch  $s^2$  geschätzt werden, was zur geschätzten Varianz führt:

$$\widehat{\text{Var}(\widehat{\bar{Y}}_{ES})} = \frac{s^2}{n} \frac{N-n}{N}.$$

### Gütekriterien für die Mittelwertschätzung

---

Gegeben sei ein Schätzer  $\hat{Y}$  für den Mittelwert  $\bar{Y}$ .  
Der **Bias** ist der **systematische Fehler** des Schätzers

$$\text{Bias}(\hat{Y}) = E(\hat{Y}) - \bar{Y}.$$

Ein Schätzer mit  $\text{Bias}(\hat{Y}) = 0$  heißt **unverzerrt** oder **erwartungstreu**.  
Der **zufällige Fehler** des Schätzers ist die **Varianz**

$$\text{Var}(\hat{Y}) = E\left[\hat{Y} - E(\hat{Y})\right]^2.$$

Anschaulicher ist die **Standardabweichung**

$$\text{STD}(\hat{Y}) = \sqrt{\text{Var}(\hat{Y})}.$$

Insgesamt wird die Schätzung durch den **mittleren quadratischen Fehler**  
(engl. „mean square error“) bewertet

$$\text{MSE}(\hat{Y}) := E(\hat{Y} - \bar{Y})^2.$$

Als intuitives Maß verwendet man die Wurzel des MSE  
(engl. „root mean square error“)

$$\text{RMSE} := \sqrt{\text{MSE}}.$$

Es gilt allgemein:

$$\text{MSE}(\hat{Y}) = \text{Var}(\hat{Y}) + [\text{Bias}(\hat{Y})]^2.$$

Insbesondere sind also  $\text{Var}(\hat{Y})$  und  $\text{MSE}(\hat{Y})$  für unverzerrte  
Schätzer identisch.

Aus den Formeln ist ersichtlich, dass die Standardabweichung des Schätzers direkt proportional zur Standardabweichung des Merkmals in der Grundgesamtheit ist. Weiterhin ist bei der Standardabweichung die Abhängigkeit vom Stichprobenumfang im Wesentlichen durch den Faktor  $1/\sqrt{n}$  gegeben. Im Gegensatz zur konventionellen Statistik taucht in den Formeln ein zusätzlicher Faktor der Form  $(N - n)/N = 1 - (n/N)$  auf. Dieser Faktor wird auch als **Korrekturfaktor für endliche Populationen** bezeichnet und  $n/N$  nennt man auch **Auswahlsatz**. Insbesondere bewirkt der Korrekturfaktor, dass für  $n = N$  die Varianz des Schätzers 0 ist. Das macht Sinn, bedeutet doch  $n = N$ , dass alle Elemente der Population in die Stichprobe aufgenommen werden, was inhaltlich einer Vollerhebung gleichkommt. Somit folgt für  $n = N$ , dass  $\bar{y} = \bar{Y}$  ist, also weist  $\hat{\bar{Y}}_{ES}$  eine Varianz von 0 auf. Ganz allgemein wird die Varianz des Mittelwertschätzers mit steigendem Auswahlsatz  $n/N$  kleiner. Der Korrekturfaktor ist notwendig, da jeder Merkmalsträger höchstens einmal in die Stichprobe gezogen wird, was wir als Ziehen ohne Zurücklegen bezeichnen. Wir greifen diesen Punkt in Abschn. 2.7 nochmal auf.

Es ist an dieser Stelle wichtig zu bemerken, dass gebräuchliche Softwarepakete den Korrekturfaktor vernachlässigen. Dies ist gerechtfertigt, wenn der Stichprobenumfang im Vergleich zum Populationsumfang klein ist, das heißt, wenn  $n \ll N$ . In diesem Fall ist  $(N - n)/N \approx 1$ . Möglichkeiten, diese Korrektur softwaretechnisch einzubauen, werden in Abschn. 2.12 aufgezeigt.

### Mittelwertschätzung bei einer einfachen Zufallsstichprobe (arithmetisches Mittel)

Gegeben sei eine Stichprobe  $y_1, \dots, y_n$  vom Umfang  $n$ , gezogen als einfache Zufallsstichprobe (ohne Zurücklegen) aus einer Population vom Umfang  $N$ .

Ein unverzerrter Schätzer für den Mittelwert  $\bar{Y}$  ist

$$\hat{\bar{Y}}_{ES} = \bar{y} = \frac{1}{n} \sum_{k=1}^n y_k.$$

Die Varianz von  $\hat{\bar{Y}}_{ES}$  kann erwartungstreu geschätzt werden durch

$$\text{Var}(\hat{\bar{Y}}_{ES}) = \frac{N-n}{N} \frac{1}{n(n-1)} \sum_{k=1}^n (y_k - \hat{\bar{Y}}_{ES})^2.$$

*Herleitung:* Nachfolgend leiten wir den Erwartungswert und die Varianz des Schätzers einer einfachen Zufallsstichprobe her. Um die folgenden Berechnungen sinnvoll durchführen zu können, gehen wir zu einer nach der Reihenfolge der Ziehung geordneten Stichprobe  $(y_1, \dots, y_n)$

über. Diese entspricht genau der Vorstellung vom Ziehen ohne Zurücklegen. Da die Reihenfolge berücksichtigt wird, hat jede geordnete Stichprobe die Wahrscheinlichkeit

$$\frac{1}{N(N-1)\dots(N-n+1)} = \frac{1}{\binom{N}{n} n!}.$$

Wir beginnen mit der Wahrscheinlichkeitsverteilung der ersten gezogenen Einheit  $y_1$ . Da alle Einheiten der Population die Wahrscheinlichkeit  $1/N$  haben, im ersten Zug gezogen zu werden, gilt

$$P(y_1 = y) = \frac{1}{N} \sharp\{i | Y_i = y\}.$$

Dabei bedeutet das Zeichen  $\sharp$  die Anzahl der Elemente der entsprechenden Menge. Die Wahrscheinlichkeit für  $y$  ist also die relative Häufigkeit von  $y$  in der Grundgesamtheit. Damit entspricht die Wahrscheinlichkeitsverteilung von  $y_1$  der festen, im Allgemeinen unbekannten empirischen Verteilung des Merkmals  $Y$  in der Population. Dieser einfache Zusammenhang bildet die Basis für die statistische Analyse der einfachen Zufallsstichprobe. Als erste Folgerung ergibt sich

$$E(y_1) = \frac{1}{N} \sum_{i=1}^N Y_i = \bar{Y}.$$

Im zweiten Schritt zeigen wir, dass die Wahrscheinlichkeitsverteilung von  $y_k$  für alle  $k$  identisch ist. Das heißt insbesondere, dass die Wahrscheinlichkeitsverteilung des  $k$ -ten Zuges gleich der Wahrscheinlichkeitsverteilung des ersten Zuges ist. Man beachte, dass es sich dabei um die Verteilung von  $y_k$  ohne Betrachtung der vorherigen Ziehungen handelt. Wenn die Ziehungen vor  $k$  bekannt sind, ist die obige Aussage nicht mehr gültig. Intuitiv lässt sich die Aussage damit begründen, dass das Ziehen ohne Zurücklegen im Prinzip auch in einem Schritt erfolgen kann. Damit ist die Nummerierung der gezogenen Elemente eigentlich unerheblich und hat daher keinen Einfluss auf die Verteilung. Dieses Argument kann wie folgt formalisiert werden:

$$P(y_k = y) = \frac{\sharp\{(i_1, i_2, \dots, i_n) | y_{i_k} = y, i_k, i_l \in \{1, \dots, N\}, i_k \neq i_l \text{ für } k \neq l\}}{\binom{N}{n} n!}.$$

Die Wahrscheinlichkeit entspricht also der Zahl aller Stichproben, bei denen an der  $k$ -ten Stelle ein Element mit der Ausprägung  $y$  gezogen wird, geteilt durch die Gesamtzahl aller Stichproben. Die Anzahl im Zähler ist offensichtlich nicht vom Index  $k$  abhängig. Daher gilt  $P(y_k = y) = P(y_1 = y)$ .

Damit entsprechen auch die Verteilungen der anderen Züge der empirischen Verteilung des Merkmals in der Population. Somit können wir auf einfache Weise den Erwartungswert des Stichprobenmittels berechnen:

$$E(\hat{Y}_{ES}) = E\left(\frac{1}{n} \sum_{k=1}^n y_k\right) = \frac{1}{n} \sum_{k=1}^n E(y_k) = \frac{1}{n} n E(y_1) = \bar{Y}.$$

Die Varianz von  $\hat{Y}_{ES}$  ergibt sich wie folgt:

$$\text{Var}(\hat{Y}_{ES}) = \text{Var}\left(\frac{1}{n} \sum_{k=1}^n y_k\right) = \frac{1}{n^2} \sum_{k=1}^n \sum_{\substack{l=1 \\ k \neq l}}^n \text{Cov}(y_k, y_l) + \frac{1}{n^2} \sum_{k=1}^n \text{Var}(y_k).$$

Für die Einzelvarianzen  $\text{Var}(y_k)$  gilt:

$$\text{Var}(y_k) = \text{Var}(y_1) = E\left(y_1^2\right) - (E(y_1))^2 = \frac{1}{N} \sum_{i=1}^N Y_i^2 - \bar{Y}^2 = S^2.$$

Zur Berechnung der Kovarianzen  $\text{Cov}(y_k, y_l)$  benutzen wir wie oben die Symmetrieeigenschaft des Ziehens ohne Zurücklegen. Es gilt für  $k \neq l$ :

$$P(y_k = y^{(1)}, y_l = y^{(2)}) = \frac{\#\{(i_1, \dots, i_n) | y_{i_k} = y^{(1)}, y_{i_l} = y^{(2)}, i_k, i_l \in \{1, \dots, N\}, i_k \neq i_l \text{ für } k \neq l\}}{\binom{N}{n} n!}.$$

Auch hier ist diese Wahrscheinlichkeit nicht von den Indizes  $k$  und  $l$  abhängig. Damit ist die gemeinsame Wahrscheinlichkeitsverteilung von  $y_k$  und  $y_l$  identisch mit der von  $y_1$  und  $y_2$  und es gilt

$$\text{Cov}(y_k, y_l) = \text{Cov}(y_1, y_2).$$

Es bleibt somit die Kovarianz  $\text{Cov}(y_1, y_2)$  zu berechnen. Für diese gilt

$$\begin{aligned} E(y_1 y_2) &= \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{\substack{j=1 \\ i \neq j}}^N Y_i Y_j \\ &= \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N Y_i Y_j - \frac{1}{N(N-1)} \sum_{i=1}^N Y_i^2 \\ &= \bar{Y}^2 \frac{N^2}{N(N-1)} - \frac{1}{N(N-1)} \sum_{i=1}^N Y_i^2. \\ \text{Cov}(y_1, y_2) &= E(y_1 y_2) - E(y_1) E(y_2) \\ &= \bar{Y}^2 \frac{1}{N-1} - \frac{1}{N-1} \frac{1}{N} \sum_{i=1}^N Y_i^2 \\ &= -\frac{1}{N-1} S^2. \\ \text{Var}(\widehat{Y}_{ES}) &= \frac{1}{n^2} \cdot n(n-1) \cdot \left(-\frac{1}{N-1} S^2\right) + \frac{1}{n^2} n S^2 \\ &= \frac{S^2}{n} \left(1 - \frac{n-1}{N-1}\right) \\ &= \frac{S^2}{n} \frac{N-n}{N-1}. \end{aligned}$$

$$\begin{aligned}
E(s^2) &= E \left[ \frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2 \right] \\
&= \frac{1}{n-1} E \left[ \left( \sum_{k=1}^n y_k^2 \right) - n \bar{y}^2 \right] \\
&= \frac{1}{n-1} \left\{ n(\bar{Y}^2) - n \left[ (E(\bar{y}))^2 + \text{Var}(\bar{y}) \right] \right\} \\
&= \frac{1}{n-1} \left\{ n(\bar{Y}^2) - n \bar{Y}^2 - n \frac{1}{n} \frac{N-n}{N-1} S^2 \right\} \\
&= \frac{1}{n-1} \left\{ n S^2 - \frac{N-n}{N-1} S^2 \right\} \\
&= S^2 \frac{N}{N-1}.
\end{aligned}$$

Aus  $E(s^2) = S^2 \frac{N}{N-1}$  folgt nun unmittelbar

$$E \left[ \widehat{\text{Var}}(\hat{Y}_{ES}) \right] = \text{Var}(\hat{Y}_{ES}).$$

■

### 2.6.3 Konfidenzintervalle

Die Varianz (bzw. die Standardabweichung) ist nur ein mögliches Maß, um die Unsicherheit des Schätzers anzugeben. Eine anschauliche und in der Praxis verbreitete Alternative dazu ist die Angabe eines Bereiches (eines Intervalls), in dem der wahre Wert liegen soll. Man spricht im Allgemeinen von „Intervallschätzung“ und verlangt, dass dieser Bereich den wahren Wert mit einer vorgegebenen Wahrscheinlichkeit von 95 bzw. 99% (auch andere Werte sind möglich) enthält. Um ein solches Intervall zu erhalten, benötigt man die Verteilung des Schätzers. Hierzu benutzt man das Konzept des **Zentralen Grenzwertsatzes**.

In seiner einfachsten Form besagt der Zentrale Grenzwertsatz, dass die Summe von unabhängigen und identisch verteilten Zufallsgrößen für wachsenden Stichprobenumfang approximativ normalverteilt ist. Diese Aussage lässt sich auf den Fall einer Stichprobe aus einer endlichen Grundgesamtheit übertragen. Hier haben wir jedoch mit der zusätzlichen Hürde zu kämpfen, dass der Ziehungsprozess ohne Zurücklegen erfolgt und somit die gezogenen Elemente (und damit die Zufallsgrößen  $y_k$ ) nicht unabhängig sind. Allerdings lässt sich unter weiteren technischen Voraussetzungen eine asymptotische Theorie für Stichprobenszenarien entwickeln, aus der auch die approximative Normalverteilung des Stichprobenmittels für große Stichproben und große Grundgesamtheiten folgt, siehe dazu etwa Thompson (2002). Weiter wurde in verschiedenen Simulationsstudien gezeigt, dass die asymptotische Normalverteilung in vielen praktischen Fällen angemessen erscheint, siehe dazu etwa Cochran (1977). Wir veranschaulichen die asymptotische Normalität an einem kleinen Simulationsbeispiel.

*Beispiel 2.12:* Nehmen wir an, unsere Population bestehe aus den 100 Elementen

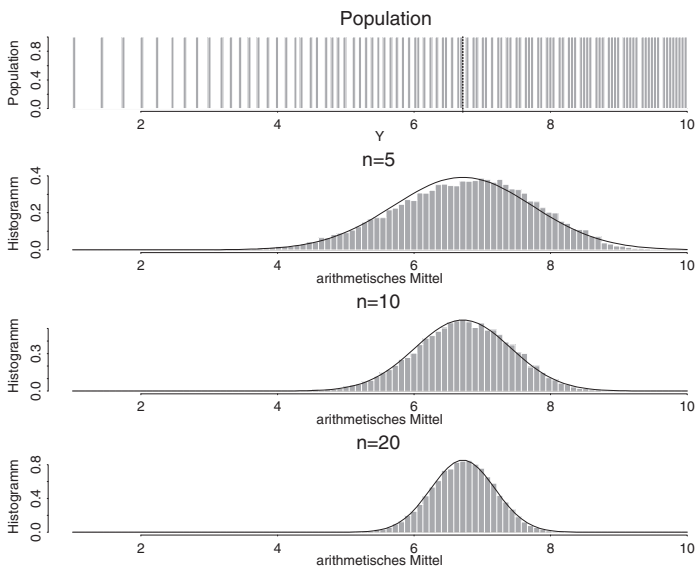
$$Y_1 = \sqrt{1}, \quad Y_2 = \sqrt{2}, \quad \dots, \quad Y_{99} = \sqrt{99}, \quad Y_{100} = \sqrt{100}.$$

Wir ziehen nun eine einfache Zufallsstichprobe vom Umfang  $n$ , wobei wir exemplarisch  $n = 5, 10$  bzw.  $20$  setzen, was einem Auswahlsatz von  $5, 10$  bzw.  $20\%$  entspricht. In Abb. 2.4 zeigen wir die Verteilung von  $Y_i$  in der Population und von  $\hat{Y}_{ES}$ , basierend auf  $20\,000$  simulierten Stichproben. Es ist ersichtlich, dass die Verteilung von  $\hat{Y}_{ES}$  schon bei kleinem Stichprobenumfang einer Normalverteilung folgt, welche in den Graphiken als klassische Glockenkurve eingezeichnet ist.

Wir erweitern die Simulation auf den Fall einer kleineren Population und damit auf einen veränderten Auswahlsatz. Hierzu reduzieren wir  $N$  auf  $25$  mit den Werten

$$Y_1 = \sqrt{1}, \quad Y_2 = \sqrt{2}, \quad \dots, \quad Y_{24} = \sqrt{24}, \quad Y_{25} = \sqrt{25}.$$

Wieder ziehen wir  $n = 5, 10$  bzw.  $20$  Individuen durch eine einfache Zufallsstichprobe. Abbildung 2.5 zeigt die Verteilung von  $\hat{Y}_{ES}$ , wieder basierend auf  $20\,000$  Simulationen. Ist der Auswahlsatz groß (hier jetzt  $20, 40$  bzw.  $80\%$ ), so ist ersichtlich, dass die Glockenkurve, also die Normalverteilung, die Verteilung von  $\hat{Y}_{ES}$  ebenfalls gut approximiert. Außerdem ist die Varianz des Schätzers kleiner. In Abb. 2.5 ist nämlich zu beachten, dass die  $x$ -Achse jetzt nur noch bis  $5$  anstatt bis  $10$  reicht. ◀



**Abb. 2.4** Population und Verteilung von  $\hat{Y}_{ES}$  für verschiedene Stichprobenumfänge ( $N = 100$ )

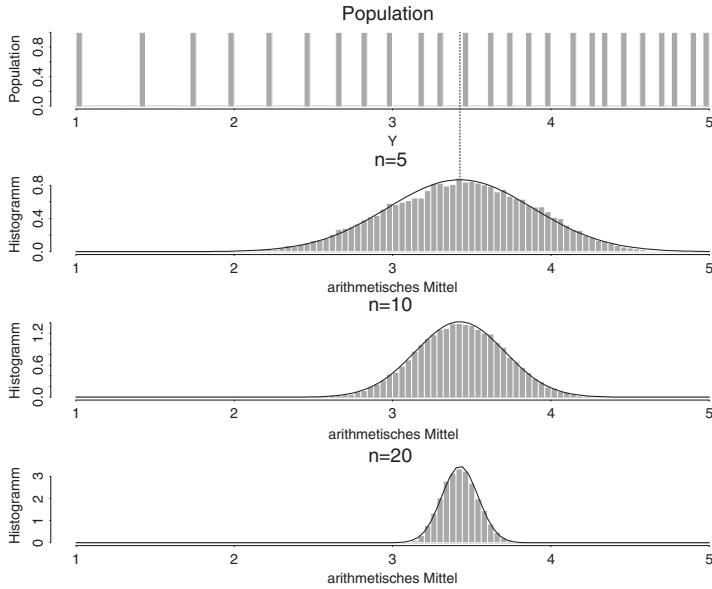


Abb. 2.5 Population und Verteilung von  $\hat{Y}_{ES}$  für verschiedene Stichprobenumfänge ( $N = 25$ )

Wir notieren die approximative Normalverteilung durch

$$\hat{Y} \stackrel{a}{\sim} N\left(\bar{Y}, \text{Var}(\hat{Y})\right).$$

Mit dieser Verteilungsannahme berechnen wir nun ein Intervall, das den gesuchten Parameter mit vorgegebener Wahrscheinlichkeit enthält. Dieses sogenannte **Konfidenzintervall** ist bestimmt durch

$$\left[ \hat{Y} - 1,96\sqrt{\text{Var}(\hat{Y})}, \hat{Y} + 1,96\sqrt{\text{Var}(\hat{Y})} \right].$$

Der Faktor 1,96 wird aus der Normalverteilungsannahme und der geforderten Überdeckungswahrscheinlichkeit abgeleitet, die konventionell auf 95% gesetzt ist. Somit können wir sagen, dass

$$P\left(\bar{Y} \in \left[ \hat{Y} \pm 1,96\sqrt{\text{Var}(\hat{Y})} \right]\right) \approx 0,95.$$

Das Konfidenzintervall beinhaltet also mit einer Wahrscheinlichkeit von 95% den unbekannten Parameter der Population. Allgemein wird das  $(1 - \frac{\alpha}{2})$ -Quantil  $z_{1-\frac{\alpha}{2}}$

der Standardnormalverteilung für ein Konfidenzintervall mit der Überdeckungswahrscheinlichkeit  $1 - \alpha$  verwendet, welches für  $\alpha = 0,05$  den Wert 1,96 liefert und für  $\alpha = 0,01$  den Wert 2,58. Da  $\text{Var}(\hat{Y})$  nicht bekannt ist, schätzt man diese durch  $\widehat{\text{Var}}(\hat{Y})$ .

### Konfidenzintervall

Unter der Annahme, dass ein Schätzer  $\hat{Y}$  approximativ normalverteilt ist, ergibt sich ein 95% Konfidenzintervall durch

$$\left[ \hat{Y} - 1,96\sqrt{\widehat{\text{Var}}(\hat{Y})}, \hat{Y} + 1,96\sqrt{\widehat{\text{Var}}(\hat{Y})} \right].$$

Ein  $(1 - \alpha)$ -Konfidenzintervall hat die Form

$$\left[ \hat{Y} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}(\hat{Y})}, \hat{Y} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{Var}}(\hat{Y})} \right].$$

wobei  $z_{1-\frac{\alpha}{2}}$  das  $(1 - \frac{\alpha}{2})$ -Quantil der Standardnormalverteilung ist.

Generelle und allgemeingültige Faustregeln, ab welchem Stichprobenumfang die approximative Normalität greift, können nicht gegeben werden. Die Faustregel  $n \geq 30$ , die an verschiedenen Stellen der statistischen Literatur geliefert wird, ist im Stichprobenfall sicher nur dann sinnvoll anzuwenden, wenn  $N$  hinreichend groß ist. Asymptotische Aussagen beruhen auf der Annahme, dass  $n$ ,  $N$  und  $N - n$  ansteigen, was aus praktischen Gesichtspunkten fraglich ist, da  $N$  der feste, gegebene Populationsumfang ist. Die Interpretation, dass mit 95% Überdeckungswahrscheinlichkeit das Intervall (den unbekannten Parameter)  $\bar{Y}$  überdeckt, kann daher fragwürdig sein, falls  $n$  klein oder  $n/N$  groß ist. Dies gilt insbesondere dann, wenn in der Grundgesamtheit extreme Werte vorliegen.

In der Literatur wird manchmal vorgeschlagen, statt des Quantils  $z_{1-\frac{\alpha}{2}}$  der Standardnormalverteilung das entsprechende Quantil der  $t$ -Verteilung zu verwenden (siehe z.B. Cochran 1977; An & Watts 2000), da die Unsicherheit bei der Schätzung der Varianz damit berücksichtigt wird. Da aber die Verwendung der  $t$ -Verteilung aus der Normalverteilungsannahme für die Verteilung des Merkmals in der Grundgesamtheit hergeleitet wird und diese in unserem Fall meist verletzt ist, ist auch dieses Vorgehen problematisch. Außerdem sind die Unterschiede für Stichproben vom Umfang  $n \geq 30$  praktisch vernachlässigbar.

### 2.6.4 Schätzung von Anteilen

In vielen Fragestellungen ist man an der Bestimmung eines Anteils interessiert. Beispielsweise soll durch eine Umfrage der Anteil der Studierenden bestimmt werden, die gegen die Einführung von Studiengebühren prinzipiell keine Einwände haben. In völliger Analogie zu den obigen Formeln definieren wir nun  $Y_i$  als die Antwort der Person  $i$ . Wir kodieren mit  $Y_i = 1$  die Personen, die auf die gestellte Frage mit „ja“ antworten und mit  $Y_i = 0$  die negativen Antworten. Für die in die Stichprobe gezogenen Merkmalsträger beobachten wir  $y_k$  mit  $y_k = 1$  oder  $y_k = 0$ . Interessiert sind wir an dem Anteilswert

$$P = \frac{1}{N} \sum_{i=1}^N Y_i,$$

der dem arithmetischen Mittel der  $Y_i$  entspricht. Die Größe  $P$  ist somit der (unbekannte) Anteil der Studierenden, die keine Einwände gegen die Einführung von Studiengebühren haben.

In der Grundgesamtheit sind also  $N \cdot P$  Personen mit  $Y_i = 1$  und  $N \cdot (1 - P)$  Personen mit  $Y_i = 0$ . Da wir bisher keine Voraussetzungen an die möglichen Werte  $Y$  gestellt haben, können die bisherigen Überlegungen auch auf diesen Fall angewendet werden. Durch die besonders einfache Struktur ergeben sich zusätzliche Möglichkeiten der Inferenz. Für die Varianz in der Grundgesamtheit gilt:

$$S^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{1}{N} \left[ NP(1 - P)^2 + N(1 - P)(0 - P)^2 \right] = P(1 - P).$$

Basierend auf einer Stichprobe schätzen wir  $P$  durch

$$\hat{P}_{ES} = p = \frac{1}{n} \sum_{k=1}^n y_k.$$

Als Varianz von  $p$  ergibt sich

$$\text{Var}(\hat{Y}_{ES}) = \text{Var}(p) = \frac{S^2}{n} \cdot \frac{N - n}{N - 1} = \frac{P(1 - P)}{n} \cdot \frac{N - n}{N - 1}.$$

Insbesondere hängt die Varianz von  $p$  vom unbekannten Anteil  $P$  ab. Dieser kann wiederum geschätzt werden und man erhält die geschätzte Varianz

$$\widehat{\text{Var}}(\hat{Y}_{ES}) = \widehat{\text{Var}}(p) = \frac{p(1 - p)}{n - 1} \cdot \frac{N - n}{N}. \quad (2.1)$$

Das Vorgehen bei der Anteilsschätzung entspricht dem Vorgehen bei der Mittelwertschätzung eines mit den Werten 0 und 1 kodierten Merkmals. Die Varianzschät-

zung entspricht ebenso dem Vorgehen bei der Mittelwertschätzung. Entsprechend kann man das 95%-Konfidenzintervall durch  $\left[ p - 1,96\sqrt{\widehat{\text{Var}}(p)}; p + 1,96\sqrt{\widehat{\text{Var}}(p)} \right]$  bestimmen.

### Anteilsschätzung bei einer einfachen Zufallsstichprobe

Gegeben sei die Stichprobe  $y_1, \dots, y_n$  vom Umfang  $n$ , gezogen mit dem Design der einfachen Zufallsstichprobe aus einer Population vom Umfang  $N$ . Dabei ist  $y_k \in \{0, 1\}$ .

Ein unverzerrter Schätzer für den Anteil  $P = \bar{Y}$  in der Grundgesamtheit ist

$$\widehat{P}_{ES} = p = \frac{1}{n} \sum_{k=1}^n y_k.$$

Die Varianz von  $\widehat{P}_{ES}$  kann geschätzt werden durch

$$\widehat{\text{Var}}(\widehat{P}_{ES}) = \frac{p(1-p)}{n-1} \frac{N-n}{N}.$$

In den meisten Fällen weist das so konstruierte Konfidenzintervall zufriedenstellende Eigenschaften auf. Dies ist jedoch nicht der Fall für hohe oder niedrige Anteilswerte, das heißt für  $P$  in der Nähe von 0 oder 1. Dann empfiehlt es sich, sogenannte **exakte Konfidenzintervalle** zu berechnen. Dazu benutzen wir die spezielle Struktur der Stichprobe. Wir bezeichnen die Anzahl der Personen mit  $Y_i = 1$  mit  $M$ , das heißt  $M = \sum_{i=1}^N Y_i = N \cdot P$ . Die Anzahl der Beobachtungen mit  $y_k = 1$  in der Stichprobe notieren wir der Konvention folgend mit  $m = \sum_{k=1}^n y_k$ . Die Wahrscheinlichkeitsverteilung von  $m$  kann nun durch die sogenannte hypergeometrische Verteilung charakterisiert werden. Die Wahrscheinlichkeitsfunktion lautet

$$P \left( \sum_{k=1}^n y_k = m \right) = \frac{\binom{M}{m} \binom{N-M}{n-m}}{\binom{N}{n}}.$$

Im Zähler steht genau die Anzahl der Stichproben, die zum Wert  $\sum_{k=1}^n y_k = m$  führt. Für Erwartungswert und Varianz der hypergeometrischen Verteilung gilt

$$E \left( \sum_{k=1}^n y_k \right) = n P,$$

$$\text{Var} \left( \sum_{k=1}^n y_k \right) = n P (1 - P) \frac{N - n}{N - 1}.$$

Nach Division durch  $n$  erhält man die entsprechenden Werte für die Anteilsschätzung. Um ein  $(1 - \alpha)$ -Konfidenzintervall zu erhalten, setzen wir  $\sum_{k=1}^n y_k = m$  und wählen nun als untere beziehungsweise obere Grenze für das unbekannte  $M$  die Größen  $U$  und  $O$ , so dass gilt

$$\sum_{r=0}^m \frac{\binom{O}{r} \binom{N-O}{n-r}}{\binom{N}{n}} = \alpha_1, \quad (2.2)$$

$$(2.3)$$

$$\sum_{r=m}^n \frac{\binom{U}{r} \binom{N-U}{n-r}}{\binom{N}{n}} = \alpha_2. \quad (2.4)$$

$$\alpha_1 + \alpha_2 \leq \alpha.$$

Damit ist  $[U; O]$  ein  $(1 - \alpha)$ -Konfidenzintervall für den Parameter  $M$ . Entsprechend ist  $[U/N; O/N]$  ein  $(1 - \alpha)$ -Konfidenzintervall für den Anteil  $P$ .

### Konfidenzintervall für Anteile

Gegeben sei die Stichprobe  $y_1, \dots, y_n$  vom Umfang  $n$ , gezogen mit dem Design der einfachen Zufallsstichprobe aus einer Population vom Umfang  $N$ . Dabei ist  $y_k \in \{0, 1\}$ .

Ein unverzerrter Schätzer für den Anteil  $P = \bar{Y}$  in der Grundgesamtheit ist

$$\hat{P}_{ES} = p = \frac{1}{n} \sum_{k=1}^n y_k.$$

Für große Stichprobenumfänge und mittlere Anteile  
 $p$  (wobei  $n \cdot p \cdot (1 - p) > 10$ )  
 hat das  $(1 - \alpha)$ -Konfidenzintervall die Form

$$\left[ p - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{n-1} \cdot \frac{N-n}{N}}; p + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{n-1} \cdot \frac{N-n}{N}} \right],$$

wobei  $z_{1-\frac{\alpha}{2}}$  das  $(1 - \frac{\alpha}{2})$ -Quantil der Standardnormalverteilung ist.

Ein exaktes  $(1 - \alpha)$ -Konfidenzintervall ist gegeben durch  $[\frac{U}{N}; \frac{O}{N}]$ , falls gilt

$$\sum_{r=0}^m \frac{\binom{O}{r} \binom{N-O}{n-r}}{\binom{N}{n}} = \alpha_1,$$

$$\sum_{r=m}^n \frac{\binom{U}{r} \binom{N-U}{n-r}}{\binom{N}{n}} = \alpha_2,$$

$$\alpha_1 + \alpha_2 \leq \alpha.$$

Auch wenn das Konzept der Bestimmung des exakten Konfidenzintervalls relativ einfach ist (siehe dazu auch Thompson 2002), ist die konkrete Umsetzung nur in einfachen Beispielen ohne den Einsatz eines Computers realisierbar. Zur numerischen Umsetzung siehe Abschn. 2.12.3. Weitere Probleme liegen in der Wahl von  $\alpha_1$  und  $\alpha_2$ . Zunächst wäre die Wahl von  $\alpha_1 = \alpha_2 = \alpha/2$  eine sinnvolle Wahl. Allerdings lassen sich auf Grund der Beschränkung auf ganze Zahlen  $U$  und  $O$  nicht so bestimmen, dass die beiden obigen Gleichungen exakt erfüllt sind. Also wird in den meisten Fällen  $\alpha_1 + \alpha_2 < \alpha$  gewählt, was letztendlich zu einem Konfidenzintervall führt, dass ein höheres Konfidenzniveau hat. Wir illustrieren dies an einem Beispiel.

*Beispiel 2.13:* In einem Betrieb mit  $N = 300$  Mitarbeitern möchte die Firmenleitung von ihren Angestellten wissen, welche Maßnahmen diese zur Verbesserung des Betriebsklimas für geeignet halten. Dabei wurden unter anderem die folgenden zwei Fragen gestellt, welche mit ja oder nein beantwortet werden konnten: „Wünschen Sie sich flexiblere Arbeitszeiten?“ und „Wünschen Sie sich einen Betriebskindergarten?“. Es wurden  $n = 100$  Personen befragt. Die Firmenleitung interessiert sich für die Anteilswerte  $P_1$  und  $P_2$  der zwei Fragen. Frage 1 wurde von  $m_1 = 45$  Personen und Frage 2 von  $m_2 = 2$  Personen mit „Ja“ beantwortet. Die Anteilsschätzungen sind nun

$$p_1 = \frac{45}{100} = 0,45$$

$$p_2 = \frac{2}{100} = 0,02.$$

Die entsprechenden Konfidenzintervalle nach der **approximativen Methode** werden wie folgt berechnet

$$\widehat{\text{Var}}(p_1) = \frac{0,45(1-0,45)}{99} \frac{300-100}{300} = 0,00167$$

$$\widehat{\text{Var}}(p_2) = \frac{0,02(1-0,02)}{99} \frac{300-100}{300} = 0,00013$$

und es ergibt sich das 95%-Konfidenzintervall für  $p_1$  zu

$$\left[ 0,45 - 1,96 \sqrt{0,00167}; 0,45 + 1,96 \sqrt{0,00167} \right] = [0,370; 0,530],$$

wohingegen das 95%-Konfidenzintervall für  $p_2$  die Werte

$$\left[ 0,02 - 1,96 \sqrt{0,000132}; 0,02 + 1,96 \sqrt{0,000132} \right] = [-0,003; 0,043]$$

annimmt. Das zweite Konfidenzintervall hat eine negative untere Grenze, was klarerweise wenig informativ ist. Eine Korrektur dieser Grenze auf 0 ist insofern

auch nicht sehr hilfreich, da in der Stichprobe bereits 2 Personen sind, die die Frage 2 mit „Ja“ beantworten. Daher ist der gesuchte Anteil  $p_2$  in der Population mindestens  $2/300$ , also größer als 0.

Die Berechnung des **exakten Konfidenzintervalls** liefert (siehe dazu numerische Umsetzung in 2.12) das exakte 95%-Konfidenzintervall für  $p_1$  zu

$$[0, 366; 0, 537],$$

wohingegen das exakte 95%-Konfidenzintervall für  $p_2$  nun lautet

$$[0, 006; 0, 064].$$

Die Grenzen der beiden exakten Konfidenzintervalle entsprechen den Anzahlen [110; 161] bzw. [2; 19] in der Grundgesamtheit. Für  $p_1$  sind exaktes und approximatives Konfidenzintervall praktisch identisch. Für  $p_2$  hingegen ergeben sich Unterschiede und das exakte Konfidenzintervall ist klar zu bevorzugen. ◀

Wie das Beispiel zeigt, ist es sinnvoll, insbesondere für kleine Anteile exakte Konfidenzintervalle zu berechnen. Ist der Umfang der Grundgesamtheit im Vergleich zur Stichprobe groß, so kann man auch Konfidenzintervalle für das Ziehen mit Zurücklegen verwenden, welche im folgenden Abschnitt behandelt werden. Manchmal kann es bei kleinen Wahrscheinlichkeiten auch von Interesse sein, nur eine obere Grenze anzugeben. Das führt zu sogenannten einseitigen Konfidenzintervallen mit Untergrenze 0. Dies kann bei den exakten Konfidenzintervallen durch Wahl von  $\alpha_1 \leq 0, 05$  und  $\alpha_2 = 0$  realisiert werden.

*Beispiel 2.14:* Bei einem Tierbestand von  $N = 10\,000$  Tieren soll eine obere Grenze für den Anteil der mit einer seltenen Krankheit infizierten Tiere angegeben werden. Dazu wurde eine einfache Zufallsstichprobe vom Umfang  $n = 500$  gezogen. Dabei war ein Tier infiziert, was einem Anteil von 0, 2% entspricht. Das einseitige Konfidenzintervall ergibt sich zu  $[0; 0, 0093]$ . Die obere (95%-Konfidenz-) Grenze liegt also bei 0, 93%.

Bei der Benutzung der approximativen Normalverteilung (siehe Kasten S. 28) erhält man einseitige Konfidenzintervalle (obere bzw. untere Grenze) durch Anwendung des  $(1 - \alpha)$ -Quantils statt des  $(1 - \frac{\alpha}{2})$ -Quantils der Normalverteilung. Es lautet dann

$$\left[ 0; \hat{P}_{ES} + z_{1-\alpha} \cdot \sqrt{\widehat{Var}(\hat{P}_{ES})} \right]$$

bzw.

$$\left[ \hat{P}_{ES} - z_{1-\alpha} \cdot \sqrt{\widehat{Var}(\hat{P}_{ES})}; 1 \right].$$

◀

## 2.7 Ziehen mit Zurücklegen

Wir wollen nun die Stichprobenziehung leicht verändern und erlauben, dass ein Individuum auch mehrmals in die Stichprobe gezogen werden kann. Dies wird im statistischen Jargon „Ziehen mit Zurücklegen“ genannt. Dieses Vorgehen führt dazu, dass die einzelnen Ziehungen voneinander unabhängig sind. Jedes Individuum hat unabhängig von den vorherigen Zügen die Wahrscheinlichkeit  $1/N$ , im  $k$ -ten Zug gezogen zu werden. Wir befinden uns also im klassischen Fall der Statistik, der sogenannten **unabhängig und identisch verteilten Stichprobe** (engl.: i.i.d. für independent and identically distributed). Wir wollen die Konsequenzen dieses Ansatzes anhand des Beispiels in Abschn. 2.6.2 betrachten.

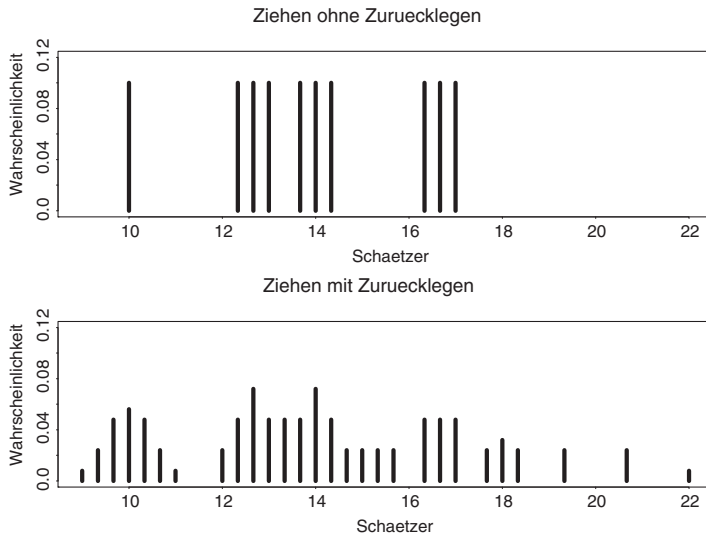
*Beispiel 2.15:* Aus der Population mit

$$Y_1 = 9, \quad Y_2 = 10, \quad Y_3 = 11, \quad Y_4 = 18, \quad Y_5 = 22$$

ziehen wir nun **mit Zurücklegen**, d.h. ist ein Individuum gezogen, so wird der Wert von  $Y$  notiert und das Individuum wird in die Population „zurückgelegt“. Im nächsten Zug kann es somit noch einmal gezogen werden. Im Falle von Ziehen mit Zurücklegen ergeben sich damit die folgenden Stichproben, wobei die Reihenfolge der Ziehung berücksichtigt wird:

Gezogene Individuen			Mittelwert	Wahrscheinlichkeit
1	1	1	9,00	1/125
1	1	2	9,33	1/125
1	1	3	9,67	1/125
1	1	4	12,00	1/125
1	1	5	13,33	1/125
	...		...	...
5	5	1	17,67	1/125
5	5	2	18,00	1/125
5	5	3	18,33	1/125
5	5	4	20,67	1/125
5	5	5	22,00	1/125

Wir bezeichnen den entsprechenden Mittelwertschätzer als  $\hat{Y}_{MZ}$ , wobei der Index als Abkürzung für **Mit Zurücklegen** steht. Die Wahrscheinlichkeitsverteilung des Schätzers ist in Abb. 2.6 dargestellt. Zum Vergleich ist die Wahrscheinlichkeitsverteilung von  $\hat{Y}_{ES}$  dargestellt, wie sie in Abschn. 2.6 hergeleitet wurde, also die Verteilung bei einem Ziehungsprozess ohne Zurücklegen. Es ist deutlich zu erkennen, dass die Varianz beim Ziehen mit Zurücklegen größer ist als beim Ziehen ohne Zurücklegen, die Wahrscheinlichkeitsverteilung weist also eine höhere Streuung auf. ◀



**Abb. 2.6** Verteilung von  $\hat{Y}$  beim Ziehen ohne und mit Zurücklegen

Berechnet man nun Erwartungswert und Varianz des Schätzers, so ergibt sich als Erwartungswert  $E(\hat{Y}_{MZ}) = 14$  und als Varianz  $\text{Var}(\hat{Y}_{MZ}) = 8,67$ . Allgemein ist  $\hat{Y}_{MZ}$  erwartungstreu und hat die Varianz

$$\text{Var}(\hat{Y}_{MZ}) = \frac{S^2}{n}.$$

Diese kann geschätzt werden durch

Mittelwertschätzung bei einer einfachen Zufallsstichprobe  
mit Zurücklegen

Gegeben sei eine Stichprobe  $y_1, \dots, y_n$  vom Umfang  $n$  mit Zurücklegen.  
Ein unverzerrter Schätzer für den Mittelwert der Population ist

$$\hat{Y}_{MZ} = \bar{y} = \frac{1}{n} \sum_{k=1}^n y_k.$$

Die Varianz von  $\hat{Y}_{MZ}$  kann geschätzt werden durch

$$\widehat{\text{Var}}(\hat{Y}_{MZ}) = \frac{1}{n(n-1)} \sum_{k=1}^n (y_k - \hat{Y}_{MZ})^2.$$

$$\widehat{\text{Var}}(\widehat{Y}_{MZ}) = \frac{1}{n(n-1)} \sum_{k=1}^n (y_k - \widehat{Y}_{MZ})^2.$$

*Herleitung:* Da alle  $y_k$ ,  $k = 1, \dots, n$ , die gleiche Verteilung haben und unabhängig sind, ergeben sich die obigen Formeln aus den elementaren Rechenregeln für Mittelwert und Varianz und aus der Verteilung von  $y_1$  bei der Stichprobe ohne Zurücklegen. ■

Ein Vergleich der Varianzen mit und ohne Zurücklegen zeigt, dass die Varianz beim „Ziehen mit Zurücklegen“ größer ist. Allgemein ist sie um den Faktor  $(N-1)/(N-n)$  größer, der sich aus den zugrunde liegenden Verteilungsmodellen ergibt. Im Fall der Schätzung eines Anteils erhält man eine Binomialverteilung.

### Ziehen mit Zurücklegen für binäre Merkmale (Binomialverteilung)

---

Aus einer Grundgesamtheit von  $N$  Elementen werden  
 $n$  Elemente mit Zurücklegen gezogen.

Wir betrachten ein binäres Merkmal mit Werten 0 oder 1.

In der Grundgesamtheit sind  $M$  Einsen vorhanden.

Der Anteil der Einsen in der Grundgesamtheit beträgt folglich  $P = M/N$ .

Die Wahrscheinlichkeit, dass  $m$  von den  $n$  gezogenen Elementen  
den Wert 1 haben ist

$$P(y = m | P, n) = \binom{n}{m} P^m (1 - P)^{n-m}.$$

Die Anzahl der Einsen  $y = \sum_{i=1}^n y_i$  in der Stichprobe  
ist binomialverteilt mit den Parametern

$$\begin{aligned} E(y) &= nP, \\ \text{Var}(y) &= nP(1 - P). \end{aligned}$$

Ein unverzerrter Schätzer für den Anteil  $P$  ist

$$\widehat{P}_{MZ} = \frac{1}{n} \sum_{k=1}^n y_k.$$

Als Varianzschätzung ergibt sich

$$\widehat{\text{Var}}(\widehat{P}_{MZ}) = \frac{1}{n} \widehat{P}_{MZ}(1 - \widehat{P}_{MZ}).$$

Die Konfidenzintervalle ergeben sich analog zu dem Fall ohne Zurücklegen als

$$\left[ \hat{P}_{MZ} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\text{Var}}(\hat{P}_{MZ})} ; \hat{P}_{MZ} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\text{Var}}(\hat{P}_{MZ})} \right].$$

Als Beispiel wollen wir ein Ergebnis der Sonntagsfrage aus dem Jahr 2009 analysieren.

*Beispiel 2.16:* Zur Sonntagsfrage vom 3.7.2009 wurden  $n = 1\,206$  wahlberechtigte Personen gefragt, welche Partei sie wählen würden, wenn am kommenden Sonntag Bundestagswahl wäre. Dabei gaben  $m = 302$  Personen an, dass sie die SPD wählen würden. Damit ergibt sich ein exaktes Konfidenzintervall für den Anteil von  $[0,2262; 0,2759]$ , d.h. das für die SPD ein Anteil zwischen 22 und 28% zu erwarten ist. ◀

Es werden in der Literatur auch andere Verfahren zur Bestimmung von Konfidenzintervallen diskutiert, siehe dazu z.B. Held (2008). Für kleine Anteile und/oder kleine Stichproben können auch die exakten Konfidenzintervalle nach Clopper-Pearson genutzt werden, siehe dazu z.B. Fleiss, Levin und Paile (2003). Eine gute Alternative ist das Konfidenzintervall nach Wilson, das auf dem Score-Test basiert und auch für kleine Anteile gut geeignet ist, siehe dazu auch Agresti und Coull (1998). In der Praxis wird das Ziehen mit Zurücklegen kaum angewendet. Allerdings sind bei großen Grundgesamtheiten Ziehen mit und ohne Zurücklegen praktisch identisch. Da insbesondere bei modellbasierten Schätzmethoden, wie wir sie später im Buch behandeln, die Berechnung der Varianz im Falle des Ziehens mit Zurücklegen wesentlich einfacher ist, wird diese bei entsprechend großen Populationen auch der Einfachheit halber angewendet. Man beachte, dass bei der Parameterschätzung der Umfang  $N$  der Grundgesamtheit nicht eingeht.

## 2.8 Bestimmung des Stichprobenumfangs

Betrachtet man die Varianz des Schätzers  $\hat{Y}_{ES}$ , so stellt man fest, dass diese mit wachsendem Stichprobenumfang abnimmt. Dies bedeutet insbesondere, dass Konfidenzintervalle mit wachsendem Stichprobenumfang kleiner werden. Inhaltlich lässt sich dies folgendermaßen interpretieren: Das Ergebnis der Stichprobe wird mit wachsendem Stichprobenumfang genauer. Wir können nun anders herum fragen, wie groß eine Stichprobe mindestens sein muss, um eine gewisse Genauigkeit zu erfüllen. Wir wollen also nun eine gewünschte Genauigkeit vorgeben und damit den erforderlichen Stichprobenumfang berechnen. Es soll somit der Stichprobenumfang  $n$  so gewählt werden, dass der Schätzwert mit einer vorgegebenen Wahrscheinlichkeit  $1 - \alpha$  einen Abstand kleiner als  $e$  vom wahren Wert hat. Dabei ist  $e$  eine vorgegebene Genauigkeit und  $1 - \alpha$  das Sicherheitsniveau bzw.  $\alpha$  die Fehlerwahrscheinlichkeit. Als Formel geschrieben heißt das

$$P\left(\left|\widehat{Y}_{ES} - \bar{Y}\right| < e\right) \geq 1 - \alpha,$$

anders ausgedrückt, die Wahrscheinlichkeit, dass der Schätzer  $\widehat{Y}_{ES}$  um mehr als  $e$  vom unbekannten Populationsmittel  $\bar{Y}$  abweicht, soll höchstens  $\alpha$  betragen. Die obige Formel lässt sich umformen zu

$$P\left(\frac{\left|\widehat{Y}_{ES} - \bar{Y}\right|}{\sqrt{\text{Var}\left(\widehat{Y}_{ES}\right)}} < \frac{e}{\sqrt{\text{Var}\left(\widehat{Y}_{ES}\right)}}\right) \geq 1 - \alpha.$$

Daraus ergibt sich mit Hilfe der Normalverteilungsannahme für  $\widehat{Y}_{ES}$

$$\frac{e}{\sqrt{\text{Var}\left(\widehat{Y}_{ES}\right)}} \geq z_{1-\frac{\alpha}{2}},$$

wobei  $z_{1-\frac{\alpha}{2}}$  das  $(1 - \frac{\alpha}{2})$ -Quantil der Standardnormalverteilung ist. Für die vorgegebene Fehlerwahrscheinlichkeit  $\alpha = 0,05$  erhalten wir  $z_{1-0,05/2} = z_{0,975} = 1,96$ .

Einsetzen der Formel für  $\text{Var}\left(\widehat{Y}_{ES}\right)$  und quadrieren liefert unter Verwendung von

$$\frac{N-n}{N-1} \approx \frac{N-n}{N}$$

$$\frac{e^2}{\frac{S^2}{n} \frac{N-n}{N}} \geq z_{1-\frac{\alpha}{2}}^2 \cdot \alpha. \quad (2.5)$$

Lösen wir (2.5) nach  $n$  auf, so erhalten wir

$$n \geq \frac{S^2}{e^2/z_{1-\frac{\alpha}{2}}^2 + S^2/N}. \quad (2.6)$$

Ist die Population im Vergleich zur Stichprobe groß, so können wir den Korrekturfaktor  $\frac{N-n}{N}$  für endliche Populationen ignorieren und erhalten die Näherungslösung

$$n \geq z_{1-\frac{\alpha}{2}}^2 \frac{S^2}{e^2}. \quad (2.7)$$

Die übliche Wahl  $\alpha = 0,05$  ergibt mit  $z_{1-\alpha/2} = 1,96 \approx 2$  die Faustregel  $n \geq 4 \cdot \left(\frac{S}{e}\right)^2$ . Man benötigt also bei großen Populationen nur das Verhältnis  $S/e$ , um den Stichprobenumfang näherungsweise zu bestimmen.

Ein anderer Ansatz zur Bestimmung des notwendigen Stichprobenumfangs ist es, die erwartete Länge des Konfidenzintervalls vorzugeben. Aus der Form

$$\left[ \hat{Y}_{ES} - z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{Y}_{ES})}, \hat{Y}_{ES} + z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{Y}_{ES})} \right]$$

des Konfidenzintervalls ergibt sich dessen Länge durch

$$2 z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{Y}_{ES})} = 2 z_{1-\frac{\alpha}{2}} \sqrt{\frac{s^2 N - n}{n} \frac{N - n}{N}}.$$

Da die Größe  $s^2$  aus der Stichprobe berechnet wird, ist die Länge des Konfidenzintervalls zufällig und nicht im Vorfeld bestimmbar. Daher kann man sinnvollerweise auch nur die erwartete Länge  $l$  des Konfidenzintervalls vorgeben. Unter Berücksichtigung von  $E\left(s^2 \frac{N-n}{N}\right) = S^2 \frac{N-n}{N-1}$  erhält man

$$l \approx 2 z_{1-\frac{\alpha}{2}} \sqrt{\frac{S^2 N - n}{n} \frac{N - n}{N - 1}}.$$

Damit berechnen wir bei vorgegebener maximaler erwarteter Länge  $l$  den notwendigen Stichprobenumfang mit der Formel

$$n \geq \frac{S^2}{\left(\frac{l}{2}\right)^2 / z_{1-\frac{\alpha}{2}}^2 + S^2 / N}, \quad (2.8)$$

wobei wir wiederum die Approximation  $\frac{N-n}{N-1} \approx \frac{N-n}{N}$  verwenden. Ein Vergleich mit der obigen Strategie zeigt, dass wir für  $l = 2e$  identische Stichprobenumfänge erhalten.

Zur konkreten Bestimmung des Stichprobenumfangs mit Formel (2.6) bzw. (2.8) ist die Kenntnis der Standardabweichung  $S$  des Merkmals in der Population nötig. Vor der Stichprobenziehung ist  $S$  jedoch üblicherweise unbekannt. Auch die Schätzung von  $S$  ist hier unmöglich, denn es soll mit (2.6) bzw. (2.8) ja gerade der Stichprobenumfang geplant werden, d.h. die Stichprobe ist noch nicht gezogen und somit können keine Größen geschätzt werden. Als Ausweg aus diesem Dilemma bieten sich zwei Möglichkeiten an:

1. Man ersetzt  $S$  in (2.6) bzw. (2.8) durch einen geschätzten Wert aus vorherigen Erhebungen. In einigen Anwendungen ist das Merkmal in vorherigen Stichproben schon einmal erhoben worden. In diesem Fall kann man auf solches Wissen zurückgreifen, um eine Größenvorstellung von  $S$  zu bekommen.
2. Man zieht eine sogenannte Pilotstichprobe von kleinerem Umfang  $n'$ , wobei  $n' \leq n$  angenommen ist. Aus dieser berechnet man einen Schätzer für  $S$  und plant damit den benötigten Stichprobenumfang.

Beide Methoden sind nicht notwendigerweise immer praktikabel. Oftmals lässt sich jedoch für  $S$  eine obere Schranke festlegen, dass heißt ein „worst case“

Szenario. Dieses kann wiederum genutzt werden, um eine Obergrenze des Stichprobenumfangs festzulegen. Aus der Formel (2.7) ist ersichtlich, dass der Stichprobenumfang umgekehrt proportional zum Quadrat der geforderten Genauigkeit  $e$  ist. Das bedeutet, dass man eine Halbierung der Länge des Konfidenzintervalls mit einer Vervierfachung des Stichprobenumfangs „bezahlen“ muss.

#### Bestimmung des Stichprobenumfangs bei einer einfachen Zufallsstichprobe

Gegeben sei die Genauigkeit  $e$  (erwartete halbe Länge  $l$  des Konfidenzintervalls), das Konfidenzniveau  $1 - \alpha$  und die Standardabweichung  $S$  in der Grundgesamtheit.  
Für den nötigen Stichprobenumfang gilt dann:

$$n \geq \frac{S^2}{e^2/z_{1-\frac{\alpha}{2}}^2 + S^2/N}.$$

Beim Ziehen mit Zurücklegen und bei großen Grundgesamtheiten benutzt man:

$$n \geq z_{1-\frac{\alpha}{2}}^2 \frac{S^2}{e^2}.$$

Zur Bestimmung von  $S$  sind in der Regel eine Pilotstichprobe oder andere externe Zusatzinformationen erforderlich.

Ist die interessierende Größe ein Anteilswert, so ergibt sich der Stichprobenumfang durch:

$$n \geq \frac{P(1-P)}{e^2/z_{1-\frac{\alpha}{2}}^2 + P(1-P)/N},$$

bzw.

$$n \geq z_{1-\frac{\alpha}{2}}^2 \frac{P(1-P)}{e^2}.$$

Für  $P$  verwendet man bei fehlendem Vorwissen den worst case 0,5.  
Ist hingegen bekannt, dass  $P$  einen Wert  $P_{\text{priori}} < 0,5$  unterschreitet  
bzw. einen Wert  $P_{\text{priori}} > 0,5$  überschreitet,  
so kann man diesen Wert als a priori bekannte Grenze in die Formel einsetzen.

Ist die interessierende Größe ein Anteilswert  $P$ , so ergibt sich der Stichprobenumfang ebenfalls aus Formel (2.6). In diesem Falle ersetzt man jedoch  $S^2$  durch  $P(1 - P)$ , so dass sich ergibt

$$n \geq \frac{P(1 - P)}{e^2/z_{1-\frac{\alpha}{2}}^2 + P(1 - P)/N},$$

(2.9)

und für  $N$  hinreichend groß  $n \geq z_{1-\frac{\alpha}{2}}^2 \cdot P(1 - P)/e^2$ . Wir wollen Formel (2.9) genauer betrachten. Dazu nehmen wir exemplarisch an, dass wir eine Genauigkeit von  $e = 0,1$  fordern. In Tabelle 2.2 sind die Stichprobenumfänge, wie sie sich durch Formel (2.9) ergeben, für verschiedene Werte von  $P$  und  $N$  bestimmt. Weiterhin geben wir in Tabelle 2.3 die benötigten Stichprobenumfänge für große Grundgesamtheiten ( $N \geq 10\,000$ ) für verschiedene Genauigkeiten  $e$  an. Hier unterscheiden wir die Fälle  $P = 0,5$  und  $P = 0,1$ . Bemerkenswert ist, dass für eine Genauigkeit von einem Prozentpunkt ( $e = 0,01$ ) bei Anteilswerten mit  $P = 0,5$  der hohe Stichprobenumfang von fast 10 000 Personen erforderlich ist. Bei der Befragung von 1 000 Personen ergibt sich eine Genauigkeit von etwa drei Prozentpunkten. Das bedeutet z.B. für Befragungen zur Wahlabsicht, dass Stichprobenumfänge von unter 2 000 ohne weitere methodische Verbesserungen nur ungenaue Ergebnisse liefern.

In Abb. 2.7 ist der geforderte Stichprobenumfang gemäß Formel (2.9) für verschiedene Populationsgrößen aufgetragen. Es zeigt sich, dass der geforderte Stichprobenumfang für große Populationen nicht vom Populationsumfang abhängt. Das bedeutet inhaltlich, fordert man in einer Stichprobe eine gewisse Genauigkeit, so wird diese mit einer Stichprobe mit Mindeststichprobenumfang wie in Formel (2.9) bestimmt erreicht, egal wie groß die Population ist. In der Praxis bedeutet dies zum Beispiel, dass es bei Überlegungen zur Stichprobengröße und der damit verbundenen Schätzgenauigkeit unerheblich ist, ob die zugehörige Population die Münchner Bevölkerung oder die Bevölkerung Deutschlands ist. Die häufig verwendete Strategie

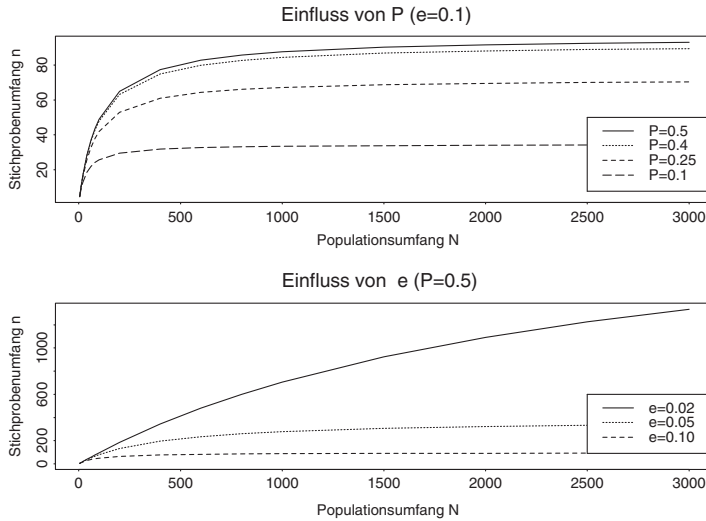
**Tabelle 2.2** Notwendiger Stichprobenumfang bei einer geforderten Genauigkeit von  $e = 0,1$  für den Anteilswert

$P$	$N = 10$	$N = 100$	$N = 1\,000$	$N = 10\,000$
0,2	9	39	58	62
0,3	9	45	75	81
0,4	10	48	85	92
0,5	10	49	88	96

**Tabelle 2.3** Notwendiger Stichprobenumfang für große Grundgesamtheiten ( $N \geq 10\,000$ ), ausgehend von einem Sicherheitsniveau von  $1 - \alpha = 95\%$

$e$	$P = 0,5$	$P = 0,1$
0,1	97	*
0,05	385	139
0,03	1068	385
0,02	2401	865
0,01	9604	3458

\*  $P = 0,1$  bedeutet, dass a priori bekannt ist, dass der wahre Anteil kleiner als 0,1 ist. Daher ist  $e = 0,1$  offensichtlich keine sinnvolle Vorgabe.



**Abb. 2.7** Stichprobenumfang in Abhängigkeit von der Populationsgröße für verschiedene Werte von  $P$  und  $e$

gie, den Auswahlssatz  $n/N$  ohne Berücksichtigung der Populationsgröße festzulegen (z.B. „3%-Stichprobe“) ist damit in vielen Fällen unsinnig.

Wie bei einer Stichprobe zu einem metrischen Merkmal besteht auch für den Anteilswert  $P$  bei Stichprobenplanung das Problem, dass der explizite Wert von  $P$  unbekannt ist. Wie aus Tabelle 2.2 ersichtlich, ist der benötigte Stichprobenumfang für  $P = 0,5$  am größten, was daran liegt, dass die Funktion  $P \cdot (1 - P)$  ihr Maximum bei  $P = 0,5$  hat. Daher ist die Wahl von  $P = 0,5$  das „Worst Case“ Szenario, womit man immer auf der sicheren Seite ist. Falls aber  $P = 0,5$  unrealistisch ist, ist es sinnvoll, zur Stichprobenplanung eine obere Schranke von  $P$  zur Grundlage der Planung zu machen. Ansonsten wird der Stichprobenumfang unnötig groß, was aus Kostengründen zu vermeiden ist.

Wenn man bei einer ausführlichen Befragung mit vielen Merkmalen vor der Frage steht, wie groß der Stichprobenumfang gewählt werden sollte, sind die Überlegungen für Anteilsschätzungen oft hilfreich. Man kann sich überlegen, mit welcher Genauigkeit man Anteile von Antwortkategorien auf die Fragen schätzen will. Bei großer Grundgesamtheit ergibt sich aus Tabelle 2.3 für  $n = 97$  eine Genauigkeit von  $e = 0,1$  (10 Prozentpunkte) im „Worst Case“ von  $P = 0,5$ . Diese ist dann für alle Fragen gültig. Bei einer Anforderung von  $e = 0,05$  (5 Prozentpunkte) ist ein Stichprobenumfang von  $n = 385$  nötig.

## 2.9 Systematische Stichprobe

Bei der einfachen Zufallsstichprobe wird für jeden Zug eine neue Zufallsvariable gezogen. Dies kann aufwendig sein und ebenso zu zufälligen, vielleicht aber unerwünschten Gruppierungen führen. Es zeigt sich, dass eine systematische Stichprobe

insbesondere dann von Vorteil ist, wenn die Population mit einer Ordnung oder Abhängigkeitsstruktur versehen ist. Wir ziehen in diesem Falle nicht mehr  $n$  Elemente zufällig aus der Population, sondern wenden eine Systematik an, indem wir jedes  $p$ -te Individuum der Population in die Stichprobe aufnehmen. Zur Ziehung einer systematischen Stichprobe muss damit eine Zufallszahl aus den Zahlen 1 bis  $p$  mit  $p = N/n$  gezogen werden, wobei der Einfachheit halber angenommen wird, dass  $N/n$  ganzzahlig ist. Diese Zufallszahl gibt das erste zu ziehende Element an und entsprechend ist die zu ziehende Stichprobe vollständig bestimmt. Wir wollen dazu ein einfaches Beispiel betrachten.

*Beispiel 2.17:* Nehmen wir an, eine Population bestehe aus den Werten

$$Y_1 = 10, \quad Y_2 = 30, \quad Y_3 = 80, \quad Y_4 = 20, \quad Y_5 = 70, \quad Y_6 = 90$$

und wir ziehen eine systematische Stichprobe vom Umfang  $n = 2$ . Wir wählen dabei  $p = 3$  und ziehen eine Zufallszahl  $j$  aus den Werten 1 bis  $p$  mit einer Gleichverteilung. Die Stichprobe ergibt sich dann gemäß  $Y_j, Y_{j+p}$  (im Allgemeinen gemäß  $Y_j, Y_{j+p}, \dots, Y_{j+(n-1)p}$ ). Wir erhalten somit die folgenden möglichen Stichproben

Gezogene Individuen		$\bar{y}$	Wahrscheinlichkeit
1	4	15	1/3
2	5	50	1/3
3	6	85	1/3



Eine wichtige Eigenschaft bei systematischen Stichproben ist, dass die Anzahl der möglichen Stichproben klein ist, nämlich genau  $p$ . Als Schätzer für den Mittelwert ergibt sich bei der systematischen Stichprobe in allgemeiner Form

$$\widehat{Y}_{\text{syst}} = \frac{1}{n} \sum_{k=1}^n Y_{j+(k-1)p},$$

wobei wir, wie schon gesagt, der Einfachheit halber annehmen, dass  $N/n$  eine ganze Zahl ist, so dass  $n = N/p$  gilt. Zur Bestimmung der Varianz von  $\widehat{Y}_{\text{syst}}$  gehen wir weiter davon aus, dass die Ordnung der Elemente zufällig ist. In diesem Fall können wir die Varianz schätzen durch

$$\text{Var}(\widehat{Y}_{\text{syst}}) = \frac{s^2}{n} \frac{N-n}{N},$$

was der Varianz einer einfachen Zufallsstichprobe entspricht. Die Voraussetzung, dass die Elemente der Population zufällig geordnet sind, ist dabei essentiell. Ist diese Voraussetzung verletzt, hängt es entscheidend von der tatsächlichen Ordnung ab, ob die oben gegebene Varianz eine Über- oder Unterschätzung liefert. Da wir aber im

Allgemeinen keine Information über die Ordnung haben, ist die obige Varianzrechnung durchaus gerechtfertigt. Im Prinzip kann eine systematische Stichprobe auch als eine Cluster-Stichprobe aufgefasst werden, wie sie in Abschn. 5.2 vorgestellt wird. Wir werden diesen Punkt dort noch einmal aufgreifen.

### Systematische Stichprobe

Die Population bestehe aus  $N$  Elementen, von denen  $n$  zufällig ausgewählt werden. Es sei  $p = N/n$  ganzzahlig und  $j$  eine Zufallszahl aus  $\{1, 2, \dots, p\}$ .

Ein unverzerrter Schätzer für den Mittelwert  $\bar{Y}$  ist

$$\hat{Y}_{\text{syst}} = \frac{1}{n} \sum_{k=1}^n Y_{j+(k-1)p}.$$

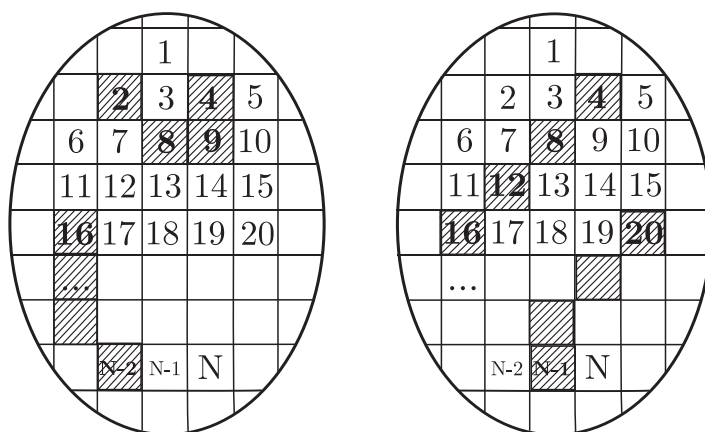
Die Varianz kann geschätzt werden durch

$$\widehat{\text{Var}}(\hat{Y}_{\text{syst}}) = \frac{N-n}{N} \frac{1}{n(n-1)} \sum_{k=1}^n \left( Y_{j+(k-1)p} - \hat{Y}_{\text{syst}} \right)^2.$$

*Beispiel 2.18:* Ein Wirtschaftsprüfer möchte die Rechnungen eines Unternehmens prüfen. Dabei entscheidet er sich für eine systematische Stichprobe mit folgendem Vorgehen: Von den in Frage kommenden Rechnungen wird jede  $p$ -te in die Stichprobe gezogen und einer genauen Prüfung unterzogen. Die erste zu ziehende Rechnung wird jedoch zufällig ausgewählt, indem eine Zufallsvariable zwischen 1 und  $p$  gezogen wird, welche die erste Rechnung, die in die Stichprobe aufgenommen wird, bestimmt. ◀

*Beispiel 2.19:* Ein großes Gewässer soll bezüglich seines Schadstoffgehalts untersucht werden. Dazu wird das Gewässer in Planquadrate eingeteilt, die von 1 bis  $N$  durchnummeriert werden. Der Einfachheit halber betrachten wir nur die Planquadrate, die vollständig über dem See verteilt liegen. Aus diesen Planquadraten werden  $n$  zufällig ausgewählt und anschließend per Wasserprobe untersucht. Schematisch lässt sich dies wie in der linken Graphik in Abb. 2.8 darstellen. Die zufällig gewählten Planquadrate sind dabei schraffiert dargestellt. Die Wahl der Planquadrate ist zufällig, und ebenso zufällig kann es zu einer Gruppierung von benachbarten Planquadraten kommen. Die gezogene Stichprobe erscheint unvoreilhaft, insbesondere da davon auszugehen ist, dass die Wasserqualität in einem Planquadrat ähnlich ist zu denen der Nachbarplanquadrate, das heißt zwischen den Messungen in benachbarten Planquadraten kann eine Abhängigkeit bestehen. Um diesen Punkt zu berücksichtigen, wollen wir eine systematische Stichprobe ziehen. Hierbei wählt man die Planquadrate in einer systematischen Form aus.

Die Zufälligkeit besteht nun darin, das erste Planquadrat auszuwählen, die verbleibenden Planquadrate sind damit durch das systematische Muster bestimmt. Schematisch ist dies in der rechten Skizze von Abb. 2.8 gezeigt, wobei wir hier jedes 4-te Planquadrat in die Stichprobe aufnehmen. Je nachdem welches Planquadrat der ersten vier Planquadrate gezogen wird (1, 2, 3 oder 4) ist der Rest der Stichprobe systematisch bestimmt. Die Zufälligkeit besteht also in der Auswahl des ersten Quadrats. ◀



**Abb. 2.8** Einfache Zufallsstichprobe (*links*) und systematische (*zufällige*) Stichprobe (*rechts*), gezogene Planquadrate sind jeweils schraffiert.

## 2.10 Beispiel

Wir besprechen nun ein Beispiel aus der Praxis, um die Methoden des Kapitels zu veranschaulichen. Das Beispiel bezieht sich auf eine große Population und wir werten es nach dem Prinzip einer Ziehung mit Zurücklegen aus.

**Beispiel 2.20:** Im Rahmen der Diskussion um die Finanzierung der baye-rischen Hochschulen kam es 2004 nach den Kürzungsbeschlüssen der Baye-rischen Staatsregierung zu Demonstrationen und anderen Formen des Protests von Studierenden in Bayern. Um die Wirkung der Proteste und die Meinung der Münchener Bevölkerung zu einigen damit verbundenen Themen in Erfahrung zu bringen, wurde von der Fachschaft Statistik der Ludwig-Maximilians-Universität München eine telefonische Befragung durchgeführt. Die Auswahl wurde mit zu-fälliger Wahl von Telefonnummern realisiert. Insgesamt wurden 251 Personen befragt.

Damit lassen sich zu den einzelnen Fragen die entsprechenden Anteile in der Münchener Bevölkerung hochrechnen. In der folgenden Tabelle sind die Er-gebnisse für die einzelnen Fragen zusammengestellt. Dabei wurden jeweils die Konfidenzintervalle mit der Normalverteilungsapproximation verwendet. Da die zugrunde liegende Population groß genug ist, kann auf die Korrektur für endliche

Populationen verzichtet werden. Konkret handelt es sich hier ungefähr um den Faktor  $\frac{10^6 - 251}{10^6} = 0,999749$ , wenn man berücksichtigt, dass die Bevölkerung von München etwa eine Million Personen umfasst.

In diesem Beispiel werden die ersten drei Fragen der telefonischen Befragung betrachtet. Diese lauten

**Frage 1:** Haben Sie schon von den Studentenprotesten in Bayern gehört?

**Frage 2:** Halten Sie die Proteste der Studenten für gerechtfertigt?

**Frage 3:** Von welchen Protestaktionen haben Sie gehört? (Mehrfachnennungen möglich)

Dabei wurde Frage 2 lediglich den Personen gestellt, welche Frage 1 mit „ja“ beantwortet haben, sprich bereits von den Studentenprotesten gehört haben. Die Antwortmöglichkeiten für Frage 3 sind in die Kategorien Großdemonstrationen, Lichterkette, Trauermarsch, 24 h Vorlesung und öffentliche Vorlesungen unterteilt. Bei dieser Frage wurde wieder die gesamte Stichprobe berücksichtigt, wobei die Individuen, die Frage 1 mit „nein“ beantwortet haben, auch hier überall mit „nein“ bewertet wurden. Die Ergebnisse sind in Tabelle 2.4 dargestellt.

**Tabelle 2.4** Auswertung der Telefonumfrage zu den Studentenprotesten in Bayern

Frage	„Ja“ Anteil in Stichprobe	95%-Konfidenzintervall
Frage 1	0,849	[0,7982; 0,8906]
Frage 2	0,756	[0,6925; 0,8120]
Frage 3		
Großdemonstrationen	0,7251	[0,6654; 0,7794]
Lichterkette	0,2590	[0,2059; 0,3178]
Trauermarsch	0,2550	[0,2023; 0,3136]
24 h Vorlesung	0,2669	[0,2133; 0,3262]
Öffentliche Vorlesungen	0,2829	[0,2280; 0,3429]



Die einfache Zufallsstichprobe ist das sicherlich am häufigsten verwendete Stichprobendesign. Bei Anwendung und Dokumentation ist immer darauf zu achten, dass Konfidenzintervalle anzugeben sind, wenn ein Anspruch auf Übertragung des Ergebnisses auf die Grundgesamtheit besteht. Weiter ist die Umsetzung der zufälligen Ziehung in vielen Fällen nur ansatzweise oder durch Ersatzverfahren möglich. Daher gehört die Angabe der konkreten Ziehungsstrategie auch immer zu der Dokumentation der Ergebnisse.

## 2.11 Literatur

Eine Einführung in die nötigen Kenntnisse der Statistik findet man in Fahrmeir, Künstler, Pigeot, und Tutz (2009) oder Mosler und Schmid (2004). Eine umfangreiche Einführung in Stichprobenverfahren liefert Cochran (1977), der generell als

Standardwerk herangezogen werden kann. (Man beachte, dass das Buch in deutscher Übersetzung als Cochran (1972) vorliegt). Ebenfalls als deutschsprachige Literatur verweisen wir auf Kreienbrock (2004) oder Schwarz (1975). Empfehlenswert ist außerdem Scheaffer, Mendenhall und Ott (1995) und Leiner (1989). Dort wird eine elementare Einführung in die wichtigen Kapitel der Stichprobenverfahren gegeben. Umfassendes Material wird außerdem bereitgestellt in Thompson (2002) oder Levy und Lemeshow (1999), wobei Levy und Lemeshow praktische Aspekte deutlicher in den Vordergrund stellen.

## 2.12 Numerische Umsetzung

Die numerische Umsetzung von einfachen Zufallsstichproben ist mit vielen Softwarepaketen zu bewerkstelligen. Unsere Hinweise beziehen sich jedoch ausschließlich auf das Programmpaket **R**. Eine kurze Einführung und weitere Informationen zu diesem Programm sind in Anhang A zu finden. Das zu diesem Buch erstellte **R**-Paket `samplingbook` ist von der Homepage des R-Projekts ([www.r-project.org](http://www.r-project.org)) herunterladbar. Nach dem Herunterladen und Installieren kann es mit

```
> library(samplingbook)
```

geladen und auf diese Weise die darin enthaltenen Funktionen und Datensätze verfügbar gemacht werden.

In den folgenden Abschnitten werden das Ziehen einer einfachen Zufallsstichprobe, sowohl aus einem Vektor als auch aus einem Datensatz, die Mittelwert- und Anteilsschätzung für einfache Zufallsstichproben und die Bestimmung des Stichprobenumfangs zur Mittelwert- bzw. Anteilsschätzung dargestellt.

### 2.12.1 Ziehen einer einfachen Zufallsstichprobe

Bei der konkreten Durchführung einer einfachen Zufallsstichprobe muss aus der Populationsliste eine Stichprobe vom Umfang  $n$  ohne Zurücklegen gezogen werden. Dies ist in **R** mit der Funktion `sample(·)` realisierbar, welche ohne vorheriges Laden eines Pakets direkt zur Verfügung steht.

```
> sample(x, size, replace = FALSE, prob = NULL)
```

```
# x      Either a (numeric, complex, character or logical)
#        vector of more than one element from which to
#        choose, or a positive integer.
# size    non-negative integer giving the number of items to
#        choose.
# replace Should sampling be with replacement?
```

```
# prob      A vector of probability weights for obtaining the
#           elements of the vector being sampled.
```

Die Funktion `sample(.)` zieht eine Stichprobe festgelegter Größe `size` aus den Elementen des Vektors `x`, je nach Bedarf mit oder ohne Zurücklegen. Die Funktion zieht standardmäßig ohne Zurücklegen (`replace=FALSE`). Mit der Option `replace=TRUE` kann aber auch eine Stichprobe mit Zurücklegen gezogen werden. Mit `prob` können den einzelnen Elementen unterschiedliche Ziehungswahrscheinlichkeiten zugewiesen werden. Dies ist bei dieser Prozedur nur für das Ziehen mit Zurücklegen sinnvoll. Für das Ziehen ohne Zurücklegen liefert diese Option keine sinnvollen Ergebnisse, siehe dazu Kap. 4.

Da wir eine einfache Zufallsstichprobe ziehen wollen, ist für uns die Standardeinstellung des Ziehens ohne Zurücklegen bereits richtig voreingestellt. Außerdem sollen alle Elemente der Grundgesamtheit die gleiche Wahrscheinlichkeit besitzen, in die Stichprobe zu gelangen, weshalb wir auch die Option `prob=NULL` unverändert übernehmen können. Diese beiden Optionen können somit in der Syntax weggelassen werden.

Folglich müssen wir nur den Vektor `x`, aus dem die Zufallsstichprobe gezogen werden soll, und die Stichprobengröße `size` spezifizieren. Wollen wir beispielsweise eine Stichprobe vom Umfang  $n = 100$  aus einer Grundgesamtheit vom Umfang  $N = 2\,000$  ziehen, kann dies mit folgender Syntax realisiert werden:

```
> N <- 2000
> n <- 100
> populationlist <- 1:N
> set.seed(67399)
> sample(x=populationlist, size=n)

[1] 1093  538  932  110  796  322  187 1947 1981  740 1045  30
[13] 494 1846 1883 1446 1667 1322 1219 1712 1576 1874  261  904
[25] 1291 1033 1150  323  125 1135  687 1844  659 1145  936 1163
[37]  398  914 1544 1614 1248  850  861 1752  680  973  463 1198
[49] 1526 1905  848 1196  600 1513  146  862 1102 1162  392  531
[61]  817  769  708  841  665 1201 1413  337 1281  646 1984 1206
[73] 1414 1270  240  858  550 1059  652  778  795 1994 1224  27
[85] 1359  199  719 1493 1719  982  430  955  438 1042  382  672
[97]  662  728  855 1754
```

Wir nehmen dabei an, dass die Populationsliste von 1 bis  $N$  läuft und die gezogenen Werte den Identifikationsnummern der gezogenen Individuen entsprechen. Um das Ergebnis reproduzierbar zu machen, setzen wir unter Verwendung der Funktion `set.seed(.)` einen Startwert für den Algorithmus. Das Ergebnis kann in einen neuen Vektor, beispielsweise mit Namen `sample1` gespeichert werden. Außerdem können die gezogenen Zahlen mit `sort(.)` sortiert werden. Durch Eingabe von `sample1` kann das Ergebnis am Bildschirm ausgegeben werden.

```
> sample1 <- sort(sample(x=populationlist, size=n))
> sample1
```

[1]	31	32	96	121	152	165	205	211	215	251	262	265
[13]	281	282	285	292	308	310	338	388	394	410	418	419
[25]	431	443	456	485	487	513	534	543	558	559	565	622
[37]	645	666	717	719	750	761	766	776	791	809	824	839
[49]	867	875	879	894	924	946	986	993	1020	1027	1030	1031
[61]	1077	1095	1104	1122	1141	1197	1233	1240	1242	1258	1270	1273
[73]	1328	1361	1364	1383	1402	1409	1433	1460	1478	1538	1568	1578
[85]	1591	1592	1603	1631	1679	1701	1715	1718	1742	1768	1813	1861
[97]	1863	1932	1942	1948								

Der Aufruf von `sample(·)` liefert jedoch bei jedem Durchlauf ein anderes, zufälliges Ergebnis. Um die Ziehung reproduzierbar zu machen, empfiehlt es sich, mit einem sogenannten „seed“, einem Startwert für den Algorithmus (in diesem Fall zum Ziehen von Zufallszahlen), zu arbeiten. Dadurch wird sichergestellt, dass bei einem erneuten Ausführen die gleiche Stichprobe gezogen wird. Dieser Startwert kann festgelegt werden, indem man eine beliebig gewählte Zahl als Startwert definiert.

```
> start <- 13072008
> set.seed(start)
```

Im Anschluss daran liefert eine Funktion, die einen Zufallszahlengenerator enthält, wie z.B. `sample(·)`, immer dasselbe Ergebnis. Weitere Informationen zum Setzen eines Startwertes sind in Anhang A.8 zu finden.

Manchmal will man jedoch nicht eine Zufallsstichprobe aus einer geordneten Liste, sondern aus einem vorhandenen Datensatz ziehen. Zur Illustration verwenden wir als Datenbasis eine imaginäre Liste, welche vom Paket `samplingbook` bereitgestellt wird. Über die folgenden Befehle wird der Datensatz geladen und am Bildschirm ausgegeben. Der Datensatz ist nun im Objekt `pop` gespeichert.

```
> data(pop)
> print(pop)
```

	id	X	Y
1	1	11	9
2	2	11	10
3	3	11	11
4	4	21	18
5	5	21	22

Die Populationsliste ist bestimmt durch die Spalte `id`, denn jede ID ist exakt einer Person zugeordnet und umgekehrt. Nun gilt es, eine Stichprobe vom Umfang  $n$  zu ziehen, das heißt es sollen  $n$  ID-Nummern gewählt werden, um die zugehörigen Personen beispielsweise in eine Umfrage einzuschließen. Die entsprechenden zufällig gezogenen ID-Nummern erhält man mit

```
> n <- 3
> set.seed(93456)
> idsample <- sample(x=pop$id, size=n)
```

wobei die Stichprobengröße hier exemplarisch auf 3 festgelegt ist. Die ausgewählten ID-Nummern werden mit

```
> idsample
```

```
[1] 1 2 4
```

ausgegeben. Die Elemente der Stichprobe sind damit gezogen und können somit befragt werden. Auch hier kann man unter Verwendung eines Startwertes für den Algorithmus die Stichprobenziehung reproduzierbar gestalten.

Einen Datensatz mit den bereits vorhandenen Informationen zu den Personen in der Stichprobe erhält man beispielsweise mit folgender Syntax, welche sehr ausführlich gehalten ist, um die einzelnen, dazu notwendigen Schritte zu verdeutlichen. Man definiert zuerst eine Matrix, in die die Daten der Stichprobe geschrieben werden sollen, und initialisiert diese zunächst mit NA, d.h. mit fehlenden Werten. Unter Verwendung einer Zählvariable  $j$  füllen wir diese Matrix Zeile für Zeile mit den Daten der Stichprobe. Eine Schleife durchläuft die gezogenen ID-Nummern. In jedem Durchlauf werden die Daten des nächsten Stichprobenelements in die  $j$ -te Zeile der vordefinierten Matrix geschrieben und  $j$  um 1 erhöht.

```
> sample1 <- as.data.frame(matrix(data=NA, nrow=n,
+ ncol=ncol(pop)))
> j <- 1
> for(i in idsample)
+ {
+   sample1[j,] <- pop[pop$id==i,]
+   j <- j+1
+ }
```

Alternativ kann die bereits in **R** zur Verfügung gestellte Funktion `subset(.)` verwendet werden, welche aus einem gegebenen Datensatz eine Teilmenge von Zeilen auswählt.

```
> sample2 <- subset(x=pop, subset=(pop$id %in% idsample))
```

Weiterhin ist es möglich, direkt eine Stichprobe aus den Zeilen zu ziehen. In diesem Fall wird die bisher vorangegangene, separate Stichprobenziehung direkt integriert.

```
> sample3 <- pop[sample(1:nrow(pop), size=n),]
```

### 2.12.2 Mittelwertschätzung

In dem zu dem Buch gehörigen **R**-Paket `samplingbook` sind die Formeln zur Mittelwertschätzung und zur Bestimmung der Konfidenzintervalle umgesetzt. Die Funktion `Smean(.)` erlaubt diese Berechnungen bei Bedarf inklusive der Korrektur für endliche Populationen.

```
> Smean(y, N = Inf, level = 0.95)

# y      vector of sample data
# N      positive integer specifying population size.
#        Default is N=Inf, which means that
#        calculations are carried out
#        without finite population correction.
# level  coverage probability for confidence intervals.
#        Default is level=0.95.
```

Der Datenvektor `y` muss übergeben werden, die anderen beiden Angaben sind optional. Mit `N` kann der Populationsumfang übergeben werden, mit `level` die Überdeckungswahrscheinlichkeit des Konfidenzintervalls. Wird der Funktion kein Populationsumfang übergeben, wird `N=Inf` gesetzt und somit bei der Berechnung auf den Korrekturfaktor für endliche Populationen verzichtet. Dies ist sinnvoll, falls der Umfang der Grundgesamtheit sehr groß ist. Bei kleinen Grundgesamtheiten sollte der Populationsumfang übergeben werden. In letzterem Fall wird die Korrektur mit dem Faktor  $(N - n)/N$  bei der Berechnung der Varianz durchgeführt. Die Überdeckungswahrscheinlichkeit für das Konfidenzintervall ist mit 95% voreingestellt, kann aber bei Bedarf geändert werden.

Zur Illustration verwenden wir die Daten aus Beispiel 2.11:

$$Y_1 = 9, \quad Y_2 = 10, \quad Y_3 = 11, \quad Y_4 = 18, \quad Y_5 = 22.$$

Daraus wurde eine einfache Zufallsstichprobe vom Umfang  $n = 3$  gezogen. Anschließend können der geschätzte Mittelwert  $\hat{Y}_{ES}$ , die geschätzte Standardabweichung von  $\hat{Y}_{ES}$ , d.h.  $\sqrt{\text{Var}(\hat{Y}_{ES})}$  und das zugehörige Konfidenzintervall berechnet werden. Die dazu benötigte Variable  $Y$  ist im Datensatz `pop` enthalten. Dieser wird zuerst geladen und anschließend wird die relevante Information im Vektor `Y` gespeichert. Nach Festlegen eines Startwerts wird eine einfache Zufallsstichprobe vom Umfang  $n = 3$  gezogen und im Vektor `y` gespeichert. Zuletzt wird der Popu-

lationsmittelwert unter Berücksichtigung der Korrektur für endliche Populationen geschätzt.

```
> data(pop)
> Y <- pop$Y
> Y
```

```
[1]  9 10 11 18 22
```

```
> set.seed(93456)
> y <- sample(x=Y, size=3)
> y
```

```
[1]  9 10 18
```

```
> est <- Smean(y=y, N=length(Y))
> est
```

Smean object: Sample mean estimate  
With finite population correction: N=5

Mean estimate: 12.3333  
Standard error: 1.8012  
95% confidence interval: [8.803,15.8637]

### 2.12.3 Anteilsschätzung

Die Funktion `Sprop(·)`, ebenfalls aus dem Paket `samplingbook`, gibt die entsprechenden Werte bei der Anteilsschätzung zurück. Hierbei werden unter anderem die in Abschn. 2.6.4 dargestellten Prozeduren realisiert.

```
> Sprop(y, m, n = length(y), N = Inf, level = 0.95)

# y      vector of sample data containing values 0 and 1
# m      an optional non-negative integer for number of
#        positive events
# n      an optional positive integer for sample size.
#        Default is n = length(y).
# N      positive integer for population size. Default
#        is N=Inf, which means calculations are carried out
#        without finite population correction.
# level  coverage probability for confidence intervals.
#        Default is level=0.95.
```

Mit  $y$  kann der Datenvektor übergeben werden. Dieser muss dazu dummykodiert sein, d.h. er darf nur Nullen und Einsen enthalten, wobei die „ja“-Antworten typischerweise mit „1“ kodiert sind. Alternativ können  $m$ , die Anzahl an „ja“-Antworten bzw. „positiven Ereignissen“, und der Stichprobenumfang  $n$  übergeben werden. Werden  $m$  und  $y$  gleichzeitig angegeben, so muss  $m$  der Anzahl an Einsen im Datenvektor  $y$  entsprechen. Mit  $N$  kann wiederum die Größe der Grundgesamtheit angegeben werden. Die Angabe der Überdeckungswahrscheinlichkeit des Konfidenzintervalls ist mit 95% vorbelegt.

Die Ausgabe hängt von der Vorgabe für  $N$  ab. Bei einer endlichen Grundgesamtheit werden neben der Schätzung zwei Konfidenzintervalle ausgegeben. Das Erste basiert auf der approximativen Normalverteilung und benutzt die Varianz  $\widehat{\text{Var}}(\widehat{P}_{ES})$  aus Formel (2.1). Weiter wird das exakte Konfidenzintervall aus den Formeln (2.2) und (2.4) berechnet. Im Fall von großen Grundgesamtheiten (ab  $N > 100\,000$ ) wird  $N = \text{Inf}$  gesetzt, da eine exakte Berechnung sehr aufwendig wäre. Für diese Einstellung werden die in Abschn. 2.7 diskutierten Konfidenzintervalle berechnet. Das erste Intervall wird basierend auf der Normalverteilungsannahme berechnet, das zweite Intervall basiert ebenfalls auf der Normalverteilungsannahme, benutzt aber eine von Agresti und Coull (1998) vorgeschlagene Korrektur. Das Konfidenzintervall nach Clopper-Pearson stellt das exakte Konfidenzintervall dar.

Zur Illustration verwenden wir die Daten aus Beispiel 2.13. Dort wurden  $n = 100$  von  $N = 300$  Beschäftigten eines Betriebes zwei Fragen gestellt, wobei die erste Frage von  $m_1 = 45$  Personen und die zweite Frage von  $m_2 = 2$  Personen mit „Ja“ beantwortet wurde. Mit diesen Angaben können die geschätzten Anteile und die zugehörigen asymptotischen und exakten Konfidenzintervalle berechnet werden. Für die erste Frage ergibt sich:

```
> Sprop(m=45, n=100, N=300)
```

```
Sprop object: Sample proportion estimate
With finite population correction: N= 300
```

```
Proportion estimate: 0.45
Standard error: 0.0408
```

```
95% approximate hypergeometric confidence interval:
  proportion: [0.37,0.53]
  number in population: [111,159]
95% exact hypergeometric confidence interval:
  proportion: [0.3667,0.5367]
  number in population: [110,161]
```

Für die zweite Frage ergibt sich:

```
> Sprop(m=2, n=100, N=300)
```

```
Sprop object: Sample proportion estimate
With finite population correction: N= 300
```

```
Proportion estimate: 0.02
Standard error: 0.0115
95% approximate hypergeometric confidence interval:
  proportion: [-0.0025,0.0425]
  number in population: [0,12]
95% exact hypergeometric confidence interval:
  proportion: [0.0067,0.0633]
  number in population: [2,19]
```

Um die Vorgehensweise an realen Daten zu demonstrieren, verwenden wir die Ergebnisse zur Sonntagsfrage vom 3.7.2009 aus Beispiel 2.16. Dabei wurden  $n = 1206$  wahlberechtigte Personen gefragt, welche Partei sie wählen würden, wenn am kommenden Sonntag Bundestagswahl wäre. Es gaben  $m = 302$  Personen an, dass sie die SPD wählen würden. Weiterhin gaben  $m = 133$  Personen an, dass sie die Grünen wählen würden. Die geschätzten Anteile und die zugehörigen Konfidenzintervalle werden wie folgt berechnet, wobei  $N=\text{Inf}$  gesetzt wurde. Die Berechnung mit der exakten Anzahl an wahlberechtigten Personen von ca. 61 Mio liefert die gleichen Ergebnisse. Für die SPD ergibt sich:

```
> Sprop(m=302, n=1206, N=Inf)
```

```
Sprop object: Sample proportion estimate
Without finite population correction: N= Inf
```

```
Proportion estimate: 0.2504
Standard error: 0.0125

95% asymptotic confidence interval:
  proportion: [0.226,0.2749]
95% asymptotic confidence interval with correction by Wilson:
  proportion: [0.2268,0.2756]
95% exact confidence interval by Clopper-Pearson:
  proportion: [0.2262,0.2759]
```

Für die Grünen ergibt sich:

```
> Sprop(m=133, n=1206, N=Inf)
```

```
Sprop object: Sample proportion estimate
Without finite population correction: N= Inf
```

```
Proportion estimate: 0.1103
```

Standard error: 0.009

95% asymptotic confidence interval:

proportion: [0.0926,0.128]

95% asymptotic confidence interval with correction by Wilson:

proportion: [0.0938,0.1292]

95% exact confidence interval by Clopper-Pearson:

proportion: [0.0932,0.1293]

Somit liegt der erwartete Stimmenanteil nach der Methode von Wilson für die SPD zwischen 22, 6 und 27, 6% und für die Grünen zwischen 9, 3 und 13, 0%.

In Beispiel 2.14 wollten wir ein einseitiges Konfidenzintervall zur Abschätzung der maximalen Anzahl an Tieren, die an einer seltenen Krankheit leiden, erhalten. Dazu kann man bei Verwendung von `Sprop(·)` die Überdeckungswahrscheinlichkeit dementsprechend anpassen. Um also z.B. eine obere Grenze für ein einseitiges 95%-Konfidenzintervall zu erhalten, kann man die Funktion mit `level=0.9` aufrufen und aus dem Output die obere Grenze ablesen.

```
> Sprop(m = 1, n = 500, N=10000, level = 0.9)
```

Sprop object: Sample proportion estimate

With finite population correction: N= 10000

Proportion estimate: 0.002

Standard error: 0.0019

90% approximate hypergeometric confidence interval:

proportion: [-0.0012,0.0052]

number in population: [-12,52]

90% exact hypergeometric confidence interval:

proportion: [1e-04,0.0093]

number in population: [1,93]

Als obere Grenze des einseitigen exakten 95%-Konfidenzintervalls ergibt sich somit 0.0093 für den Anteil bzw. 93 für die Anzahl. Die ausgegebene untere Grenze wird nicht weiter berücksichtigt. Man beachte, dass das approximative Konfidenzintervall hier nicht sinnvoll ist.

### ***2.12.4 Bestimmung des Stichprobenumfangs bei Mittelwertschätzung***

Die Funktion `sample.size.mean(·)` im Paket `samplingbook` berechnet den notwendigen Stichprobenumfang, um einen Mittelwert mit vorgegebener Genauigkeit  $e$  (halbe Länge des Konfidenzintervalls, siehe S. 40) zu schätzen.

```
> sample.size.mean(e, S, N = Inf, level = 0.95)

# e      positive number specifying the precision which is half
#        width of confidence interval
# S      standard deviation in population
# N      positive integer for population size. Default is N=Inf,
#        which means that calculations are carried out
#        without finite population correction.
# level  coverage probability for confidence intervals.
#        Default is level=0.95.
```

Die Genauigkeit  $e$  und die Standardabweichung  $S$  müssen der Funktion übergeben werden. Die Angaben zum Populationsumfang  $N$  und zur Überdeckungswahrscheinlichkeit des Konfidenzintervalls sind optional.

Als Beispiel betrachten wir den Fall, dass bei einer endlichen Grundgesamtheit von  $N = 300$  und einer Standardabweichung von  $S = 10$  eine Genauigkeit von  $e = 4$  (bzw.  $e = 1$ ) erreicht werden soll. Das Konfidenzniveau soll jeweils 95% betragen. Für eine Genauigkeit von  $e = 4$  ergibt sich:

```
> sample.size.mean(e=4, S=10, N=300)
```

```
sample.size.mean object: Sample size for mean estimate
With finite population correction: N=300, precision e=4
and standard deviation S=10
```

```
Sample size needed: 23
```

Für eine Genauigkeit von  $e = 1$  ergibt sich:

```
> sample.size.mean(e=1, S=10, N=300)
```

```
sample.size.mean object: Sample size for mean estimate
With finite population correction: N=300, precision e=1
and standard deviation S=10
```

```
Sample size needed: 169
```

Der benötigte Stichprobenumfang beträgt bei einer Genauigkeit von  $e = 4$  somit 23 Personen und bei einer Genauigkeit von  $e = 1$  sogar 169 Personen.

### 2.12.5 Bestimmung des Stichprobenumfangs bei Anteilsschätzung

Die Funktion `sample.size.prop(.)` aus dem Paket `samplingbook` ist das Analogon zur Funktion `sample.size.mean(.)` und berechnet den notwendigen Stichprobenumfang, um einen Anteil mit vorgegebener Genauigkeit  $e$  zu schätzen.

```
> sample.size.prop(e, P = 0.5, N = Inf, level = 0.95)
```

```
# e      positive number specifying the precision which is
#        half width of confidence interval
# P      expected proportion of events with domain between
#        values 0 and 1. Default is P=0.5.
# N      positive integer for population size. Default
#        is N=Inf, which means that calculations are carried
#        out without finite population correction.
# level  coverage probability for confidence intervals.
#        Default is level=0.95.
```

Die Genauigkeit  $e$  muss der Funktion übergeben werden, die Angaben zum Anteil an „Ja“-Antworten bzw. positiven Ereignissen  $P$ , zum Populationsumfang  $N$  und zur Überdeckungswahrscheinlichkeit des Konfidenzintervalls sind optional. Für den Anteil an „Ja“-Antworten bzw. positiven Ereignissen ist das „Worst Case“ Szenario von  $P=0.5$  voreingestellt. Alternativ kann man für  $P$  eine obere Abschätzung für Anteile kleiner 0.5 bzw. eine untere Abschätzung für Anteile größer 0.5 angeben.

Zunächst wollen wir uns mit der benötigten Stichprobengröße für Wahlprognosen wie in Beispiel 2.16 beschäftigen. Kurz vor der Bundestagswahl 2005 ist eine größere Genauigkeit der Prognosen für die einzelnen Parteien äußerst wichtig. Deshalb möchte ein Meinungsforschungsinstitut eine Stichprobe ziehen, mit der die Anteile der einzelnen Parteien mit einer Genauigkeit von  $e = 0.01$  geschätzt werden können. Will man die Anteile für alle Parteien mit dieser Genauigkeit schätzen, sollte man wieder  $P = 0.5$  wählen.

```
> sample.size.prop(e=0.01, P=0.5, N=Inf)
```

```
sample.size.prop object: Sample size for proportion estimate
Without finite population correction: N=Inf, precision e=0.01
and expected proportion P=0.5
```

```
Sample size needed: 9604
```

Will man nur den Anteil einer bestimmten Partei abschätzen, kann man Vorwissen aus früheren Wahlergebnissen nutzen. Die Tendenzen für die SPD lassen beispielsweise erkennen, dass die Partei das Wahlergebnis bei der letzten Wahl 2002, nämlich 39%, nicht überschreiten wird. Man erhält somit

```
> sample.size.prop(e=0.01, P=0.39, N=Inf)
```

```
sample.size.prop object: Sample size for proportion estimate
Without finite population correction: N=Inf, precision e=0.01
and expected proportion P=0.39
Sample size needed: 9139
```

Weiterhin betrachten wir nochmal das Beispiel 2.13 auf S. 32 zur Umfrage zur Verbesserung des Betriebsklimas in einem Betrieb mit  $N = 300$  Mitarbeitern. Die flexibleren Arbeitszeiten wurden erfolgreich umgesetzt. Nach einem Jahr sollen diese mit der Frage „Sind Sie mit der neuen Arbeitszeitregelung zufrieden?“ (Ja/Nein) evaluiert werden.

Die Umfrage soll dabei möglichst effizient sein, weshalb man den Stichprobenumfang diesmal im Vorhinein berechnen will. Deshalb werden die zwei Genauigkeiten  $e = 0.05$  und  $e = 0.1$  zur Auswahl gestellt, wobei als Wahrscheinlichkeit  $P = 0.5$  gewählt wurde, da zu der aktuellen Einschätzung noch kein Vorwissen vorhanden ist. Für eine Genauigkeit von  $e = 0.05$  ergibt sich:

```
> sample.size.prop(e=0.05, P=0.5, N=300)
```

```
sample.size.prop object: Sample size for proportion estimate
With finite population correction: N=300, precision e=0.05
and expected proportion P=0.5
```

```
Sample size needed: 169
```

Für eine Genauigkeit von  $e = 0.1$  ergibt sich:

```
> sample.size.prop(e=0.1, P=0.5, N=300)
```

```
sample.size.prop object: Sample size for proportion estimate
With finite population correction: N=300, precision e=0.1
and expected proportion P=0.5
```

```
Sample size needed: 73
```

Entsprechend lassen sich die Werte aus den Tabellen 2.2 und 2.3 mit der Funktion `sample.size.prop(.)` erzeugen. Die erste Spalte von Tabelle 2.2 ergibt sich durch:

```
> sample.size.prop(e=0.1, P=0.2, N=10)
```

```
sample.size.prop object: Sample size for proportion estimate
With finite population correction: N=10, precision e=0.1
```

and expected proportion  $P=0.2$

Sample size needed: 9

```
> sample.size.prop(e=0.1, P=0.3, N=10)
```

```
sample.size.prop object: Sample size for proportion estimate  
With finite population correction: N=10, precision e=0.1  
and expected proportion P=0.3
```

Sample size needed: 9

```
> sample.size.prop(e=0.1, P=0.4, N=10)
```

```
sample.size.prop object: Sample size for proportion estimate  
With finite population correction: N=10, precision e=0.1  
and expected proportion P=0.4
```

Sample size needed: 10

```
> sample.size.prop(e=0.1, P=0.5, N=10)
```

```
sample.size.prop object: Sample size for proportion estimate  
With finite population correction: N=10, precision e=0.1  
and expected proportion P=0.5
```

Sample size needed: 10

Die Werte in den weiteren Spalten und in Tabelle 2.3 erhält man analog.

Stichproben

Methoden und praktische Umsetzung mit R

Kauermann, G.; Küchenhoff, H.

2011, X, 261 S. 16 Abb., Softcover

ISBN: 978-3-642-12317-7