2-Punkte-Verteilung. Die größte Streuung resultiert, wenn jeweils die Hälfte aller Merkmalsträger die Werte X_{max} und X_{min} annehmen (Abb. 7.7a). Sie hat dann den Wert

$$\hat{\sigma} = \frac{X_{max} - X_{min}}{2} = \frac{R}{2} = 0{,}5 \cdot R. \tag{7.11}$$

Sind die beiden extremen Merkmalsausprägungen nicht gleich häufig besetzt, reduziert sich die Streuung. Sie lässt sich bestimmen, wenn zusätzlich die Größenordnung des Mittelwertes $\bar{X}$ der Verteilung bekannt ist:

$$\hat{\sigma} = \sqrt{(X_{max} - \bar{X}) \cdot (\bar{X} - X_{min})}\,. \tag{7.12}$$

Gleichverteilung. Merkmale, die zwischen X_{min} und X_{max} in etwa gleichverteilt sind (Abb. 7.7b) haben eine Streuung von

$$\hat{\sigma} = \frac{1}{\sqrt{3}} \cdot \frac{R}{2} = \frac{R}{\sqrt{12}} = 0{,}289 \cdot R. \tag{7.13}$$

Die Streuung von Merkmalen, die in zwei Bereichen unterschiedlich gleichverteilt sind (konstante Dichte in einem Bereich und konstante, aber andere Dichte in einem anderen Bereich), lässt sich ermitteln, wenn sich für $\bar{X}$ ein plausibler Wert angeben lässt:

$$\hat{\sigma} = \frac{1}{\sqrt{3}} \cdot \sqrt{(X_{max} - \bar{X}) \cdot (\bar{X} - X_{min})}\,. \tag{7.14}$$

Dreiecksverteilung. Merkmale, deren Dichte von einem Merkmalsextrem zum anderen kontinuierlich sinkt (oder steigt), heißen Dreiecksverteilung (Abb. 7.7c). Für Merkmale mit dieser Verteilungsform kann die Streuung nach folgender Gleichung geschätzt werden:

$$\hat{\sigma} = \frac{R}{\sqrt{18}} = 0{,}236 \cdot R\,. \tag{7.15}$$

Gleichseitige Dreiecksverteilung. Bei einem häufig anzutreffenden Verteilungsmodell strebt die Dichte vom Merkmalszentrum aus nach beiden Seiten gegen null (Abb. 7.7d). Für diese Verteilungsform lässt sich die Streuung in folgender Weise schätzen:

$$\hat{\sigma} = \frac{R}{\sqrt{24}} = 0{,}204 \cdot R\,. \tag{7.16}$$

Normalverteilung. Ist es realistisch, für das untersuchte Merkmal eine Normalverteilung anzunehmen, ermöglicht die folgende Gleichung eine brauchbare Streuungsschätzung:

$$\hat{\sigma} = \frac{R}{5{,}15} = 0{,}194 \cdot R\,. \tag{7.17}$$

Die Streubreite R entspricht hierbei einem Intervall, in dem sich etwa 99% aller Werte befinden (Abb. 7.7e).

! Um die Populationsstreuung σ als Punktschätzung zu schätzen, nutzt man Transformationen der Streubreite (Range: Maximalwert – Minimalwert). Wie der Range in die Streuung zu transformieren ist, hängt von der Verteilungsform des interessierenden Merkmals ab.

Liegen überhaupt keine Angaben über die mutmaßliche Größe von σ oder R vor, bleibt letztlich nur die Möglichkeit, den endgültigen Stichprobenumfang erst während der Datenerhebung festzulegen. Man errechnet z. B. aus den ersten 20 Messwerten eine vorläufige Streuungsschätzung, die für eine erste Schätzung des erforderlichen Stichprobenumfanges herangezogen wird. Liegen weitere Messwerte (z. B. insgesamt 40 Messwerte) vor, wird die Streuung erneut berechnet und die Stichprobengröße ggf. korrigiert. Dieser Korrekturvorgang setzt sich so lange fort, bis sich die Streuungsschätzung stabilisiert oder der zuletzt errechnete Stichprobenumfang erreicht ist.

7.2 Möglichkeiten der Präzisierung von Parameterschätzungen

Bisher erfolgte die Beschreibung von Populationen bzw. die Schätzung von Populationsparametern aufgrund einfacher Zufallsstichproben, was die Methode der Wahl ist, wenn man tatsächlich über eine vollständige Liste aller Objekte der Grundgesamtheit verfügt, sodass alle Objekte mit gleicher Wahrscheinlichkeit Mitglied der Stichprobe werden können (▶ S. 398 ff.). Die Praxis lehrt uns jedoch, dass für Umfragen in größeren Referenz-

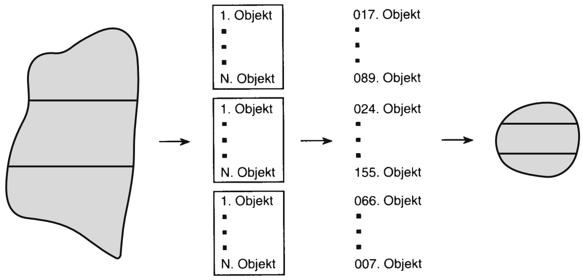
J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

populationen derartige Listen nicht existieren bzw. nur mit einem unzumutbaren Aufwand erstellt werden können. Will man dennoch auf eine probabilistische Stichprobe nicht verzichten (was für die seriöse wissenschaftliche Umfrageforschung unabdingbar ist), kann auf eine der Stichprobentechniken zurückgegriffen werden, die in diesem Abschnitt zu behandeln sind (zu Qualitätskriterien sozialwissenschaftlicher Umfrageforschung vgl. auch Kaase, 1999). Diese Stichprobentechniken haben gegenüber der einfachen Zufallsstichprobe zudem den Vorteil einer genaueren Parameterschätzung. Dies setzt allerdings voraus, dass man bereits vor der Untersuchung weiß, welche Merkmale mit dem interessierenden Merkmal zusammenhängen bzw. wie dieses Merkmal ungefähr verteilt ist. Wenn derartige Vorkenntnisse geschickt im jeweiligen Stichprobenplan eingesetzt werden, kann sich dieses Wissen in Form von präziseren Parameterschätzungen mehr als bezahlt machen.

Von den verschiedenen Möglichkeiten einer **verbesserten Parameterschätzung** werden hier nur die wichtigsten Ansätze aufgegriffen: Besonderheiten einer geschichteten Stichprobe (▶ Abschn. 7.2.1), einer Klumpenstichprobe (▶ Abschn. 7.2.2), einer mehrstufigen Stichprobe (▶ Abschn. 7.2.3), der wiederholten Untersuchung von Teilen einer Stichprobe (▶ Abschn. 7.2.4) sowie die Nutzung von Vorinformationen nach dem Bayes'schen Ansatz (▶ Abschn. 7.2.5). Im ▶ Abschn. 7.2.6 werden wir kurz auf Resamplingtechniken eingehen.

7.2.1 Geschichtete Stichprobe

Zur Erläuterung des Begriffs geschichtete Stichprobe (oder stratifizierte Stichprobe, »Stratified Sample«) möge das in ◘ Box 7.3 beschriebene Beispiel dienen, in dem es um die Schätzung der durchschnittlichen Seitenzahl von Diplomarbeiten anhand einer Zufallsstichprobe ging. Die dort beschriebene Vorgehensweise ließ die Art der Diplomarbeit außer Acht, obwohl bekannt ist, dass z. B. theoretische Literaturarbeiten in der Regel umfangreicher sind als empirische Arbeiten. Das Merkmal »Art der Diplomarbeit« korreliert – so können wir annehmen – mit dem untersuchten Merkmal »Anzahl der Seiten«. Dieser Zusammenhang lässt sich für eine präzisere Parameterschätzung nutzen. Die untersuchten Arbeiten werden nach der Art der Themenstellung in



◘ **Abb. 7.8.** Ziehung einer geschichteten Stichprobe

einzelne »Schichten« oder »Strata« eingeteilt, aus denen sich – wie im Folgenden gezeigt wird – eine verbesserte Schätzung der durchschnittlichen Seitenzahl aller Diplomarbeiten ableiten lässt. Eine schematische Darstellung einer geschichteten Stichprobe zeigt ◘ Abb. 7.8.

Für die meisten human- und sozialwissenschaftlichen Forschungen, die Personen als Merkmalsträger untersuchen, erweisen sich biografische und soziodemografische Merkmale (Alter, Geschlecht, soziale Schicht, Bildung etc.) als günstige Schichtungsmerkmale. Hat ein Schichtungsmerkmal k Ausprägungen (im oben erwähnten Beispiel wäre k=2), wird für jede Ausprägung eine Zufallsstichprobe des Umfanges n_j benötigt (j=1,2…,k). Bei einem gegebenen Schichtungsmerkmal entscheidet allein die Aufteilung der Gesamtstichprobe auf die einzelnen Schichten, also die Größe n_j der Teilstichproben, über die Präzision der Parameterschätzung. Die folgenden Abschnitte diskutieren Vor- und Nachteile verschiedener Aufteilungsstrategien, wobei sich die Ausführungen erneut nur auf die Schätzung des Populationsparameters μ (Mittelwert) und π (Populationsanteil) beziehen. Ferner werden die Konsequenzen für Stichprobenumfänge diskutiert, wenn anstelle einer Zufallsstichprobe eine geschichtete Stichprobe eingesetzt wird.

! Man zieht eine geschichtete Stichprobe, indem man die Zielpopulation auf der Basis einer oder mehrerer Merkmale in Teilpopulationen (Schichten) einteilt – pro Merkmalsausprägung bzw. Merkmalskombination entsteht eine Teilpopulation – und aus jeder dieser Schichten eine Zufallsstichprobe entnimmt.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Hier und im Folgenden gehen wir erneut davon aus, dass der »Auswahlsatz« in jeder Schicht $(n_j/N_j)<0{,}05$ ist (▶ S. 415).

Wenn die Stichprobenumfänge n_j zu ihren jeweiligen Teilpopulationen N_j proportional sind, sprechen wir von einer **proportional geschichteten Stichprobe**. In diesem Falle haben alle Objekte der Gesamtpopulation – wie bei der einfachen Zufallsstichprobe – dieselbe Auswahlwahrscheinlichkeit n_j/N_j=const. Bei disproportional geschichteten Stichproben sind die Auswahlwahrscheinlichkeiten für die einzelnen Teilpopulationen unterschiedlich. Die Berechnung erwartungstreuer Schätzwerte für Mittel- und Anteilswerte setzt hier voraus, dass die schichtspezifischen Auswahlwahrscheinlichkeiten bekannt sind (oder zumindest geschätzt werden können), sodass Ergebnisverzerrungen durch eine geeignete Gewichtung kompensiert werden können. Hiermit beschäftigt sich der folgende Abschnitt.

Schätzung von Populationsmittelwerten

Zur Schätzung des Populationsparameters μ_j der Teilpopulation j verwenden wir das arithmetische Mittel $\bar{X}_j$ der Teilstichprobe j:

$$\bar{X}_j = \frac{\sum_{i=1}^{n_j} x_i}{n_j}. \tag{7.18}$$

Wie lässt sich nun aus den einzelnen Teilstichprobenmittelwerten $\bar{X}_j$ eine Schätzung des Populationsparameters μ der gesamten Population ableiten?

Beliebige Aufteilung. Für beliebige Stichprobenumfänge n_j stellt die folgende gewichtete Summe der einzelnen Teilstichprobenmittelwerte eine erwartungstreue Schätzung des Populationsparameters μ dar:

$$\bar{X}_{bel} = \sum_{j=1}^{k} g_j \cdot \bar{X}_j, \tag{7.19}$$

wobei

$$\sum_{j=1}^{k} g_j = 1$$

(»bel« steht für »beliebige Stichprobenumfänge«. Zur Herleitung dieser und der folgenden Gleichungen wird z. B. auf Schwarz, 1975, verwiesen.)

Die Gewichte g_j reflektieren die relative Größe einer Schicht (Teilpopulation im Verhältnis zur Größe der Gesamtpopulation). Bezeichnen wir den Umfang der Teilpopulation j mit N_j und den Umfang der Gesamtpopulation mit N, ist ein Gewicht g_j durch

$$g_j = \frac{N_j}{N} \tag{7.20}$$

definiert.

Man beachte, dass diese Art der Zusammenfassung einzelner Mittelwerte nicht mit der Zusammenfassung von Mittelwerten identisch ist, die alle erwartungstreue Schätzungen ein und desselben Parameters μ sind. Werden aus einer Population mehrere Zufallsstichproben des Umfanges n_j gezogen, ergibt sich der Gesamtmittelwert $\bar{X}$ aller Teilmittelwerte $\bar{X}_j$ nach der Beziehung

$$\bar{X} = \frac{\sum_{j=1}^{k} n_j \cdot \bar{X}_j}{\sum_{j=1}^{k} n_j}.$$

Die Zufälligkeit der Teilstichprobenmittelwerte $\bar{X}_j$ bedingt, dass auch $\bar{X}_{bel}$ eine Zufallsvariable darstellt. Der Stichprobenkennwert $\bar{X}_{bel}$ ist annähernd normalverteilt, wenn in jeder Stichprobe j $n_j \cdot g_j \geq 10$ ist. Die Streuung der Mittelwertverteilung $\bar{X}_{bel}$ bzw. der Standardfehler von $\bar{X}_{bel}$ lautet:

$$\hat{\sigma}_{\bar{x}(bel)} = \sqrt{\sum_{j=1}^{k} g_j^2 \cdot \hat{\sigma}_{\bar{x}_j}^2} = \sqrt{\sum_{j=1}^{k} g_j^2 \cdot \frac{\hat{\sigma}_j^2}{n_j}}. \tag{7.21}$$

Hierbei ist $\hat{\sigma}_j^2$ die aufgrund der Teilstichprobe j geschätzte Varianz der Teilpopulation j. Mit diesem Standardfehler ergibt sich folgendes Konfidenzintervall des Populationsparameters μ:

$$\bar{X}_{bel} \pm z \cdot \hat{\sigma}_{\bar{x}(bel)} \tag{7.22}$$

Für das 95%ige Konfidenzintervall wählen wir z=1,96 und für das 99%ige Intervall z=2,58.

Die Gleichungen verdeutlichen, dass das Konfidenzintervall kleiner wird, wenn sich die Streuungen in den Teilstichproben verringern. Niedrige Streuungen in den einzelnen Schichten bedeuten **Homogenität** des untersuchten Merkmals in den einzelnen Schichten, die wiederum um so höher ausfällt, je deutlicher das Schich-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

tungsmerkmal mit dem untersuchten Merkmal korreliert. Besteht zwischen diesen beiden Merkmalen kein Zusammenhang, schätzt jedes $\hat{\sigma}_j$ die Streuung σ der Gesamtpopulation, d. h., die Schichtung ist bedeutungslos. Eine Zufallsstichprobe schätzt dann den Populationsparameter genauso gut wie eine geschichtete Stichprobe. Bei ungleichen $\hat{\sigma}_j$-Werten wird die Parameterschätzung genauer, wenn stark streuende Teilstichproben mit einem niedrigen Gewicht versehen sind.

Die Informationen, die beim Einsatz geschichteter Stichproben bekannt sein müssen, beziehen sich damit nicht nur auf Merkmale, die mit dem untersuchten Merkmal korrelieren, sondern auch auf die Größen der Teilpopulationen bzw. die Gewichte der Teilstichproben. Auch wenn man sicher ist, dass Merkmale wie Geschlecht, Alter, Wohngegend usw. mit dem untersuchten Merkmal zusammenhängen, nützt dies wenig, wenn nicht zusätzlich auch die Größen der durch das Schichtungsmerkmal definierten Teilpopulationen bekannt sind.

Hierin liegt der entscheidende Nachteil geschichteter Stichproben. Zwar informieren die vom Statistischen Bundesamt in regelmäßigen Abständen herausgegebenen amtlichen Statistiken über die Verteilung vieler wichtiger Merkmale; dennoch interessieren häufig Schichtungsmerkmale, bei denen man nicht weiß, wie groß die Teilpopulationen sind.

In diesen Fällen muss man sich mit Schätzungen begnügen, die entweder auf Erfahrung bzw. ähnlichen Untersuchungen beruhen oder die aus den in einer Zufallsstichprobe angetroffenen Größen der Teilstichproben abgeleitet werden (**Ex-post-Stratifizierung**, Box 7.4). In jedem Falle ist zu fordern, dass bei geschätzten Stichprobengewichten das Konfidenzintervall der geschichteten Stichprobe demjenigen Konfidenzintervall gegenüber gestellt wird, das resultiert, wenn man die Schichtung außer acht lässt (d. h., wenn man die Stichprobe als einfache Zufallsstichprobe behandelt und das Konfidenzintervall nach den auf ▶ S. 414 ff. beschriebenen Regeln bestimmt). Der Leserschaft einer solchen Untersuchung bleibt es dann überlassen, ob sie die Gewichtsschätzungen und damit auch das Konfidenzintervall aufgrund der geschichteten Stichprobe akzeptiert oder ob sie die in der Regel ungenauere, aber unproblematischere Schätzung aufgrund der Zufallsstichprobe für angemessener hält.

Gleiche Aufteilung. Wenn die Zufallsstichproben, die den einzelnen Teilpopulationen entnommen werden, gleich groß sind, sprechen wir von einer gleichen Aufteilung. In diesem Falle ist $n_1=n_2=\ldots=n/k$, sodass für den Standardfehler des Mittelwertes $\hat{\sigma}_{\bar{x}(\text{gleich})}$ resultiert:

$$\hat{\sigma}_{\bar{x}(\text{gleich})} = \sqrt{\frac{k}{n} \cdot \sum_{j=1}^{k} g_j^2 \cdot \hat{\sigma}_j^2} \qquad (7.23)$$
$$(n = n_1 + n_2 + \ldots + n_k).$$

Für den Mittelwert $\bar{X}_{(\text{gleich})}$ ist es unerheblich, ob die Teilstichproben gleich groß oder beliebig groß sind, d. h., $\bar{X}_{(\text{gleich})}$ wird ebenfalls über ▶ Gl. (7.19) berechnet. In der Gleichung zur Bestimmung des Konfidenzintervalls ist deshalb nur der Standardfehler des Mittelwertes für geschichtete Stichproben mit beliebigen Umfängen ($\hat{\sigma}_{\bar{x}(\text{bel})}$) durch den hier aufgeführten Standardfehler $\hat{\sigma}_{\bar{x}(\text{gleich})}$ zu ersetzen. Auch dieser Ansatz wird in Box 7.4 an einem Beispiel erläutert.

Proportionale Aufteilung. Stichprobenumfänge, die im gleichen Verhältnis zueinander stehen wie die entsprechenden Teilpopulationen, heißen proportionale Stichprobenumfänge. In diesem Falle ist

$$\frac{n_j}{n} = \frac{N_j}{N} = g_j$$

bzw.

$$n_j = n \cdot \frac{N_j}{N} = n \cdot g_j.$$

Für den Mittelwert $\bar{X}_{\text{prop}}$ resultiert dann

$$\bar{X}_{\text{prop}} = \sum_{j=1}^{k} g_j \cdot \bar{X}_j = \sum_{j=1}^{k} \frac{n_j}{n} \cdot \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \qquad (7.24)$$
$$= \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}.$$

Für proportionale Stichprobenumfänge entfallen bei der Berechnung des Stichprobenmittelwertes die Gewichte der einzelnen Teilstichproben. Man bezeichnet deshalb eine geschichtete Stichprobe mit proportionalen Stichprobenumfängen auch als **selbstgewichtende Stich-**

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

7

Box 7.4

Wie umfangreich sind Diplomarbeiten?
II: Die geschichtete Stichprobe

Box 7.3 demonstrierte ein Konfidenzintervall für μ an einem Beispiel, bei dem es um die durchschnittliche Seitenzahl von Diplomarbeiten ging. Die Überprüfung von 100 zufällig ausgewählten Diplomarbeiten führte zu einem Mittelwert von $\bar{x}$=92 Seiten und einer Standardabweichung von $\hat{\sigma} = \sqrt{1849} = 43$ Seiten. Damit ist $\hat{\sigma}_{\bar{x}} = 4{,}3$. Für das 99%ige Konfidenzintervall resultierte der Bereich 92±11 Seiten.

Es soll nun geprüft werden, ob sich dieses Konfidenzintervall verkleinern lässt, wenn die Zufallsstichprobe aller Diplomarbeiten als »geschichtete« Stichprobe behandelt wird, die sich aus theoretischen Literaturarbeiten und empirischen Arbeiten zusammensetzt. Für diese zwei Kategorien mögen sich folgende Häufigkeiten, Mittelwerte und Standardabweichungen ergeben haben:

theoretische Literaturarbeiten:

$n_1 = 32, \bar{x}_1 = 132, \hat{\sigma}_1 = 51;$

empirische Arbeiten:

$n_2 = 68, \bar{x}_2 = 73, \hat{\sigma}_2 = 19.$

Zunächst ermitteln wir das Konfidenzintervall für μ nach den Gleichungen für eine geschichtete Stichprobe mit beliebigen Stichprobenumfängen. Hierbei gehen wir davon aus, dass die relativen Größen der zwei Stichproben den Gewichten g_j entsprechen. Sie lauten damit: g_1=0,32; g_2=0,68. Es ergeben sich dann:

$$\begin{aligned}\bar{x}_{bel} &= \sum_{j=1}^{k} g_j \cdot \bar{x}_j = 0{,}32 \cdot 132 + 0{,}68 \cdot 73\\ &= 91{,}88 \approx 92\end{aligned}$$

und

$$\begin{aligned}\hat{\sigma}_{\bar{x}(bel)} &= \sqrt{\sum_{j=1}^{k} g_j^2 \cdot \frac{\hat{\sigma}_j^2}{n_j}}\\ &= \sqrt{0{,}32^2 \cdot \frac{51^2}{32} + 0{,}68^2 \cdot \frac{19^2}{68}}\\ &= \sqrt{10{,}78} = 3{,}28.\end{aligned}$$

Das 99%ige Konfidenzintervall heißt also:

$$\bar{x}_{bel} \pm 2{,}58 \cdot \hat{\sigma}_{\bar{x}(bel)} = 92 \pm 2{,}58 \cdot 3{,}28 \approx 92 \pm 8.$$

Das Konfidenzintervall hat sich durch die Berücksichtigung des Schichtungsmerkmals »Art der Diplomarbeit« deutlich verkleinert. Es hat nun die Grenzen 84 Seiten und 100 Seiten.

Da die Populationsanteile (bzw. die Gewichte g_j) der beiden Schichtungskategorien aus der Stichprobe geschätzt wurden, sind die Teilstichprobenumfänge n_j zwangsläufig proportional zu den Gewichten g_j. Wir können damit das Konfidenzintervall auch nach den einfacheren Regeln für proportional geschichtete Stichproben bestimmen. Die Resultate sind, wie die folgenden Berechnungen zeigen, natürlich mit den oben genannten identisch.

$$\bar{x}_{prop} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} = 91{,}88 \approx 92$$

(Die Gesamtsumme aller Seitenzahlen wurde im Beispiel nicht vorgegeben. Sie lässt sich jedoch nach der Beziehung

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} = \sum_{j=1}^{k} n_j \cdot \bar{x}_j$$

einfach bestimmen.)

$$\begin{aligned}\hat{\sigma}_{\bar{x}(prop)} &= \sqrt{\frac{1}{n} \sum_{j=1}^{k} g_j \cdot \hat{\sigma}_j^2}\\ &= \sqrt{\frac{1}{100} \cdot (0{,}32 \cdot 51^2 + 0{,}68 \cdot 19^2)} = 3{,}28.\end{aligned}$$

▼

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Die hier vorgenommene Stichprobenaufteilung bezeichnet man als Ex-post-Schichtung oder **Ex-post-Stratifizierung**: Die zufällig ausgewählten Untersuchungseinheiten werden erst nach der Stichprobenentnahme den Stufen des Schichtungsmerkmals zugeordnet (zum Vergleich von ex-post-stratifizierten Stichproben mit geschichteten Stichproben, bei denen das Schichtungsmerkmal als Selektionskriterium für die einzelnen Untersuchungseinheiten eingesetzt wird, s. Kish, 1965, Kap. 3.4c).

Bei der Ex-post-Schichtung schätzen wir die relative Größe der Teilpopulationen über die relative Größe der Teilstichproben, d. h., wir behaupten, dass $n_j/n \approx N_j/N$ ist. Diese Behauptung ist korrekt, wenn die Gesamtstichprobe (n) tatsächlich zufällig aus der Gesamtpopulation (N) gezogen wurde. Es resultiert eine proportional geschichtete Stichprobe mit einer für alle Diplomarbeiten identischen Auswahlwahrscheinlichkeit von $n_j/N_j = n/N = \text{const}$.

Eine geschichtete Stichprobe mit gleicher Aufteilung liegt vor, wenn $n_1=50$ theoretische Literaturarbeiten und $n_2=50$ empirische Arbeiten untersucht werden. Auch in diesem Falle seien $\bar{x}_1=132$, $\bar{x}_2=73$, $\hat{\sigma}_1=51$ und $\hat{\sigma}_2=19$. Bleiben wir zusätzlich bei den Gewichten der beiden Teilpopulationen $g_1=0{,}32$ und $g_2=0{,}68$ (die bei gleicher Aufteilung natürlich nicht aus den Stichprobendaten, sondern aufgrund externer Informationen geschätzt werden müssen), resultieren:

$$\bar{x}_{\text{gleich}} = \bar{x}_{\text{bel}} = 92$$

$$\hat{\sigma}_{\bar{x}(\text{gleich})} = \sqrt{\frac{k}{n} \cdot \sum_{j=1}^{k} g_j^2 \cdot \hat{\sigma}_j^2}$$

$$= \sqrt{\frac{2}{100} \cdot (0{,}32^2 \cdot 51^2 + 0{,}68^2 \cdot 19^2)}$$

$$= 2{,}94.$$

Der Standardfehler $\hat{\sigma}_{\bar{x}(\text{gleich})}$ ist also in diesem Beispiel kleiner als der Standardfehler $\hat{\sigma}_{\bar{x}(\text{prop})}$ für proportionale Stichprobenumfänge. Am (ganzzahlig gerundeten) Konfidenzintervall ändert sich dadurch jedoch nichts:

$$\bar{x}_{\text{gleich}} \pm 2{,}58 \cdot \hat{\sigma}_{\bar{x}(\text{gleich})} = 92 \pm 2{,}58 \cdot 2{,}94 \approx 92 \pm 8.$$

Sind sowohl die Gewichte g_1 und g_2 als auch die Streuungen σ_1 und σ_2 bereits vor der Stichprobenentnahme bekannt, gewährleisten die folgenden Stichprobenumfänge eine bestmögliche Schätzung von μ (optimale Stichprobenumfänge):

$$n_1 = \frac{g_1 \cdot \sigma_1}{\sum_{j=1}^{k} g_j \cdot \sigma_j} \cdot n = \frac{0{,}32 \cdot 51}{(0{,}32 \cdot 51 + 0{,}68 \cdot 19)} \cdot 100 \approx 56$$

$$n_2 = \frac{g_2 \cdot \sigma_2}{\sum_{j=1}^{k} g_j \cdot \sigma_j} \cdot n = \frac{0{,}68 \cdot 19}{(0{,}32 \cdot 51 + 0{,}68 \cdot 19)} \cdot 100 \approx 44.$$

(Bei diesen Berechnungen gingen wir davon aus, dass die Streuungsschätzungen $\hat{\sigma}_j$ den tatsächlichen Streuungen σ_j entsprechen.)

Die gleiche Aufteilung kommt damit der optimalen Aufteilung recht nahe. Unter Verwendung dieser Stichprobenumfänge resultiert das folgende Konfidenzintervall:

$$\bar{x}_{\text{opt}} = \sum_{j=1}^{k} g_j \cdot \bar{x}_j = 0{,}32 \cdot 132 + 0{,}68 \cdot 73 = 92$$

$$\sigma_{\bar{x}(\text{opt})} = \frac{1}{\sqrt{n}} \cdot \sum_{j=1}^{k} g_j \cdot \sigma_j = \frac{1}{\sqrt{100}} \cdot (0{,}32 \cdot 51 + 0{,}68 \cdot 19) = 2{,}92$$

$$\bar{x}_{(\text{opt})} \pm 2{,}58 \cdot \sigma_{\bar{x}(\text{opt})} = 92 \pm 2{,}58 \cdot 2{,}92 \approx 92 \pm 8.$$

Folgerichtig führt die Aufteilung der Gesamtstichprobe in Teilstichproben mit optimalen Umfängen in unserem Beispiel nur zu einer geringfügigen Verbesserung der Schätzgenauigkeit gegenüber einer Aufteilung in gleich große Stichprobenumfänge.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

probe. Hier entspricht der Mittelwert $\bar{X}_{prop}$ der Summe aller Messwerte, dividiert durch n.

Ersetzen wir n_j durch $n \cdot g_j$ in ► Gl. (7.21) für $\hat{\sigma}_{\bar{x}(bel)}$, resultiert

$$\hat{\sigma}_{\bar{x}(prop)} = \sqrt{\frac{1}{n}\sum_{j=1}^{k} g_j \cdot \hat{\sigma}_j^2}. \tag{7.25}$$

Dieser Standardfehler geht in ► Gl. (7.22) zur Bestimmung des Konfidenzintervalls von μ ein (■ Box 7.4).

Die proportionale Schichtung wird wegen ihrer rechnerisch einfachen Handhabung relativ häufig angewandt. Will man jedoch die Teilstichprobenmittelwerte $\bar{X}_j$ gleichzeitig zur Schätzung der Teilpopulationsparameter μ_j verwenden, ist zu beachten, dass diese Art der Schichtung bei unterschiedlich großen Teilpopulationen (und damit auch unterschiedlich großen Teilstichproben) zu Schätzungen mit unterschiedlichen Genauigkeiten führt.

7

! Wenn die prozentuale Verteilung der Schichtungsmerkmale in der Stichprobe mit der Verteilung in der Population identisch ist, sprechen wir von einer proportional geschichteten Stichprobe.

Optimale Aufteilung. Die drei bisher behandelten Modalitäten für die Berechnung des Standardfehlers $\hat{\sigma}_{\bar{x}}$ verdeutlichen, dass die Größe des Standardfehlers davon abhängt, wie die Gesamtstichprobe auf die einzelnen Schichten verteilt wird, d. h., wie groß die einzelnen Teilstichproben sind. Damit eröffnet sich die interessante Möglichkeit, allein durch die Art der Aufteilung der Gesamtstichprobe auf die einzelnen Schichten bzw. durch geeignete Wahl der Anzahl der aus den einzelnen Teilpopulationen zu entnehmenden Untersuchungsobjekte, den Standardfehler zu minimieren und damit die Schätzgenauigkeit zu maximieren.

Gesucht wird also eine Aufteilung des Gesamtstichprobenumfanges n in einzelne Teilstichproben n_j, die bei gegebenem σ_j und g_j-Werten $\sigma_{\bar{x}}$ minimieren. Dies ist ein Problem der Differenzialrechnung (Minimierung von $\sigma_{\bar{x}}$ in Abhängigkeit von n_j unter der Nebenbedingung $n=n_1+n_2+\ldots+n_k$), dessen Lösung z. B. bei Cochran (1972, Kap. 5.5) behandelt wird. Für eine Schicht j resultiert als optimaler Stichprobenumfang

$$n_j = \frac{N_j \cdot \sigma_j}{\sum_{j=1}^{k} N_j \cdot \sigma_j} \cdot n$$

bzw. unter Verwendung der Beziehung $N_j=g_j \cdot N$ (► Gl. 7.20):

$$n_j = \frac{g_j \cdot \sigma_j}{\sum_{j=1}^{k} g_j \cdot \sigma_j} \cdot n. \tag{7.26}$$

Für die Ermittlung optimaler Stichprobenumfänge müssen damit nicht nur die Gewichte g_j für die Schichten (bzw. die Populationsumfänge N_j) bekannt sein, sondern auch die Standardabweichungen in den einzelnen Teilpopulationen. Letztere sind vor Durchführung der Untersuchung in der Regel unbekannt. Man wird deshalb – ggf. unter Zuhilfenahme der in ■ Abb. 7.7 zusammengefassten Regeln – die Standardabweichung schätzen müssen und erst nach der Datenerhebung (wenn die $\hat{\sigma}_j$-Werte berechnet werden können) feststellen, wie stark die vorgenommene Stichprobeneinteilung von der optimalen abweicht.

Den Gesamtstichprobenmittelwert $\bar{X}$ ermitteln wir auch bei optimalen Stichprobenumfängen nach der Beziehung

$$\bar{X}_{opt} = \sum_{j=1}^{k} g_j \cdot \bar{X}_j.$$

Für den Standardfehler ergibt sich

$$\sigma_{\bar{x}(opt)} = \frac{1}{\sqrt{n}}\sum_{j=1}^{k} g_j \cdot \sigma_j. \tag{7.27}$$

Die Verwendung dieses Standardfehlers bei der Ermittlung des Konfidenzintervalles für μ veranschaulicht wiederum ■ Box 7.4.

Sind die Gesamtkosten für die Stichprobenerhebung vorkalkuliert, können sie bei der Ermittlung der optimalen Aufteilung der Gesamtstichprobe auf die einzelnen Schichten berücksichtigt werden. Dies führt jedoch nur dann zu Abweichungen von der hier behandelten optimalen Aufteilung, wenn die Erhebungskosten pro Untersuchungsobjekt in den verschiedenen Schichten unterschiedlich sind (was z. B. der Fall wäre, wenn bei einer regionalen Schichtung die Befragung von Perso-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

nen in verschiedenen Regionen unterschiedliche Reisekosten erfordert). Es könnte dann von Interesse sein, die Aufteilung so vorzunehmen, dass bei gegebenem Standardfehler (Schätzgenauigkeit) und festliegendem Gesamtstichprobenumfang n die Erhebungskosten minimiert werden (weitere Informationen z. B. bei Cochran, 1972, Kap. 5).

Vergleichende Bewertung. Geschichtete Stichproben erfordern gegenüber einfachen Zufallsstichproben einen höheren organisatorischen und rechnerischen Aufwand, der nur zu rechtfertigen ist, wenn sich durch die Berücksichtigung eines Schichtungsmerkmals die Schätzgenauigkeit deutlich verbessert. Generell gilt, dass sich eine Schichtung umso vorteilhafter auswirkt, je kleiner die Streuung in den Teilstichproben im Vergleich zur Streuung in der Gesamtstichprobe ist (homogene Teilstichproben). Eine Schichtung ist sinnlos, wenn die Teilstichproben genauso heterogen sind wie die Gesamtstichprobe.

Hat man weder Angaben über die Gewichte g_j der Teilstichproben noch über die Streuungen σ_j, wird man beide aus den Stichprobendaten schätzen müssen. Dies führt zu einer ex post stratifizierten Stichprobe mit proportionalen Schichtanteilen, deren Schätzgenauigkeit dann gegenüber einer reinen Zufallsstichprobe verbessert ist, wenn n_j/n die Gewichte $g_j = N_j/N$ und die Streuungen in den Teilstichproben die Streuungen in den Teilpopulationen akzeptabel schätzen, was nur bei großen, ex post stratifizierten Stichproben der Fall sein dürfte. Sind zwar die Gewichte g_j, aber nicht die Streuungen σ_j bekannt, und unterscheiden sich zudem die Gewichte nur unerheblich, führen geschichtete Stichproben mit gleich großen Teilstichproben zu einer guten Schätzgenauigkeit. Teilstichproben mit proportionalen Umfängen sind vorzuziehen, wenn die bekannten Gewichte sehr unterschiedlich sind. In diesem Falle ist vor allem bei schwach gewichteten Teilstichproben auf die Bedingung $g_j \cdot n_j \geq 10$ zu achten. Kennt man sowohl die Gewichte als auch die Streuungen der Teilpopulationen, führen optimale Stichprobenumfänge zu einer bestmöglichen Schätzgenauigkeit.

Stichprobenumfänge. Tab. 7.6 (▶ S. 422) zeigte, welche Stichprobenumfänge zu wählen sind, wenn ein Parameter μ mit einer vorgegebenen Genauigkeit (Konfidenzintervall) geschätzt werden soll. Diese Stichprobenumfänge lassen sich erheblich reduzieren, wenn es möglich ist, statt einer einfachen Zufallsstichprobe eine sinnvoll geschichtete Stichprobe zu ziehen. Gelingt es, homogene Schichten zu finden, reduziert sich der Standardfehler $\hat{\sigma}_{\bar{x}}$, d. h., kleinere geschichtete Stichproben erreichen die gleiche Schätzgenauigkeit wie größere Zufallsstichproben.

Die Berechnungsvorschriften für den Umfang geschichteter Stichproben erhält man einfach durch Auflösen der auf ▶ S. 425 ff. genannten Bestimmungsgleichungen des Standardfehlers $\hat{\sigma}_{\bar{x}}$ nach n. Wir errechnen für:

- gleiche Aufteilungen

$$n = \frac{k \cdot \sum_{j=1}^{k} g_j^2 \cdot \sigma_j^2}{\hat{\sigma}_{\bar{x}}^2}, \tag{7.28}$$

- proportionale Aufteilungen

$$n = \frac{\sum_{j=1}^{k} g_j \cdot \sigma_j^2}{\hat{\sigma}_{\bar{x}}^2}, \tag{7.29}$$

- optimale Aufteilungen

$$n = \frac{\left(\sum_{j=1}^{k} g_j \cdot \sigma_j\right)^2}{\hat{\sigma}_{\bar{x}}^2}; \tag{7.30}$$

mit

g_j = Gewicht der Schicht j (N_j/N),
σ_j = Standardabweichung des untersuchten Merkmals in der Schicht j und
$\hat{\sigma}_{\bar{x}}$ = geschätzter Standardfehler des Mittelwertes $\bar{x}$.

Der Gesamtumfang einer geschichteten Stichprobe mit beliebiger Schichtung ist nicht kalkulierbar. Dies verdeutlicht die Bestimmungsgleichung (7.21) für $\hat{\sigma}_{\bar{x}(\text{bel})}$, in der der Gesamtstichprobenumfang n nicht vorkommt.

Den drei Gleichungen ist zu entnehmen, dass für die Kalkulation des Umfanges einer geschichteten Stichprobe die Schichtgewichte g_j sowie die Standardabweichungen in den einzelnen Schichten σ_j bekannt sein müssen. Erneut wird man sich bei der Planung von Stichprobenerhebungen häufig mit Schätzungen dieser Kennwerte begnügen müssen.

Die Schätzgenauigkeit hängt zudem davon ab, welchen Wert man für $\hat{\sigma}_{\bar{x}}$ einsetzt. Die Wahl von $\hat{\sigma}_{\bar{x}}$

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

wird durch den Umstand erleichtert, dass die Zufallsvariable $\bar{X}$ bei genügend großen Stichproben normalverteilt ist (▶ S. 411 f.). Der Bereich $\pm 3 \cdot \hat{\sigma}_{\bar{x}}$ umschließt damit praktisch alle denkbaren Werte für $\bar{X}$. Wir legen deshalb einen sinnvoll erscheinenden Wertebereich (Range) für $\bar{X}$ fest und dividieren diesen durch 6. Das Resultat ist in der Regel ein brauchbarer Wert für $\hat{\sigma}_{\bar{x}}$. Einen genaueren Wert erhält man nach ▶ Gl. (7.17).

Beispiel: Die Handhabung der Bestimmungsgleichungen für Stichprobenumfänge sei im Folgenden an einem Beispiel demonstriert. Es interessiert die Frage, wieviel Geld 16- bis 18-jährige Lehrlinge monatlich im Durchschnitt für Genußmittel (Zigaretten, Süßigkeiten, alkoholische Getränke etc.) ausgeben. Der Parameter μ soll mit einer Genauigkeit von $\hat{\sigma}_{\bar{x}}$=€ 2,– geschätzt werden, d. h., das 95%ige Konfidenzintervall lautet $\bar{X}$±1,96×€ 2,–= $\bar{X}$±€ 3,92 bzw. ≈$\bar{X}$±€ 4,–. Man vermutet, dass die Ausgaben vom Alter der Lehrlinge abhängen und plant deshalb eine nach dem Alter (16-, 17- und 18-jährige Lehrlinge) geschichtete Stichprobe. Altersstatistiken von Lehrlingen legen die Annahme folgender Schichtgewichte nahe:
16-jährige: g_1=0,40,
17-jährige: g_2=0,35,
18-jährige: g_3=0,25.

Um die Streuungen der Ausgaben in den einzelnen Altersklassen schätzen zu können, befragt man einige 16-, 17- und 18-jährige Lehrlinge, wieviel Geld gleichaltrige Lehrlinge höchstens bzw. wenigstens ausgeben. Diese Angaben führen unter Verwendung der in ◘ Abb. 7.7 genannten Schätzformeln zu folgenden Werten:
16-jährige: $\hat{\sigma}_1$=20,
17-jährige: $\hat{\sigma}_2$=24,
18-jährige: $\hat{\sigma}_3$=28,
Gesamtstreuung: $\hat{\sigma}$=36.

Die folgenden Berechnungen zeigen, welche Stichprobenumfänge in Abhängigkeit von der Art der Schichtung erforderlich sind, um den Parameter μ mit der vorgegebenen Genauigkeit schätzen zu können. Zum Vergleich wird zunächst der Stichprobenumfang für eine einfache, nichtgeschichtete Zufallsstichprobe kalkuliert (▶ S. 421 ff.).

- Ungeschichtete Zufallsstichprobe (mit $\hat{\sigma}_{\bar{x}}$=2):

$$n = \frac{\hat{\sigma}^2}{\hat{\sigma}_{\bar{x}}^2} = \frac{36^2}{2^2} = 324.$$

Man erhält diese Gleichung, wenn man in ▶ Gl. (7.10) $z^2 = (\bar{x}-\mu)^2/\sigma^2$ und $e^2 = (\bar{x}-\mu)^2$ setzt.

- Gleichmäßig geschichtete Zufallsstichprobe:

$$n = \frac{k \cdot \sum_{j=1}^{k} g_j^2 \cdot \hat{\sigma}_j^2}{\hat{\sigma}_{\bar{X}}^2}$$
$$= \frac{3 \cdot (0{,}40^2 \cdot 20^2 + 0{,}35^2 \cdot 24^2 + 0{,}25^2 \cdot 28^2)}{2^2}$$
$$= 137{,}7 \approx 138.$$

- Proportional geschichtete Zufallsstichprobe:

$$n = \frac{\sum_{j=1}^{k} g_j \cdot \hat{\sigma}_j^2}{\hat{\sigma}_{\bar{X}}^2}$$
$$= \frac{0{,}40 \cdot 20^2 + 0{,}35 \cdot 24^2 + 0{,}25 \cdot 28^2}{2^2}$$
$$= 139{,}4 \approx 139.$$

- Optimal geschichtete Zufallsstichprobe:

$$n = \frac{\left(\sum_{j=1}^{k} g_j \cdot \hat{\sigma}_j\right)^2}{\hat{\sigma}_{\bar{X}}^2}$$
$$= \frac{(0{,}40 \cdot 20 + 0{,}35 \cdot 24 + 0{,}25 \cdot 28)^2}{2^2}$$
$$= 136{,}9 \approx 137.$$

Wie zu erwarten, erfordert die gewünschte Schätzgenauigkeit eine sehr viel größere Stichprobe, wenn keine Schichtung vorgenommen wird. Berücksichtigt man das Schichtungsmerkmal Alter, ist es (in diesem Beispiel) für die Schätzgenauigkeit praktisch unerheblich, ob die Stichprobe gleichmäßig, proportional oder optimal aufgeteilt wird. Die gegenüber der ungeschichteten Stichprobe erheblich reduzierten Stichprobenumfänge unterscheiden sich nur unbedeutend. Auch wenn die eingangs genannten Gewichte und Standardabweichungen der Schichten nur ungefähr richtig

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

sind, dürfte ein Stichprobenumfang von n=150 bei allen Schichtungsarten eine ausreichende Schätzgenauigkeit gewährleisten. Wird dieser optimal aufgeteilt, wären nach ▶ Gl. (7.26)

$$n_1 = \frac{0{,}40 \cdot 20}{23{,}4} \cdot 150 = 51{,}3 \approx 51 \quad 16-\text{jährige},$$

$$n_2 = \frac{0{,}35 \cdot 24}{23{,}4} \cdot 150 = 53{,}8 \approx 54 \quad 17-\text{jährige}$$

und

$$n_3 = \frac{0{,}25 \cdot 28}{23{,}4} \cdot 150 = 44{,}9 \approx 45 \quad 18-\text{jährige}$$

Lehrlinge zu befragen.

Schätzung von Populationsanteilen

Im Unterschied zum Mittelwertparameter μ, der mit geschichteten Stichproben in der Regel erheblich genauer geschätzt werden kann als mit einfachen Zufallsstichproben, führt die Berücksichtigung eines Schichtungsmerkmals bei der Schätzung eines Anteilsparameters π meistens nur zu einer unwesentlichen Genauigkeitserhöhung. Dennoch soll dieser Weg einer Parameterschätzung kurz beschrieben werden.

Wie bereits erwähnt, stellt die relative Häufigkeit bzw. der Anteil p der in einer Zufallsstichprobe angetroffenen Untersuchungsobjekte mit dem untersuchten Merkmal A eine erwartungstreue Schätzung des Populationsparameters π_A dar (▶ S. 407):

$$p_A = \frac{n_A}{n} = \text{Anteil der Untersuchungsteilnehmer mit dem Merkmal A.}$$

Setzt man eine Stichprobe des Umfanges n aus k Teilstichproben zusammen, die sich bezüglich eines Schichtungsmerkmals unterscheiden, ist der Anteil p_j in jeder Teilstichprobe j eine erwartungstreue Schätzung von π_j, wobei

$$p_j = \frac{n_{A(j)}}{n_j}$$

$$\left(\sum_{j=1}^{k} n_{A(j)} = n_A \text{ und} \sum_{j=1}^{k} n_j = n\right).$$

Bekannte Schichtgewichte g_j ($g_j=N_j/N$; ▶ S. 426) vorausgesetzt, führt die folgende Gleichung zu einer die Schichten zusammenfassenden, erwartungstreuen Schätzung von π:

$$p = \sum_{j=1}^{k} g_j \cdot p_j. \tag{7.31}$$

Diese Gleichung gilt unabhängig davon, wie viele Untersuchungseinheiten den einzelnen Schichten entnommen wurden.

Bei mehrfacher Ziehung geschichteter Stichproben streut dieser p-Wert mit

$$\hat{\sigma}_p = \sqrt{\sum_{j=1}^{k}\left(g_j^2 \cdot \frac{p_j \cdot (1-p_j)}{n_j}\right)}, \tag{7.31a}$$

wobei auch diese Gleichung für beliebige Aufteilung gilt.

Das folgende **Beispiel** zeigt, wie man mit diesen Gleichungen zu einem Konfidenzintervall für den Parameter π gelangt. Bei einer Befragung von n=1000 zufällig ausgewählten, wahlberechtigten Personen einer Großstadt gaben 350 (also 35%) an, bei der nächsten Wahl Partei A wählen zu wollen. Ohne Berücksichtigung eines Schichtungsmerkmals resultiert für den Standardfehler des Schätzwertes p=0,35 (▶ S. 418):

$$\hat{\sigma}_p = \sqrt{\frac{p \cdot (1-p)}{n}} = \sqrt{\frac{0{,}35 \cdot 0{,}65}{1000}}$$
$$= \sqrt{0{,}000228} = 0{,}0151.$$

Für das 99%ige Konfidenzintervall (mit z=2,58) ergibt sich also nach ▶ Gl. (7.8)

$$0{,}35 \pm 2{,}58 \cdot 0{,}0151 = 0{,}35 \pm 0{,}0390.$$

Das Konfidenzintervall lautet 35%±3,9%.

Nun sei jedoch bekannt, dass die Attraktivität der Partei A vom Bildungsniveau der Wähler abhängt. Man stellt fest, dass 37% der befragten Personen eine weiterführende Schule besuchten; 32% haben einen Hauptschulabschluss mit Lehre und 31% einen Hauptschulabschluss ohne Lehre. Diese Zahlen werden als Schätzungen der Schichtgewichte bzw. Populationsanteile verwendet. Die Anzahl der Personen innerhalb

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

dieser Schichten, die Partei A zu wählen beabsichtigen, lauten:

weiterführende Schule: $n_{A(1)}=185$; $p_1=0{,}500$; $n_1=370$
Hauptschule mit Lehre: $n_{A(2)}=96$; $p_2=0{,}300$; $n_2=320$
Hauptschule ohne Lehre: $n_{A(3)}=69$; $p_3=0{,}223$; $n_3=310$

Als Schätzwert für p ergibt sich damit

$$\begin{aligned}p &= \sum_{j=1}^{k} g_j \cdot p_j \\ &= 0{,}37 \cdot 0{,}500 + 0{,}32 \cdot 0{,}300 + 0{,}31 \cdot 0{,}223 \\ &= 0{,}35.\end{aligned}$$

(Dieser p-Wert ist natürlich mit dem p-Wert der ungeschichteten Stichprobe identisch, da die Schichtgewichte g_j den relativen Häufigkeiten entsprechen. Die Stichprobe ist **selbstgewichtend**; ▶ S. 427).

Für $\hat{\sigma}_p$ erhält man nach ▶ Gl. (7.31a):

$$\begin{aligned}\hat{\sigma}_p &= \sqrt{\sum_{j=1}^{k} g_j^2 \cdot \frac{p_j \cdot (1-p_j)}{n_j}} \\ &= \sqrt{0{,}37^2 \cdot \frac{0{,}5 \cdot 0{,}5}{370} + 0{,}32^2 \cdot \frac{0{,}3 \cdot 0{,}7}{320} + 0{,}31^2 \cdot \frac{0{,}223 \cdot 0{,}777}{310}} \\ &= 0{,}0146.\end{aligned}$$

Damit beträgt das Konfidenzintervall:

$$0{,}35 \pm 2{,}58 \cdot 0{,}0146 = 0{,}35 \pm 0{,}0377.$$

Das Konfidenzintervall hat sich also nur geringfügig (um 0,13%) verkleinert, obwohl ein Schichtungsmerkmal berücksichtigt wurde, das offensichtlich eng mit dem untersuchten Merkmal zusammenhängt.

Die Vorhersage wird auch nicht viel präziser, wenn die Gesamtstichprobe nicht – wie im Beispiel – proportional, sondern nach folgender Gleichung **optimal** aufgeteilt wird:

$$n_j = n \cdot \frac{g_j \cdot \sqrt{p_j \cdot (1-p_j)}}{\sum_{j=1}^{k} g_j \cdot \sqrt{p_j \cdot (1-p_j)}}. \qquad (7.32)$$

Im Beispiel ergeben sich unter Verwendung der empirischen p_j-Werte:

$$n_1 = 1000 \cdot \frac{0{,}37 \cdot \sqrt{0{,}5 \cdot 0{,}5}}{0{,}4607} = 401{,}6 \approx 402,$$

$$n_2 = 1000 \cdot \frac{0{,}32 \cdot \sqrt{0{,}3 \cdot 0{,}7}}{0{,}4607} = 318{,}3 \approx 318,$$

$$n_3 = 1000 \cdot \frac{0{,}31 \cdot \sqrt{0{,}223 \cdot 0{,}777}}{0{,}4607} = 280{,}1 \approx 280$$

Mit diesen Stichprobenumfängen (und unter Beibehaltung der übrigen Werte) resultiert für p ebenfalls eine Standardabweichung von $\hat{\sigma}_p = 0{,}0146$, d. h., die optimale Aufteilung führt in diesem Beispiel (zumindest für die ersten 4 Nachkommastellen) zu keiner verbesserten Schätzung.

Stichprobenumfänge. Auch für Untersuchungen von Populationsanteilen ist es ratsam, den erforderlichen Umfang der geschichteten Stichprobe vor Untersuchungsbeginn zu kalkulieren. Wiederum benötigen wir hierfür Angaben über die Schichtgewichte, über die mutmaßlichen p-Werte innerhalb der Schichten sowie über eine maximal tolerierbare Fehlergröße (Konfidenzintervall). Die Stichprobenumfänge ergeben sich nach folgenden Gleichungen (vgl. Schwarz, 1975):

- Gleiche Aufteilungen

$$n = \frac{k \cdot \sum_{j=1}^{k} g_j^2 \cdot p_j \cdot (1-p_j)}{\hat{\sigma}_p^2}, \qquad (7.33)$$

- Proportionale Aufteilungen

$$n = \frac{\sum_{j=1}^{k} g_j \cdot p_j \cdot (1-p_j)}{\hat{\sigma}_p^2}, \qquad (7.34)$$

- Optimale Aufteilungen

$$n = \frac{\left(\sum_{j=1}^{k} g_j \cdot \sqrt{p_j \cdot (1-p_j)}\right)^2}{\hat{\sigma}_p^2}. \qquad (7.35)$$

Wir wollen diese Gleichungen am oben erwähnten Beispiel verdeutlichen. Gesucht wird derjenige Stichprobenumfang, der mit 99%iger Wahrscheinlichkeit eine maximale Fehlertoleranz von 1% gewährleistet, d. h.,

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

das 99%ige Konfidenzintervall soll p±1% betragen. $\hat{\sigma}_p$ errechnet sich dann in folgender Weise:

$$2{,}58 \cdot \hat{\sigma}_p = 0{,}01$$

$$\text{bzw.}\ \hat{\sigma}_p = 0{,}00388.$$

Als p_j- und g_j-Werte verwenden wir die bereits bekannten Angaben. Es resultieren die folgenden Stichprobenumfänge:

- Zufallsstichprobe ohne Schichtung:

$$n = \frac{p \cdot (1-p)}{\hat{\sigma}_p^2} = \frac{0{,}35 \cdot 0{,}65}{0{,}00388^2} = 15111{,}9 \approx 15112.$$

- Geschichtete Stichprobe mit gleichmäßiger Aufteilung:

$$n = \frac{k \cdot \sum_{j=1}^{k} g_j^2 \cdot p_j \cdot (1-p_j)}{\hat{\sigma}_p^2}$$
$$= \frac{3 \cdot (0{,}37^2 \cdot 0{,}5 \cdot 0{,}5 + 0{,}32^2 \cdot 0{,}3 \cdot 0{,}7 + 0{,}31^2 \cdot 0{,}223 \cdot 0{,}777}{0{,}00388^2}$$
$$= 14423{,}8 \approx 14424.$$

- Geschichtete Stichprobe mit proportionaler Aufteilung:

$$n = \frac{\sum_{j=1}^{k} g_j \cdot p_j \cdot (1-p_j)}{\hat{\sigma}_p^2}$$
$$= \frac{0{,}37 \cdot 0{,}5 \cdot 0{,}5 + 0{,}32 \cdot 0{,}3 \cdot 0{,}7 + 0{,}31 \cdot 0{,}223 \cdot 0{,}777}{0{,}00388^2}$$
$$= 14176{,}2 \approx 14176.$$

- Geschichtete Stichprobe mit optimaler Aufteilung:

$$n = \frac{\left(\sum_{j=1}^{k} g_j \cdot \sqrt{p_j \cdot (1-p_j)}\right)^2}{\hat{\sigma}_p^2}$$
$$= \frac{\left(0{,}37 \cdot \sqrt{0{,}5 \cdot 0{,}5} + 0{,}32 \cdot \sqrt{0{,}3 \cdot 0{,}7} + 0{,}31 \cdot \sqrt{0{,}223 \cdot 0{,}777}\right)^2}{0{,}00388^2}$$
$$= 14097{,}4 \approx 14097.$$

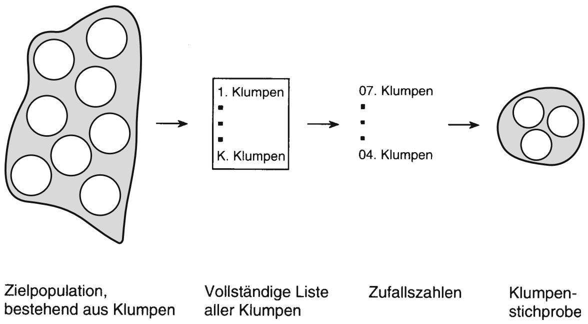


Abb. 7.9. Ziehung einer Klumpenstichprobe

7.2.2 Klumpenstichprobe

Parameterschätzungen – so zeigte der vergangene Abschnitt – werden genauer, wenn die Gesamtstichprobe keine einfache Zufallsstichprobe ist, sondern sich aus mehreren Teilstichproben, so genannten Schichten, zusammensetzt, welche die Ausprägungen eines mit dem untersuchten Merkmal möglichst hoch korrelierenden Schichtungsmerkmals repräsentieren. Die Gesamtstichprobe besteht aus mehreren Teilstichproben, die zufällig aus den durch das Schichtungsmerkmal definierten Teilpopulationen entnommen sind.

Die Klumpenstichprobe (**»Cluster Sample«**) erfordert, dass die Gesamtpopulation aus vielen Teilpopulationen oder Gruppen von Untersuchungsobjekten besteht, von denen eine zufällige Auswahl von Gruppen vollständig erhoben wird (Abb. 7.9). Bei einer Untersuchung von Schülern würde man beispielsweise die Stichprobe aus allen Schülern mehrerer zufällig ausgewählter Schulklassen zusammensetzen und bei einer Untersuchung von Betriebsangehörigen einzelne Betriebe oder Abteilungen vollständig erheben. Untersucht man Krankenhauspatienten, könnte man z. B. alle Patienten einiger zufällig ausgewählter Krankenhäuser zu einer Stichprobe zusammenfassen. Wann immer eine Population aus vielen Gruppen oder natürlich zusammenhängenden Teilkollektiven besteht (für die in der deutschsprachigen Literatur die Bezeichnung »Klumpen« üblich ist), bietet sich die Ziehung einer Klumpenstichprobe an. Wir werden später erläutern, unter welchen Umständen diese Stichprobentechnik zu genaueren Parameterschätzungen führt als eine einfache Zufallsauswahl.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Die Klumpenstichprobe erfordert weniger organisatorischen Aufwand als die einfache Zufallsstichprobe. Während eine einfache Zufallsstichprobe strenggenommen voraussetzt, dass alle Untersuchungsobjekte der Population einzeln erfasst sind, benötigt die Klumpenstichprobe lediglich eine vollständige Liste aller in der Population enthaltenen Klumpen. Diese ist in der Regel einfacher anzufertigen als eine Zusammenstellung aller einzelnen Untersuchungsobjekte.

Sämtliche Untersuchungsobjekte, die sich in den zufällig ausgewählten Klumpen befinden, bilden die Klumpenstichprobe. Der Auswahlvorgang bezieht sich hier nicht, wie bei der einfachen Zufallsstichprobe, auf die einzelnen Untersuchungsobjekte, sondern auf die Klumpen, wobei sämtliche ausgewählten Klumpen vollständig, d. h., mit allen Untersuchungsobjekten erfasst werden ($n_j=N_j$). Die Auswahlwahrscheinlichkeit ist für jeden Klumpen gleich. Es ist darauf zu achten, dass jedes Untersuchungsobjekt nur einem Klumpen angehört, dass sich also die Klumpen nicht wechselseitig überschneiden.

Im Folgenden wird gezeigt, wie Populationsmittelwerte (μ) und Populationsanteile (π) mit Klumpenstichproben geschätzt werden können.

! Man zieht eine Klumpenstichprobe, indem man aus einer in natürliche Gruppen (Klumpen) gegliederten Population nach dem Zufallsprinzip eine Anzahl von Klumpen auswählt und diese Klumpen dann vollständig untersucht.

Schätzung von Populationsmittelwerten

Eine Population möge aus K Klumpen bestehen, von denen k Klumpen zufällig ausgewählt werden. Im Unterschied zur Zufallsstichprobe, bei der die geforderte Relation von Stichprobenumfang (n) zu Populationsumfang (N) ($n/N<0{,}05$) in der Regel als gegeben angesehen wurde, ist der Auswahlsatz $f=k/K$ bei der Untersuchung einer Klumpenstichprobe häufig nicht zu vernachlässigen. In vielen Untersuchungen besteht die zu beschreibende Population nur aus einer begrenzten Anzahl von Klumpen (z. B. Großstädte, Universitäten, Wohnareale in einer Stadt, Häuserblocks in einem Wohnareal, Wohnungen in einem Häuserblock), sodass die Berücksichtigung des Auswahlsatzes f die Präzision der Parameterschätzung erheblich verbessern kann.

Der **Auswahlsatz** f stellt die Wahrscheinlichkeit dar, mit der ein Klumpen der Population in die Stichprobe aufgenommen wird. Da jedes Untersuchungsobjekt nur einem Klumpen angehören darf, ist dies gleichzeitig die Auswahlwahrscheinlichkeit eines beliebigen Untersuchungsobjektes.

Mittelwert. Bezeichnen wir den Messwert des i-ten Untersuchungsobjektes ($i=1,2...N_j$) im j-ten Klumpen ($j=1,2\ldots k$) mit x_{ij}, ist der Mittelwert eines Klumpens j ($\bar{x}_j$) durch

$$\bar{x}_j = \frac{\sum_{i=1}^{N_j} x_{ij}}{N_j}$$

definiert. (Da die ausgewählten Klumpen vollständig erhoben werden, verwenden wir N_j für den Umfang des Klumpens j.) Hierbei gehen wir von der realistischen Annahme aus, dass die einzelnen Klumpen unterschiedliche Umfänge aufweisen.

Den Gesamtmittelwert der Untersuchungsobjekte aller ausgewählten Klumpen bezeichnen wir mit $\bar{\bar{x}}$. Er ergibt sich zu

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j} x_{ij}}{\sum_{j=1}^{k} N_j} \qquad (7.36)$$

An dieser Stelle zeigt sich bereits eine Besonderheit von Klumpenstichproben: Der Mittelwert $\bar{\bar{x}}$ ist nur dann eine erwartungstreue Schätzung von μ, wenn alle Klumpen den gleichen Umfang aufweisen und/oder die einzelnen Klumpen Zufallsstichproben der Gesamtpopulation darstellen. Dies trifft natürlich nicht zu, denn zum einen wurde jeder zufällig ausgewählte Klumpen unbeschadet seiner Größe vollständig erhoben, und zum anderen muss man davon ausgehen, dass natürlich zusammenhängende Untersuchungsobjekte (Schulklassen, Abteilungen etc.) gegenüber der Gesamtpopulation klumpenspezifische Besonderheiten aufweisen. Werden nur wenige Klumpen mit sehr unterschiedlichen Umfängen untersucht, können diese zu erheblichen Fehlschätzungen von μ führen (**Klumpeneffekt**; im einzelnen vgl. hierzu Böltken, 1976, S. 305 ff., oder genauer Kish, 1965, Kap. 6). Klumpenstichproben eignen sich deshalb nur dann für die Beschreibung von Populationen, wenn man annehmen kann, dass alle Klumpen

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Box 7.5

Wie umfangreich sind Diplomarbeiten? III: Die Klumpenstichprobe

Nach der Schätzung der durchschnittlichen Seitenzahl von Diplomarbeiten im Fach Psychologie aufgrund einer Zufallsstichprobe (Box 7.3) und aufgrund einer geschichteten Stichprobe (Box 7.4) wird die gleiche Fragestellung nun aufgegriffen, um den Einsatz einer Klumpenstichprobe zu veranschaulichen. Zunächst ist die Frage zu erörtern, in welche Klumpen die Gesamtpopulation der Diplomarbeiten aufgeteilt werden soll. Hier bieten sich z. B. die an den einzelnen psychologischen Instituten pro Semester abgeschlossenen Arbeiten an, die von einer Hochschullehrerin bzw. einem Hochschullehrer (HSL) in einem Jahr betreuten Arbeiten oder eine Gruppierung aller Arbeiten nach verwandten Themen.

Da die Aufstellung aller HSL relativ wenig Mühe bereitet, entschließt man sich für die zweite Art der Klumpenbildung. Aus der Liste aller HSL werden 15 zufällig ausgewählt und um Angaben über die Seitenzahlen der von ihnen im vergangenen Jahr betreuten Diplomarbeiten gebeten. Das Jahr, über das ein HSL zu berichten hat, wird ebenfalls aus den vergangenen 10 Jahren, die der Populationsdefinition zugrunde liegen (vgl. Box 7.3), pro HSL zufällig ausgewählt. Diese Erhebungen führen zu den unten folgenden statistischen Daten. (Es wurde bewusst darauf geachtet, dass die statistischen Angaben für die Klumpenstichprobe mit den entsprechenden Daten für die Zufallsstichprobe in Box 7.3 und für die geschichtete Stichprobe in Box 7.4 weitgehend übereinstimmen.)

Anzahl aller HSL: K=450

Anzahl der ausgewählten HSL: k=15

$$n = \sum_{j=1}^{k} N_j = 100;\ \sum_{j=1}^{k}\sum_{i=1}^{N_j} x_{ij} = 9200;\ \bar{\bar{x}} = 92$$

Wie die folgende Berechnung zeigt, sind die Unterschiede zwischen den Klumpenumfängen N_j nach ► Gl. (7.37) tolerierbar.

Nr. d. HSL (j)	Anzahl der Arbeiten (N_j)	Summe der Seitenzahlen	Durchschnittliche Seitenzahl $\bar{x}_j$
1	8	720	90
2	2	210	105
3	10	950	95
4	9	837	93
5	7	658	94
6	9	819	91
7	6	552	92
8	1	124	124
9	7	616	88
10	11	946	86
11	5	455	91
12	9	801	89
13	3	291	97
14	6	570	95
15	7	651	93

$$V = \frac{\hat{\sigma}_{\bar{N}}}{\bar{N}} = \frac{\sqrt{\sum_{j=1}^{k}(N_j - \bar{N})^2}}{\sqrt{k \cdot (k-1)} \cdot \bar{N}} = \frac{\sqrt{119{,}33}}{\sqrt{15 \cdot 14} \cdot 6{,}67} = 0{,}113,$$

wobei

$$\bar{N} = \frac{\sum_{j=1}^{k} N_j}{k} = \frac{100}{15} = 6{,}67.$$

Der Wert ist kleiner als 0,20, d. h., die gezogene Klumpenstichprobe ist nach Kish (1965) für die Schätzung des Parameters μ akzeptabel.

Für den Standardfehler $\hat{\sigma}_{\bar{\bar{x}}}$ resultiert:

$$\hat{\sigma}_{\bar{\bar{x}}} = \sqrt{\frac{1-f}{n^2} \cdot \frac{k}{k-1} \cdot \sum_{j=1}^{k}\left(\sum_{i=1}^{N_j} x_{ij} - \bar{\bar{x}} \cdot N_j\right)^2}$$

$$= \sqrt{\frac{1 - \frac{15}{450}}{100^2} \cdot \frac{15}{14} \cdot 9706} = 1{,}003.$$

▼

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Der Korrekturfaktor 1–f hat in diesem Beispiel den Wert $1-\frac{15}{450}=0{,}967$ und verkleinert damit den Standardfehler nur unwesentlich.

Aus Tab. F3 ist zu entnehmen, dass die Werte t=±2,977 von der t-Verteilung mit 14 Freiheitsgraden an den Extremen jeweils 0,005% der Fläche abschneiden. Für das 99%ige Konfidenzintervall resultiert damit nach ▶ Gl. (7.7a):

$$92 \pm 2{,}977 \cdot 1{,}003 = 92 \pm 2{,}99.$$

Diese Klumpenstichprobe schätzt damit die durchschnittliche Seitenzahl von Diplomarbeiten erheblich genauer als die Zufallsstichprobe bzw. die geschichtete Stichprobe der ◘ Boxen 7.3 und 7.4.

die Gesamtpopulation annähernd gleich gut repräsentieren. Dies wiederum bedeutet, dass die Klumpen – im Unterschied zu den Schichten einer geschichteten Stichprobe – untereinander sehr ähnlich, die einzelnen Klumpen in sich aber heterogen sind. Je unterschiedlicher die Untersuchungsobjekte eines jeden Klumpens in bezug auf das untersuchte Merkmal sind, desto genauer schätzt die Klumpenstichprobe den unbekannten Parameter. Diese Aussage betrifft nicht nur das Kriterium der Erwartungstreue, sondern – wie weiter unten gezeigt wird – auch das Konfidenzintervall.

Für die Praxis schätzt der Stichprobenkennwert $\bar{\bar{x}}$ nach Cochran (1972, Kap. 6.5) den Parameter μ hinreichend genau, wenn der Variationskoeffizient V der einzelnen Klumpenumfänge den Wert 0,1 nicht überschreitet (nach Kish, 1965, Kap. 6.3, sind auch Werte V<0,2 noch tolerierbar). Dieser Variationskoeffizient relativiert den Standardfehler des durchschnittlichen Klumpenumfangs am Durchschnitt aller Klumpenumfänge:

$$V = \frac{\hat{\sigma}_{\bar{N}}}{\bar{N}} \leq 0{,}2. \tag{7.37}$$

Diese Überprüfung wird in ◘ Box 7.5 numerisch verdeutlicht.

Standardfehler. Durch die zufällige Auswahl von Klumpen ist nicht nur die Summe aller Messwerte (Zähler in ▶ Gl. 7.36), sondern auch der Stichprobenumfang (Nenner in ▶ Gl. 7.36) eine Zufallsvariable, was bei der Ermittlung des Standardfehlers von $\bar{\bar{x}}$ zu berücksichtigen ist. Für V≤0,2 stellt die folgende Gleichung eine brauchbare Approximation des Standardfehlers $\hat{\sigma}_{\bar{x}}$ dar.

$$\hat{\sigma}_{\bar{\bar{x}}} = \sqrt{\frac{1-f}{n^2} \cdot \frac{k}{k-1} \cdot \sum_{j=1}^{k} \left(\sum_{i=1}^{N_j} x_{ij} - \bar{\bar{x}} \cdot N_j \right)^2} \tag{7.38}$$

$$\text{mit } n = \sum_{j=1}^{k} N_j$$

(zur Herleitung dieser Gleichung s. Kish, 1965, Kap. 6.3). Diese Gleichung enthält implizit die Annahme, dass $\bar{N}$, der Durchschnitt **aller** Klumpenumfänge, durch $\sum_j N_j/k$ hinreichend genau geschätzt wird, was um so eher zutrifft, je größer k ist (vgl. Cochran, 1972, Gl. 11.12). Sie verdeutlicht ferner, dass die Unterschiedlichkeit der Werte innerhalb der Klumpen den Standardfehler zumindest direkt nicht beeinflusst. Dadurch, dass alle ausgewählten Klumpen vollständig erhoben werden, sind die Klumpenmittelwerte $\bar{x}_j$ frei von Stichprobenfehlern, sodass der Standardfehler ausschließlich auf der Unterschiedlichkeit zwischen den Klumpen basiert.

Die Unterschiedlichkeit der Klumpen wird in ▶ Gl. (7.38) durch den Klammerausdruck erfasst. Er vergleicht die Summe der Messwerte pro Klumpen mit derjenigen Summe, die zu erwarten wäre, wenn sich die Klumpen nicht (bzw. nur in ihren Umfängen) unterscheiden würden.

Sind die Klumpensummen nur wenig voneinander verschieden, resultiert ein kleiner Standardfehler.

Die Abweichung eines Messwertes x_{ij} vom Gesamtmittel $\bar{\bar{x}}$ lässt sich in die Abweichung des Messwertes x_{ij} vom Klumpenmittel $\bar{x}_j$ und die Abweichung des Klumpenmittels $\bar{x}_j$ vom Gesamtmittel $\bar{\bar{x}}$ zerlegen:

$$(x_{ij} - \bar{\bar{x}}) = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{\bar{x}}).$$

Sind nun bei gegebenen Abweichungen $(x_{ij} - \bar{x}_j)$ die Abweichungen der Klumpenmittel $\bar{x}_j$ vom Gesamtmittel

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

$\bar{\bar{x}}$ klein (diese Abweichungen entsprechen dem Klammerausdruck in ▶ Gl. 7.38), müssen die Abweichungen $(x_{ij} - \bar{x}_j)$ zwangsläufig groß sein. Hier zeigt sich erneut, wann die Ziehung einer Klumpenstichprobe empfehlenswert ist: Die Unterschiedlichkeit zwischen den Klumpen sollte klein, die Unterschiedlichkeit innerhalb der Klumpen jedoch groß sein.

Der Ausdruck 1–f korrigiert (verkleinert) den Standardfehler in Abhängigkeit von der Größe des Auswahlsatzes f=k/K. Für praktische Zwecke ist er wirkungslos, wenn f<0,05.

! Bei der Klumpenstichprobe sollte jeder einzelne Klumpen die Population annähernd gleich gut repräsentieren, d. h., die Klumpen sollten in sich heterogen, aber untereinander möglichst ähnlich sein. Demgegenüber sind bei einer gut geschichteten Stichprobe die einzelnen Schichten in sich homogen, aber untereinander sehr unterschiedlich.

Konfidenzintervall. Unter Verwendung von ▶ Gl. (7.38) lässt sich das Konfidenzintervall für μ in üblicher Weise berechnen (▶ Gl. 7.7a oder ◘ Box 7.5). Man muss jedoch beachten, dass die Verteilung von $\bar{\bar{x}}$ (normal- oder t-verteilt) nicht vom Stichprobenumfang n, sondern von der Anzahl der Klumpen k abhängt. Ist k≤30, verwendet man für die Bestimmung des Konfidenzintervalls die t-Verteilung mit k–1 Freiheitsgraden unter der Voraussetzung normalverteilter Klumpenmittelwerte. Wegen des zentralen Grenzwerttheorems (▶ S. 411 f.), das bei ungleich großen Stichproben allerdings nur bedingt gültig ist, dürfte diese Voraussetzung auf die Verteilung der Mittelwerte eher zutreffen als auf die Verteilung des untersuchten Merkmals.

◘ Box 7.5 zeigt den Rechengang zur Bestimmung eines Konfidenzintervalls auf der Basis einer Klumpenstichprobe. Um diese Stichprobentechnik mit den bisher behandelten Stichprobentechniken besser vergleichen zu können, wird hierfür erneut das Beispiel der ◘ Boxen 7.3 und 7.4 verwendet.

Ist eine Klumpenstichprobe zur Schätzung von μ unbrauchbar, weil die Klumpenumfänge nach ▶ Gl. (7.37) zu heterogen sind, muss die Anzahl der Klumpen erhöht werden. Wie man sich leicht überzeugen kann, verringert sich dadurch der Variationskoeffizient V.

Wenn eine Population nur aus sehr großen Klumpen besteht, ist es häufig zu aufwändig, eine Klumpenauswahl vollständig zu erheben. In diesem Fall wird aus den zufällig ausgewählten Klumpen jeweils nur eine begrenzte Auswahl zufällig ausgewählter Untersuchungsobjekte untersucht. Wir werden hierüber ausführlich in ▶ Abschn. 7.2.3 (mehrstufige Stichproben) berichten.

Schätzung von Populationsanteilen

Für die Schätzung von Populationsanteilen können wir auf die Überlegungen des letzten Abschnittes (Schätzung von Populationsmittelwerten) zurückgreifen. Man behandelt die in den einzelnen Klumpen registrierten Merkmalsanteile wie eine stetige Variable (die natürlich nur Werte zwischen 0 und 1 annimmt) und ersetzt den Durchschnittswert $\bar{x}_j$ einfach durch p_j. Die üblicherweise für Anteilsschätzungen einschlägige Binomialverteilung ist hier nicht zu verwenden. Cochran (1972, Kap. 3.12) macht anhand einiger Beispiele auf häufig in diesem Zusammenhang begangene Fehler aufmerksam.

Die für Anteilsschätzungen modifizierten Gleichungen lauten damit:

$$p_j = \frac{N_{A(j)}}{N_j}.$$

$N_{A(j)}$ ist hierbei die Anzahl der Untersuchungsobjekte mit dem Merkmal A im Klumpen j. Der Wert p_j stellt also den Anteil der Untersuchungsobjekte mit dem Merkmal A im Klumpen j dar. Eine alle Klumpen zusammenfassende Schätzung für π liefert folgende Gleichung:

$$\bar{p} = \frac{\sum_{j=1}^{k} N_{A(j)}}{\sum_{j=1}^{k} N_j}. \tag{7.38a}$$

Der Wert $\bar{p}$ schätzt π für praktische Zwecke hinreichend genau, wenn ▶ Gl. (7.37) erfüllt ist. Mit dieser Symbolik resultiert für den Standardfehler $\hat{\sigma}_{\bar{p}}$:

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{1-f}{n^2} \cdot \frac{k}{k-1} \cdot \sum_{j=1}^{k} \left(N_{A(j)} - \bar{p} \cdot N_j\right)^2}. \tag{7.39}$$

Beispiel: Zur Erläuterung dieser Gleichungen wählen wir erneut das auf ▶ S. 433 erwähnte Beispiel, das den Anteil der Wahlberechtigten einer Großstadt prüfte, die beabsichtigen, eine Partei A zu wählen. Statt einer Zufallsstichprobe von N=1000 Personen soll nun eine Klumpe nstichprobe mit ungefähr gleichem Umfang

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Tab. 7.7. Ergebnis einer Umfrage unter den Bewohnern von 10 Stadtgebieten

Nummer des Stadtgebietes (j)	Anzahl der Bewohner (N_j)	Anzahl der Wähler von A ($N_{A(j)}$)	Wähleranteil (P_j)
1	109	28	0,26
2	88	39	0,44
3	173	50	0,29
4	92	23	0,25
5	28	16	0,57
6	114	34	0,30
7	55	19	0,35
8	163	70	0,43
9	77	21	0,27
10	101	50	0,50
	1000	350	0,35

gezogen werden. Hierfür teilt man – unter Ausschluss von Grünflächen und gewerblich genutzten Flächen – das Stadtgebiet z. B. in 10.000 gleich große Flächenareale auf und wählt aus diesen 10 Flächenareale aus. Alle wahlberechtigten Personen dieser 10 Gebiete (Klumpen) werden befragt. Die Befragung führt zu den in Tab. 7.7 aufgeführten Werten. (Die Zahlen wurden so gewählt, dass der Rechengang überschaubar bleibt und die Ergebnisse mit den auf ▶ S. 433 f. berichteten Resultaten verglichen werden können.)

Insgesamt beabsichtigen 35% ($\bar{p}$=0,35) der befragten Personen, Partei A zu wählen. Mit ▶ Gl. (7.37) prüfen wir, ob dieser Wert als Schätzer des Populationsparameters π akzeptabel ist.

$$V = \frac{\hat{\sigma}_{\bar{N}}}{\bar{N}} = \frac{\hat{\sigma}_N}{\sqrt{k}\cdot\bar{N}} = \frac{44{,}123}{\sqrt{10}\cdot 100} = 0{,}14.$$

(Mit $\hat{\sigma}_N$=Standardabweichung der Klumpenumfänge N_j.)

Trotz der recht beachtlichen Unterschiede in den Klumpenumfängen bleibt der Variationskoeffizient unter der kritischen Grenze von 0,2, was damit zusammenhängt, dass der durchschnittliche Klumpenumfang mit $\bar{N}$=100 relativ groß ist.

Für den Standardfehler von $\bar{p}$ ermitteln wir nach ▶ Gl. (7.39)

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{1-f}{n^2}\cdot\frac{k}{k-1}\cdot\sum_{j=1}^{k}\left(N_{A(j)} - \bar{p}\cdot N_j\right)^2}$$

$$= \sqrt{\frac{1-0{,}001}{1000^2}\cdot\frac{10}{9}\cdot 857{,}25} = 0{,}0308.$$

Der Standardfehler ist in diesem Beispiel ungefähr doppelt so groß wie der Standardfehler einer entsprechenden Zufallsstichprobe (▶ S. 433). Dieses Ergebnis reflektiert die deutlichen Unterschiede in den Wähleranteilen der einzelnen Klumpen (Tab. 7.7, letzte Spalte). Bei dieser Unterschiedlichkeit der Klumpen erweist sich die Ziehung einer Klumpenstichprobe als äußerst ungünstig.

Für die Bestimmung des 99%igen Konfidenzintervalls benötigen wir die t-Verteilung mit 9 Freiheitsgraden (▶ Anhang F, Tab. F3). Der Wert t=±3,25 schneidet an den Extremen dieser Verteilung jeweils 0,005% der Verteilung ab, d. h., wir erhalten als Konfidenzintervall:

$$0{,}35 \pm 3{,}25\cdot 0{,}0308 = 0{,}35 \pm 0{,}1003.$$

Obwohl der Gesamtstichprobenumfang mit n= 1000 vergleichsweise groß ist, resultiert für den Wähleranteil ein Konfidenzintervall von 25–45%. Dieses Konfidenzintervall dürfte den Untersuchungsaufwand in keiner Weise rechtfertigen. Das Beispiel demonstriert damit eine Untersuchungssituation, die für die Ziehung einer Klumpenstichprobe (mit Wohnarealen als Klumpen) denkbar ungeeignet ist.

7.2.3 Die mehrstufige Stichprobe

Eine Klumpenstichprobe – so zeigte der vorige Abschnitt – setzt sich aus mehreren vollständig erhobenen Klumpen zusammen. In der Praxis kommt es jedoch häufig vor, dass die natürlich angetroffenen Klumpen zu groß sind, um sie vollständig erheben zu können. In dieser Situation wird man statt einer Klumpenstichprobe eine 2- oder mehrstufige Stichprobe ziehen (»**Multi-stage Sampling**«). Die erste Stufe betrifft die Zufallsauswahl der Klumpen und die zweite die Zufallsauswahl der Untersuchungsobjekte innerhalb der Klumpen. Damit erfasst eine 2-stufige Stichprobe im Unterschied zur Klumpenstichprobe die einzelnen Klumpen nicht vollständig, sondern nur in zufälligen Ausschnitten.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

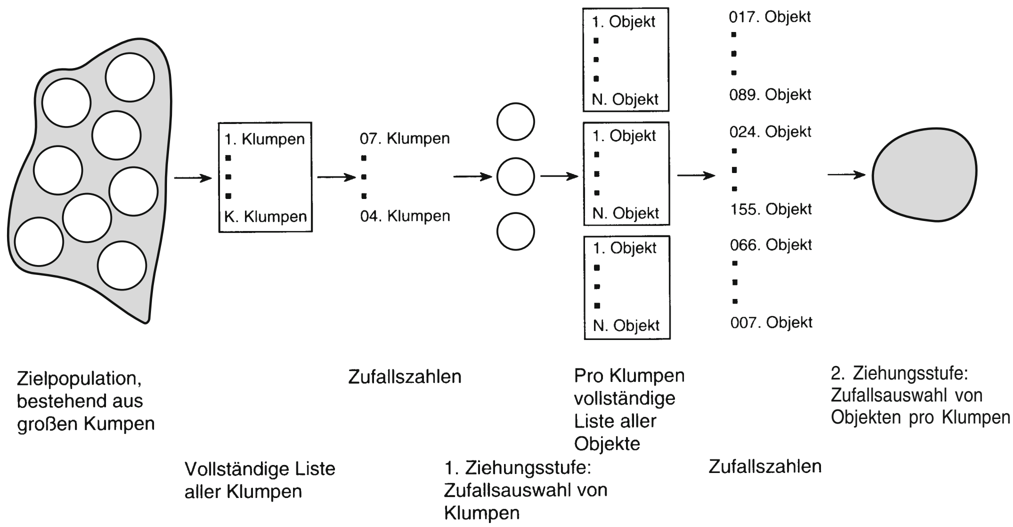


Abb. 7.10. Ziehung einer zweistufigen Stichprobe

Die Klumpenstichprobe stellt einen Spezialfall der 2-stufigen Stichprobe dar. Aber auch die geschichtete Stichprobe ist ein Spezialfall der 2-stufigen Stichprobe. Hier werden auf der ersten Auswahlstufe alle Schichten (statt einer Auswahl von Klumpen) berücksichtigt, aus denen man jeweils eine Zufallsauswahl entnimmt.

Die 2-stufige Stichprobe bereitet – wie bereits die Klumpenstichprobe – erheblich weniger organisatorischen Aufwand als eine einfache Zufallsstichprobe. Für die Ziehung einer Zufallsstichprobe benötigen wir streng genommen eine komplette Liste aller Untersuchungsobjekte der Population, während für die 2-stufige Stichprobe lediglich eine vollständige Liste aller Klumpen sowie Listen der Untersuchungsobjekte in den ausgewählten Klumpen erforderlich sind (Abb. 7.10).

Wollte man beispielsweise eine Befragung unter Mitgliedern von Wohngemeinschaften in einer Stadt durchführen, wäre hierfür eine Liste aller Wohngemeinschaften und – nach erfolgter Zufallsauswahl einiger Wohngemeinschaften – eine Liste der Mitglieder dieser Wohngemeinschaften erforderlich. Die aus dieser Liste zufällig ausgewählten Personen konstituieren die 2-stufige Stichprobe.

Eine **3-stufige Stichprobe** von Gymnasiasten erhält man beispielsweise, wenn aus der Liste aller zur Population gehörenden Gymnasien eine Zufallsauswahl getroffen wird (1. Stufe), aus diesen Schulen zufällig Schulklassen (2. Stufe) und aus den Klassen wiederum einige Schüler ausgewählt werden (3. Stufe). Ein Beispiel für ein 3-stufiges Stichprobensystem, das Repräsentativität für die Gesamtbevölkerung anstrebt, ist das Stichprobensystem des Arbeitskreises Deutscher Marktforschungsinstitute (**ADM-Mastersample**, ▶ S. 484, Box 7.9). Dieses Stichprobensystem kommt u.a. beim ALLBUS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften) zum Einsatz, mit dem seit 1980 in zweijährigem Abstand Mehrthemenumfragen (Omnibusumfragen) zu Themen wie Arbeit, Soziales, Umwelt, Politik etc., durchgeführt werden (vgl. z. B. Lipsmeier, 1999, S. 102 f. bzw. Anhang C zum Stichwort ZUMA).

! Man zieht eine mehrstufige Stichprobe, indem man zunächst zufällig eine Klumpenstichprobe mit großen Klumpen zieht (1. Ziehungsstufe). Diese Klumpen werden nicht vollständig untersucht, sondern aus ihnen wird eine Zufallsstichprobe der Untersuchungsobjekte gezogen (2. Ziehungsstufe). Zieht man auf der zweiten Stufe wieder eine Klumpenstichprobe, ergibt sich durch Ziehung einer Zufallsstichprobe aus diesen Klumpen eine 3. Ziehungsstufe usw.

Im Folgenden wird gezeigt, wie aus mehrstufigen Stichproben Schätzungen von Mittelwertparametern μ und Populationsanteilen π abgeleitet werden können.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Schätzung von Populationsmittelwerten

Für eine 2-stufige Stichprobe benötigen wir eine Zufallsauswahl von k Klumpen aus den K Klumpen der zu beschreibenden Population. Als ersten Auswahlsatz definieren wir $f_1=k/K$. Für jeden ausgewählten Klumpen j ziehen wir aus den N_j Untersuchungsobjekten eine Zufallsstichprobe des Umfanges n_j. Der zweite Auswahlsatz heißt damit $f_{2(j)}=n_j/N_j$.

Wenn die Stichprobenumfänge zur Größe der Klumpen proportional sind, resultiert pro Objekt folgende Auswahlwahrscheinlichkeit: In jedem ausgewählten Klumpen j beträgt die Auswahlwahrscheinlichkeit n_j/N_j=const. Jeder Klumpen wiederum hat eine Auswahlwahrscheinlichkeit von k/K. Da beide Wahrscheinlichkeiten voneinander unabhängig sind, erhält man die Auswahlwahrscheinlichkeit für ein Objekt als Produkt dieser Wahrscheinlichkeiten: $(n_j/N_j)\cdot(k/K)$. Ersetzt man n_j durch $\bar{n}$ (durchschnittliche Stichprobengröße) und N_j durch $\bar{N}$ (durchschnittliche Größe der Teilpopulationen), erkennt man, dass dies gleichzeitig die Auswahlwahrscheinlichkeit n/N für ein Objekt ist, wenn aus einer Population des Umfanges $K\cdot\bar{N}=N$ eine einfache Zufallsstichprobe des Umfanges $k\cdot\bar{n}=n$ gezogen wird.

Gelegentlich ist man daran interessiert, aus allen ausgewählten Klumpen gleichgroße Stichproben des Umfanges n_c zu ziehen. In diesem Falle ist die Auswahlwahrscheinlichkeit eines Objektes in einem großen Klumpen natürlich kleiner als in einem kleinen Klumpen. Um dies zu kompensieren, werden für die Klumpen keine konstanten Auswahlwahrscheinlichkeiten angesetzt, sondern Auswahlwahrscheinlichkeiten, die proportional zur Größe der Klumpen sind: N_j/N. Dieses Auswahlverfahren wird **PPS-Design** genannt (»Probability Proportional to Size«).

Werden aus einem ausgewählten Klumpen n_c Objekte zufällig gezogen, ergibt sich eine Ziehungswahrscheinlichkeit von n_c/N_j, was insgesamt zu einer Auswahlwahrscheinlichkeit von $(N_j/N)\cdot(n_c/N_j)=n_c/N$ führt. Dies ist die Auswahlwahrscheinlichkeit für ein Objekt, wenn nur ein Klumpen gezogen wird. Zieht man eine Stichprobe von k Klumpen, erhöht sich diese Wahrscheinlichkeit um das k-fache: $k\cdot n_c/N$. Dies wiederum ist die Auswahlwahrscheinlichkeit für ein Objekt, wenn aus einer Population des Umfangs N eine einfache Zufallsstichprobe des Umfanges $k\cdot n_c=n$ gezogen wird (Einzelheiten zur Technik des PPS-Designs findet man z. B. bei Schnell et al., 1999, Anhang F).

Die folgenden Ausführungen gehen davon aus, dass die Stichprobenumfänge n_j proportional zu den Klumpenumfängen N_j sind, dass also $n_j/N_j=\text{const.}=f_2$ ist. Dies setzt voraus, dass die Umfänge der ausgewählten Klumpen zumindest ungefähr bekannt sind. (Cochran, 1972, Kap. 11, beschreibt Varianten mehrstufiger Stichproben, die diese Voraussetzung nicht machen.)

Bezeichnen wir die i-te Messung im j-ten Klumpen mit x_{ij} (mit $i=1,2\ldots n_j$ und $j=1,2\ldots k$), schätzt der folgende Stichprobenmittelwert den Parameter μ erwartungstreu:

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{\sum_{j=1}^{k} n_j}. \tag{7.40}$$

Wegen der zur Klumpengröße proportionalen Stichprobenumfänge kann auf eine Gewichtung der einzelnen Klumpenmittelwerte verzichtet werden (**selbstgewichtende Stichprobe**, ▶ S. 427). Den Standardfehler von $\bar{\bar{x}}$ ermitteln wir nach ▶ Gl. (7.41).

$$\hat{\sigma}_{\bar{\bar{x}}} = \sqrt{\frac{1-f_1}{n^2}\cdot\frac{k}{k-1}\cdot\sum_{j=1}^{k}\left(\sum_{i=1}^{n_j} x_{ij}-\bar{\bar{x}}\cdot n_j\right)^2 + f_1\cdot(1-f_2)\cdot\frac{1}{n}\cdot\sum_{j=1}^{k} g_j\cdot\hat{\sigma}_j^2} \tag{7.41}$$

mit

$$\hat{\sigma}_j^2 = \frac{1}{n_j-1}\cdot\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2$$

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j},$$

$$g_j = \frac{N_j}{N}\left(\sum_{j=1}^{k} g_j = 1\right)$$

$$N = \sum_{j=1}^{k} N_j$$

und

$$n = \sum_{j=1}^{k} n_j.$$

Die Bestimmung der Gewichte g_j setzt – wie bereits die Festlegung des Auswahlsatzes f_2 – voraus, dass die Klumpenumfänge bekannt sind oder doch zumindest ge-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Box 7.6

Wie umfangreich sind Diplomarbeiten?
IV: Die 2-stufige Stichprobe

Bezugnehmend auf das Beispiel der Boxen 7.3–7.5 verdeutlicht diese Box die Ermittlung der durchschnittlichen Seitenzahl von Diplomarbeiten (einschließlich eines Konfidenzintervalls) anhand einer 2-stufigen Stichprobe. Die erste Auswahlstufe umfasst sämtliche psychologischen Institute der Bundesrepublik Deutschland als Klumpen, von denen k=10 zufällig ausgewählt werden. Die in diesen Instituten in den letzten 10 Jahren angefertigten Diplomarbeiten bilden die 2. Auswahlstufe. Es werden aus den Arbeiten dieser Institute Zufallsstichproben gezogen, deren Größen proportional zur Anzahl aller Arbeiten des jeweiligen Instituts sind, die im vorgegebenen Zeitraum angefertigt wurden. Es wird ein Gesamtstichprobenumfang von ungefähr 100 Arbeiten angestrebt. (Die Stichprobe wäre 3-stufig, wenn man zusätzlich aus den ausgewählten Instituten z. B. Zufallsstichproben von Hochschullehrern gezogen hätte.)

Die Gesamtzahl aller psychologischen Institute mit einem Diplomstudiengang möge K=46 betragen. Für zehn auszuwählende Institute resultiert für den ersten Auswahlsatz also $f_1=k/K=10/46=0{,}217$. Von den ausgewählten Instituten werden Listen aller Diplomarbeiten der letzten 10 Jahre angefordert. Die Anzahl der Diplomarbeiten entspricht den Umfängen N_j der Klumpen. Die Größe der Stichproben n_j wird so gewählt, dass sie – bei einer Gesamtstichprobe von $n\approx100$ – zu diesen N_j-Werten proportional sind.

Im Durchschnitt wurden pro Institut während des angegebenen Zeitraumes etwa 750 Diplomarbeiten abgegeben, d. h., der zweite Auswahlsatz lautet bei durchschnittlich zehn auszuwählenden Arbeiten $f_2=\frac{10}{750}=0{,}013$. Dieser Wert ist kleiner als 0,05; man könnte ihn deshalb in ▶ Gl. (7.41) vernachlässigen. Um den Rechengang vollständig zu demonstrieren, soll er jedoch nicht entfallen.

Die folgende Aufstellung enthält die für die Berechnungen erforderlichen Angaben. Auf die Wiedergabe der einzelnen x_{ij}-Werte, die bekannt sein müssen, um die $\hat{\sigma}_j^2$-Werte zu bestimmen, wurde verzichtet.

$$k=10;\ n=\sum_{j=1}^{k} n_j=100;\ \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}=9200;\ \bar{\bar{x}}=92$$

Im Durchschnitt haben die 100 untersuchten Arbeiten also 92 Seiten. Für den Standardfehler dieses Mittelwertes erhalten wir nach ▶ Gl. (7.41)

Nr. der Stichprobe (j)	Größe der Stichprobe (n_j)	Anzahl der Seiten in Stichprobe j	Durchschnittl. Seitenzahl in Stichprobe j ($\bar{x}_j$)	Varianz in Stichprobe j ($\hat{\sigma}_j^2$)
1	16	1488	93	1016
2	20	1700	85	970
3	4	360	90	840
4	8	752	94	1112
5	11	979	89	763
6	9	783	87	906
7	9	891	91	1004
8	11	1012	92	895
9	7	735	105	642
10	5	500	100	930

▼

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

$$\hat{\sigma}_{\bar{\bar{x}}} = \sqrt{\frac{1-f_1}{n^2}\cdot\frac{k}{k-1}\cdot\sum_{j=1}^{k}\left(\sum_{i=1}^{n_j}x_{ij} - \bar{\bar{x}}\cdot n_j\right)^2 + f_1\cdot(1-f_2)\cdot\frac{1}{n}\cdot\sum_{j=1}^{k}g_j\cdot\hat{\sigma}_j^2}$$

$$= \sqrt{\frac{1-0{,}217}{100^2}\cdot\frac{10}{9}\cdot 37140 + 0{,}217\cdot(1-0{,}013)\cdot\frac{1}{100}\cdot 924{,}84}$$

$$= \sqrt{3{,}23+1{,}98} = \sqrt{5{,}21} = 2{,}28.$$

Da die Stichprobenumfänge n_j proportional zu den Klumpenumfängen N_j festgelegt wurden, entsprechen die Gewichte g_j den durch n dividierten Stichprobenumfängen (also z. B. g_1=16/100=0,16).

Für die Bestimmung des Konfidenzintervalls legen wir in diesem Beispiel die t-Verteilung mit 9 Freiheitsgraden zugrunde. Für das 99%ige Konfidenzintervall gilt t=±3,25, d. h., das Konfidenzintervall ergibt sich zu

$$92 \pm 3{,}25\cdot 2{,}28 = 92 \pm 7{,}41.$$

Die Schätzgenauigkeit der 2-stufigen Stichprobe entspricht damit in diesem Beispiel der Schätzgenauigkeit der geschichteten Stichprobe in ◘ Box 7.4.

Betrachten wir abschließend noch die Auswahlwahrscheinlichkeit für eine Diplomarbeit. Diese errechnet sich bei einer einfachen Zufallsstichprobe zu 100 (= Anzahl der ausgewählten Arbeiten)/(46·750) (= Anzahl aller Arbeiten) =0,0029. Für die 2-stufige Stichprobe (mit gleicher Wahrscheinlichkeit für die Institute) wird ein Institut mit einer Wahrscheinlichkeit von 10/46 gezogen und innerhalb eines ausgewählten Institutes eine Arbeit mit einer Wahrscheinlichkeit von n_j/N_j=const (z. B. 10/750), was zusammengenommen zu der bereits bekannten Auswahlwahrscheinlichkeit von 0,0029 führt: (10/46)·(10/750)=0,0029.

Dieser Wert wird auch für das PPS-Design errechnet: Ein Institut mit N_j Arbeiten hat eine Auswahlwahrscheinlichkeit von N_j/46·750 und eine Arbeit in diesem Institut die Wahrscheinlichkeit 10/N_j. Man erhält also (N_j/46·750)·(10/N_j)= 10/46·750=0,00029. Dies ist die Auswahlwahrscheinlichkeit für eine Diplomarbeit, falls nur ein Institut ausgewählt wird. Werden 10 Institute mit PPS gezogen, ergibt sich wieder 10·0,00029=0,0029.

schätzt werden können. ◘ Box 7.6 erläutert, wie aus den Daten einer 2-stufigen Stichprobe ein Konfidenzintervall für $\bar{x}$ zu berechnen ist.

Vergleichende Bewertung. ▶ Gl. (7.41) verdeutlicht, unter welchen Umständen eine 2-stufige Stichprobe den Parameter μ besonders genau schätzt. Abgesehen von der Gesamtstichprobe, die bei allen Stichprobenarten mit wachsendem Umfang den Standardfehler verkleinert, beeinflussen sowohl die Unterschiedlichkeit zwischen den Klumpen als auch die Unterschiedlichkeit innerhalb der Klumpen die Schätzgenauigkeit. Beide Unterschiedlichkeiten vergrößern den Standardfehler, wobei die Heterogenität der Messungen innerhalb der Klumpen keine Rolle spielt, wenn f_1 zu vernachlässigen ist, wenn also die Anzahl aller Klumpen in der Population im Vergleich zur Anzahl der ausgewählten Klumpen sehr groß ist. Unter diesen Umständen entfällt der zweite Teil unter der Wurzel von ▶ Gl. (7.41). Der Standardfehler wird dann ausschließlich von der Varianz zwischen den Klumpen bestimmt. Es empfiehlt sich deshalb, eine Population in möglichst viele (und damit kleine) Klumpen zu zerlegen.

Darüber hinaus bestätigt ▶ Gl. (7.41) die eingangs formulierte Behauptung, Klumpenstichproben und geschichtete Stichproben seien Spezialfälle von 2-stufigen Stichproben. Bei einer geschichteten Stichprobe wird die Gesamtpopulation in Schichten (z. B. nach dem Bildungsniveau, dem Alter oder dem Geschlecht, ▶ Abschn. 7.2.1) eingeteilt, und aus jeder Schicht wird eine Zufallsstichprobe gezogen. Bezeichnen wir die Anzahl der Schichten mit K, ist – da jede Schicht stichprobenartig untersucht wird – k=K bzw. f_1=1. Dadurch entfällt der erste Summand unter der Wurzel von ▶ Gl. (7.41). Der verblei-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

bende Teil entspricht ▶ Gl. (7.25), dem Standardfehler von $\bar{x}$ für geschichtete Stichproben mit proportionalen Stichprobenumfängen.

Bei Klumpenstichproben werden die ausgewählten Klumpen vollständig untersucht, d. h. $n_j=N_j$. Dadurch wird in ▶ Gl. (7.41) $f_2=1$, d. h., der zweite Summand unter der Wurzel entfällt. Der Rest der Gleichung entspricht ▶ Gl. (7.38), dem Standardfehler von $\bar{\bar{x}}$ für Klumpenstichproben.

Drei- und mehrstufige Stichproben. Bei dreifach gestuften Stichproben wird auf der ersten Auswahlstufe eine Zufallsstichprobe von **Primäreinheiten** gezogen (z. B. eine Stichprobe von Ausgaben eines Wochenmagazins), auf der zweiten Auswahlstufe jeweils eine Zufallsstichprobe von **Sekundäreinheiten** innerhalb der Primäreinheiten (z. B. eine Stichprobe von Seiten aus jedem ausgewählten Magazin), und auf der dritten Auswahlstufe entnimmt man schließlich den Sekundäreinheiten die eigentlichen Untersuchungsobjekte (z. B. eine Zufallsstichprobe von Zeilen jeder ausgewählten Seite). Die so gefundenen Untersuchungsobjekte werden hinsichtlich des interessierenden Merkmals untersucht (z. B. durchschnittliche Wortlänge als Beispiel für die Schätzung eines Mittelwertparameters oder die relative Häufigkeit von Substantiven als Beispiel für eine Anteilsschätzung).

Bezeichnen wir die Anzahl aller Primäreinheiten mit L und die Anzahl der ausgewählten Primäreinheiten mit l, die Anzahl aller Sekundäreinheiten einer jeden Primäreinheit mit K und die Anzahl der pro Primäreinheit ausgewählten Sekundäreinheiten mit k sowie die Anzahl aller Untersuchungsobjekte einer jeden Sekundäreinheit mit N und die Anzahl der pro Sekundäreinheit ausgewählten Untersuchungsobjekte mit n, resultiert als Schätzwert für μ

$$\bar{\bar{\bar{x}}} = \frac{\sum_{m=1}^{l}\sum_{j=1}^{k}\sum_{i=1}^{n} x_{ijm}}{l \cdot k \cdot n}. \tag{7.42}$$

Diese und die folgende Gleichung für den Standardfehler setzen also auf jeder Stufe eine konstante Anzahl von Auswahleinheiten voraus. (Im Beispiel der ◘ Box 7.6 wären also l Institute auszuwählen, pro Institut k Hochschullehrer und pro Hochschullehrer n Diplomarbeiten, sodass sich die Gesamtzahl aller Arbeiten zu l·k·n ergibt.) Ist diese Bedingung (zumindest annähernd) erfüllt, ergibt sich für den Standardfehler

$$\hat{\sigma}_{\bar{\bar{\bar{x}}}} = \sqrt{\frac{1-f_1}{1}\cdot\hat{\sigma}_1^2 + \frac{f_1\cdot(1-f_2)}{l\cdot k}\cdot\hat{\sigma}_2^2 + \frac{f_1\cdot f_2\cdot(1-f_3)}{l\cdot k\cdot n}\cdot\hat{\sigma}_3^2} \tag{7.43}$$

wobei:

$$f_1 = \frac{l}{L},$$

$$f_2 = \frac{k}{K},$$

$$f_3 = \frac{n}{N},$$

$$\hat{\sigma}_1^2 = \frac{\sum_{m=1}^{l}(\bar{\bar{x}}_m - \bar{\bar{\bar{x}}})^2}{l-1},$$

$$\hat{\sigma}_2^2 = \frac{\sum_{m=1}^{l}\sum_{j=1}^{k}(\bar{x}_{mj} - \bar{\bar{x}}_m)^2}{l\cdot(k-1)},$$

$$\hat{\sigma}_3^2 = \frac{\sum_{m=1}^{l}\sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ijm} - \bar{x}_{mj})^2}{l\cdot k\cdot(n-1)},$$

$$\bar{\bar{x}}_m = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n} x_{ijm}}{k\cdot n},$$

$$\bar{x}_{mj} = \frac{\sum_{i=1}^{n} x_{ijm}}{n}.$$

(Zur Herleitung dieser Gleichung vgl. Cochran, 1972, Kap. 10.8). Dem Aufbau dieser Gleichung ist leicht zu entnehmen, wie Mittelwert- und Standardfehlerbestimmungen auf Stichproben mit mehr als drei Stufen zu erweitern sind.

Schätzung von Populationsanteilen

Will man den Anteil aller Untersuchungsobjekte einer Population, die durch ein Merkmal A gekennzeichnet sind, aufgrund einer mehrstufigen Stichprobe schätzen, ist bei den bisherigen Überlegungen die Summe der Merkmalsausprägungen durch die Anzahl der Untersuchungseinheiten mit dem Merkmal A zu ersetzen. Die Stichprobenumfänge n_j seien erneut proportional zu den Populationsumfängen N_j. Für den Parameter π (Anteil aller Untersuchungseinheiten mit dem

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Merkmal A) resultiert bei 2-stufigen Stichproben folgender Schätzwert:

$$\bar{p} = \frac{\sum_{j=1}^{k} n_{A(j)}}{\sum_{j=1}^{k} n_j}.$$

In Analogie zu ▶ Gl. (7.41) erhalten wir als Standardfehler von $\bar{p}$:

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{1-f_1}{n^2} \cdot \frac{k}{k-1} \cdot \sum_{j=1}^{k}(n_{A(j)} - \bar{p} \cdot n_j)^2 + f_1 \cdot (1-f_2) \cdot \frac{1}{n} \cdot \sum_{j=1}^{k} g_j \cdot p_j \cdot (1-p_j)} \tag{7.44}$$

$$\text{mit } p_j = \frac{n_{A(j)}}{n_j} \text{ und } g_j = \frac{n_j}{n}$$

(Zur Erläuterung der übrigen Symbole siehe ▶ Gl. 7.41.)

Beispiel: Wollen wir mit Hilfe einer zweifach geschichteten Stichprobe den Anteil aller Wähler einer Partei A in einer Großstadt (vgl. Beispiel ▶ S. 433 und ▶ S. 439 f.) schätzen, sind folgende Überlegungen und Berechnungen erforderlich. Zunächst wird das gesamte Stadtgebiet in Areale (Klumpen) aufgeteilt. Die Gesamtzahl aller Klumpen sei K=1000, und aus diesen werden k=15 zufällig ausgewählt. Damit erhalten wir f_1=15/1000=0,015. Man schätzt (oder ermittelt anhand von Karteien des Einwohnermeldeamtes) die Anzahl wahlberechtigter Personen in den ausgewählten Arealen (N_j) und zieht aus diesen Teilpopulationen Zufallsstichproben des Umfanges n_j. Der Gesamtstichprobenumfang n soll sich in diesem Beispiel auf etwa 1000 belaufen. Es wird darauf geachtet, dass das Verhältnis $n_j/N_j=f_2$ in allen Arealen konstant ist (proportionale Stichprobenumfänge). Bei einer Million wahlberechtigter Personen lautet der Auswahlsatz f_2=1000/1000000=0,001, d. h., in jedem Areal sind 1 Promill der dort ansässigen Wahlberechtigten zu befragen. In ◻ Tab. 7.8 sind die Resultate dieser Befragung aufgeführt.

Die Daten wurden so gewählt, dass sich für $\bar{p}$ wiederum 0,35 ergibt. Die Berechnung des Standardfehlers führt zu folgendem Resultat:

$$\hat{\sigma}_{\bar{p}} = \sqrt{\frac{1-0{,}015}{1000^2} \cdot \frac{15}{14} \cdot 291{,}97 + 0{,}015 \cdot (1-0{,}001) \cdot \frac{1}{1000} \cdot 0{,}224}$$

$$= \sqrt{0{,}00031 + 0{,}000003} = 0{,}0177.$$

Als Gewichte g_j verwendet diese Berechnung die an n=1000 relativierten Stichprobenumfänge n_j (also g_1=0,051, g_2=0,142 etc.).

Mit einem Standardfehler von 0,0177 und einem t-Wert von 2,977 (df=14) resultiert als 99%iges Konfidenzintervall

$$0{,}35 \pm 2{,}977 \cdot 0{,}0177 = 0{,}35 \pm 0{,}053.$$

Gegenüber der einfachen Klumpenstichprobe, die in ◻ Tab. 7.7 beschrieben wird, hat sich das Konfidenzintervall damit um etwa die Hälfte verkleinert.

3-stufige Stichprobe. Anteilsschätzungen, die aufgrund einer dreifach gestuften Stichprobe vorgenommen werden, haben entsprechend ▶ Gl. (7.43) folgenden Standardfehler:

$$\hat{\sigma}_{\bar{\bar{p}}} = \sqrt{\frac{1-f_1}{1} \cdot \hat{\sigma}_1^2 + \frac{f_1 \cdot (1-f_2)}{l \cdot k} \cdot \hat{\sigma}_2^2 + \frac{f_1 \cdot f_2 \cdot (1-f_3)}{l \cdot k \cdot n} \cdot \hat{\sigma}_3^2} \tag{7.45}$$

◻ Tab. 7.8. Ergebnis einer Befragung von 15 Zufallsstichproben aus 15 zufällig ausgewählten Stadtgebieten

Nummer des Stadtgebietes (j)	Größe der Stichprobe (n_j)	Anzahl der Wähler von A ($n_{A(j)}$)	Wähleranteil (P_j)
1	51	16	0,31
2	142	50	0,35
3	90	29	0,32
4	22	8	0,36
5	70	21	0,30
6	68	20	0,29
7	65	22	0,34
8	49	16	0,33
9	14	4	0,29
10	112	53	0,47
11	62	28	0,45
12	83	25	0,30
13	68	22	0,32
14	40	15	0,38
15	64	21	0,33
	n=1000	n_A=350	$\bar{p}=0{,}35$

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

mit

$$\hat{\sigma}_1^2 = \frac{\sum_{m=1}^{l}(\bar{p}_m - \bar{\bar{p}})^2}{l-1},$$

$$\hat{\sigma}_2^2 = \frac{\sum_{m=1}^{l}\sum_{j=1}^{k}(p_{jm} - \bar{p}_m)^2}{l \cdot (k-1)},$$

$$\hat{\sigma}_3^2 = \frac{\sum_{m=1}^{l}\sum_{j=1}^{k}p_{jm} \cdot (1-p_{jm})}{l \cdot k \cdot (n-1)}.$$

7.2.4 Wiederholte Stichprobenuntersuchungen

Viele Fragestellungen beinhalten die Evaluation verhaltens- oder einstellungsändernder Maßnahmen wie z. B. die Beeinflussung des Essverhaltens durch ein Diätprogramm, die Veränderung des Kaufverhaltens in Abhängigkeit von der Anzahl der Werbekontakte, Einstellungswandel gegenüber Ausländern durch gezielte Pressemitteilungen oder den Abbau innerbetrieblicher Konflikte durch Einführung eines neuen Führungsstils. Im Vordergrund derartiger Untersuchungen stehen Veränderungen des geprüften Merkmals zwischen zwei (oder mehreren) Zeitpunkten, die am günstigsten durch die wiederholte Verwendung einer Stichprobe zu erfassen sind. Es handelt sich um hypothesenprüfende Untersuchungen (es wird z. B. die Hypothese überprüft, dass die durchgeführte Maßnahme positiv wirkt), die ausführlich in ▶ Abschn. 8.2.5 erörtert werden.

Hier befassen wir uns nach wie vor mit der Frage, wie die Präzision einer populationsbeschreibenden Untersuchung durch die Nutzung bereits vorhandener Kenntnisse über den Untersuchungsgegenstand gesteigert werden kann. Will man beispielsweise die Zufriedenheit der Bewohner einer großen Neubausiedlung mit ihren Wohnverhältnissen mittels einer Stichprobe schätzen, lässt sich die Genauigkeit dieser Untersuchung oftmals erheblich verbessern, wenn man auf frühere stichprobenartige Befragung der gleichen Bewohner zu einer ähnlichen Thematik zurückgreifen kann. Ein weiteres Beispiel: Der durchschnittliche Alkoholkonsum der Bevölkerung eines ländlichen Gebietes lässt sich genauer schätzen, wenn sich in der Stichprobe einige Personen befinden, die zur gleichen Thematik bereits früher (z. B. anlässlich einer ärztlichen Untersuchung) Angaben machten.

! Werden eine Stichprobe oder Teile einer Stichprobe wiederholt untersucht, führt dies in der Regel zu einem deutlichen Genauigkeitsgewinn für die Parameterschätzung.

Panelforschung. Viele politische und wirtschaftliche Entscheidungen erfordern aktuelle Planungsunterlagen, die kostengünstig und kurzfristig nur zu beschaffen sind, wenn wiederholt auf eine bereits eingerichtete, repräsentative Stichprobe zurückgegriffen werden kann. Eine Stichprobe, die wiederholt zu einer bestimmten Thematik (Fernsehgewohnheiten, Konsumgewohnheiten etc.) oder auch zu verschiedenen Themen befragt wird, bezeichnet man als ein Panel. Die wiederholte Befragung der Panelmitglieder findet typischerweise mündlich oder schriftlich statt (▶ Kap. 4.4), wobei schriftliche Befragungen postalisch oder elektronisch (Online-Panel) durchgeführt werden.

! Ein Panel ist eine Stichprobe, die wiederholt untersucht wird.

Den mit Kosten- und Zeitersparnis verbundenen Vorteilen stehen bei Paneluntersuchungen jedoch einige gravierende Nachteile gegenüber. Man muss damit rechnen, dass ein Panel im Laufe der Zeit seine Aussagekraft bzw. Repräsentativität verliert, weil die einzelnen Panelmitglieder durch die Routine, die sie während vieler Befragungen allmählich gewinnen, nicht mehr »naiv« und unvoreingenommen reagieren. Das Bewusstsein, Mitglied eines Panels zu sein, kann sowohl das alltägliche Verhalten als auch das Verhalten in der Befragungssituation entscheidend beeinträchtigen.

Auf der anderen Seite können sich wiederholte Befragungen auch positiv auf die Qualität eines Interviews (oder einer schriftlichen Befragung) auswirken. Die anfängliche Unsicherheit, sich in einer ungewohnten Befragungssituation zu befinden, verliert sich im Verlaufe der Zeit, die Befragten lernen, ihre eigenen Ansichten und Meinungen treffsicherer und genauer zu formulieren. Die anfängliche Skepsis, persönliche Angaben könnten trotz zugesicherter Anonymität missbräuchlich verwendet werden, schwindet, das Panelmitglied entwickelt so etwas wie ein Verantwortungsbewusstsein dafür, dass die zu treffenden

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Entscheidungen auf korrekten Planungsunterlagen beruhen etc.

Angesichts der Komplexität der mit Paneluntersuchungen verbundenen Probleme entwickelte man aufwändige Austausch- und Rotationspläne, denen zu entnehmen ist, in welchen zeitlichen Abständen und in welchem Umfang »alte« Panelmitglieder auszuscheiden und durch »neue« Panelmitglieder zu ersetzen sind (vgl. z. B. Kish, 1965, Kap. 12.5). Es resultieren Untersuchungen von Stichproben, die sich mehr oder weniger überschneiden. Verbindliche Angaben über eine vertretbare Dauer der Panelzugehörigkeit bzw. über die maximale Anzahl von Befragungen, die mit einem Panelmitglied durchgeführt werden können, sind – zumindest im Hinblick auf die oben erwähnten Konsequenzen wiederholter Befragungen (»**Paneleffekte**«) – nicht möglich, denn diese Werte hängen in starkem Maße von der jeweiligen Befragungssituation und den Inhalten der Befragung ab. Im Zweifelsfalle wird man nicht umhin können, die Brauchbarkeit der Befragungsergebnisse durch **Panelkontrollstudien** zu überprüfen, in denen das Antwortverhalten »alter« Panelmitglieder dem Antwortverhalten erstmalig befragter Personen gegenübergestellt wird. Dieser Vergleich muss natürlich auf »Matched Samples« (▶ S. 527) basieren.

Betrachtet man die Vor- und Nachteile wiederholter Befragungen unter statistischen Gesichtspunkten, lässt sich die Frage, ob bzw. in welchem Ausmaß die Zusammensetzung der Stichprobe beibehalten oder geändert werden soll, eindeutiger beantworten. Dies belegen die folgenden Abschnitte, die den Einsatz wiederholter Stichprobenuntersuchungen zur Schätzung von Populationsmittelwerten (μ) und Populationsanteilen (π) behandeln. (Eine ausführliche Behandlung der Panelforschung findet man bei Arminger & Müller, 1990; Faulbaum, 1988; Kasprzyk et al., 1989; Petersen, 1993.)

Schätzung von Populationsmittelwerten

Wir wollen zunächst den einfachen Fall betrachten, dass zu zwei Zeitpunkten t_1 und t_2 Stichproben des Umfanges n_1 und n_2 untersucht wurden und dass sich von den Untersuchungsobjekten, die zum Zeitpunkt t_1 (dies ist in der Regel der frühere Zeitpunkt) untersucht wurden, s Untersuchungsobjekte auch in der zweiten Untersuchung befinden. Die n_2 Untersuchungsobjekte der zweiten Untersuchung setzen sich damit aus s »alten« Untersuchungsobjekten und $n_2-s=u$ »neuen« Untersuchungsobjekten zusammen. Wir nehmen ferner an, dass die geschätzten Streuungen $\hat{\sigma}_1$ und $\hat{\sigma}_2$ in beiden Untersuchungen annähernd gleich sind und dass die Auswahlsätze ($f_1=n_1/N$ und $f_2=n_2/N$) kleiner als 0,05 sind. Der Mittelwert $\bar{x}_2$ der zweiten Untersuchung soll zur Schätzung des Parameters μ herangezogen werden. Er basiert auf u neuen Messungen mit einem Mittelwert $\bar{x}_{2u}$ und auf s wiederholten Messungen mit dem Mittelwert $\bar{x}_{2s}$. Bekannt seien ferner $\bar{x}_1$ als Mittelwert aller Messungen zum Zeitpunkt t_1 und $\bar{x}_{1s}$ als Mittelwert der s Messungen zum Zeitpunkt t_1.

Kombination alter und neuer Messungen. Zunächst stellt sich das Problem, wie die Mittelwerte $\bar{x}_{2u}$ und $\bar{x}_{2s}$ zu einem gemeinsamen Mittelwert $\bar{x}_2$ zusammengefasst werden können, wenn wir zusätzlich die Ergebnisse der ersten Untersuchung berücksichtigen. Cochran (1972, Kap. 12.10) schlägt folgende Vorgehensweise vor:

Man beginnt mit der Korrektur des Mittelwertes $\bar{x}_{2s}$ bezüglich der Ergebnisse der ersten Untersuchung. Dies geschieht mit folgender Gleichung:

$$\bar{x}'_{2s} = \bar{x}_{2s} + b \cdot (\bar{x}_1 - \bar{x}_{1s}). \qquad (7.46)$$

In dieser Gleichung stellt b den Regressionskoeffizienten (▶ Anhang B) zur Vorhersage der s Messungen zum Zeitpunkt t_2 aufgrund der s Messungen zum Zeitpunkt t_1 dar. Seine Bestimmungsgleichung lautet

$$b = \frac{\sum_{i=1}^{s}(x_{1i} - \bar{x}_{1s}) \cdot (x_{2i} - \bar{x}_{2s})}{\sum_{i=1}^{s}(x_{1i} - \bar{x}_{1s})^2} = \frac{s \cdot \sum_{i=1}^{s} x_{1i}x_{2i} - \sum_{i=1}^{s} x_{1i} \cdot \sum_{i=1}^{s} x_{2i}}{s \cdot \sum_{i=1}^{s} x_{1i}^2 - \left(\sum_{i=1}^{s} x_{1i}\right)^2}. \qquad (7.47)$$

Die Zusammenfassung des korrigierten Mittelwertes $\bar{x}'_{2s}$ mit $\bar{x}_{2u}$ zu einem korrigierten Schätzwert $\bar{x}'_2$ für den gesuchten Parameter μ geschieht in folgender Weise:

$$\bar{x}'_2 = v \cdot \bar{x}_{2u} + (1 - v) \cdot \bar{x}'_{2s}. \qquad (7.48)$$

In ▶ Gl. (7.48) ist v so zu wählen, dass die beiden unabhängigen Schätzwerte $\bar{x}'_{2s}$ und $\bar{x}_{2u}$ mit den Reziprokwerten ihrer quadrierten Standardfehler gewichtet wer-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

den. Dadurch erhält der unsichere Schätzwert (also der Schätzwert mit dem größeren Standardfehler) ein kleineres Gewicht als der sichere Schätzwert. Wir setzen

$$v = \frac{w_{2u}}{w_{2u} + w_{2s}}. \tag{7.49}$$

w_{2u} ist aus dem Standardfehler des Mittelwertes von u-Messwerten $\left(\frac{\sigma}{\sqrt{u}}\right)$ abzuleiten. Die hierfür benötigte Populationsvarianz σ^2 muss in der Regel aus den Daten geschätzt werden. Die beste Schätzung erhalten wir, wenn die Daten der zweiten Erhebung mit den einmalig erhobenen Daten der ersten Erhebung (dies seien $n_1 - s = q$ Daten) zu einer gemeinsamen Schätzung $\hat{\sigma}^2$ vereint werden:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{q}(x_{1i} - \bar{x}_{1q})^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{(q-1) + (n_2 - 1)}. \tag{7.50}$$

Hieraus folgt dann für w_{2u}

$$w_{2u} = \frac{u}{\hat{\sigma}^2}. \tag{7.51}$$

Bei der Herleitung von w_{2s} ist zu beachten, dass s Untersuchungsobjekte zweimal bezüglich desselben Merkmals untersucht wurden. Je ähnlicher diese Messungen sind, desto stabiler (reliabler, ▶ S. 196 ff.) kann das Merkmal offenbar erfasst werden. Die Reliabilität der Messungen wiederum beeinflusst den Standardfehler. Je reliabler die Messungen, desto sicherer schätzt der Mittelwert $\bar{x}_{2s}$ den Parameter μ. Dieser Sachverhalt ist in der folgenden Bestimmungsgleichung für w_{2s} berücksichtigt.

$$w_{2s} = \frac{1}{\frac{\hat{\sigma}^2(1-r^2)}{s} + r^2 \cdot \frac{\hat{\sigma}^2}{n_2}}. \tag{7.52}$$

Hierin ist r die Korrelation (▶ Anhang B) zwischen den s Messwerten zum Zeitpunkt t_1 und den s Messwerten zum Zeitpunkt t_2. Sie wird nach folgender Beziehung berechnet:

$$r = \frac{\hat{\sigma}_{1s} \cdot b}{\hat{\sigma}_{2s}}. \tag{7.53}$$

Standardfehler. Unter Verwendung der Gleichungen (7.48) bis (7.52) resultiert für μ ein Schätzwert $\bar{x}_2'$ mit folgendem Standardfehler:

$$\hat{\sigma}_{\bar{x}_2'} = \sqrt{\frac{\hat{\sigma}^2 \cdot (n_2 - u \cdot r^2)}{n_2^2 - u^2 \cdot r^2}}. \tag{7.54}$$

Mit diesem Standardfehler wird nach den bereits bekannten Regeln ein Konfidenzintervall bestimmt (▶ S. 410 ff.).

Die Handhabung dieser etwas komplizierten Berechnungen wird in ◘ Box 7.7 an einem Beispiel verdeutlicht. Wie auch in den vorangegangenen Beispielen sind die Zahlen so gewählt, dass der Rechenweg möglichst einfach nachvollziehbar ist.

Man beachte, dass der Standardfehler nach ▶ Gl. (7.54) dem Standardfehler einfacher Zufallsstichproben (▶ Gl. 7.5) entspricht, wenn $u=0$ oder $u=n_2$ ist. Der Fall $u=n_2$ ist trivial: Wenn keines der Untersuchungsobjekte bereits untersucht wurde, entspricht die hier besprochene Stichprobentechnik einer normalen Zufallsstichprobe, d. h., es muss auch der entsprechende Standardfehler resultieren. Interessant ist der Fall $u=0$, der besagt, dass sämtliche Untersuchungsobjekte wiederholt gemessen werden. In diesem Fall trägt die Tatsache, dass von allen Untersuchungsobjekten bereits eine Messung vorliegt, nicht dazu bei, die Präzision der zweiten Untersuchung zu erhöhen, und zwar unabhängig von der Höhe der Korrelation beider Messwertreihen. Diese ist für den Standardfehler nur maßgebend, wenn $u>0$ und $u<n_2$. In diesem Fall verringert sich der Standardfehler, wenn $r\neq0$. Für $r=0$ entspricht der nach ▶ Gl. (7.54) bestimmte Standardfehler dem Standardfehler einfacher Zufallsstichproben.

Optimale Mischungen. Um wiederholte Untersuchungen möglichst vorteilhaft einzusetzen, kommt es offenbar darauf an, bei einer gegebenen Korrelation zwischen den Messungen des ersten und des zweiten Zeitpunktes für die zweite Untersuchung das richtige Mischungsverhältnis aus »alten« und »neuen« Untersuchungsobjekten zu finden.

Das **optimale Mischungsverhältnis** ergibt sich, wenn $\hat{\sigma}_{\bar{x}_2'}$ gem. ▶ Gl. (7.54) in Abhängigkeit von u minimiert wird. Das Resultat lautet:

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Box 7.7

Studierende/r	1. Befragung	2. Befragung
1	260	
2	180	
3	190	
4	210	
5	80	
6	120	
7	190	
8	400	
9	170	
10	210	
11	110	150
12	0	0
13	90	130
14	180	180
15	140	150
16	220	240
17	290	300
18	190	200
19	320	350
20	380	400
21		170
22		0
23		500
24		160
25		360
26		290
27		330
28		360
29		360
30		290

Was zahlen Studierende für ihre Literatur?
Eine große Universität (die Anzahl der Studierenden liegt über 10.000) befragt eine Zufallsauswahl von 20 Studierenden nach den Ausgaben, die sie pro Semester für ihre Literatur aufbringen. Da die Buch- und Kopierpreise zunehmend steigen, entschließt man sich nach Ablauf eines Jahres erneut zu einer Umfrage. Für diese Umfragen werden 10 Adressen wieder verwendet
und weitere 10 neue Adressen zufällig ausgewählt. Die beiden Befragungen führten zu den in der Tabelle links dargestellten Euro-Angaben.

$$\sum_{i=1}^{20} x_{1i} = 3930 \quad \sum_{i=11}^{30} x_{2i} = 4920$$

$$\bar{x}_1 = 196{,}5 \quad \bar{x}_2 = 246$$

Die Studierenden 11 bis 20 wurden – wie die Tabelle zeigt – wiederholt befragt (s=10). Damit sind n_1=20, n_2=20, q=10 und u=10. Die übrigen für den Rechengang benötigten Größen lauten:

$$\bar{x}_{1q} = \frac{\sum_{i=1}^{q} x_{1i}}{q} = 201 \quad \text{(nicht wiederbefragte Studierende der 1. Erhebung)}$$

$$\bar{x}_{1s} = \frac{\sum_{i=1}^{s} x_{1i}}{s} = 192 \quad \text{(wiederbefragte Studierende der 1. Erhebung)}$$

$$\bar{x}_{2u} = \frac{\sum_{i=1}^{u} x_{2i}}{u} = 282 \quad \text{(neue Studierende der 2. Erhebung)}$$

$$\bar{x}_{2s} = \frac{\sum_{i=1}^{s} x_{2s}}{s} = 210 \quad \text{(wiederbefragte Studierende der 2. Erhebung)}$$

$$\hat{\sigma}_{1s} = \sqrt{\frac{\sum_{i=1}^{s} (x_{1i} - \bar{x}_{1s})^2}{s-1}} = 114{,}97$$

$$\hat{\sigma}_{2s} = \sqrt{\frac{\sum_{i=1}^{s} (x_{2i} - \bar{x}_{2s})^2}{s-1}} = 117{,}09$$

▼

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

$$\hat{\sigma}_{2u} = \sqrt{\frac{\sum_{i=1}^{u}(x_{2i} - \bar{x}_{2u})^2}{u-1}} = 139{,}50$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{q}(x_{1i} - \bar{x}_{1q})^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{(q-1)+(n_2-1)}$$

$$= \frac{66090 + 324480}{9+19} = 13948$$

$$b = \frac{\sum_{i=1}^{s}(x_{1i} - \bar{x}_{1s}) \cdot (x_{2i} - \bar{x}_{2s})}{\sum_{i=1}^{s}(x_{1i} - \bar{x}_{1s})^2}$$

$$= \frac{120200}{118960} = 1{,}010$$

$$r = \frac{\hat{\sigma}_{1s} \cdot b}{\hat{\sigma}_{2s}} = \frac{114{,}97 \cdot 1{,}010}{117{,}09} = 0{,}992.$$

Zunächst berechnen wir den korrigierten Mittelwert der s=10 wiederbefragten Studierenden in der zweiten Erhebung.

$$\bar{x}'_{2s} = \bar{x}_{2s} + b \cdot (\bar{x}_1 - \bar{x}_{1s})$$
$$= 210 + 1{,}010 \cdot (196{,}5 - 192)$$
$$= 210 + 4{,}545$$
$$= 214{,}545.$$

Unser nächstes Ziel ist die Berechnung des korrigierten Gesamtmittelwertes der zweiten Erhebung $\bar{x}'_2$ nach ► Gl. (7.48). Hierfür sind die folgenden Zwischenberechnungen erforderlich:

$$w_{2u} = \frac{u}{\hat{\sigma}^2} = \frac{10}{13948} = 0{,}000717$$

$$w_{2s} = \frac{1}{\frac{13948 \cdot (1 - 0{,}992^2)}{10} + 0{,}992^2 \cdot \frac{13948}{20}}$$

$$= \frac{1}{22{,}23 + 686{,}29} = 0{,}0014.$$

Damit ist

$$v = \frac{0{,}000717}{0{,}000717 + 0{,}0014} = 0{,}338.$$

Nach ► Gl. (7.48) erhalten wir für $\bar{x}'_2$:

$$\bar{x}'_2 = 0{,}338 \cdot 282 + (1 - 0{,}338) \cdot 214{,}545$$
$$= 95{,}32 + 142{,}03 = 237{,}35.$$

Für den Standardfehler dieses korrigierten Mittelwertes resultiert:

$$\hat{\sigma}_{\bar{x}'_2} = \sqrt{\frac{\hat{\sigma}^2 \cdot (n_2 - u \cdot r^2)}{n_2^2 - u^2 \cdot r^2}}$$

$$= \sqrt{\frac{13948 \cdot (20 - 10 \cdot 0{,}992^2)}{20^2 - 10^2 \cdot 0{,}992}}$$

$$= \sqrt{\frac{141702{,}75}{301{,}59}} = 21{,}68.$$

Sind die Aufwendungen für Literatur in der Population normalverteilt (diese Annahme ist nicht erforderlich, wenn – wie in vielen Fällen – $n_2 \geq 30$), lautet das 95%ige Konfidenzintervall (mit t=2,093 bei 19 Freiheitsgraden)

$$237{,}35 \pm 2{,}093 \cdot 21{,}68 = 237{,}35 \pm 45{,}38.$$

Befänden sich in der Stichprobe keine Studierenden, die zu einem früheren Zeitpunkt bereits untersucht wurden, ergäbe sich der Standardfehler gem. ► Gl. (7.5) zu

$$\hat{\sigma}_{\bar{x}_2} = \frac{\hat{\sigma}}{\sqrt{n_2}} = \frac{118{,}10}{\sqrt{20}} = 26{,}41.$$

Die Tatsache, dass zehn Studierende wiederholt befragt wurden, führt damit gem. ► Gl. (7.56a) zu einem Genauigkeitsgewinn von ca. 48%.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Tab. 7.9. Verbesserung von Schätzungen des Parameters μ (in Prozent) in Abhängigkeit vom Mischungsverhältnis wiederbefragter und neuer Untersuchungsobjekte sowie von der Höhe der Korrelation (Erläuterungen ▶ Text)

Korrelation	Anteil neuer Untersuchungsobjekte in der 2. Erhebung in Prozent										
	0	10	20	30	40	50	60	70	80	90	100
0,00	0	0	0	0	0	0	0	0	0	0	0
0,10	0	0,1	0,2	0,2	0,2	**0,3**	0,2	0,2	0,2	0,1	0
0,20	0	0,4	0,7	0,9	1,0	**1,0**	1,0	0,9	0,7	0,4	0
0,30	0	0,8	1,5	1,9	2,2	**2,4**	2,3	2,0	1,6	0,9	0
0,40	0	1,5	2,6	3,5	4,1	**4,4**	4,3	3,8	2,9	1,7	0
0,50	0	2,3	4,2	5,7	6,7	7,1	**7,1**	6,4	5,0	2,9	0
0,60	0	3,4	6,2	8,5	10,1	11,0	**11,0**	10,1	8,1	4,8	0
0,70	0	4,6	8,7	12,1	14,6	16,2	**16,7**	15,6	12,9	7,9	0
0,80	0	6,2	11,7	16,6	20,7	23,5	**24,9**	24,4	21,0	13,6	0
0,90	0	7,9	15,5	22,5	28,8	34,0	37,8	**39,3**	36,8	26,9	0
0,95	0	8,9	17,6	26,0	33,9	41,1	47,2	51,5	**51,9**	43,3	0
0,99	0	9,8	19,5	29,2	38,7	48,1	57,1	65,6	72,6	**74,8**	0

$$\frac{u}{n_2} = \frac{1}{1+\sqrt{1-r^2}}. \tag{7.55}$$

Bezogen auf das Beispiel in ▣ Box 7.7 mit $u/n_2=0{,}5$ wäre das Mischungsverhältnis

$$\frac{u_{opt}}{n_2} = \frac{1}{1+\sqrt{1-0{,}992^2}} = \frac{1}{1{,}126} \approx \frac{18}{20}$$

optimal.

Wird das optimale u in ▶ Gl. (7.54) eingesetzt, resultiert (nach einigen nicht ganz einfachen Umformungen) für den Standardfehler folgende vereinfachte Form:

$$\hat{\sigma}^2_{\text{opt}\,\bar{x}'_2} = \frac{\hat{\sigma}^2}{2\cdot n_2}\cdot\left(1+\sqrt{1-r^2}\right). \tag{7.56}$$

Unter Verwendung der optimalen Anzahl neuer Untersuchungsobjekte ($u_{opt}=18$) ergäbe sich (bei sonst gleichen Werten) folgender Standardfehler:

$$\hat{\sigma}^2_{\text{opt}\,\bar{x}'_2} = \frac{13948}{2\cdot 20}\cdot\left(1+\sqrt{1-0{,}992^2}\right) = 392{,}72$$

$$\hat{\sigma}_{\text{opt}\,\bar{x}'_2} = \sqrt{392{,}72} = 19{,}82.$$

Dieser Standardfehler mit einem optimalen Mischungsverhältnis ist kleiner als der in ▣ Box 7.7 ermittelte Standardfehler mit einem ungünstigen Mischungsverhältnis.

Verbesserung der Schätzgenauigkeit bei unterschiedlichen Mischungsverhältnissen. Aus ▣ Tab. 7.9 ist zu entnehmen, wie sich das Mischungsverhältnis aus neuen und wiederverwendeten Untersuchungsobjekten sowie die Korrelation zwischen der ersten und der zweiten Erhebung (die natürlich nur für die wiederverwendeten Untersuchungsobjekte ermittelt werden kann) auf die Präzision der Parameterschätzung auswirken.

Die in ▣ Tab. 7.9 aufgeführten Werte geben an, um wieviel Prozent die Schätzung von μ für ein bestimmtes Mischungsverhältnis und eine bestimmte Korrelation gegenüber einer einfachen Schätzung aufgrund einer Zufallsstichprobe präziser ist. Die Prozentwerte wurden nach folgender Beziehung bestimmt:

$$\text{Verbesserung}\,(\%) = \left(\frac{\hat{\sigma}^2_{\bar{x}_2}}{\hat{\sigma}^2_{\text{opt}\,\bar{x}'_2}} - 1\right)\cdot 100\%, \tag{7.56a}$$

wobei $\hat{\sigma}^2_{\bar{x}_2} = \dfrac{\hat{\sigma}^2}{n_2}$.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Der Wert 24,9% für r=0,8 und u=60% besagt beispielsweise, dass eine Stichprobe, die zu 60% aus neuen und 40% aus wiederverwendeten Untersuchungsobjekten besteht, den Parameter μ um 24,9% genauer schätzt als eine einfache Zufallsstichprobe gleichen Umfanges. Die Korrelation zwischen der ersten Messung und der zweiten Messung der 40% wiederverwendeten Untersuchungsobjekte muss hierbei r=0,8 betragen. Liegen vor der Untersuchung einigermaßen verlässliche Angaben über die Höhe der Korrelation vor, kann man der Tabelle entnehmen, wie viele neue Untersuchungsobjekte günstigerweise in die Stichprobe aufgenommen werden sollten. Weiß man beispielsweise aus vergangenen Untersuchungen, dass mit einer Korrelation von r≈0,9 zu rechnen ist, sollten etwa 70% neue Untersuchungsobjekte in die Stichprobe aufgenommen werden. (Der genaue Wert lässt sich nach ▶ Gl. 7.55 bestimmen.)

In ◘ Tab. 7.9 wird deutlich, dass der Anteil neuer Untersuchungsobjekte unabhängig von der Höhe der Korrelation niemals unter 50% liegen sollte. (Die fett gedruckten Werte geben für jede Korrelation ungefähr den maximalen Präzisionsgewinn wieder.)

Mehr als zwei Messungen. Die Ausführungen bezogen sich bis jetzt nur auf die einmalige Wiederverwendung von Untersuchungsobjekten. In der Praxis (insbesondere bei Forschungen mit einem Panel) kommt es jedoch nicht selten vor, dass Untersuchungsobjekte mehrmals wiederholt befragt werden. Im Prinzip besteht damit die Möglichkeit, für eine aktuelle Parameterschätzung die Daten mehrerer, weiter zurückliegender Untersuchungen zu berücksichtigen. In der Regel nimmt jedoch die Korrelation zweier Erhebungen mit wachsendem zeitlichen Abstand ab, sodass der Präzisionsgewinn zu vernachlässigen ist.

Unabhängig von der Höhe der Korrelation empfiehlt sich bei mehreren wiederholten Schätzungen eines Parameters eine Austauschstrategie, bei der in jeder Untersuchung ca. 50% der Untersuchungsobjekte der vorangegangenen Untersuchungen wieder verwendet werden. Cochran (1972, Kap. 12.11) zeigt, dass das optimale Verhältnis neuer und wiederverwendeter Untersuchungsobjekte im Laufe der Zeit (etwa nach der fünften Untersuchung) für beliebige Korrelationen gegen den Grenzwert von 1:1 strebt (weiterführende Literatur: Finkner & Nisselson, 1978; Patterson, 1950; Smith, 1978; Yates, 1965).

Schätzung von Populationsanteilen

Die Wiederverwendung von Untersuchungsobjekten aus vergangenen Stichprobenerhebungen kann sich auch auf die Schätzung eines aktuellen Populationsparameters π günstig auswirken. Es gelten hierfür die gleichen Prinzipien wie bei der Schätzung eines Mittelwertparameters μ. Da der Rechengang – wie ◘ Box 7.7 zeigte – bei diesem Stichprobenverfahren etwas komplizierter ist, wollen wir keine neuen Gleichungen einführen, sondern die bereits erläuterten Gleichungen analog verwenden. Hierbei machen wir wiederholt von einem Kunstgriff Gebrauch, der die kontinuierlichen Messungen x_i für Mittelwertschätzungen durch dichotome Messungen (0 und 1) ersetzt (vgl. Cochran, 1972, Kap. 3.2). Für jedes Untersuchungsobjekt i ist x_i=1, wenn es das Merkmal A aufweist. Ein Untersuchungsobjekt erhält den Wert 0, wenn es nicht durch das Merkmal A gekennzeichnet ist. Damit ergeben sich folgende Vereinfachungen:

$$\begin{aligned}
&\sum_{i=1}^{n} x_i = n_A,\\
&\sum_{i=1}^{n} x_i^2 = n_A,\\
&\left(\sum_{i=1}^{n} x_i\right)^2 = n_A^2,\\
&\bar{x} = \frac{n_A}{n} = p,\\
&\hat{\sigma} = \sqrt{\frac{n \cdot p \cdot (1-p)}{n-1}},\\
&\hat{\sigma}_p = \sqrt{\frac{p \cdot (1-p)}{n-1}}.
\end{aligned} \tag{7.57}$$

Der weitere Rechengang für die Schätzung eines Populationsanteils π sei im Folgenden an einem Beispiel verdeutlicht.

Beispiel: Die Landesregierung Bayern beauftragt ein Marktforschungsinstitut, den Bevölkerungsanteil Bayerns zu ermitteln, der im vergangenen Jahr Urlaub im Ausland machte. Das Marktforschungsinstitut be-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

fragt daraufhin eine Zufallsstichprobe des Umfanges n_2=1000. In dieser Stichprobe befinden sich 500 Personen, die bereits im letzten Jahr die gleiche Frage beantworteten. Auch in jenem Jahr befragte das Institut insgesamt n_1=1000 Personen. Die Stichprobe besteht damit aus s=500 wiederbefragten Personen und u=500 neuen Personen.

In der ersten Befragung berichteten n_{1A}=580 Personen, sie hätten ihren Urlaub im Ausland verbracht. Damit ist

$$\sum_{i=1}^{n_1} x_{1i} = n_{1A} = 580$$

bzw. $\bar{x}_1 = p_{1A} = 0{,}58$.

Der entsprechende Wert für die s=500 wiederbefragten Personen möge lauten:

$$\sum_{i=1}^{s} x_{1i} = n_{1s(A)} = 280$$

bzw. $\bar{x}_{1s} = p_{1s(A)} = 0{,}56$.

Für diejenigen Personen, die nicht wiederholt befragt wurden, folgt daraus

$$\sum_{i=1}^{q} x_{1i} = n_{1q(A)} = 300$$

bzw. $\bar{x}_{1q} = p_{1q(A)} = 0{,}60$.

Für die zweite Befragung ermittelt man

$$\sum_{i=1}^{n_2} x_{2i} = n_{2A} = 500$$

bzw. $\bar{x}_2 = p_{2A} = 0{,}50$

$$\sum_{i=1}^{s} x_{2i} = n_{2s(A)} = 240$$

bzw. $\bar{x}_{2s} = p_{2s(A)} = 0{,}48$ und

$$\sum_{i=1}^{u} x_{2i} = n_{2u(A)} = 260$$

bzw. $\bar{x}_{2u} = p_{2u(A)} = 0{,}52$.

Von den 280 wiederbefragten Personen, die bei der ersten Erhebung die Frage nach einem Auslandsurlaub bejahten, gaben 200 an, sie hätten auch im letzten Jahr ihren Urlaub im Ausland verbracht.

$$\sum_{i=1}^{s} x_{1i} \cdot x_{2i} = n_{12s(A)} = 200.$$

Die für den weiteren Rechengang benötigten Standardabweichungen haben folgende Werte:

$$\hat{\sigma}_{1s} = \sqrt{\frac{s \cdot p_{1s(A)} \cdot (1 - p_{1s(A)})}{s-1}} = \sqrt{\frac{500 \cdot 0{,}56 \cdot 0{,}44}{499}} = 0{,}497,$$

$$\hat{\sigma}_{2s} = \sqrt{\frac{s \cdot p_{2s(A)} \cdot (1 - p_{2s(A)})}{s-1}} = \sqrt{\frac{500 \cdot 0{,}48 \cdot 0{,}52}{499}} = 0{,}500,$$

$$\hat{\sigma}_{2u} = \sqrt{\frac{u \cdot p_{2u(A)} \cdot (1 - p_{2u(A)})}{u-1}} = \sqrt{\frac{500 \cdot 0{,}52 \cdot 0{,}48}{499}} = 0{,}500,$$

$$\hat{\sigma} = \sqrt{\frac{q \cdot p_{1q(A)} \cdot (1 - p_{1q(A)}) + n \cdot p_{2A} \cdot (1 - p_{2A})}{(q-1)+(n-1)}} = \sqrt{\frac{500 \cdot 0{,}60 \cdot 0{,}40 + 1000 \cdot 0{,}50 \cdot 0{,}50}{499 + 999}} = 0{,}499.$$

Für b und r erhalten wir nach den Gleichungen (7.47) und (7.53):

$$b = \frac{s \cdot \sum_{i=1}^{s} x_{1i} \cdot x_{2i} - \sum_{i=1}^{s} x_{1i} \cdot \sum_{i=1}^{s} x_{2i}}{s \cdot \sum_{i=1}^{s} x_{1i}^2 - \left(\sum_{i=1}^{s} x_{1i}\right)^2} = \frac{500 \cdot 200 - 280 \cdot 240}{500 \cdot 280 - 280^2} = 0{,}532$$

$$r = \frac{\hat{\sigma}_{1s} \cdot b}{\hat{\sigma}_{2s}} = \frac{0{,}497 \cdot 0{,}532}{0{,}500} = 0{,}529.$$

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Der korrigierte Anteil derjenigen wiederbefragten Personen, die bei der zweiten Erhebung die Frage bejahten ($p'_{2s(A)}$), lautet damit

$$\begin{aligned}\bar{x}'_{2s} &= p'_{2s(A)} = \bar{x}_{2s} + b \cdot (\bar{x}_1 - \bar{x}_{1s}) \\ &= p_{2s(A)} + b \cdot (p_{1A} - p_{1s(A)}) \\ &= 0{,}48 + 0{,}532 \cdot (0{,}58 - 0{,}56) \\ &= 0{,}49.\end{aligned}$$

Im Folgenden berechnen wir die korrigierte Parameterschätzung aufgrund der zweiten Untersuchung (p'_{2A} nach ▶ Gl. 7.48). Die hierfür benötigten Zwischengrößen (▶ Gl. 7.49 bis 7.52) lauten

$$w_{2u} = \frac{u}{\hat{\sigma}^2} = \frac{500}{0{,}499^2} = 2008{,}02$$

$$\begin{aligned}w_{2s} &= \frac{1}{\frac{\hat{\sigma}^2 \cdot (1 - r^2)}{s} + r^2 \cdot \frac{\hat{\sigma}^2}{n_2}} \\ &= \frac{1}{\frac{0{,}499^2 \cdot (1 - 0{,}529^2)}{500} + 0{,}529^2 \cdot \frac{0{,}499^2}{1000}} \\ &= 2334{,}70\end{aligned}$$

$$v = \frac{w_{2u}}{w_{2u} + w_{2s}} = 0{,}462.$$

Eingesetzt in ▶ Gl. (7.48) ergibt sich

$$\begin{aligned}\bar{x}'_2 &= p'_{2(A)} = v \cdot \bar{x}_{2u} + (1 - v) \cdot \bar{x}'_{2s} \\ &= v \cdot p_{2u(A)} + (1 - v) \cdot p'_{2s(A)} \\ &= 0{,}462 \cdot 0{,}52 + (1 - 0{,}462) \cdot 0{,}49 \\ &= 0{,}504.\end{aligned}$$

Als Standardfehler ermitteln wir nach ▶ Gl. (7.54)

$$\begin{aligned}\hat{\sigma}_{\bar{x}'_2} &= \hat{\sigma}_{p'_{2A}} = \sqrt{\frac{\hat{\sigma}^2 \cdot (n_2 - u \cdot r^2)}{n_2^2 - u^2 \cdot r^2}} \\ &= \sqrt{\frac{0{,}499^2 \cdot (1000 - 500 \cdot 0{,}529^2)}{1000^2 - 500^2 \cdot 0{,}529^2}} \\ &= 0{,}01517.\end{aligned}$$

Hieraus folgt für das 95%ige Konfidenzintervall (mit z±1,96)

$$0{,}504 \pm 1{,}96 \cdot 0{,}01517 = 0{,}504 \pm 0{,}0297.$$

Das Konfidenzintervall, in dem sich diejenigen Anteilsparameter befinden, die den Stichprobenkennwert p=0,504 mit 95%iger Wahrscheinlichkeit »erzeugt« haben können, hat die Grenzen 0,474 und 0,534.

Wiederum soll vergleichend derjenige Standardfehler bestimmt werden, der sich ergeben hätte, wenn keine Person wiederholt befragt, sondern eine einfache Zufallsstichprobe mit n=1000 gezogen worden wäre. Hätten in dieser Stichprobe ebenfalls n_A=500 Personen die Frage bejaht, ergäbe sich nach ▶ Gl. (7.57) als Standardfehler

$$\hat{\sigma}_p = \sqrt{\frac{p \cdot (1 - p)}{n - 1}} = \sqrt{\frac{0{,}5 \cdot 0{,}5}{999}} = 0{,}01582.$$

Die Schätzung wird also nur unwesentlich schlechter, wenn eine reine Zufallsstichprobe verwendet wird – ein Befund, der in diesem Beispiel vor allem auf die mäßige Korrelation zwischen der ersten und der zweiten Befragung der 500 wiederbefragten Personen (r=0,529) zurückgeht.

7.2.5 Der Bayes'sche Ansatz

Die Genauigkeit von Parameterschätzungen, die man mit einfachen Zufallsstichproben erzielt, lässt sich – so zeigten die letzten vier Abschnitte – verbessern,

- wenn Merkmale bekannt sind, die mit der untersuchten Variablen zusammenhängen (geschichtete Stichprobe, ▶ Abschn. 7.2.1),
- wenn sich die Population aus vielen homogenen Teilgesamtheiten zusammensetzt, von denen jede zufällig ausgewählte Teilgesamtheit vollständig untersucht wird (Klumpenstichprobe, Abschn. ▶ 7.2.2)
- wenn sich die Population aus großen Teilgesamtheiten zusammensetzt und jede zufällig ausgewählte Teilgesamtheit stichprobenartig untersucht wird (mehrstufige Stichprobe, ▶ Abschn. 7.2.3)
- oder wenn es möglich ist, einige Untersuchungsobjekte mehrmals in eine Stichprobe einzubeziehen (wiederholte Stichprobenuntersuchungen, ▶ Abschn. 7.2.4).

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Bei diesen Ansätzen regulieren die Kenntnisse, über die man bereits vor Durchführung der Untersuchung verfügt, die Art der zu erhebenden Stichprobe. Die Qualität der Parameterschätzung hängt davon ab, ob es gelingt, einen **Stichprobenplan** aufzustellen, der möglichst viele Vorkenntnisse berücksichtigt.

Anders funktioniert die Nutzung von Vorinformationen nach dem Bayes'schen Ansatz: Dieser Ansatz vereinigt das Vorwissen bzw. die Erwartung der Forschenden über mögliche Untersuchungsergebnisse und die tatsächlichen Ergebnisse einer beliebigen Stichprobenuntersuchung zu einer gemeinsamen Schätzung des unbekannten Populationsparameters. Vorwissen und Stichprobenergebnis sind zwei voneinander unabhängige Informationsquellen, die zumindest formal als zwei gleichwertige Bestimmungsstücke der Parameterschätzung genutzt werden.

Allerdings erfordert der Bayes'sche Ansatz – wenn er zu einer substantiellen Verbesserung der Parameterschätzung führen soll – sehr spezifische Kenntnisse über den Untersuchungsgegenstand. Sie beziehen sich nicht auf Merkmale, die mit der untersuchten Variablen zusammenhängen, oder auf sonstige Besonderheiten der Zusammensetzung der Population, sondern betreffen direkt den zu schätzenden Parameter. Man muss also in der Lage sein, Angaben über die mutmaßliche Größe des gesuchten Parameters zu machen.

Bereits an dieser Stelle sei auf eine Besonderheit des Bayes'schen Ansatzes gegenüber den bisher behandelten »klassischen« Parameterschätzungen hingewiesen. Die klassische Parameterschätzung geht davon aus, dass der unbekannte Parameter irgendeinen bestimmten Wert aufweist und dass bei gegebenem Parameter verschiedene Stichprobenergebnisse (wie z. B. Stichprobenmittelwerte) unterschiedlich wahrscheinlich sind (vgl. hierzu die Ausführungen auf ▶ S. 410 ff.).

Der Bayes'sche Ansatz argumentiert hier anders. Er behauptet nicht, dass der Parameter einen bestimmten Wert aufweist, sondern behandelt den Parameter als eine **Zufallsvariable**. Dies hat zur Folge, dass mehrere Schätzungen des Parameters möglich sind und dass diese Schätzungen (meistens) unterschiedlich sicher oder »glaubwürdig« sind. Im Unterschied zur Wahrscheinlichkeit als **relative Häufigkeit** verwendet der Bayes'sche Ansatz **subjektive Wahrscheinlichkeiten**, die den Grad der inneren Überzeugung von der Richtigkeit einer Aussage bzw. deren Glaubwürdigkeit kennzeichnen. Auch dieser Sachverhalt wird – zumindest umgangssprachlich – meistens als »Wahrscheinlichkeit« (bzw. »subjektive Wahrscheinlichkeit«) bezeichnet. Wir wollen diesem begrifflichen Unterschied zukünftig dadurch Rechnung tragen, dass wir den Bayes'schen »Wahrscheinlichkeits«-Begriff in Anführungszeichen setzen.

! Bei Parameterschätzungen nach dem Bayes'schen Ansatz werden Stichprobeninformationen und das Vorwissen der Forscher integriert.

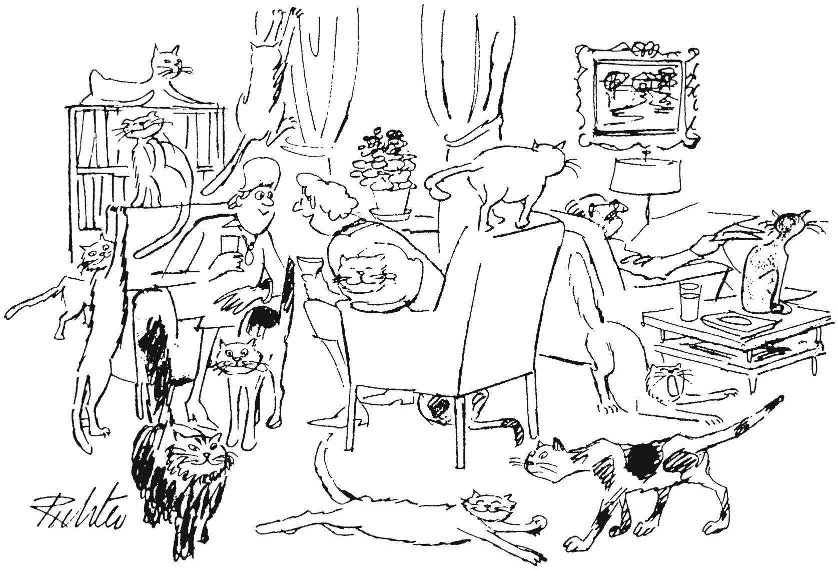
Für die Forschungspraxis bedeutet dies, dass man, um die Vorteile des Bayes'schen Ansatzes nutzen zu können, nicht nur Vorstellungen über denjenigen Parameter haben muss, der am »wahrscheinlichsten« erscheint, sondern dass man zusätzlich Angaben darüber machen muss, für wie »wahrscheinlich« oder glaubwürdig man alle übrigen denkbaren Ausprägungen des Parameters hält. Kurz formuliert: Man muss bei diskreten Zufallsvariablen Informationen über die Wahrscheinlichkeitsfunktion des Parameters und bei stetigen Zufallsvariablen Informationen über dessen Dichtefunktion (vgl. ▫ Box 7.2) haben.

Ein Beispiel soll die Erfordernisse, die mit einer Parameterschätzung nach dem Bayes'schen Ansatz verbunden sind, verdeutlichen. Eine Studentin möge sich für die Frage interessieren, wie viele Semester Psychologiestudenten durchschnittlich bis zum Abschluss ihres Studiums benötigen. Nach Durchsicht einiger Studien- und Prüfungsordnungen hält sie 10 Semester als durchschnittliche Studienzeit für am plausibelsten. Dass die durchschnittliche Studienzeit weniger als 8 Semester und mehr als 12 Semester betragen könnte, kommt für sie nicht in Betracht. Die »Wahrscheinlichkeit«, dass eine dieser extremen Semesterzahlen dem wahren Durchschnitt entspricht, wird null gesetzt. Für die übrigen Semester erscheinen folgende »Wahrscheinlichkeitsangaben« realistisch:

$$p(\mu_1 = 8\,\text{Semester}) = 2\%,$$
$$p(\mu_2 = 9\,\text{Semester}) = 18\%,$$
$$p(\mu_3 = 10\,\text{Semester}) = 60\%,$$
$$p(\mu_4 = 11\,\text{Semester}) = 18\%,$$
$$p(\mu_5 = 12\,\text{Semester}) = 2\%.$$

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Schätzfehler sind im Alltag verbreiteter als man meint. Aus *The New Yorker*: Die schönsten Katzen-Cartoons (1993). München: Knaur, S. 102–103



»Wir haben jetzt vierzehn, aber Kevin glaubt immer noch, es wären nur zwölf.«

Damit ist die Wahrscheinlichkeitsfunktion hinreichend spezifiziert, um im weiteren den gesuchten Parameter nach dem Bayes'schen Ansatz schätzen zu können. Diese Schätzung berücksichtigt neben den subjektiven Informationen das Ergebnis einer Stichprobenuntersuchung, für die alle bisher behandelten Stichprobenpläne in Frage kommen. (Techniken zur Spezifizierung des Vorwissens werden bei Molenaar und Lewis, 1996, behandelt.)

Man wird sich vielleicht fragen, ob diese Art der Berücksichtigung subjektiver Überzeugungen wissenschaftlich zu rechtfertigen ist. Sind dadurch, dass man mehr oder weniger gesicherte subjektive Überzeugungen in die Parameterschätzung einfließen lässt, der Willkür nicht Tür und Tor geöffnet?

Dass diese Bedenken nur teilweise berechtigt sind, wird deutlich, wenn man sich vergegenwärtigt, dass viele Untersuchungen von Fachleuten durchgeführt werden, die aufgrund ihrer Erfahrungen durchaus in der Lage sind, bereits vor Durchführung der Untersuchung realistische Angaben über den Ausgang der Untersuchung zu machen. Diese Kenntnisse bleiben üblicherweise bei Stichprobenuntersuchungen ungenutzt.

Auf der anderen Seite trägt der Bayes'sche Ansatz auch der Tatsache Rechnung, dass bei einzelnen Fragestellungen nur sehr vage Vorstellungen über das untersuchte Merkmal vorhanden sind. Für diese Fälle sind Vorkehrungen getroffen, die dafür sorgen, dass wenig gesicherte subjektive Vorinformationen die Parameterschätzung nur geringfügig oder auch gar nicht beeinflussen. Die Parameterschätzung nach dem Bayes'schen Ansatz geht dann in eine »klassische«, ausschließlich stichprobenabhängige Parameterschätzung über.

Diese Flexibilität des Bayes'schen Ansatzes bietet grundsätzlich die Möglichkeit, eine Parameterschätzung mit Berücksichtigung subjektiver Informationen einer Parameterschätzung ohne subjektive Informationsanteile gegenüberzustellen, sodass die Beeinflussung der Parameterschätzung durch die subjektiven Informationen transparent wird. Hierauf wird später ausführlich einzugehen sein.

Zuvor jedoch sollen diejenigen Bestandteile des Bayes'schen Ansatzes erläutert werden, die zum Verständnis von Parameterschätzungen erforderlich sind. Ausführlichere, über die Erfordernisse von Parameterschätzungen hinausgehende Darstellungen des Bayes'schen Ansatzes findet man z. B. bei Berger (1980), Edwards et al. (1963), Koch (2000), Lindley (1965), Philips (1973), Schmitt (1969) und Winkler (1972, 1993).

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Skizze der Bayes'schen Argumentation

Das Kernstück der Bayes'schen Statistik geht auf den englischen Pfarrer Thomas Bayes zurück, der seine Ideen im Jahre 1763 veröffentlichte. Zur Erläuterung dieses Ansatzes greifen wir ein Beispiel von Hays und Winkler (1970) auf, das für unsere Zwecke geringfügig modifiziert wurde.

Beispiel: Jemand wacht nachts mit starken Kopfschmerzen auf. Schlaftrunken geht die Person zum Medizinschrank und nimmt eine von drei sehr ähnlich aussehenden Flaschen heraus, in der festen Überzeugung, die Flasche enthalte Aspirin. Nachdem sie sich mit Tabletten dieser Flasche versorgt hat, stellen sich bald heftige Beschwerden in Form von Übelkeit und Erbrechen ein. Daraufhin überprüft die Person erneut den Medizinschrank und stellt fest, dass sich unter den drei Flaschen eine Flasche mit einer giftigen Substanz befindet. Die übrigen beiden Flaschen enthalten Aspirin. Sie kann sich nicht mehr daran erinnern, ob sie irrtümlicherweise zur falschen Flasche gegriffen hat. Es stellt sich damit die Frage, ob die Beschwerden (B) auf das eventuell eingenommene Gift (G) zurückzuführen sind, oder ob das Aspirin diese Nebenwirkungen verursachte.

Soweit die Situation, die wir nun formalisieren wollen. Hierbei gehen wir von folgenden Annahmen aus: Die Wahrscheinlichkeit, dass die Person fälschlicherweise zur Giftflasche gegriffen hat, lautet bei drei mit gleicher Wahrscheinlichkeit in Frage kommenden Flaschen

$$p(G) = 0{,}33.$$

Die Wahrscheinlichkeit, dass die ausgewählte Flasche kein Gift, sondern Aspirin enthielt [$p(\overline{G})$ lies: Wahrscheinlichkeit für »non G«] beträgt also

$$p(\overline{G}) = 0{,}67.$$

Ferner sei bekannt, mit welcher Wahrscheinlichkeit das Gift zu den oben genannten Beschwerden führt.

$$p(B|G) = 0{,}80.$$

Dies ist eine **bedingte Wahrscheinlichkeit**; $p(B|G)$ (lies: die Wahrscheinlichkeit von B unter der Voraussetzung, dass G eingetreten ist) symbolisiert die Wahrscheinlichkeit von Beschwerden für den Fall, dass Gift eingenommen wurde. Die gleichen Beschwerden können jedoch auch bei Einnahme von Aspirin auftreten. Für dieses Ereignis sei in pharmazeutischen Handbüchern der Wert

Tab. 7.10. Vierfeldertafel zur Herleitung des Bayes'schen Theorems

	Gift (G)	Kein Gift ($\overline{G}$)	
Beschwerden (B)	80	10	90
Keine Beschwerden ($\overline{B}$)	20	190	210
	100	200	300

$$p(B|\overline{G}) = 0{,}05$$

aufgeführt.

Gesucht wird nun die Wahrscheinlichkeit $p(G|B)$, also die Wahrscheinlichkeit, dass Gift eingenommen wurde, wenn die einschlägigen Beschwerden vorliegen. Die Ermittlung dieser Wahrscheinlichkeit ist für den Fall, dass nur die genannten Wahrscheinlichkeitswerte bekannt sind, mit Hilfe des Bayes'schen Theorems möglich.

Für das Verständnis der folgenden Ausführungen ist es hilfreich, wenn man sich den Unterschied zwischen den bedingten Wahrscheinlichkeiten $p(G|B)$ und $p(B|G)$ klar macht: Die Wahrscheinlichkeit, Gift eingenommen zu haben, wenn man Beschwerden hat, ist nicht zu verwechseln mit der Wahrscheinlichkeit von Beschwerden, wenn man Gift eingenommen hat!

Das Bayes'sche Theorem. Für die Herleitung des Bayes'schen Theorems nehmen wir zunächst an, die Vierfeldertafel in Tab. 7.10 sei vollständig bekannt. 300 Personen, von denen 100 Gift und 200 kein Gift (sondern Aspirin) einnahmen, wurden untersucht (pragmatische und ethische Gründe ließen eine derartige Untersuchung freilich nur als Gedankenexperiment zu). Damit schätzen wir – wie vorgesehen – $p(G)$ mit $100{:}300 = 0{,}33$ und $p(\overline{G})$ mit $200{:}300 = 0{,}67$. Für die Ermittlung der Wahrscheinlichkeit $p(B|G)$ betrachten wir nur die 100 Fälle, die Gift eingenommen haben. Von denen haben 80 Beschwerden, d. h. $p(B|G) = 80{:}100 = 0{,}8$. Entsprechend ergibt sich für $p(B|\overline{G}) = 10{:}200 = 0{,}05$.

Mit den Informationen aus Tab. 7.10 bereitet die Bestimmung der gesuchten Wahrscheinlichkeit $p(G|B)$

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

keine Schwierigkeiten. Wir betrachten nur die 90 Fälle mit Beschwerden und stellen fest, dass hiervon 80 Gift einnahmen, d. h. $p(G|B)=80{:}90=0{,}89$. Wie aber lässt sich diese Wahrscheinlichkeit bestimmen, wenn die Vierfeldertafel nicht vollständig bekannt ist, sondern nur die drei oben genannten Wahrscheinlichkeiten?

Der folgende Gedankengang führt zu der gesuchten Berechnungsvorschrift. Wir betrachten zunächst die Beziehung:

$$p(B|G)=\frac{p(B \text{ und } G)}{p(G)}. \tag{7.58}$$

Die Wahrscheinlichkeit von Beschwerden unter der Voraussetzung der Gifteinnahme entspricht der Wahrscheinlichkeit des gemeinsamen Ereignisses »Beschwerden und Gifteinnahme« $p(B \text{ und } G)=80{:}300$, dividiert durch die Wahrscheinlichkeit für das Ereignis »Gifteinnahme« $p(G)=100{:}300$ (vgl. z. B. Bortz, 2005, Kap. 2.1.2). Für diesen Quotienten resultiert der bereits bekannte Wert von 0,8 (80:300/100:300=0,8).

Für $p(G|B)$ schreiben wir entsprechend

$$p(G|B)=\frac{p(B \text{ und } G)}{p(B)}. \tag{7.59}$$

Aus ► Gl. (7.58) und ► Gl. (7.59) folgt

$$p(B|G)\cdot p(G)=p(G|B)\cdot p(B) \tag{7.60}$$

und

$$p(G|B)=\frac{p(B|G)\cdot p(G)}{p(B)}. \tag{7.61}$$

Sind nicht nur die Wahrscheinlichkeiten $p(B|G)$ und $p(G)$ bekannt, sondern auch die Wahrscheinlichkeit $p(B)$ (d. h. die Wahrscheinlichkeit für das Ereignis »Beschwerden«), dienen ► Gl. (7.61) und ► Gl. (7.59) zur Ermittlung der gesuchten Wahrscheinlichkeit $p(G|B)$. Schätzen wir $p(B)$ nach ◘ Tab. 7.10 mit $p(B)=90{:}300=0{,}3$, resultiert für $p(G|B)$ der bereits bekannte Wert:

$$p(G|B)=\frac{0{,}8\cdot 0{,}33}{0{,}3}=0{,}89.$$

Nun hatten wir jedoch eingangs nicht vereinbart, dass die Wahrscheinlichkeit $p(B)$ bekannt sei. Wie man sich jedoch anhand der Vierfeldertafel in ◘ Tab. 7.10 leicht überzeugen kann, ist $p(B)$ bestimmbar durch

$$p(B)=p(B \text{ und } G)+p(B \text{ und } \bar{G}). \tag{7.61a}$$

Die Wahrscheinlichkeit des Ereignisses »Beschwerden« $[p(B)=90{:}300]$ ist gleich der Summe der Wahrscheinlichkeiten für die gemeinsamen Ereignisse »Beschwerden und Gift« $[p(B \text{ und } G)= 80{:}300]$ und »Beschwerden und kein Gift« $[p(B \text{ und } \bar{G}=10{:}300]$. Die Ausdrücke $p(B$ und $G)$ sowie $p(B$ und $\bar{G})$ ersetzen wir unter Bezugnahme auf ► Gl. (7.58) durch

$$p(B \text{ und } G)=p(B|G)\cdot p(G),$$

und

$$p(B \text{ und } \bar{G})=p(B|\bar{G})\cdot p(\bar{G}),$$

d. h., wir erhalten

$$p(B)=p(B|G)\cdot p(G)+p(B|\bar{G})\cdot p(\bar{G}). \tag{7.61b}$$

Ersetzen wir $p(B)$ in ► Gl. (7.61) durch ► Gl. (7.61b), resultiert das **Bayes'sche Theorem**.

$$p(G|B)=\frac{p(B|G)\cdot p(G)}{p(B|G)\cdot p(G)+p(B|\bar{G})\cdot p(\bar{G})}. \tag{7.62}$$

Für diese Bestimmungsgleichung sind alle benötigten Wahrscheinlichkeiten bekannt. Mit den eingangs genannten Werten erhalten wir für das Ereignis $p(G|B)$ (»Gifteinnahme unter der Voraussetzung von Beschwerden«) erneut den Wert 0,89.

$$p(G|B)=\frac{0{,}80\cdot 0{,}33}{0{,}80\cdot 0{,}33+0{,}05\cdot 0{,}67}=0{,}89.$$

Die Wahrscheinlichkeit für das Ereignis $p(\bar{G}|B)$ (»keine Gifteinnahme unter der Voraussetzung von Beschwerden«) ergibt sich als Komplementärwahrscheinlichkeit einfach zu

$$p(\bar{G}|B)=1-p(G|B)=1-0{,}89=0{,}11.$$

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Diese bedingte Wahrscheinlichkeit besagt, dass mit einer Wahrscheinlichkeit von 0,11 kein Gift eingenommen wurde, wenn Beschwerden vorliegen. Dieser Wert lässt sich anhand von Tab. 7.10 ebenfalls einfach nachvollziehen:

$$p(\overline{G}|B) = 10 : 90 = 0{,}11.$$

(Ein weiteres Beispiel, das die medizinischen Begriffe »Sensitivität« und »Spezifität« unter Bezugnahme auf das Bayes'sche Theorem erklärt, findet man z. B. bei Bortz, 2005, S. 58, oder auch bei Bortz & Lienert, 2003, Kap. 5.1.2.)

Diskrete Zufallsvariablen

Die hier geprüften Beschwerdeursachen können durch andere Ursachen ergänzt werden: übermäßiger Alkoholgenuß, Darminfektion, verdorbene Nahrungsmittel etc. Bezeichnen wir die möglichen Ursachen (einschließlich Gift und einer Restkategorie »Ursache unbekannt«) allgemein mit A_i (i=1,2…k), führt dies zu der folgenden verallgemeinerten Version des Bayes'schen Theorems:

$$p(A_i|B) = \frac{p(B|A_i) \cdot p(A_i)}{\sum_{i=1}^{k} p(B|A_i) \cdot p(A_i)}. \tag{7.63}$$

Im einleitenden Beispiel waren A_1=G und $A_2=\overline{G}$. Aus Gl. (7.63) ergibt sich die Wahrscheinlichkeit einer Ursache A_i, wenn Beschwerden registriert werden. Sie setzt voraus, dass nicht nur die Wahrscheinlichkeiten $p(A_i)$ für die einzelnen Ursachen bekannt sind, sondern auch die Werte $p(B|A_i)$, d. h. die Wahrscheinlichkeiten, mit denen Beschwerden auftreten, wenn die einzelnen Ursachen A_i zutreffen. Kurz: Es muss sowohl die Wahrscheinlichkeitsverteilung der diskreten Zufallsvariablen »Beschwerdeursachen« als auch die bedingte Wahrscheinlichkeitsverteilung des Ereignisses »Beschwerden unter der Bedingung aller möglichen Beschwerdeursachen« bekannt sein. Kennt man diese Wahrscheinlichkeiten, lässt sich mit Hilfe des Bayes'schen Theorems die bedingte Wahrscheinlichkeitsverteilung für die Beschwerdeursachen bestimmen, d. h. die Wahrscheinlichkeitsverteilung aller möglicher Beschwerdeursachen für den Fall, dass Beschwerden vorliegen.

An dieser Stelle wollen wir uns vom Beispiel lösen und versuchen, die Tragweite dieses Ansatzes auszuloten.

Gegeben sei eine (vorerst diskrete) Zufallsvariable $X=A_i$, über deren Wahrscheinlichkeitsverteilung $p(X=A_i)$ mehr oder weniger präzise Vorstellungen bestehen mögen. Diese Verteilung repräsentiert die sog. **Priorverteilung** von X. Es wird eine empirische Untersuchung durchgeführt, deren Resultat wir mit B bezeichnen wollen. Die bedingte Wahrscheinlichkeit $p(B|A_i)$ bzw. die Wahrscheinlichkeit dafür, dass B eintritt, wenn die Zufallsvariable X den Wert A_i annimmt, sei ebenfalls bekannt (oder – wie wir später sehen werden – errechenbar). Das Bayes'sche Theorem gem. Gl. (7.63) korrigiert nun die Priorverteilung $p(X=A_i)$ [oder kurz: $p(A_i)$] angesichts der Tatsache, dass B eingetreten ist. Die resultierenden Wahrscheinlichkeiten $p(X=A_i|B)$ [kurz: $p(A_i|B)$] konstituieren die sog. **Posteriorverteilung** der Zufallsvariablen X.

Beispiel: Ein Beispiel (nach Winkler, 1972, S. 44) soll diesen verallgemeinerten Ansatz verdeutlichen. Bei einem zufällig ausgewählten Bewohner einer Stadt mögen anlässlich einer Röntgenuntersuchung Schatten auf der Lunge (positiver Röntgenbefund = B) festgestellt worden sein. Einfachheitshalber nehmen wir an, dass dies ein Zeichen für Lungenkrebs (A_1) oder für Tuberkulose (A_2) sein kann, bzw. dass die Schatten im Röntgenbild durch keine der beiden Krankheiten (A_3) verursacht werden. (Der Fall, dass die Person sowohl Lungenkrebs als auch Tuberkulose hat, wird hier ausgeschlossen.) Ferner möge man die folgenden Wahrscheinlichkeiten kennen:

1. Die Wahrscheinlichkeit eines positiven Röntgenbefundes bei Personen mit Lungenkrebs: $p(B|A_1)=0{,}90$.
2. Die Wahrscheinlichkeit eines positiven Röntgenbefundes bei Personen mit Tuberkulose: $p(B|A_2)=0{,}95$.
3. Die Wahrscheinlichkeit eines positiven Röntgenbefundes bei Personen, die weder Lungenkrebs noch Tuberkulose haben: $p(B|A_3)=0{,}07$.
4. In der Stadt, in der die Untersuchung durchgeführt wurde, haben 2% aller Bewohner Lungenkrebs: $p(A_1)=0{,}02$.
5. In der Stadt, in der die Untersuchung durchgeführt wurde, haben 1% aller Bewohner Tuberkulose: $p(A_2)=0{,}01$.
6. In der Stadt, in der die Untersuchung durchgeführt wurde, haben 97% weder Lungenkrebs noch Tuberkulose: $p(A_3)=0{,}97$.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Die unter den Punkten 4–6 genannten Wahrscheinlichkeiten stellen die Priorwahrscheinlichkeiten dar. Die Wahrscheinlichkeiten, dass die untersuchte Person mit positivem Röntgenbefund Lungenkrebs, Tuberkulose oder keine dieser beiden Krankheiten hat (Posteriorwahrscheinlichkeit), lassen sich unter Verwendung von ▶ Gl. (7.63) in folgender Weise berechnen:

$$p(A_1|B) = \frac{0{,}90 \cdot 0{,}02}{0{,}90 \cdot 0{,}02 + 0{,}95 \cdot 0{,}01 + 0{,}07 \cdot 0{,}97}$$
$$= \frac{0{,}018}{0{,}0954} = 0{,}1887,$$
$$p(A_2|B) = \frac{0{,}95 \cdot 0{,}01}{0{,}0954} = 0{,}0996,$$
$$p(A_3|B) = \frac{0{,}07 \cdot 0{,}97}{0{,}0954} = 0{,}7117.$$

Die Summe der Posteriorwahrscheinlichkeiten ergibt – wie auch die Summe der Priorwahrscheinlichkeiten – den Wert 1.

Vor der Röntgenuntersuchung betrug das Risiko, an Lungenkrebs erkrankt zu sein, 2% und das Risiko, Tuberkulose zu haben, 1%. Nach dem positiven Röntgenbefund erhöhen sich diese Wahrscheinlichkeiten auf ca. 19% für Lungenkrebs und auf ca. 10% für Tuberkulose. Für die Wahrscheinlichkeit, dass der Röntgenbefund bedeutungslos ist, verbleiben damit ca. 71%.

Es ist offenkundig, dass die Posteriorwahrscheinlichkeiten nur dann verlässlich sind, wenn sowohl für die Priorwahrscheinlichkeiten $p(A_i)$ als auch für die bedingten Wahrscheinlichkeiten $p(B|A_i)$ brauchbare Schätzungen vorliegen. Auf eine Schätzung der bedingten Wahrscheinlichkeiten $p(B|A_i)$ kann man indes verzichten, wenn die untersuchte Zufallsvariable einem mathematisch bekannten Verteilungsmodell folgt (z. B. Binomialverteilung oder Poisson-Verteilung). Die Priorverteilung $p(A_i)$ spezifiziert dann Annahmen über die Art des Verteilungsmodells (z. B. Binomialverteilung mit den Parametern π=0,15, π=0,20 etc.). Erhält man aufgrund einer Untersuchung zusätzlich einen empirischen Schätzwert B für den unbekannten Parameter (z. B. p=0,12), können die Likelihoods dieses Schätzwertes bei Gültigkeit der möglichen Parameter ermittelt werden (▶ S. 407 ff.). Mit Hilfe des Bayes'schen Theorems gem. ▶ Gl. (7.63) wird dann die Priorverteilung im Hinblick auf das empirische Ergebnis B korrigiert, d. h., man erhält die Posteriorverteilung $p(A_i|B)$, mit der die »Wahrscheinlichkeits«-Verteilung der Zufallsvariablen bei gegebenem B charakterisiert wird. Die folgenden Beispiele verdeutlichen diese Anwendungsvariante.

Binomialverteilung. Eine Berufsberaterin möchte wissen, wieviel Prozent eines Abiturientenjahrganges sich für das Studienfach Psychologie interessieren. Als Prozentzahlen (Parameterschätzungen) kommen für sie nur die folgenden Werte in Frage: π_1=1%, π_2=3%, π_3=8% und π_4=15%. (Die π_i-Werte ersetzen die A_i-Werte in ▶ Gl. 7.63). Das Beispiel ist natürlich unrealistisch, da angenommen wird, dass alle übrigen Prozentzahlen überhaupt nicht zutreffen können. Richtiger wäre es, kontinuierliche Prozentzahlen anzunehmen. Wir müssen diese Variante jedoch bis zur Behandlung stetiger Variablen zurückstellen (▶ S. 467 ff.). Allerdings erscheinen der Berufsberaterin nicht alle Prozentzahlen gleich »wahrscheinlich«. Ausgehend von ihrer Erfahrung ordnet sie den vier Parameterschätzungen folgende »Wahrscheinlichkeiten« zu:

$$p(\pi_1 = 1\%) \;\; = 0{,}30,$$
$$p(\pi_2 = 3\%) \;\; = 0{,}50,$$
$$p(\pi_3 = 8\%) \;\; = 0{,}15,$$
$$p(\pi_4 = 15\%) = 0{,}05.$$

Dies ist die Priorverteilung der Berufsberaterin. Nach ihrer Ansicht sind 3% Psychologieaspiranten unter den Abiturienten am »wahrscheinlichsten« und 15% am »unwahrscheinlichsten«. Der Erwartungswert der Priorverteilung lautet

$$E(\pi_i) = \sum_{i=1}^{k} p_i \cdot \pi_i$$
$$= 0{,}30 \cdot 0{,}01 + 0{,}50 \cdot 0{,}03 + 0{,}15 \cdot 0{,}08 + 0{,}05 \cdot 0{,}15$$
$$= 0{,}0375 \text{ bzw. } 3{,}75\%.$$

Die Berufsberaterin befragt nun eine Zufallsstichprobe von 20 Abiturienten hinsichtlich ihrer Berufswünsche. Eine der Befragten beabsichtigt, Psychologie zu studieren, d. h., die Stichprobe führt zu dem Resultat B=5% (oder 0,05). Mit Hilfe des Bayes'schen Theorems gem. ▶ Gl. (7.63) lässt sich nun eine Posteriorverteilung berechnen, der zu entnehmen ist, welche der vier Parameterschätzungen bei Berücksichtigung der subjektiven

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Tab. 7.11. Ermittlung der Posteriorwahrscheinlichkeiten nach dem Bayes'schen Theorem bei einem binomial verteilten Merkmal

π_i	Priorwahrscheinlichkeit $p(\pi_i)$	Likelihood $p(B\|\pi_i)$	$p(\pi_i)\cdot p(B\|\pi_i)$	Posteriorwahrscheinlichkeit $p(\pi_i\|B)$
0,01	0,30	0,165	0,0495	0,18
0,03	0,50	0,336	0,1680	0,61
0,08	0,15	0,328	0,0492	0,18
0,15	0,05	0,137	0,0069	0,03
	1,00		0,2736	1,00

Vermutungen der Berufsberaterin und des objektiven Stichprobenergebnisses am »wahrscheinlichsten« ist.

Gemäß ▶ Gl. (7.63) benötigen wir hierfür neben den $p(\pi_i)$-Werten die $p(B|\pi_i)$-Werte, d. h. die Wahrscheinlichkeiten des Stichprobenergebnisses bei Gültigkeit der verschiedenen Annahmen über den Parameter π. Diese (bislang geschätzten) Werte bezeichneten wir auf ▶ S. 408 f. als Likelihoods. Wir fragen nach der Likelihood des Stichprobenergebnisses k=1 Psychologieaspirant unter n=20 Abiturienten, wenn der Anteil der Psychologieaspiranten in der Population aller Abiturienten z. B. π_1=0,01 (oder 1%) beträgt. Wie bereits gezeigt wurde (▶ S. 409), kann diese Likelihood über die Binomialverteilung errechnet werden.

Hierbei setzen wir voraus, dass die Antworten der Abiturienten einem **»Bernoulli-Prozess«** entsprechen, d. h., dass sie den Anforderungen der Stationarität und Unabhängigkeit genügen. Stationarität besagt in diesem Zusammenhang, dass die Wahrscheinlichkeit, sich für ein Psychologiestudium zu entscheiden, für alle Befragten konstant ist, und Unabhängigkeit meint, dass die Art der Antwort eines Abiturienten – für oder gegen ein Psychologiestudium – nicht durch die Antworten anderer Abiturienten beeinflusst wird. Bei einer echten Zufallsauswahl von 20 Abiturienten dürften beide Anforderungen erfüllt sein.

Wir erhalten nach ▶ Gl. (7.4)

$$p(B|\pi_1) = p(X = 1|\pi = 0{,}01, n = 20)$$
$$= \binom{20}{1} \cdot 0{,}01^1 \cdot 0{,}99^{19} = 0{,}165.$$

In ▣ Tab. 7.11 sind die zur Bestimmung der Posteriorverteilung erforderlichen Schritte zusammengefasst.

Die $p(B|\pi_i)$-Werte werden – analog zu der oben aufgeführten Rechnung – unter Verwendung des jeweiligen Parameters π_i bestimmt. Die Produkte $p(\pi_i)\cdot p(B|\pi_i)$ führen, relativiert an der Summe dieser Produkte, zu den gesuchten Posteriorwahrscheinlichkeiten. (Rechenkontrolle: die Priorwahrscheinlichkeiten und die Posteriorwahrscheinlichkeiten müssen sich jeweils zu 1 addieren.) Diese Werte zeigen, dass die ursprüngliche Vermutung der Berufsberaterin, π_2=0,03 sei der »wahrscheinlichste« Parameter, durch das Stichprobenergebnis untermauert wird. Die Priorwahrscheinlichkeit dieses Parameters hat sich nach Berücksichtigung des Stichprobenergebnisses auf 0,61 erhöht.

Der Erwartungswert der Posteriorverteilung lautet

$$E(\pi_1) = \sum_{i=1}^{k} \pi_i \cdot p(\pi_i|B) = 0{,}039 \text{ bzw. } 3{,}9\%.$$

Gegenüber der Priorverteilung (mit $E(\pi_i)$=3,75%) hat sich der Erwartungswert durch die Berücksichtigung der Stichprobeninformation also nur geringfügig vergrößert.

Nach Ermittlung der Posteriorwahrscheinlichkeiten könnte die Berufsberaterin erneut eine Stichprobe ziehen und die Posteriorwahrscheinlichkeiten, die jetzt als Priorwahrscheinlichkeiten eingesetzt werden, aufgrund des neuen Stichprobenergebnisses korrigieren. Die Priorwahrscheinlichkeiten berücksichtigen in diesem Falle also sowohl die subjektive Einschätzung der Berufsberaterin als auch das erste Stichprobenergebnis. Dieser Vorgang, die Korrektur der alten Posteriorwahrscheinlichkeiten als neue Priorwahrscheinlichkeiten aufgrund eines weiteren Stichprobenergebnisses, lässt sich beliebig häufig fortsetzen. Statt der wiederholten Korrektur der jeweils neuen Priorwahrscheinlichkeiten können die ursprünglichen Priorwahrscheinlichkeiten jedoch auch nur einmal durch die zusam-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

mengefassten Stichproben korrigiert werden. Beide Wege – die wiederholte Korrektur aufgrund einzelner Stichprobenergebnisse und die einmalige Korrektur durch die zusammengefassten Stichprobenergebnisse – führen letztlich zu identischen Posteriorwahrscheinlichkeiten.

Das Beispiel demonstrierte, wie unter der Annahme einer Binomialverteilung die Likelihoods $p(B|A_i)$ ermittelt werden. Im Folgenden wird gezeigt, wie die für das Bayes'sche Theorem benötigten Likelihoods zu ermitteln sind, wenn das interessierende Merkmal anderen Verteilungsmodellen folgt.

Poisson-Verteilung. Die Annahme, dass in jedem der n durchgeführten »Versuche« entweder das Ereignis A oder das Ereignis $\bar{A}$ mit jeweils konstanter Wahrscheinlichkeit auftritt, rechtfertigt die Verwendung der Binomialverteilung. Interessieren uns nun Ereignisse, die über die Zeit (oder eine andere kontinuierliche Variable) verteilt sind, kann es vorkommen, dass das Ereignis in einem bestimmten Intervall (in einem »Versuch«) mehrmals auftritt (Beispiele: Anzahl der Druckfehler pro Buchseite, Anzahl der Rosinen pro Rosinenbrötchen, Anzahl der Telefonanrufe pro Stunde). Die durchschnittliche Anzahl der Ereignisse pro »Versuch« (z. B. Anzahl der Druckfehler pro Buchseite) bezeichnen wir mit c **(Intensitätsparameter)**. Gefragt wird nach der Wahrscheinlichkeit, dass eine beliebige Buchseite k=0, 1, 2… Druckfehler enthält. Unter der Voraussetzung, dass die Wahrscheinlichkeit eines Druckfehlers für alle Seiten konstant ist **(Stationaritätsannahme)** und dass die Druckfehlerwahrscheinlichkeiten seitenweise voneinander unabhängig sind **(Unabhängigkeitsannahme)**, ergibt sich die Wahrscheinlichkeit für k Druckfehler auf einer beliebigen Seite bei gegebenem c nach der Poisson-Verteilung:

$$p(k|c) = \frac{c^k}{e^c \cdot k!} (e = 2{,}7183). \qquad (7.64)$$

Sind die Voraussetzungen für einen Bernoulli-Prozess erfüllt (Eintreten des Ereignisses A pro Versuch mit konstanter Wahrscheinlichkeit), lässt sich die dann einschlägige Binomialverteilung durch die Poisson-Verteilung approximieren, falls n>10 und π<0,05 ist (vgl. Sachs, 2002, S. 228). In diesem Falle ist $c=n\cdot\pi$.

Beispiel: Wie groß ist die Wahrscheinlichkeit, dass bei 100 Roulettespielen genau einmal die Null fällt? Binomial:

$$p = (X = 1|\pi = 1/37; n = 100) = \binom{100}{1} \cdot (1/37)^1 \cdot$$
$$\cdot (36/37)^{99} = 0{,}1794.$$

Poisson:

$$p(k = 1|c = 100/37) = \frac{(100/37)^1}{e^{(100/37)} \cdot 1!} = 0{,}1811.$$

Wir wollen die Revision einer Priorverteilung für den Fall, dass das untersuchte Merkmal poissonverteilt ist, an einem Beispiel (nach Winkler, 1972, Kap. 3.4) demonstrieren.

Ein Autohändler gruppiert Autoverkäufer in drei Kategorien: Ein hervorragender Verkäufer verkauft durchschnittlich an jedem zweiten Tag, ein guter Verkäufer an jedem vierten Tag und ein schlechter Verkäufer an jedem achten Tag ein Auto. Die durchschnittlichen Wahrscheinlichkeiten für das Ereignis »ein Auto pro Tag verkauft« lauten also π_1=0,5, π_2=0,25 und π_3=0,125. Für diese π-Werte gibt der Autohändler folgende Priorwahrscheinlichkeiten an:

$$p(\pi_1 = 0{,}5) \quad = 0{,}2,$$
$$p(\pi_2 = 0{,}25) \; = 0{,}5,$$
$$p(\pi_3 = 0{,}125) = 0{,}3.$$

Vereinfachend nehmen wir an, dass alle übrigen π-Werte eine Wahrscheinlichkeit von null aufweisen. (Die angemessenere Handhabung dieses Problems, die von einem Kontinuum der π-Parameter ausgeht, wird bei Winkler, 1972, Kap. 4.7 erörtert.)

Dieser Einschätzung folgend gehört ein neu einzustellender, noch unbekannter Verkäufer mit einer »Wahrscheinlichkeit« von 20% zur Kategorie der hervorragenden Verkäufer, mit 50% »Wahrscheinlichkeit« zur Kategorie der guten Verkäufer und mit 30% »Wahrscheinlichkeit« zur Kategorie der schlechten Verkäufer.

Ein neuer Verkäufer möge nun in n=24 Tagen 10 Autos verkauft haben. Hieraus ergeben sich folgende Likelihoods $p(k=10|c_i)$ (mit $c_1=24\cdot0{,}5=12$; $c_2=24\cdot0{,}25=6$ und $c_3=24\cdot0{,}125=3$):

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Tab. 7.12. Ermittlung der Posteriorwahrscheinlichkeiten nach dem Bayes'schen Theorem bei einem Poisson-verteilten Merkmal

c_i	Priorwahrscheinlichkeit $p(c_i)$	Likelihood $p(k\|c_i)$	$p(c_i)\cdot p(k\|c_i)$	Posteriorwahrscheinlichkeit $p(c_i\|k)$
12	0,2	0,1048	0,02096	0,501
6	0,5	0,0413	0,02065	0,493
3	0,3	0,0008	0,00024	0,006
	1,0		0,04185	1,000

$$p(k=10|c_1=12) = \frac{12^{10}}{e^{12}\cdot 10!} = 0{,}1048 \quad (0{,}1169),$$

$$p(k=10|c_2=6) = \frac{6^{10}}{e^{6}\cdot 10!} = 0{,}0413 \quad (0{,}0333),$$

$$p(k=10|c_3=3) = \frac{3^{10}}{e^{3}\cdot 10!} = 0{,}0008 \quad (0{,}0003).$$

(In den Klammern stehen die über das Binomialmodell ermittelten Likelihoods. Wegen $\pi>0{,}05$ sind die Übereinstimmungen nur mäßig.)

In Tab. 7.12 sieht man die zur Revision der Priorwahrscheinlichkeiten erforderlichen Rechenschritte.

Die Priorwahrscheinlichkeit für die Kategorie »hervorragender Verkäufer« hat sich damit von 0,2 auf ca. 0,5 erhöht. Die Wahrscheinlichkeit für die Kategorie »schlechter Verkäufer« wird hingegen verschwindend klein. Die ursprüngliche Erwartung, dass ein Verkäufer pro Tag im Durchschnitt 0,26 Autos (bzw. ein Auto in 3,8 Tagen) verkaufen würde (dies ist der Erwartungswert der Zufallsvariablen »Anzahl der Verkäufe« unter Verwendung der Priorwahrscheinlichkeiten), muss nun auf 0,37 Autos pro Tag (bzw. ein Auto in 2,7 Tagen) erhöht werden (Erwartungswert der Zufallsvariablen »Anzahl der Verkäufe« mit den Posteriorwahrscheinlichkeiten).

Hypergeometrische Verteilung. Das Modell der Binomialverteilung geht davon aus, dass die Wahrscheinlichkeiten für die Ereignisse A oder $\bar{A}$ konstant sind. Diese Voraussetzung ist verletzt, wenn sich die Wahrscheinlichkeiten von Versuch zu Versuch ändern, was z. B. der Fall ist, wenn die Anzahl aller möglichen Versuche (oder die Größe der Population) begrenzt ist. Werden beispielsweise aus einem Skatspiel (N=32 Karten) nacheinander n=6 Karten gezogen (ohne Zurücklegen), verändert sich die Wahrscheinlichkeit, eine der R=8 Herzkarten zu ziehen, von Karte zu Karte. Sie beträgt für die erste Karte 8:32, für die zweite Karte 8:31, wenn die erste Karte kein Herz war, bzw. 7:31, wenn die erste Karte ein Herz war, etc. Die Häufigkeit des Auftretens für das Ereignis A bei n Versuchen (z. B. r=3 Herzkarten unter 6 gezogenen Karten) folgt einer hypergeometrischen Verteilung. (Die Binomialverteilung wäre einschlägig, wenn man die Karten jeweils zurücklegen würde.) Die Wahrscheinlichkeiten für verschiedene r-Werte lassen sich bei gegebenem n, R und N nach folgender Gleichung ermitteln:

$$p(r|n,R,N) = \frac{\binom{R}{r}\cdot\binom{N-R}{n-r}}{\binom{N}{n}}. \tag{7.65}$$

Das folgende Beispiel zeigt die Revision von Priorwahrscheinlichkeiten bei einem hypergeometrisch verteilten Merkmal. Einem Wanderer sind während eines Regens die Streichhölzer nass geworden. Er schätzt nun, wieviele der N=50 Streichhölzer, die sich in der Schachtel befinden, durch den Regen unbrauchbar geworden sind. Die folgenden Schätzungen erscheinen ihm sinnvoll: $R_1=10$, $R_2=20$ und $R_3=30$ (erneut betrachten wir nur einige Werte). Seine Priorwahrscheinlichkeiten für diese Anteile defekter Streichhölzer legt der Wanderer (dem das Missgeschick nasser Streichhölzer nicht zum ersten Mal passiert) in folgender Weise fest:

$$p(R_1=10)=0{,}4,$$
$$p(R_2=20)=0{,}5,$$
$$p(R_3=30)=0{,}1.$$

Von n=5 Streichhölzern, die er prüft, ist nur r=1 Streichholz unbrauchbar. Die Likelihood des Ereignisses unter der Annahme R=10 lautet:

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Tab. 7.13. Ermittlung der Posteriorwahrscheinlichkeiten nach dem Bayes'schen Theorem bei einem hypergeometrisch verteilten Merkmal

R	Priorwahrscheinlichkeit $p(R_i)$	Likelihood $p(r\|R_i)$	$p(R_i)\cdot p(r\|R_i)$	Posteriorwahrscheinlichkeit $p(R_i\|r)$
10	0,4	0,4313	0,1725	0,5586
20	0,5	0,2587	0,1294	0,4190
30	0,1	0,0686	0,0069	0,0224
	1,0		0,3088	1,0000

$$p(r=1|n=5,\ R=10,\ N=50) = \frac{\binom{10}{1}\cdot\binom{40}{4}}{\binom{50}{5}} = \frac{\frac{10}{1}\cdot\frac{40\cdot39\cdot38\cdot37}{4\cdot3\cdot2\cdot1}}{\frac{50\cdot49\cdot48\cdot47\cdot46}{5\cdot4\cdot3\cdot2\cdot1}} = \frac{913900}{2118760} = 0{,}4313.$$

Für R=20 und R=30 errechnen wir nach dieser Gleichung die Likelihoods 0,2587 und 0,0686. Tab. 7.13 zeigt, wie dieses empirische Ergebnis die Priorwahrscheinlichkeiten verändert.

Die Wahrscheinlichkeiten haben sich zugunsten der Hypothese R=10 verschoben. Während der Wanderer nach seiner ersten Schätzung mit 17 unbrauchbaren Streichhölzern rechnete (10·0,4+20·0,5+30·0,1=17), kann er nach den 5 geprüften Streichhölzern davon ausgehen, dass die Schachtel insgesamt nur ca. 15 unbrauchbare Streichhölzer enthält (10·0,559+20·0,419+30·0,022=14,6).

Multinominale Verteilung. Ein einfaches Urnenbeispiel erläutert die Besonderheiten einer multinomialen Verteilung. Befinden sich in einer Urne rote und schwarze Kugeln in einem bestimmten Häufigkeitsverhältnis, sind die Wahrscheinlichkeiten, bei n Versuchen z. B. k=0, 1, 2, … rote Kugeln zu ziehen, binomial verteilt, wenn die Kugeln wieder zurückgelegt werden. Befinden sich in der Urne hingegen mehr als zwei Kugelarten, wie z. B. rote Kugeln mit der Wahrscheinlichkeit π_1, schwarze Kugeln mit der Wahrscheinlichkeit π_2, grüne Kugeln mit der Wahrscheinlichkeit π_3 und gelbe Kugeln mit der Wahrscheinlichkeit π_4, wird die Wahrscheinlichkeit dafür, dass bei n Versuchen k_1 rote, k_2 schwarze, k_3 grüne und k_4 gelbe Kugeln gezogen werden (wiederum mit Zurücklegen), über die multinomiale Verteilung berechnet. Die binomiale Verteilung verwenden wir für zwei einander ausschließende Ereignisklassen und die multinomiale Verteilung für mehr als zwei oder s einander ausschließende Ereignisklassen. Die Wahrscheinlichkeitsverteilung wird durch folgende Gleichung beschrieben:

$$p(k_1,k_2\ldots k_s\,|\,n,\pi_1,\pi_2\ldots\pi_s) = \frac{n!}{k_1!\cdot k_2!\cdot\ldots k_s!}\cdot\pi_1^{k_1}\cdot\pi_2^{k_2}\cdot\ldots\pi_s^{k_s} \tag{7.66}$$

Auch diese Gleichung sei an einem Beispiel demonstriert. Eine Werbeagentur erhält den Auftrag, für ein Produkt eine möglichst werbewirksame Verpackung zu entwickeln. Drei Vorschläge (V_1, V_2 und V_3) kommen in die engere Wahl und sollen getestet werden. Vorab bittet die Agentur ihre Werbefachleute, die Werbewirksamkeit der drei Vorschläge einzuschätzen. Da man sich nicht einigen kann, werden mehrere Einschätzungen abgegeben (die Einschätzungen verdeutlichen gleichzeitig verschiedene Strategien zur Quantifizierung subjektiver Wahrscheinlichkeiten; Näheres hierzu vgl. Philips, 1973, Teil 1). Sie lauten:

1. Einschätzung: V1 ist doppelt so wirksam wie V2. V2 und V3 unterscheiden sich nicht, d. h. π_1=0,5; π_2=0,25 und π_3=0,25.
2. Einschätzung: Die Werbewirksamkeit der drei Vorschläge steht im Verhältnis 3:2:1 zueinander, d. h. π_1=3/6=0,5; π_2=2/6=0,333 und π_3=1/6=0,167.
3. Einschätzung: V1 und V3 sind gleich wirksam, und V2 ist nahezu unwirksam, d. h. π_1=0,495; π_2=0,01 und π_3=0,495.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Tab. 7.14. Ermittlung der Posteriorwahrscheinlichkeiten nach dem Bayes'schen Theorem bei einem multinomial verteilten Merkmal

Modell (i)	Priorwahrscheinlichkeit p (Modell$_i$)	Likelihood p(Ergebnis|Modell$_i$)	p(Modell$_i$) p(Ergebnis|Modell$_i$)	Posteriorwahrscheinlichkeit p(Modell$_i$|Ergebnis)
π_1=0,5; π_2=0,25; π_3=0,25	0,40	0,0263	0,01052	0,385
π_1=0,5; π_2=0,333; π_3=0,167	0,50	0,0328	0,01640	0,601
π_1=0,495; π_2=0,01; π_3=0,495	0,02	0,0000	0,00000	0,000
π_1=0,80; π_2=0,15; π_3=0,05	0,08	0,0046	0,00037	0,014
	1,00		0,02729	1,000

4. Einschätzung: Wenn für die Werbewirksamkeit der drei Vorschläge 100 Punkte zu vergeben sind, erhält der erste Vorschlag 80 Punkte, der zweite 15 Punkte und der dritte 5 Punkte, d. h. π_1=0,8; π_2=0,15 und π_3=0,05.

Abschließend geben die Werbefachleute eine Schätzung darüber ab, für wie »wahrscheinlich« sie es halten, dass die einzelnen Einschätzungen tatsächlich zutreffen (Priorwahrscheinlichkeiten). Aus den individuellen »Wahrscheinlichkeiten« resultieren folgende Durchschnittswerte (erneut wollen wir davon ausgehen, dass nur die hier aufgeführten π-Werte als Parameterschätzungen in Frage kommen):

$$p(\pi_1 = 0{,}5; \pi_2 = 0{,}25; \pi_3 = 0{,}25) = 0{,}40,$$
$$p(\pi_1 = 0{,}5; \pi_2 = 0{,}333; \pi_3 = 0{,}167) = 0{,}50,$$
$$p(\pi_1 = 0{,}495; \pi_2 = 0{,}01; \pi_3 = 0{,}495) = 0{,}02,$$
$$p(\pi_1 = 0{,}8; \pi_2 = 0{,}15; \pi_3 = 0{,}05) = 0{,}08.$$

In einem Verkaufsexperiment entscheiden sich von n=20 Käufern k_1=12 für die erste Verpackungsart (V_1), k_2=5 für die zweite Verpackungsart (V_2) und k_3=3 für die dritte Verpackungsart (V_3). (Die Ablehnung aller drei Verpackungsarten oder die Wahl mehrerer Verpackungsarten sei ausgeschlossen.) Die Likelihood dieses empirischen Ergebnisses für das erste multinomiale Modell (erste Einschätzung) lautet:

$$p(k_1 = 12; k_2 = 5; k_3 = 3 | n = 20, \pi_1 = 0{,}5;$$
$$\begin{aligned}\pi_2 &= 0{,}25; \pi_3 = 0{,}25) = p(\text{Ergebnis}|1.\ \text{Modell})\\ &= \frac{20!}{12! \cdot 5! \cdot 3!} \cdot 0{,}5^{12} \cdot 0{,}25^5 \cdot 0{,}25^3\\ &= 7054320 \cdot 3{,}72529 \cdot 10^{-9} = 0{,}0263.\end{aligned}$$

Für die weiteren Modelle ergeben sich nach dem gleichen Algorithmus:

$$p(\text{Ergebnis}|2.\ \text{Modell}) = 0{,}0328,$$
$$p(\text{Ergebnis}|3.\ \text{Modell}) = 1{,}85 \cdot 10^{-8} \approx 0,$$
$$p(\text{Ergebnis}|4.\ \text{Modell}) = 0{,}0046.$$

Damit sind die Priorwahrscheinlichkeiten gem. Tab. 7.14 zu korrigieren.

Empirische Evidenz und **subjektive Einschätzung** führen zusammengenommen zu dem Resultat, dass das dritte und das vierte Modell praktisch ausscheiden. Das Modell mit der höchsten Priorwahrscheinlichkeit (Modell 2) hat mit 0,601 auch die höchste Posteriorwahrscheinlichkeit. Offensichtlich haben die empirischen Daten (mit p_1=0,60; p_2=0,25 und p_3=0,15) dieses Modell am besten bestätigt. Die Erwartungswerte für die Parameter der multinomialen Verteilung lauten:

Vor der empirischen Untersuchung:

$$\begin{aligned}E(\pi_1) &= 0{,}40 \cdot 0{,}5 + 0{,}50 \cdot 0{,}5 + 0{,}02 \cdot 0{,}495\\ &+0{,}08 \cdot 0{,}8 = 0{,}524,\end{aligned}$$

$$\begin{aligned}E(\pi_2) &= 0{,}40 \cdot 0{,}25 + 0{,}50 \cdot 0{,}333 + 0{,}02 \cdot 0{,}01\\ &+0{,}08 \cdot 0{,}15 = 0{,}279,\end{aligned}$$

$$\begin{aligned}E(\pi_3) &= 0{,}40 \cdot 0{,}25 + 0{,}50 \cdot 0{,}167 + 0{,}02 \cdot 0{,}495\\ &+0{,}08 \cdot 0{,}05 = 0{,}197.\end{aligned}$$

Nach der empirischen Untersuchung:

$$\begin{aligned}E(\pi_1) &= 0{,}385 \cdot 0{,}5 + 0{,}601 \cdot 0{,}5 + 0{,}000 \cdot 0{,}495\\ &+0{,}014 \cdot 0{,}8 = 0{,}504,\end{aligned}$$

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

$$E(\pi_2) = 0{,}385 \cdot 0{,}25 + 0{,}601 \cdot 0{,}333 + 0{,}000 \cdot 0{,}01 + 0{,}014 \cdot 0{,}15 = 0{,}298,$$

$$E(\pi_3) = 0{,}385 \cdot 0{,}25 + 0{,}601 \cdot 0{,}167 + 0{,}000 \cdot 0{,}495 + 0{,}014 \cdot 0{,}05 = 0{,}197.$$

Stetige Zufallsvariablen

Die letzten Beispiele basieren auf der Annahme einer diskreten Verteilung des unbekannten Parameters. Im Berufsberatungsbeispiel (▶ S. 461 f.) nahmen wir an, dass nur einige ausgewählte Prozentwerte als Schätzungen für den Anteil von Abiturienten, die sich für ein Psychologiestudium interessieren, in Frage kommen.

Zweifellos wäre hier die Annahme, dass sämtliche Prozentwerte innerhalb eines plausibel erscheinenden Prozentwertebereiches als Schätzwerte in Betracht kommen, realistischer gewesen. Ähnliches gilt für die Beispiele »Autoverkauf« (▶ S. 463 f.), »unbrauchbare Streichhölzer« (▶ S. 464 f.) und »Werbewirksamkeit von Verpackungen« (▶ S. 465 f.), die ebenfalls nur ausgewählte Werte als Schätzungen des unbekannten Parameters untersuchten.

Im Folgenden wollen wir das Bayes'sche Theorem für stetige Zufallsvariablen behandeln. Zur begrifflichen Klärung sei darauf hingewiesen, dass die Bezeichnung »Bayes'sches Theorem für stetige Zufallsvariablen« besagt, dass der zu schätzende Parameter stetig verteilt ist (wie z. B. Populationsmittelwerte oder Populationsanteile). Das Stichprobenergebnis, aufgrund dessen die Priorverteilung revidiert wird, kann hingegen entweder stetig oder diskret verteilt sein. (Bei Anteilsschätzungen beispielsweise behandeln wir den gesuchten Parameter π als eine stetige Zufallsvariable. Das Stichprobenergebnis – k-mal das Ereignis A bei n Versuchen – ist jedoch diskret. Bei der Schätzung eines Mittelwertparameters ist nicht nur μ eine stetige Zufallsvariable; auch das Stichprobenergebnis $\bar{x}$ stellt die Realisierung einer stetig verteilten Zufallsvariablen dar.)

Bayes'sches Theorem für stetige Zufallsvariablen. In Analogie zu ▶ Gl. (7.63) (Bayes'sches Theorem für diskrete Zufallsvariablen) lautet das Bayes'sche Theorem für stetige Zufallsvariablen

$$f(\theta|y) = \frac{f(\theta) \cdot f(y|\theta)}{\int_{-\infty}^{\infty} f(\theta) \cdot f(y|\theta)d\theta}. \quad (7.67)$$

Hierin sind θ (griechisch: theta) der gesuchte, stetig verteilte Parameter und y das Stichprobenergebnis. Der Ausdruck $f(\theta)$ kennzeichnet die Dichtefunktion (▶ S. 404) des Parameters θ, die – analog zu den Priorwahrscheinlichkeiten $p(A_i)$ in ▶ Gl. (7.63) – die Vorkenntnisse der Untersuchenden zusammenfasst. (Auf das schwierige Problem der Umsetzung von Vorinformationen in Priordichtefunktionen werden wir später eingehen.) Die Likelihood-Funktion des Stichprobenergebnisses y bei gegebener Verteilung von θ bezeichnen wir mit $f(y|\theta)$. Das Integral im Nenner $\int_{-\infty}^{\infty} f(\theta) \cdot f(y|\theta)d\theta$ entspricht der Summe $\sum_{i=1}^{k} p(B|A_i) \cdot p(A_i)$, die im Nenner des Bayes'schen Theorems für diskrete Zufallsvariablen steht. In beiden Fällen normiert der Nenner die Posteriorwahrscheinlichkeiten (Dichten), d. h., die Summe der Posteriorwahrscheinlichkeiten (bzw. die Fläche der Posteriorverteilung) wird – wie auch die Summe (Fläche) der Priorwahrscheinlichkeiten (Dichten) – eins gesetzt. (Zur Herleitung des Bayes'schen Theorems für stetige Zufallsvariablen vgl. z. B. Winkler, 1972, Kap. 4.2.)

Beispiel: Ein kleines Beispiel (nach Winkler, 1972, S. 145 ff.) soll die Handhabung von ▶ Gl. (7.67) verdeutlichen. Es geht um die Schätzung von θ, des zukünftigen Marktanteils eines neuen Produktes. Einfachheitshalber nehmen wir die in ◻ Abb. 7.11a wiedergegebene Priorverteilung für θ an. Es handelt sich um eine Dreiecksverteilung, die besagt, dass hohe Marktanteile für unwahrscheinlicher gehalten werden als niedrige Marktanteile. (Die graue Fläche über dem Bereich $0{,}5<\theta<1$ entspricht der Wahrscheinlichkeit, dass der wahre Parameter θ in diesen Bereich fällt.) Die Dichtefunktion der Priorverteilung heißt

$$f(\theta) = 2 \cdot (1-\theta) \ \ (\text{für } \ 0 \le \theta \le 1).$$

Von n=5 Testpersonen möge eine (k=1) das neue Produkt kaufen. Wenn wir realistischerweise annehmen, dass die Anzahl der Käufer (k) binomial verteilt ist, lautet die Likelihood-Funktion gem. ▶ Gl. (7.4)

$$f(y|\theta) = p(k=1|\theta;\ n=5) = \binom{5}{1} \cdot \theta^1 \cdot (1-\theta)^4 = 5 \cdot \theta \cdot (1-\theta)^4.$$

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Abb. 7.11b zeigt diese Likelihood-Funktion. Man erkennt, dass das Stichprobenergebnis (k=1) am »wahrscheinlichsten« ist,wenn $\theta \approx 0{,}2$ ist. Eingesetzt in ▶ Gl. (7.67) resultiert für die Posteriorverteilung

$$f(\theta|y) = \frac{[2\cdot(1-\theta)]\cdot[5\theta\cdot(1-\theta)^4]}{\int_0^1[2\cdot(1-\theta)]\cdot[5\theta\cdot(1-\theta)^4]d\theta}$$

$$= \frac{10\cdot\theta(1-\theta)^5}{10\cdot\int_0^1\theta(1-\theta)^5 d\theta}$$

$$= \frac{\theta\cdot(1-\theta)^5}{\int_0^1\theta(1-\theta)^5 d\theta} \quad (\text{für } 0\le\theta\le 1).$$

Die (mathematisch nicht einfache) Auflösung des Integrals im Nenner dieser Gleichung führt zu 5!/7!=1/42, was zusammengenommen folgende Dichtefunktion für die Posteriorverteilung ergibt:

$$f(\theta|y) = 42\cdot\theta\cdot(1-\theta)^5 \quad (\text{für } 0\le\theta\le 1).$$

Abb. 7.11c stellt die Posteriorverteilung grafisch dar.

Wie man sieht, revidiert die Stichprobenuntersuchung die Priorverteilung so, dass Werte im Bereich $0{,}5<\theta<1$ unwahrscheinlicher werden. Subjektive Erwartung und empirische Evidenz führen zusammengenommen zu dem Resultat, dass Marktanteile im Bereich um 0,2 am plausibelsten sind (wie dieses Problem auch einfacher zu lösen ist, zeigt ▶ S. 478).

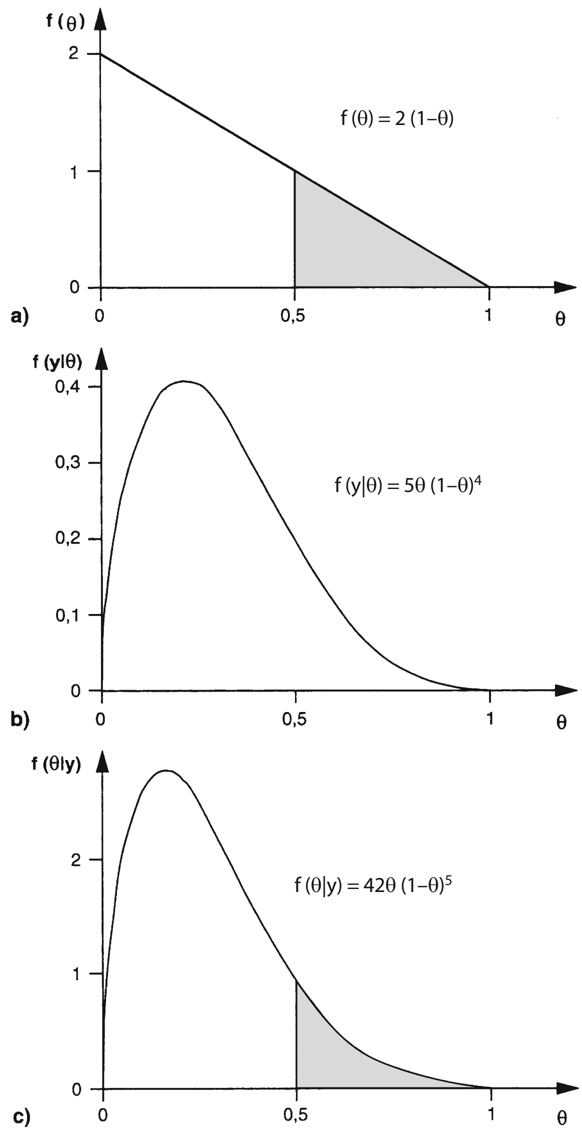


Abb. 7.11. **a** Priorverteilung; **b** Likelihood-Funktion; **c** Posteriorverteilung

Konjugierte Verteilungsfamilien. Die Ermittlung der Posteriorverteilung nach ▶ Gl. (7.67) kann bei komplizierten Priorverteilungen (Priordichten) und Likelihood-Funktionen mathematisch erhebliche Schwierigkeiten bereiten. Diesen Schwierigkeiten kann man jedoch aus dem Wege gehen, wenn man für Priorverteilungen nur ganz bestimmte, mathematisch einfach zu handhabende Funktionstypen einsetzt. Diese Einschränkung ist nicht so gravierend, wie es zunächst erscheinen mag; durch entsprechende Parameterwahl bilden diese Funktionstypen nämlich eine Vielzahl von Verteilungsformen ab, sodass es – zumindest für unsere Zwecke – meistens gelingen wird, die Vorstellungen über die Priorverteilung durch eine dieser Funktionen hinreichend genau abzubilden. Der Einsatz dieser Funktionen ist zumindest immer dann zu rechtfertigen, wenn die Posteriorverteilung für die exakte Priorverteilung praktisch genauso aussieht wie die Posteriorverteilung, die resultieren würde, wenn die Vorstellungen von der Priorverteilung durch die Verwendung einer dieser Funktionen nur ungefähr wiedergegeben werden. »Sensitivitätsanalysen« (vgl. z. B. Hays & Winkler, 1970, Kap. 8.16) belegen, dass dies – zumal, wenn größere Stichproben untersucht werden – in den meisten Fällen

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

zutrifft, d. h., die Posteriorverteilung ist weitgehend invariant gegenüber mäßigen Veränderungen der Priorverteilung.

Die Verteilungsformen, die mit einem Funktionstyp bei unterschiedlicher Festlegung ihrer Parameter erzeugt werden, bezeichnet man als **Verteilungsfamilie**. (Alle Geraden, die durch unterschiedliche Festsetzung von a und b durch die Geradengleichung y=ax+b beschrieben werden, stellen z. B. eine solche Verteilungsfamilie dar.) Ein großer Teil der Bayes'schen Literatur ist nun darauf ausgerichtet, Verteilungsfamilien zu finden, deren Likelihood-Funktion einfach bestimmbar ist. Priorverteilungen mit Likelihood-Funktionen, die auf denselben Funktionstypus zurückgeführt werden können, bezeichnet man als **konjugierte Priorverteilungen**. (Eine genauere Definition geben z. B. Berger, 1980, Kap. 4; de Groot, 1970, Kap. 9; Koch, 2000, Kap. 2.6.3.)

Wenn die Priorverteilung zu einer konjugierten Verteilungsfamilie gehört, dann fällt auch die Posteriorverteilung in diese Verteilungsfamilie. Priorverteilung und Likelihood-Funktion können dann äußerst einfach zu einer neuen Posteriorverteilung kombiniert werden. Drei Verteilungstypen haben in diesem Zusammenhang eine besondere Bedeutung: die Normalverteilung, die Betaverteilung und die Gleichverteilung.

Normalverteilung: Die Normalverteilung ist das übliche Verteilungsmodell, das wir annehmen, wenn eine Priorverteilung durch einen Mittelwertparameter zu spezifizieren ist. In der Regel wird eine bestimmte Ausprägung für diesen Parameter die höchste Plausibilität aufweisen (Modalwert), wobei Werte mit größer werdendem Abstand von diesem Modalwert zunehmend weniger plausibel sind. Je mehr man von der Richtigkeit seiner (im Modalwert festgelegten) Parameterschätzung überzeugt ist, desto kleiner wird die Streuung der Priorverteilung (Abb. 7.12).

Eine Priorverteilung über einen Mittelwertparameter wird üblicherweise als Normalverteilung spezifiziert.

Betaverteilung: Die Betaverteilung benötigen wir zur Spezifizierung der Priorverteilung für einen Populationsanteil. Die Funktionsgleichung für die Betaverteilung lautet

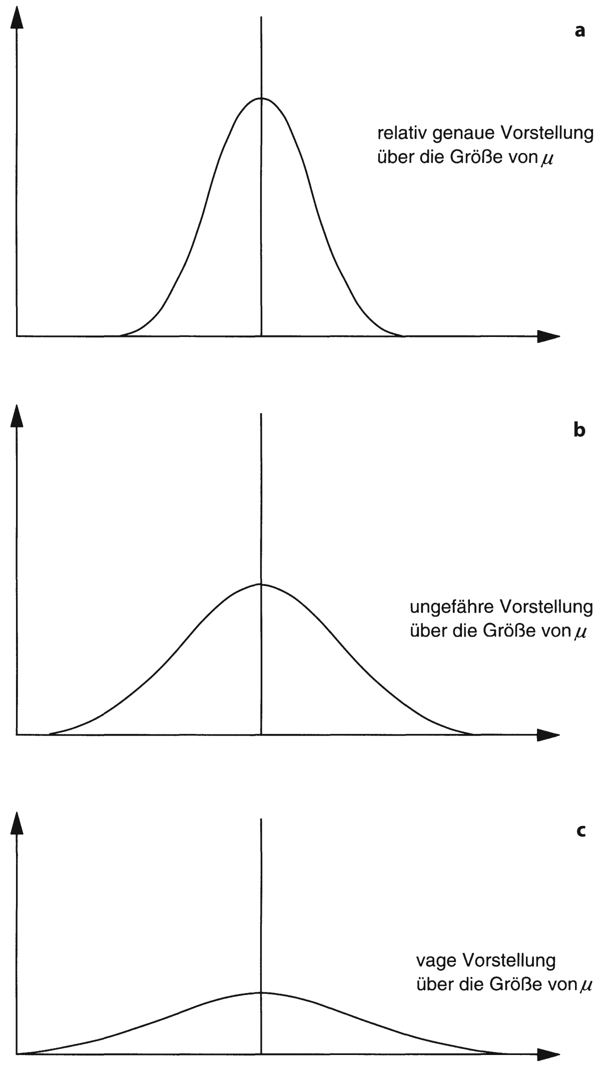


Abb. 7.12a–c. Priorverteilungen für Mittelwertparameter bei unterschiedlicher Sicherheit

$$f(x) = \frac{(k+r-1)!}{(k-1)!(r-1)!} \cdot x^{k-1} \cdot (1-x)^{r-1} \qquad (7.68)$$

$$(\text{für } k>0,\ r>0 \text{ und } 0<x<1),$$

wobei

k = Anzahl der Untersuchungsobjekte mit A,

r = Anzahl der Untersuchungsobjekte mit $\bar{A}$ (non A),

sodass

k+r=n (Stichprobenumfang).

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Durch entsprechende Wahl der Parameter k und r beschreibt diese Funktionsgleichung eine Vielzahl von Verteilungsformen. Ist k=r, resultieren symmetrische Verteilungen. Für k<r sind die Verteilungen linkssteil und für k>r rechtssteil. Mit k>1 und r>1 erhält man unimodale Verteilungen mit folgendem Modalwert:

$$\text{Modalwert} = \frac{k-1}{k+r-2}\,. \qquad (7.69)$$

Ist k≤1 oder r≤1, resultieren entweder unimodale Verteilungen, deren Modalwerte bei 0 oder 1 liegen, oder u-förmige Verteilungen mit Höchstwerten bei 0 und 1 bzw. Gleichverteilungen. (Hierbei ist zu beachten, dass für k<1 und r<1 die Gammafunktion als Verallgemeinerung der elementaren Fakultät heranzuziehen ist; vgl. z. B. Kreyszig, 1973, Abschn. 60.) ► Gl. (7.68) führt zu einer Dreiecksverteilung mit positiver Steigerung (f(x)=2x), wenn k=2 und r=1 sind und für k=1 und r=2 zu einer Dreiecksverteilung mit negativer Steigung (f(x)=2·(1–x)).

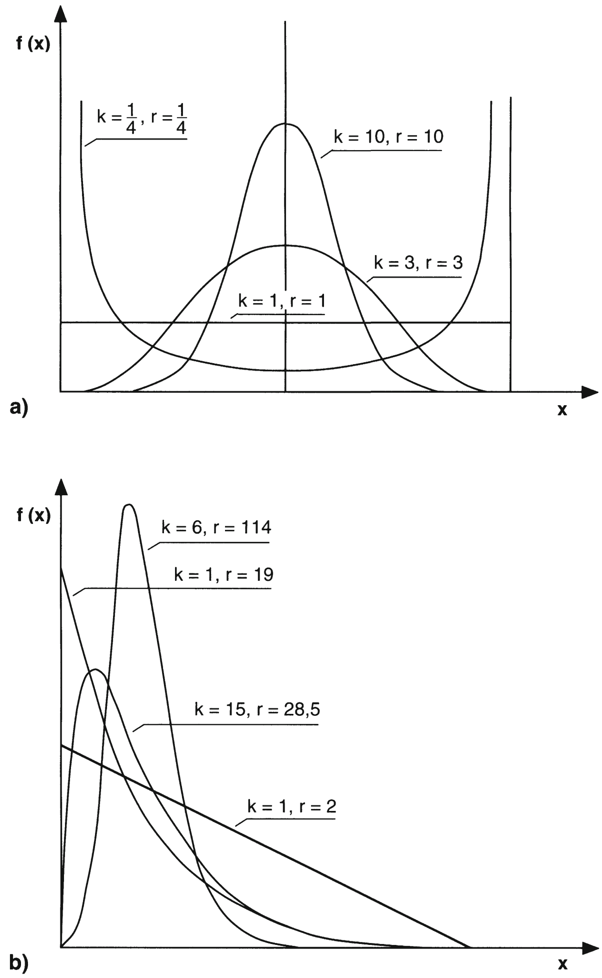
Das arithmetische Mittel einer Betaverteilung lautet

$$\mu = \frac{k}{k+r}. \qquad (7.70)$$

Für die Standardabweichung resultiert

$$\sigma = \sqrt{\frac{k \cdot r}{(k+r)^2 \cdot (k+r+1)}}\,. \qquad (7.71)$$

In ◘ Abb. 7.13 werden einige Verteilungsformen der Betafunktion für ausgewählte k- und r-Werte gezeigt. (Die für unsere Zwecke wichtigsten Betaverteilungen sind im ► Anhang F, ◘ Tab. F4 wiedergegeben.)



◘ **Abb. 7.13a,b.** Betaverteilungen. **a** symmetrisch, **b** asymmetrisch

! Eine Priorverteilung über einen Populationsanteil wird als Betaverteilung formuliert.

Lässt sich die Priorverteilung für einen Populationsanteil als eine Betaverteilung beschreiben, resultiert als Posteriorverteilung ebenfalls eine Betaverteilung. Wir werden hierauf später ausführlicher eingehen.

Gleichverteilung: Eine Gleichverteilung der Parameter repräsentiert eine sog. **diffuse Priorverteilung**, die den Zustand totaler Informationslosigkeit abbildet. Hat man keinerlei Informationen über die mutmaßliche Größe des gesuchten Parameters, kommen alle denkbaren Werte mit gleicher »Wahrscheinlichkeit« in Frage. Der Einsatz einer diffusen Priorverteilung führt zu einer Posteriorverteilung, die ausschließlich vom Stichprobenergebnis abhängt. Der Vergleich einer Posteriorverteilung, die bei Verwendung einer nichtdiffusen Priorverteilung resultiert (also einer Priorverteilung, die bereits vorhandene Kenntnisse abbildet), mit einer Posteriorverteilung für diffuse Priorverteilungen informiert somit über die Beeinflussung der Posteriorverteilung durch die subjektiven Informationen. Diffuse Priorver-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

teilungen stellen die »Schnittstelle« von »klassischer« und Bayes'scher Statistik dar.

Ohne Vorkenntnisse über die Populationsverhältnisse wird die Priorverteilung als Gleichverteilung festgelegt (sog. diffuse Priorverteilung). Bei diffusen Priorverteilungen hängt die Posteriorverteilung genau wie bei der Parameterschätzung in der »klassischen« Statistik allein vom Stichprobenergebnis ab.

Schätzung von Populationsmittelwerten

In einer populationsbeschreibenden Untersuchung wird vor Ziehung einer Stichprobe der Populationsmittelwert μ mit μ' als dem plausibelsten Wert geschätzt. Andere Schätzwerte für μ mögen – allerdings mit geringerer »Wahrscheinlichkeit« – ebenfalls in Frage kommen. Insgesamt seien die Vorstellungen über den unbekannten Parameter μ durch eine normalverteilte Priorverteilung mit dem Erwartungswert μ' und der Varianz σ'^2 beschreibbar. Diese Priorverteilung repräsentiert den Informationsstand vor der empirischen Untersuchung.

Eine Zufallsstichprobe von n Untersuchungsobjekten führt zu einem Mittelwert $\bar{x}$. Für n >30 stellt $\bar{X}$ (wegen des zentralen Grenzwerttheorems, ▶ S. 411 f.) eine normalverteilte Zufallsvariable dar, deren Streuung vom Betrag $\sigma_{\bar{x}}$ ist. Die Streuung σ des untersuchten Merkmals in der Population wird entweder aufgrund vergangener Untersuchungen oder aus den Stichprobendaten geschätzt. Eine verlässliche Schätzung von σ durch $\hat{\sigma}$ der Stichprobe setzt allerdings eine genügend große Stichprobe (n>30) voraus. Der Fall »n<30« und »σ unbekannt« wird hier nicht dargestellt. Eine anschauliche Einführung in die Besonderheiten dieses Problems findet man bei Phillips (1973, Kap. 11.3).

Die Varianz der Posteriorverteilung σ''^2 (bzw. deren Reziprokwert) ergibt sich zu

$$\frac{1}{\sigma''^2} = \frac{1}{\sigma'^2} + \frac{n}{\sigma^2}. \tag{7.72}$$

Der Erwartungswert der Posteriorverteilung lautet

$$\mu'' = \frac{\frac{1}{\sigma'^2} \cdot \mu' + \frac{n}{\sigma^2} \cdot \bar{x}}{\frac{1}{\sigma'^2} + \frac{n}{\sigma^2}}. \tag{7.73}$$

(Zur Herleitung dieser Gleichungen s. Berger, 1980, Kap. 4.2.) Die Posteriorverteilung ist wie auch die Priorverteilung normal. Damit sind die Bestimmungsstücke bekannt, um für μ'' ein Schätzintervall zu bestimmen. ◘ Box 7.8 erläutert den Rechengang an einem Beispiel.

Glaubwürdigkeitsintervalle. Das in ◘ Box 7.8 ermittelte Intervall nennen wir Glaubwürdigkeitsintervall und nicht – wie bisher – Konfidenzintervall. Diese konzeptionelle Unterscheidung lässt sich wie folgt begründen:

Im »klassischen« Ansatz stellt das ermittelte Konfidenzintervall eine Realisierung der Zufallsvariablen »Konfidenzintervalle« dar, das den Parameter μ entweder umschließt oder nicht umschließt. Da die Bestimmung dieses Konfidenzintervalls jedoch so angelegt ist, dass 95% aller denkbaren Intervalle den Parameter umschließen, ist es sehr plausibel, dass auch das gefundene Konfidenzintervall den Parameter umschließt (ausführlicher hierzu ▶ S. 410 ff.).

Im Unterschied hierzu betrachtet der Bayes'sche Ansatz den Parameter μ als eine Zufallsvariable, d. h., bei normalverteilten Posteriorverteilungen sind unterschiedliche Wertebereiche für μ mehr oder weniger plausibel oder glaubwürdig. Wir vermeiden hier erneut bewusst den durch relative Häufigkeiten definierten Wahrscheinlichkeitsbegriff, denn tatsächlich existiert für μ nur ein Wert a, d. h. $p(\mu=a)=1$ und $p(\mu\neq a)=0$. Die Priorverteilung und auch die Posteriorverteilung sind damit keine Wahrscheinlichkeitsverteilungen im engeren Sinne, sondern Verteilungen, die subjektive Vorstellungen oder Überzeugungen über mögliche Ausprägungen von μ wiederspiegeln.

Die Bereiche $\mu''\pm1{,}96\cdot\sigma''$ (für 95%) bzw. $\mu''\pm2{,}58\cdot\sigma''$ (für 99%) werden deshalb als Glaubwürdigkeitsintervalle (**»Credible Intervals«**) bezeichnet. Der gesuchte Parameter befindet sich in diesen Bereichen mit einer »Glaubwürdigkeit« von 95% bzw. 99%.

Das Glaubwürdigkeitsintervall kennzeichnet denjenigen Bereich eines Merkmals, in dem sich mit 95%iger (99%iger) Glaubwürdigkeit der gesuchte Populationsparameter befindet. Das Glaubwürdigkeitsintervall des Populationsmittelwertes berechnet sich aus Erwartungswert der Posterior-

▼

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

verteilung (μ″) und Streuung der Posteriorverteilung (σ″):

$$\text{krit. Glaub. (95\%)} = \mu'' \pm z_{(2,5\%)} \cdot \sigma'' \quad \text{und}$$
$$\text{krit. Glaub. (99\%)} = \mu'' \pm z_{(0,5\%)} \cdot \sigma''$$

Glaubwürdigkeitsintervalle sind immer kleiner als Konfidenzintervalle, die zur Parameterschätzung ausschließlich das Stichprobenergebnis heranziehen – vorausgesetzt, man verfügt über Informationen, die die Spezifizierung einer nicht diffusen Priorverteilung rechtfertigen. Damit liegt es nahe zu fragen, welchen Anteil Priorverteilung und Stichprobenergebnis am Zustandekommen der Posteriorverteilung haben bzw. mit welchen Gewichten Priorverteilung und Stichprobenergebnis in die Bestimmung der Posteriorverteilung eingehen. Der folgende Gedankengang beantwortet diese Frage.

7

Vorinformationen als Stichprobenäquivalente. Es wurde bereits darauf hingewiesen, dass die Genauigkeit der Vorstellungen über die mutmaßliche Größe des unbekannten Parameters in der Varianz der Priorverteilung ihren Niederschlag findet (vgl. 7.12). Je kleiner die Varianz, desto sicherer sind die Vorinformationen. Der folgende »**Sicherheitsindex**« S′ formalisiert diesen intuitiv einleuchtenden Sachverhalt:

$$S' = \frac{1}{\sigma'^2}. \tag{7.74}$$

Ein entsprechendes Maß definieren wir für die Populationsvarianz

$$S = \frac{1}{\sigma^2}. \tag{7.75}$$

Die Sicherheit der Vorinformationen (S′) relativ zur reziproken Populationsvarianz (S) nennen wir n′:

$$n' = \frac{S'}{S} = \frac{\frac{1}{\sigma'^2}}{\frac{1}{\sigma^2}} = \frac{\sigma^2}{\sigma'^2}. \tag{7.76}$$

Lösen wir nach σ'^2 auf, resultiert

$$\sigma'^2 = \frac{\sigma^2}{n'}. \tag{7.77}$$

Diese Gleichung stellt das Quadrat des Standardfehlers für Mittelwerte aus Stichproben des Umfanges n′ dar (▶ Gl. 7.5); n′ ist damit derjenige Stichprobenumfang, der erforderlich ist, um bei einer Populationsvarianz von σ^2 einen quadrierten Standardfehler der Größe σ'^2 zu erhalten. Der Informationsgehalt der Priorverteilung entspricht damit der Information einer Stichprobe des Umfanges n′.

Die Antwort auf die Frage nach den Gewichten von Priorverteilung und Stichprobenergebnis lässt sich hieraus einfach ableiten. Setzen wir σ'^2 gem. ▶ Gl. (7.77) in ▶ Gl. (7.73) ein, resultiert

$$\begin{aligned}\mu'' &= \frac{\frac{n'}{\sigma^2}\cdot\mu' + \frac{n}{\sigma^2}\cdot\bar{x}}{\frac{n'}{\sigma^2} + \frac{n}{\sigma^2}} \\ &= \frac{n'\cdot\mu' + n\cdot\bar{x}}{n' + n} \\ &= \frac{n'}{n'+n}\cdot\mu' + \frac{n}{n'+n}\cdot\bar{x} \\ &= \frac{n'}{n''}\cdot\mu' + \frac{n}{n''}\cdot\bar{x}\end{aligned} \tag{7.78}$$

mit $n''=n'+n$.

Der Erwartungswert der Posteriorverteilung ist eine gewichtete Summe aus μ′, dem Erwartungswert der Priorverteilung, und $\bar{x}$, dem Stichprobenmittelwert. Die Gewichte sind n′, der implizit in der Priorverteilung »verborgene« Stichprobenumfang, und n, der Umfang der tatsächlich gezogenen Stichprobe, jeweils relativiert an der Summe n″ beider Stichprobenumfänge.

Die Posteriorverteilung entspricht einer Mittelwerteverteilung mit einer Varianz, die man erhält, wenn Stichproben des Umfanges $n''=n+n'$ gezogen werden. Auch diese Zusammenhänge werden in ◻ Box 7.8 numerisch erläutert.

Diffuse Priorverteilung. Mit den oben genannten Überlegungen sind wir in der Lage, auch den Fall totaler Informationslosigkeit zu berücksichtigen. Totale Informationslosigkeit bedeutet, dass jeder beliebige Wert mit gleicher Plausibilität als Parameterschätzung in Frage kommt. Als Priorverteilung muss dann eine Gleichver-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Box 7.8

Wie umfangreich sind Diplomarbeiten? V. Bayes'scher Ansatz

Obwohl das Beispiel »Durchschnittliche Seitenzahl von Diplomarbeiten im Fach Psychologie« nun schon mehrfach der Veranschaulichung diente (vgl. Boxen 7.3–7.6), soll es – um die verschiedenen Techniken zur Erhöhung der Präzision von Parameterschätzungen besser vergleichen zu können – erneut zur Demonstration einer Parameterschätzung herangezogen werden. Die bisher behandelten Stichprobenpläne verzichteten auf die Nutzung eventuell vorhandener Vorinformationen über die mutmaßliche durchschnittliche Seitenzahl. Dies ist beim Bayes'schen Ansatz anders. Er kombiniert das bereits vorhandene Wissen mit einem Stichprobenergebnis zu einer gemeinsamen Parameterschätzung.

Umfragen unter Bekannten, Kontakte mit Betreuern und die Durchsicht einiger Diplomarbeiten veranlassen die studentische Arbeitsgruppe, die sich für diese Frage interessiert, einen Mittelwert von $\mu'=100$ Seiten als den plausibelsten Wert anzunehmen. Durchschnittswerte unter 70 Seiten oder über 130 Seiten werden für äußerst unwahrscheinlich gehalten. Die in Frage kommenden durchschnittlichen Seitenzahlen weisen damit eine Streubreite (Range) von 130–70=60 auf, wobei stärker von 100 abweichende Werte für unwahrscheinlicher gehalten werden als weniger stark abweichende Werte. (Man beachte, dass hier der Range von Durchschnittswerten und nicht von Seitenzahlen einzelner Diplomarbeiten geschätzt wird, die natürlich stärker streuen als Durchschnittswerte.) Als Verteilungsvorstellung akzeptiert man eine Normalverteilung, deren Streuung auf $\sigma'=60:6=10$ geschätzt wird. (Da sich in den Grenzen $\pm 3\sigma$ etwa 100% der Normalverteilungsfläche befinden, dividieren wir zur Schätzung der Streuung den Range durch 6; genauer hierzu ▶ S. 423 f.) Für die Priorverteilung wird damit eine Normalverteilung mit $\mu'=100$ und $\sigma'=10$ angenommen.

Wie in ◼ Box 7.3 beschrieben, zieht die studentische Arbeitsgruppe nun eine Zufallsstichprobe von n=100 Diplomarbeiten und errechnet einen Mittelwert von $\bar{x}=92$ sowie eine Standardabweichung von $\hat{\sigma}=43$. Dieser Wert wird als Schätzwert für σ herangezogen. Nach ▶ Gl. (7.72) und ▶ Gl. (7.73) resultiert damit eine Posteriorverteilung mit folgenden Parametern:

$$\frac{1}{\sigma''^2}=\frac{1}{\sigma'^2}+\frac{n}{\sigma^2}=\frac{1}{10^2}+\frac{100}{43^2}=0{,}064$$

bzw.

$$\sigma''^2=15{,}60$$

$$\mu''=\frac{\frac{1}{\sigma'^2}\cdot\mu'+\frac{n}{\sigma^2}\cdot\bar{x}}{\frac{1}{\sigma'^2}+\frac{n}{\sigma^2}}$$

$$=\frac{\frac{1}{10^2}\cdot 100+\frac{100}{43^2}\cdot 92}{\frac{1}{10^2}+\frac{100}{43^2}}=93{,}2.$$

Der Erwartungswert der Posteriorverteilung ist damit nur um 1,2 Seitenzahlen größer als der Stichprobenmittelwert. Unter Verwendung von $\sigma''=\sqrt{15{,}60}=3{,}95$ resultiert das folgende 99%ige »Glaubwürdigkeitsintervall«:

$$\mu''\pm 2{,}58\cdot\sigma''=93{,}2\pm 2{,}58\cdot 3{,}95=93{,}2\pm 10{,}2.$$

Dieses Intervall ist gegenüber dem Konfidenzintervall für eine einfache Zufallsstichprobe nur geringfügig verkleinert (◼ Box 7.3). Die in der Priorverteilung zusammengefasste Vorinformation beeinflusst die Parameterschätzung also nur unerheblich.

Nach ▶ Gl. (7.76) ermitteln wir das Stichprobenäquivalent der Vorinformationen. Es lautet

$$n'=\frac{\sigma^2}{\sigma'^2}=\frac{1849}{100}=18{,}49.$$

▼

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Die Priorverteilung enthält damit Informationen, die den Informationen einer Zufallsstichprobe des Umfanges $n \approx 18$ entsprächen.

Eine Stichprobe dieser Größenordnung kann natürlich nur vage Kenntnisse über die wahren Populationsverhältnisse vermitteln, wodurch die relativ geringe Beeinflussung der Posteriorverteilung durch die Priorverteilung erklärt ist.

Die Posteriorverteilung verbindet die beiden Informationsquellen nach ► Gl. (7.78) mit den Gewichten $n'/n''=18{,}49/118{,}49=0{,}156$ für die Priorverteilung und $n/n''=100/118{,}49=0{,}844$ für das Stichprobenergebnis. Diese Gewichte führen nach ► Gl. (7.78) zu der bereits bekannten Parametereinschätzung für μ'' von

$$\mu'' = 0{,}156 \cdot 100 + 0{,}844 \cdot 92 = 93{,}2.$$

Schließlich kontrastieren wir dieses Ergebnis mit demjenigen Ergebnis, das wir für eine diffuse Priorverteilung (keine Vorinformationen) erhalten. Wir setzen hierfür $n'=0$ (gem. ► Gl. 7.76), $\sigma''^2=\sigma^2/n$ (gem. ► Gl. 7.72) und $\mu'' = \bar{x}$ (gem. ► Gl. 7.78). Die Posteriorverteilung hat damit die gleichen Parameter wie die in ◘ Box 7.3 ermittelte Stichprobenkennwerteverteilung für $\bar{x}$, d. h., die Grenzen des »Glaubwürdigkeitsintervalls« entspechen den Grenzen des Konfidenzintervalls:

$$\mu'' = 92 \pm 2{,}58 \cdot \sqrt{\frac{43^2}{100}} = 92 \pm 11.$$

7

teilung angenommen werden, deren Streuung (theoretisch) gegen unendlich geht. Eine solche Verteilung heißt im Kontext Bayes'scher Analysen »diffuse Priorverteilung«. (Auf das Problem, dass diese Verteilung keine echte Wahrscheinlichkeitsverteilung ist, wird hier nicht eingegangen. Näheres hierzu bei Hays & Winkler, 1970, Kap. 8.17; ausführlicher Berger, 1980, S. 68 ff und S. 152 ff.) Nach ► Gl. (7.76) geht n' in diesem Falle gegen 0, d. h., der Zustand der Informationslosigkeit entspricht dem »Wissen«, das einer Stichprobe des Umfanges $n'=0$ zu entnehmen ist.

Gleichzeitig verdeutlicht ► Gl. (7.78), dass für $n'=0$ der Mittelwert der Posteriorverteilung (μ'') mit dem Stichprobenmittelwert ($\bar{x}$) identisch ist. Da dann zusätzlich $n''=n$ und $\sigma''^2=\sigma^2/n$ (der Ausdruck $1/\sigma'^2$ in ► Gl. (7.72) entfällt für $\sigma'^2 \rightarrow \infty$), resultiert eine Posteriorverteilung, die der Stichprobenmittelwerteverteilung für Stichproben des Umfanges n entspricht. Das Konfidenzintervall und das Glaubwürdigkeitsintervall sind dann identisch (◘ Box 7.8).

Bis auf die bereits erwähnten interpretativen Unterschiede führen der klassische Schätzansatz und die Parameterschätzung nach dem Bayes'schen Modell zum gleichen Ergebnis, wenn die vorhandenen Informationen zur Spezifizierung einer Priorverteilung nicht ausreichen. Aber auch wenn man über Vorinformationen verfügt, sind diese meistens subjektiv und für andere nur schwer nachprüfbar. Einer Bayes'schen Parameterschätzung sollte deshalb immer eine Schätzung unter Verwendung der diffusen Priorverteilung gegenüber gestellt werden. Dadurch wird die Subjektivität bzw. der Einfluss der subjektiven Informationen auf die Parameterschätzung transparent. Empfohlen sei ferner, die Gewichte für die Vorinformation und für das empirische Ergebnis (n'/n'' und n/n'') zu nennen, auch wenn man diese (bei vollständiger Angabe der hierfür benötigten Größen) selbst errechnen könnte. Diese Mindestforderungen sollten eingehalten werden, um einer missbräuchlichen Verwendung des Bayes'schen Ansatzes entgegenzuwirken.

! Der Einfluss subjektiver Wahrscheinlichkeiten auf die Parameterschätzung nach dem Bayes'schen Ansatz wird transparent gemacht, indem man dem Schätzergebnis mit spezifizierter Priorverteilung eine Schätzung mit diffuser Priorverteilung gegenüberstellt.

Schätzung von Populationsanteilen

Die Schätzung von Populationsanteilen unter Verwendung von Priorinformationen und Stichprobeninformation ist rechnerisch noch einfacher als die Schätzung von Populationsmittelwerten. Nehmen wir einmal an, die Vorkenntnisse über einen zu schätzenden Populationsanteil lassen sich durch eine Betaverteilung mit den

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Parametern k′ und r′ abbilden. Nehmen wir ferner an, in der untersuchten Stichprobe des Umfanges n wurde die Merkmalsalternative A k-mal und die Merkmalsalternative $\bar{A}$ r(=n–k)-mal beobachtet. Die Posteriorverteilung, die diese beiden Informationen vereint, hat dann die Parameter

$$k'' = k + k' \tag{7.79}$$

und

$$r'' = r + r'. \tag{7.80}$$

Die größte Schwierigkeit besteht in der Spezifizierung der Priorverteilung als Betaverteilung. Um dies zu erleichtern, sind im Anhang (neben den in Abb. 7.13 wiedergegebenen Verteilungen) einige wichtige Betaverteilungen grafisch dargestellt (▶ Anhang F4). Die Handhabung dieser Abbildungen (nach Philips, 1973) wird im Folgenden erläutert.

Spezifizierung einer Betaverteilung. Die plausibelste Schätzung des Populationsanteils π sei π′=0,7. In ▶ Anhang F, Tab. F4c finden sich fünf verschiedene Betaverteilungen, die alle einen Modalwert von 0,7 aufweisen. Sie unterscheiden sich lediglich in der Streuung, die – wie bereits im letzten Abschnitt erwähnt – die Sicherheit der Parameterschätzung reflektiert. Je stärker die Verteilung streut, desto unsicherer ist die Schätzung. Gibt eine dieser Verteilungen die Priorverteilung einigermaßen richtig wieder, sind der Abbildung direkt die entsprechenden Parameter k′ und r′ zu entnehmen. Zur Absicherung der getroffenen Entscheidung sind unterhalb der Abbildung für jede Verteilung drei gleich wahrscheinliche Bereiche für den unbekannten Parameter π aufgeführt. Diese drei Bereiche unterteilen das Kontinuum möglicher Anteilswerte (von 0 bis 1) in drei äquivalente Intervalle. Baut unsere Schätzung π′=0,7 auf sehr sicheren Vorkenntnissen auf, werden wir vermutlich die steilste der fünf Betaverteilungen mit den Parametern k′=50 und r′=22 wählen. Von der Richtigkeit unserer Wahl überzeugen wir uns, indem wir überprüfen, ob die Bereiche 0<π<0,67; 0,67<π<0,72 und 0,72<π<1,00 tatsächlich auch nach unseren Vorstellungen gleich wahrscheinlich sind. (Hilfsregel: Sollten wir auf einen bestimmten Bereich, in dem sich π vermutlich befindet, wetten, müsste es uns schwerfallen, für diese Wette einen der drei Bereiche auszuwählen.)

Nach dieser Festlegung erfolgt die eigentliche empirische Untersuchung. Wir ziehen eine Stichprobe des Umfanges n, in der die Ereignisalternative A k-mal und $\bar{A}$ r-mal auftritt. Als Schätzwert für den Anteil des Merkmals A in der Population (π) resultiert

$$p(A) = \frac{k}{k+r} = \frac{k}{n}.$$

Nehmen wir an, n sei 100, dann ergibt sich z. B. für k=62 (und r=38) die Wahrscheinlichkeit p(A)=0,62.

Für die Häufigkeit der Ereignisse A und $\bar{A}$ haben wir die gleichen Symbole verwendet wie für die Parameter der Betaverteilung (die allerdings zusätzlich mit einem Strich versehen sind). Dies ist kein Zufall, denn es wird damit zum Ausdruck gebracht, dass die Information, die die Priorverteilung enthält, mit der Information einer Stichprobe des Umfanges n′=k′+r′ mit den Häufigkeiten k′ für A und r′ für $\bar{A}$ äquivalent ist. Wählen wir eine Betaverteilung mit k′=50 und r′=22 als Priorverteilung, wird damit behauptet, dass unsere Vorinformationen mit dem Ergebnis einer Stichprobenuntersuchung gleichwertig sind, in der unter 50+22=72 Untersuchungseinheiten 50-mal die Merkmalsalternative A beobachtet wurde. Aufgrund dieser fiktiven Stichprobe würden wir den gesuchten Parameter mit p(A)=50:72=0,69 schätzen, d. h. mit einem Wert, der dem arithmetischen Mittel der Betaverteilung für k=50 und r=22 gem. ▶ Gl. (7.70) entspricht. (Man beachte, dass die Betaverteilungen für k≠r nicht symmetrisch sind, dass also Modalwert und arithmetisches Mittel in diesem Falle nicht identisch sind. Unsere anfängliche Entscheidung, π′=0,70 als beste Schätzung anzusehen, bezog sich auf den Modalwert als den »wahrscheinlichsten« Wert und nicht auf den Mittelwert.) Nach ▶ Gl. (7.79) und ▶ Gl. (7.80) sind die Parameter der Posteriorverteilung leicht zu ermitteln. Sie lauten k″=62+50=112 und r″=38+22=60. Die Posteriorverteilung hat damit einen Mittelwert von 112:172=0,651 und nach ▶ Gl. (7.69) einen hiervon nur geringfügig abweichenden Modalwert von 111:170=0,653. Der Unterschied dieser Werte wächst – wie ▶ Gl. (7.69) und ▶ Gl. (7.70) zeigen – mit abnehmender Summe k+r.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

In ▶ Anhang F4 sind nur Betaverteilungen mit den Modalwerten 0,5; 0,6; 0,7; 0,8 und 0,9 enthalten. Betaverteilungen mit Modalwerten unter 0,5 sind deshalb nicht aufgeführt, weil die Parameter k und r symmetrisch sind. Verteilungen mit einem Modalwert bei 0,4 z. B. sind die Spiegelbilder der Verteilungen mit einem Modalwert bei 0,6. Verteilungen, deren Modalwerte unter 0,5 liegen, erhält man also einfach durch Vertauschen der Parameter k und r.

Trifft keine der im ▶ Anhang F4 wiedergegebenen Verteilungen die Vorstellung über die Priorverteilung, wird man probeweise andere als die dort herausgegriffenen Parameter in ▶ Gl. (7.68) einsetzen und sich die dann resultierende Verteilung grafisch veranschaulichen. (In der Regel genügen hierfür einige Punkte der Verteilung.) Auf die Wiedergabe von Betaverteilungen mit $k<1$ und $r<1$ wurde verzichtet, weil diese Betaverteilungen u-förmig sind und als Priorverteilungen für einen zu schätzenden Populationsanteil nicht in Frage kommen.

Glaubwürdigkeitsintervalle. Auf ▶ S. 414 wurde bereits darauf hingewiesen, dass bei einer Normalverteilung prinzipiell beliebig viele Intervalle existieren, über denen sich 95% (99%) der Gesamtfläche befinden. Dies gilt natürlich ebenso für die Betaverteilung, auch wenn diese nur zwischen den Werten 0 und 1 definiert ist. Die formal gleichwertigen Intervalle unterscheiden sich jedoch in ihrer Länge. Als Glaubwürdigkeitsintervall wählen wir das kürzeste Intervall bzw. das Intervall mit der höchsten Wahrscheinlichkeitsdichte.

Um diese Intervalle zu finden, benötigen wir das **Integral der Betaverteilung**. Wir suchen diejenigen Grenzen, zwischen denen sich einerseits 95% (99%) der Gesamtfläche befinden und die andererseits einen minimalen Abstand voneinander haben. Diese Grenzen für die jeweilige Betaverteilung zu finden, die als Posteriorverteilung resultiert, ist rechnerisch sehr aufwändig. Sie sind deshalb für die gebräuchlichsten Betaverteilungen im ▶ Anhang F5 tabellarisch aufgeführt.

Die Tabellen enthalten die Grenzen für 95%ige und 99%ige Glaubwürdigkeitsintervalle von Betaverteilungen mit den Parametern $k \leq 60$ und $r \leq 60$. Resultiert für die Posteriorverteilung eine Betaverteilung, deren Parameter außerhalb dieser Grenzen liegen (was leicht passiert, wenn relativ große Stichproben untersucht werden), macht man sich die Tatsache zunutze, dass die Betaverteilung mit wachsendem k und r in eine Normalverteilung übergeht. Die Grenzen der Glaubwürdigkeitsintervalle können dann nach der schon bekannten Formel für normalverteilte Zufallsvariablen ermittelt werden:

95%iges Glaubwürdigkeitsintervall

$$\text{obere Grenze} = \mu'' + 1{,}96 \cdot \sigma''$$
$$\text{untere Grenze} = \mu'' - 1{,}96 \cdot \sigma''$$

99%iges Glaubwürdigkeitsintervall

$$\text{obere Grenze} = \mu'' + 2{,}58 \cdot \sigma''$$
$$\text{untere Grenze} = \mu'' - 2{,}58 \cdot \sigma''.$$

μ'' und σ'' werden nach den Gleichungen (7.70) und (7.71) bestimmt. Im oben erwähnten numerischen Beispiel resultierte als Posteriorverteilung eine Betaverteilung mit $k''=112$ und $r''=60$, deren Glaubwürdigkeitsintervalle nicht mehr tabelliert sind. Wir verwenden deshalb die Normalverteilungsapproximation und ermitteln

$$\mu'' = \frac{112}{112+60} = 0{,}65$$

sowie

$$\sigma'' = \sqrt{\frac{112 \cdot 60}{(112+60)^2 \cdot (112+60+1)}} = 0{,}036.$$

Damit hat das 95%ige Glaubwürdigkeitsintervall folgende Grenzen:

$$\text{obere Grenze} = 0{,}65 + 1{,}96 \cdot 0{,}036 = 0{,}72$$
$$\text{untere Grenze} = 0{,}65 - 1{,}96 \cdot 0{,}036 = 0{,}58.$$

Als diffuse Priorverteilung, die den Zustand totaler Informationslosigkeit charakterisiert, wählen wir eine »Betaverteilung« mit $k'=0$ und $r'=0$. (Die Betaverteilung ist für diese Parameter nicht definiert, deshalb die Anführungszeichen. Dennoch verwenden wir diese Parameter zur Kennzeichnung der Informationslosigkeit, denn die eigentlich angemessene Betaverteilung mit $k'=1$ und $r'=1$ – diese Parameter definieren eine Gleichverteilung [vgl. ◻ Abb. 7.13a] – repräsentiert keine totale Informationslosigkeit, sondern eine Stich-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

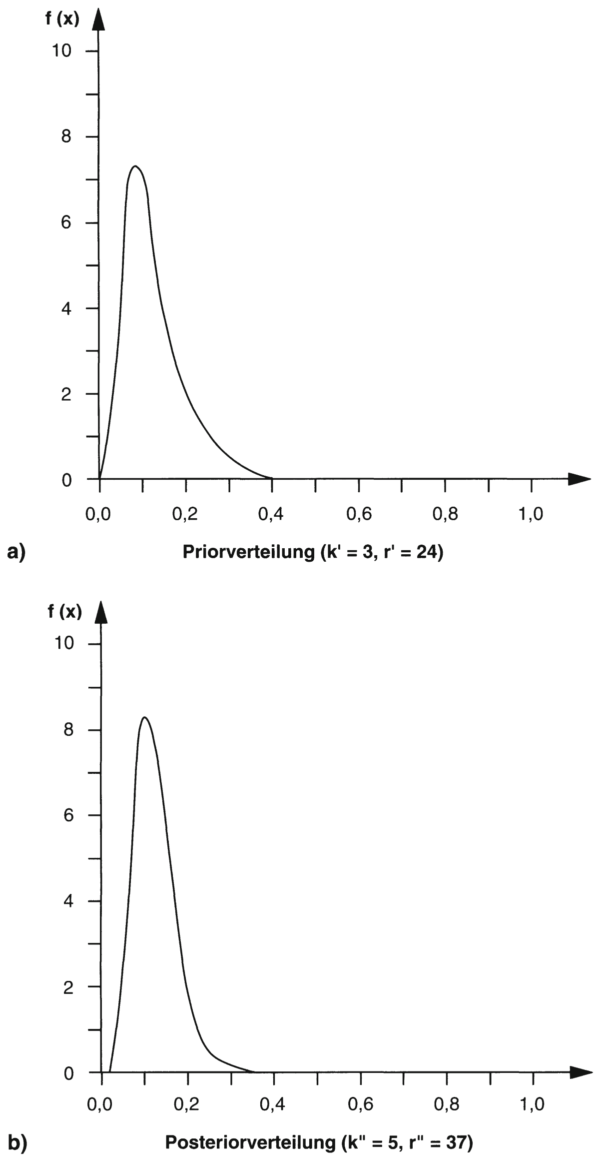


Abb. 7.14. Priorverteilung (**a**) und Posteriorverteilung (**b**) für die Wahrscheinlichkeit von Nebenwirkungen eines neuen Präparates

probe mit n′=2, k′=1 und r′=1. Näheres zu diesem Problem s. Hays und Winkler, 1970, Kap. 8.18.) Nur für k′=0 und r′=0 resultiert eine Posteriorverteilung, deren Parameter ausschließlich vom Stichprobenergebnis bestimmt wird.

Ein **Beispiel** soll die Verwendung des Bayes'schen Ansatzes zur Schätzung eines Populationsanteils verdeutlichen. Eine pharmazeutische Firma entwickelt ein neues Präparat und will dieses in einem Tierversuch mit einer Stichprobe von n=15 Tieren auf Nebenwirkungen testen. Da man mit der Wirkung ähnlicher Präparate schon viele Erfahrungen gesammelt hat, schätzt man vorab, dass ca. 8% (π′=0,08) der Population behandelter Tiere Nebenwirkungen zeigen. Als Priorverteilung wird eine Betaverteilung mit k′=3 und r′=24 spezifiziert. (Durch Festlegung des Modalwertes auf 0,08 und durch k′=3 ist r′ gem. ▶ Gl. 7.69 nicht mehr frei variierbar.) Abb. 7.14a zeigt diese Priorverteilung grafisch. Die relativ geringe Streuung dieser Verteilung belegt, dass die Untersuchenden von der Richtigkeit ihrer Parameterschätzung ziemlich fest überzeugt sind.

Bei der Überprüfung der n=15 behandelten Tiere möge sich herausstellen, dass 2 Tiere (k=2) Nebenwirkungen zeigen. Damit ist r=13. Für die Posteriorverteilung resultieren die Parameter k″=3+2=5 und r″=24+13=37. In Abb. 7.14b wird auch diese Verteilung veranschaulicht. Sie ist noch steiler als die Priorverteilung und hat einen Modalwert von 0,1. Das 95%ige Glaubwürdigkeitsintervall entnehmen wir ▶ Anhang F5. Es hat die Grenzen 0,032 und 0,217. Der wahre Anteil von Tieren, bei denen das Präparat Nebenwirkungen zeigt, liegt also mit einer »Wahrscheinlichkeit« von 95% im Bereich 3,2% bis 21,7%. In diesem Falle führt also auch die Zusammenfassung von Vorinformation und Stichprobenergebnis zu einer sehr ungenauen Parameterschätzung.

Wie bereits bei der Schätzung von Populationsmittelwerten fragen wir auch hier, mit welchen Anteilen die Vorinformationen (Priorverteilung) und das Stichprobenergebnis in die Posteriorverteilung eingehen. Die Priorverteilung ist einem Stichprobenergebnis mit k′=3, r′=24 und n′= k′+r′=27 äquivalent. Untersucht wurden n=15 Tiere, d. h., die Vorinformationen haben ein Gewicht von 27/(27+15)=0,64 und das Stichprobenergebnis von 15/(27+15)=0,36. Die Vorinformationen fallen in dieser Untersuchung stärker ins Gewicht als die empirische Evidenz.

Ohne Vorinformationen müssten wir für die Priorverteilung k′=0 und r′=0 annehmen, d. h., die Posteriorverteilung hätte die Parameter k″=2 und r″=13. Für diese Verteilung entnehmen wir ▶ Anhang F5 ein 95%iges Glaubwürdigkeitsintervall mit den Grenzen 0,0045 und 0,2988. Wie nicht anders zu erwarten, erlaubt eine Stich-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

probe mit n=15 nur eine ungenaue Schätzung des Parameters. Entsprechend wird die Priorverteilung durch die Berücksichtigung des Stichprobenergebnisses nur wenig verändert.

Datenrückgriff. Das Beispiel auf ▶ S. 467 f. (Marktanteile eines neuen Produktes) verdeutlichte die direkte Anwendung des Bayes'schen Theorems für stetige Variablen gem. ▶ Gl. (7.67). Die Lösung des dort angesprochenen Problems wird erheblich vereinfacht, wenn wir die Eigenschaften konjugierter Verteilungen (hier: Betaverteilungen) ausnutzen. Als Priorverteilung wurde eine Dreiecksverteilung mit $f(\theta)=2\cdot(1-\theta)$ spezifiziert. Diese Verteilung entspricht einer Betaverteilung mit $k'=1$ und $r'=2$:

$$
\begin{aligned}
f(\theta) &= \frac{(k+r-1)!}{(k-1)!\cdot(r-1)!}\cdot\theta^{k-1}\cdot(1-\theta)^{r-1}\\
&= \frac{(3-1)!}{(1-1)!\cdot(2-1)!}\cdot\theta^{1-1}\cdot(1-\theta)^{2-1}\\
&= \frac{2}{1\cdot 1}\cdot 1\cdot(1-\theta)\\
&= 2\cdot(1-\theta).
\end{aligned}
$$

(Man beachte: $0!=1$ und $\theta^0=1$.)

Mit $n=5$, $k=1$ und $r=4$ als Stichprobenergebnis erhalten wir eine Posteriorverteilung mit $k''=2$ und $r''=6$. Wie die folgenden Umformungen zeigen, ist diese Betaverteilung mit der Verteilung $f(\theta|y)=42\cdot\theta\cdot(1-\theta)^5$, die wir nach direkter Anwendung des Bayes'schen Theorems errechneten, identisch:

$$
\begin{aligned}
f(\theta|y) &= \frac{(2+6-1)!}{(2-1)!\cdot(6-1)!}\cdot\theta^{2-1}\cdot(1-\theta)^{6-1}\\
&= \frac{7!}{1!\cdot 5!}\cdot\theta^{1}\cdot(1-\theta)^{5}\\
&= 42\cdot\theta\cdot(1-\theta)^{5}.
\end{aligned}
$$

Hinweis. Auf ▶ S. 474 wurde im Zusammenhang mit der Schätzung von Populationsmittelwerten vor der missbräuchlichen Anwendung des Bayes'schen Ansatzes gewarnt. Die dort vorgebrachten Argumente und Empfehlungen gelten selbstverständlich auch für die Schätzung von Populationsanteilen nach dem Bayes'schen Theorem.

7.2.6 Resamplingansatz

Der bereits in den 70er Jahren des 20. Jahrhunderts entwickelte, aber erst in den 90er Jahren populär gewordene Resamplingansatz baut bei der Hypothesenprüfung und Parameterschätzung nicht auf analytischen Methoden auf, sondern arbeitet rein empirisch mit computergestützten **Simulationen**. Dazu werden aus der empirisch untersuchten Stichprobe wiederholt und systematisch (mit oder ohne Zurücklegen) weitere (Teil-)Stichproben gezogen (deswegen: re-sampling) und die dabei entstehenden Kennwerteverteilungen betrachtet.

Drei Vorteile werden dem Resamplingansatz von seinen Entwicklern zugeschrieben (vgl. Simon, o.J.; Simon & Bruce, 1991; Rietz et al., 1997):

1. Die Auswertung mit Resamplingmethoden ist auch für Personen ohne fundierte Mathematik- oder Statistikausbildung logisch gut nachvollziehbar, dementsprechend kommt es in der Forschung auch seltener zu Auswertungsfehlern.
2. Evaluationsstudien haben gezeigt, dass eine Statistikausbildung auf der Basis von Resamplingmethoden anstelle herkömmlicher analytischer Verfahren bei Studierenden zu einem besseren und schnelleren Verständnis der Materie führt.
3. Die Auswertung mit Resamplingmethoden ist für viele Datensätze möglich, teilweise auch für solche, bei denen die Voraussetzungen gängiger parametrischer oder nonparametrischer analytischer Verfahren verletzt sind.

Ein Ansatz, der Pragmatismus und Einfachheit so stark betont, steht natürlich schnell im Verdacht, letztlich unseriöse Ergebnisse zu liefern. Und tatsächlich mag es zunächst verwunderlich sein, wie man allein durch das Rechnen mit *einer* Stichprobe und daraus immer wieder neu entnommenen Teilstichproben zu verallgemeinerbaren Aussagen kommen kann. Dass man sich beim Resampling quasi am eigenen Schopf aus dem (Daten-) Sumpf zieht, deutet der auf die Arbeiten von Efron (Efron & Tibshirani, 1993) zurückgehende Begriff »**Bootstrap**«-**Verfahren** an. (Offenbar zieht sich der Held der Münchhausen-Sage in der amerikanischen Version nicht an seinem Schopf, sondern an seinen Schuhriemen = Bootstrap aus dem Sumpf). Auch wenn die Durchführung einer Bootstrapanalyse mit einem leistungsstarken Computer

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

keine besonderen Probleme bereitet – die mathematische Theorie bzw. die Beweise, die hinter diesem Ansatz stehen, sind nicht einfach.

Zur Illustration des Bootstrapverfahrens möge erneut das Diplomarbeiten-Beispiel dienen. In Box 7.3 ▶ S. 416) wurde erläutert, wie man mit einer Zufallsstichprobe von 100 Diplomarbeiten den durchschnittlichen Umfang von Diplomarbeiten schätzen kann bzw. wie die Unsicherheit in dieser Schätzung durch ein Konfidenzintervall quantifiziert wird. Beim Bootstrapverfahren würden wir aus der Stichprobe der n=100 Seitenzahlen viele (ca. 5000) neue Stichproben des Umfanges n=100 mit Zurücklegen bilden. Veranschaulicht an einem Urnenmodell haben wir es also mit einer Urne zu tun, in der sich Lose mit den Seitenzahlen der Arbeiten unserer Stichprobe befinden. Wir entnehmen der Urne zufällig ein Los, notieren die Seitenzahl und legen das Los wieder in die Urne. Die ersten so gezogenen 100 Seitenzahlen bilden die 1. sog. **Bootstrapstichprobe** (in der sich theoretisch 100-mal dieselbe Seitenzahl befinden könnte). Jede Kombination von 100 Seitenzahlen tritt mit einer Wahrscheinlichkeit von $(1/100)^{100}$ auf. Diese Prozedur wird 5000-mal wiederholt, d. h., man erzeugt 5000 Bootstrapstichproben.

Jede Bootstrapstichprobe liefert einen $\bar{x}$-Wert, dessen Verteilung der auf ▶ S. 411 erwähnten $\bar{X}$-Werteverteilung entspricht. Die Streuung dieser Verteilung schätzt den Standardfehler $\sigma_{\bar{x}} = \sqrt{\sigma^2/n}$, mit dem wir – wie auf ▶ S. 411 ff. beschrieben – ein Konfidenzintervall berechnen können.

Eine weitere zum Resamplingansatz zählende Verfahrensgruppe sind **Randomisierungstests**, auf die wir im ▶ Abschn. 8.2.6 ausführlicher eingehen.

Erwähnt werden sollte in diesem Zusammenhang auch die **Monte-Carlo-Methode**, die 1949 von Metropolis und Ulan für unterschiedliche Forschungszwecke eingeführt wurde. Wichtige Anwendungsfelder sind die Erzeugung von H_0-Verteilungen statistischer Kennwerte mit Hilfe vieler Zufallszahlen und die Überprüfung der Folgen, die mit der Verletzung der Voraussetzungen statistischer Tests verbunden sind. Zufallszahlen werden heute mit dem Computer erzeugt (zum Zufallskonzept vgl. Beltrami, 1999; Everitt, 1999), was in der Mitte des letzten Jahrhunderts nicht so ohne weiteres möglich war. Statt dessen hat man auf die Zufallszahlen von Roulettepermanenzen zurückgegriffen, die von den Spielbanken – und eben auch von der berühmten Spielbank in Monte Carlo – regelmäßig und kontinuierlich registriert werden. Viele Hinweise zu Theorie und Praxis von Monte-Carlo-Studien sowie eine ausführliche Bibliografie findet man bei Robert und Casella (2000).

Zum Thema »Resampling« empfehlen wir ferner Lunneborg (1999) und speziell für Permutationstests Good (2000). Resamplingsoftware findet man im Internet unter: http://www.resample.com/ (kostenlose Demoversion).

7.2.7 Übersicht populationsbeschreibender Untersuchungen

In ▶ Kap. 7 werden Untersuchungsarten behandelt, deren gemeinsames Ziel die Beschreibung von Populationen bzw. die Schätzung von Populationsparametern ist. Wir untersuchten die in der Praxis am häufigsten interessierenden Parameter: Populationsmittelwerte und Populationsanteile. Vollerhebungen sind hierfür in den meisten Fällen unzweckmäßig, denn sie erfordern einen zu hohen Kosten- und Zeitaufwand. Sie versagen vor allem bei der Erfassung von Merkmalen, die einem raschen zeitlichen Wandel unterliegen. Die ausschnittsweise Erfassung von Populationen durch Stichproben, die erheblich schneller und billiger zu untersuchen sind und die zu Resultaten führen, deren Präzision bei sorgfältiger Planung der einer Vollerhebung kaum nachsteht, ist deshalb ein unverzichtbares Untersuchungsinstrument.

Eine optimale Nutzung der vielfältigen Stichprobenpläne setzt eine gründliche theoretische Auseinandersetzung mit dem zu untersuchenden Merkmal bzw. mit der zu beschreibenden Population sowie das Studium von evtl. bereits durchgeführten Untersuchungen zur selben Thematik voraus. Die Berücksichtigung von Vorkenntnissen kann die Präzision einer Parameterschätzung beträchtlich erhöhen und den technischen wie auch finanziellen Untersuchungsaufwand entscheidend reduzieren. In diesem Sinne verwertbare Vorkenntnisse betreffen

1. das zu untersuchende Merkmal selbst (Art der Verteilung des Merkmals, Streuung des Merkmals, Vorstellungen über die Größe des unbekannten Parameters, »Wahrscheinlichkeitsverteilung« des Parameters),

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

2. andere, mit dem zu erhebenden Merkmal zusammenhängende Merkmale, die eine einfache Untergliederung (Schichtung) der Population gestatten (Umfang der Schichten und Streuung des zu erhebenden Merkmals in den Schichten),
3. Besonderheiten bezüglich der Zusammensetzung der Population (natürlich zusammenhängende Teilgesamtheiten oder Klumpen) sowie
4. Stichproben, die bezüglich des interessierenden Merkmals bereits untersucht wurden.

Einfache Zufallsstichprobe. Hat man sich vergewissert, dass auf keine Vorkenntnisse zurückgegriffen werden kann, dass eine Untersuchung also »wissenschaftliches Neuland« betritt, kommt als Stichprobenart nur die einfache Zufallsstichprobe in Frage. Sie setzt voraus, dass jedes einzelne Untersuchungsobjekt der Population individuell erfasst ist, sodass ein Auswahlplan erstellt werden kann, der gewährleistet, dass jedes einzelne Untersuchungsobjekt mit gleicher Wahrscheinlichkeit Teil der Stichprobe wird. Der Stichprobenumfang ist hierbei nicht willkürlich, sondern in Abhängigkeit von der gewünschten Schätzgenauigkeit (Breite des Konfidenzintervalls) festzulegen. Auf ▶ S. 419 ff. diskutierten wir Möglichkeiten, den erforderlichen Stichprobenumfang auch dann zu kalkulieren, wenn die Streuung des Merkmals in der Population unbekannt ist.

Nur wenige populationsbeschreibende Untersuchungen erfüllen die Erfordernisse einer einfachen Zufallsstichprobe perfekt. Entweder ist die vollständige Liste aller zur Population zählenden Untersuchungsobjekte unbekannt, oder man wählt die Stichprobe nach einem Verfahren aus, das keine konstante Auswahlwahrscheinlichkeit für jedes Untersuchungsobjekt garantiert. Nicht selten verletzen populationsbeschreibende Untersuchungen beide Voraussetzungen.

Dies muss nicht immer ein Zeichen für eine oberflächliche oder nachlässige Untersuchungsplanung sein. In ihrem Bemühen, beide Kriterien erfüllen zu wollen, stehen Untersuchende oft vor unüberwindlichen Schwierigkeiten, denn die Erfüllung der Kriterien erfordert einen Untersuchungsaufwand, der häufig in keinem Verhältnis zu den zu erwartenden Erkenntnissen steht. Man begnügt sich deshalb mit der Untersuchung von »Pseudozufallsstichproben«, »Bequemlichkeitsauswahlen« oder »anfallenden« Stichproben, die aus einer mehr oder weniger beliebigen, leicht zugänglichen Ansammlung von Untersuchungsobjekten bestehen.

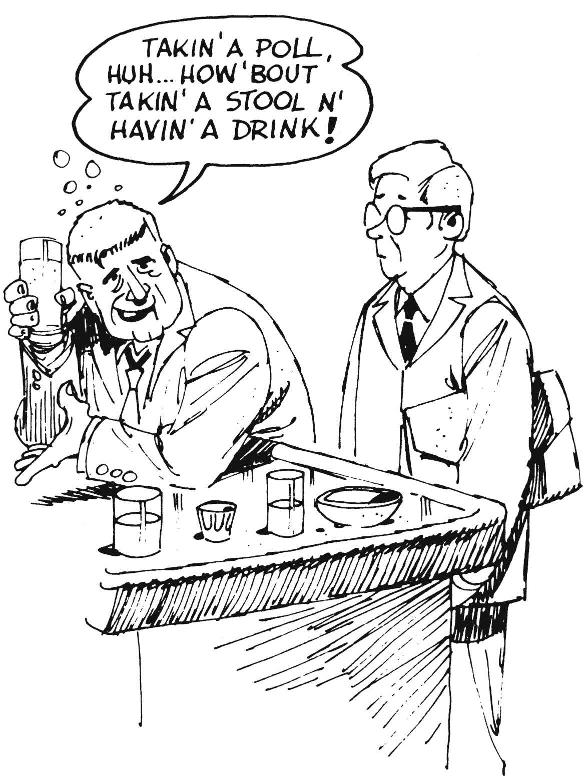
Diese Untersuchungen sind unwissenschaftlich, wenn ihre Ergebnisse leichtfertig auf Populationen verallgemeinert werden, die tatsächlich nicht einmal auszugsweise, geschweige denn nach Kriterien reiner Zufallsauswahlen, untersucht wurden. Sie haben bestenfalls den Charakter explorativer, hypothesengenerierender Untersuchungen und sollten auch als solche deklariert werden.

Wenn schon – aus welchen Gründen auch immer – bei vielen populationsbeschreibenden Untersuchungen auf die Ziehung einer reinen Zufallsstichprobe verzichtet werden muss, sollte der Untersuchungsbericht zumindest folgende Fragen diskutieren:

- Für welche Population gelten die Untersuchungsergebnisse?
- Nach welchem Verfahren wurden die Untersuchungsobjekte ausgewählt?
- Inwieweit ist die Generalisierung der Ergebnisse durch Besonderheiten des Auswahlverfahrens eingeschränkt?
- Gibt es strukturelle Besonderheiten der Population, die eine Verallgemeinerung der Ergebnisse auch auf andere, nicht untersuchte Populationen rechtfertigen?
- Welche Präzision (Konfidenzintervall) haben die Ergebnisse angesichts des untersuchten Stichprobenumfangs?
- Welche Überlegungen nahmen Einfluss auf die Festlegung des Stichprobenumfangs?

Untersuchungen, die diese Punkte kompromisslos diskutieren und Schwächen nicht verschweigen, können an Glaubwürdigkeit nur gewinnen (in vielen Punkten vorbildlich ist hierfür z. B. eine Untersuchung von Lenski, 1963, zit. nach Sudman, 1976). Dies gilt nicht nur für Untersuchungen mit einfachen Zufallsstichproben (oder »Pseudozufallsstichproben«), sondern natürlich auch für die im Folgenden zusammengefassten »fortgeschrittenen« Stichprobenpläne, die als probabilistische Stichproben auch mit dem statistischen Zufallsprinzip arbeiten. Im Vorgriff auf ▶ Kap. 8 und 9 sei bereits jetzt darauf hingewiesen, dass sich die hier geforderte freimütige Darlegung der Art der Stichprobenziehung auch auf hypothesenprüfende Untersuchungen bezieht.

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,



Die Präzision von Umfrageergebnissen hängt nicht nur von Art und Umfang der Stichprobe, sondern letztlich auch von der Zuverlässigkeit der erhaltenen Auskünfte ab. Aus Campbell, S.K. (1974). Flaws and Fallacies in Statistical Thinking. Englewood Cliffs: Prentice-Hall, S. 136

Geschichtete Stichprobe. Im Vergleich zu einfachen Zufallsstichproben wird die Parameterschätzung erheblich präziser, wenn es gelingt, die Population nach einem Merkmal zu schichten, von dem bekannt ist, dass es mit dem untersuchten Merkmal hoch korreliert. Die endgültige Stichprobe setzt sich dann aus homogenen Teilstichproben zusammen, die den Populationsschichten zufällig entnommen wurden. Der Vorteil geschichteter Stichproben gegenüber einfachen Stichproben kommt jedoch erst dann voll zum Tragen, wenn zusätzlich zum Schichtungsmerkmal die Größen der Teilpopulationen sowie deren Streuungen bekannt sind.

Schichtungsmerkmale sollten nicht nur mit dem untersuchten Merkmal hoch korrelieren, sondern zugleich einfach erhebbar sein. Intelligenz- und Einstellungsvariablen sind beispielsweise Merkmale, die zwar mit vielen sozialwissenschaftlich interessanten Merkmalen zusammenhängen; der Aufwand, der zu ihrer Erfassung erforderlich ist, macht sie jedoch als Schichtungsmerkmal praktisch unbrauchbar. Korrelieren hingegen einfache Merkmale wie Alter, Geschlecht, Einkommen, Art der Ausbildung etc. mit dem zu untersuchenden Merkmal, sind die Randbedingungen für eine geschichtete Stichprobe weitaus günstiger.

Bisher gingen wir davon aus, dass die Schichtung nur in Bezug auf ein Merkmal vorgenommen wird. Dies ist jedoch keineswegs erforderlich und bei Untersuchungen, die gleichzeitig mehrere Merkmale erheben (Omnibusuntersuchungen), auch nicht sehr sinnvoll. Hängen nämlich die einzelnen zu untersuchenden Merkmale mit jeweils anderen Schichtungsmerkmalen zusammen, kann bei einer nur nach einem Merkmal geschichteten Stichprobe natürlich nur dasjenige Merkmal genauer geschätzt werden, das mit dem Schichtungsmerkmal korreliert. Die Parameter der übrigen Merkmale werden dann genauso exakt geschätzt wie mit einer einfachen Zufallsstichprobe. In diesem Falle und im Falle eines Untersuchungsmerkmals, das mit mehreren Schichtungsmerkmalen zusammenhängt, empfiehlt es sich, die Schichtung gleichzeitig nach mehreren Merkmalen vorzunehmen.

Sind beispielsweise sowohl das Geschlecht (männlich/weiblich) als auch das Einkommen der Untersuchungsteilnehmer (geringes/mittleres/hohes Einkommen) wichtige Schichtungsmerkmale, hätte man Stichproben aus sechs Teilgesamtheiten, die als Kombinationen dieser beiden Merkmale resultieren, zu untersuchen. Dieser Aufwand lohnt sich allerdings nur, wenn auch die Umfänge und Streuungen dieser Teilpopulationen bekannt sind.

Klumpenstichprobe. Den geringsten untersuchungstechnischen Aufwand bereiten Untersuchungen von Populationen, die sich aus vielen kleinen, heterogenen (und wenn möglich gleich großen) Teilgesamtheiten (Klumpen) zusammensetzen. Hier kann auf eine vollständige Liste aller Untersuchungsobjekte der Population verzichtet werden; es genügt eine Zusammenstellung aller Klumpen, aus der eine zufällige Auswahl getroffen wird. Man untersucht jeden ausgewählten Klumpen vollständig, d. h. mit allen Untersuchungsobjekten. Gleichheit aller Klumpen und Verschiedenar-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

tigkeit der Untersuchungsobjekte innerhalb der Klumpen sind hier die besten Voraussetzungen für eine präzise Parameterschätzung.

Mehrstufige Stichprobe. Die idealen Verhältnisse für eine Klumpenstichprobe wird man in der Praxis selten antreffen. Zwar setzt sich eine Population häufig aus mehreren natürlich gewachsenen Teilgesamtheiten oder Klumpen zusammen; diese sind jedoch häufig zu groß, um einige von ihnen vollständig erheben zu können. Die in diesem Falle einschlägige 2-stufige (oder mehrstufige) Stichprobe untersucht die ausgewählten Klumpen nur stichprobenartig. Erneut benötigt man keine vollständige Liste aller Untersuchungsobjekte der Population, sondern lediglich Aufstellungen derjenigen Untersuchungsobjekte, die sich in den ausgewählten Klumpen (bzw. bei mehrstufigen Stichproben in den Einheiten der letzten Auswahlstufe) befinden.

Die geschichtete und die Klumpenstichprobe sind als Spezialfälle einer 2-stufigen Stichprobe darstellbar. Lassen wir die Klumpen einer 2-stufigen Stichprobe größer und damit deren Anzahl kleiner werden, nähern wir uns den Verhältnissen einer geschichteten Stichprobe. Besteht die Population schließlich nur noch aus wenigen, sehr großen »Klumpen«, die alle stichprobenartig untersucht werden, entspricht die 2-stufige Stichprobe einer geschichteten Stichprobe. Wächst hingegen bei einer 2-stufigen Stichprobe die Anzahl der Klumpen (was bei einer gegebenen Population einer Verkleinerung der Klumpenumfänge gleichkommt), können wir auf die Ziehung von Stichproben aus den Klumpen verzichten und einige Klumpen vollständig untersuchen. Die 2-stufige Stichprobe wäre dann einer Klumpenstichprobe gleichzusetzen.

Auf ► S. 444 f. wurde gezeigt, wann eine 2-stufige Stichprobe zu besonders präzisen Parameterschätzungen führt: Mit wachsender Klumpenanzahl (bzw. abnehmender Klumpengröße) verbessern in sich heterogene, aber untereinander homogene Klumpen die Parameterschätzung. Für wenige, aber große Klumpen hingegen sind in sich homogene, aber untereinander heterogene Klumpen vorteilhaft.

Wiederholte Stichprobenuntersuchung. Stichproben, in denen einige bereits untersuchte Personen wiederverwendet werden, gewährleisten ebenfalls präzisere Parameterschätzungen als einfache Zufallsstichproben. Dies setzt allerdings voraus, dass die zu einem früheren Zeitpunkt erhobenen Messungen mit den aktuellen Messungen korrelieren. Die Höhe dieser Korrelation bestimmt, welcher Anteil an wiederverwendeten und neuen Untersuchungsobjekten die Parameterschätzung optimiert.

Bayes'scher Ansatz. Die letzte hier behandelte Art von Vorkenntnissen zur Verbesserung einer Parameterschätzung betrifft den zu schätzenden Parameter selbst. Wenn man – sei es aufgrund von Voruntersuchungen oder anderer Informationen – eine mehr oder weniger präzise Vorstellung über den »wahrscheinlichsten« Parameter hat, wenn man bestimmte Parameterausprägungen als zu »unwahrscheinlich« ausschließen kann und wenn man für die im Prinzip in Frage kommenden Parameter eine Wahrscheinlichkeitsverteilung (Dichteverteilung) in Form einer Priorverteilung spezifizieren kann, ermöglicht die Bayes'sche Statistik eine Parameterschätzung, die sowohl die Vorkenntnisse als auch das Resultat einer Stichprobenuntersuchung berücksichtigt. Dieser für viele sozialwissenschaftliche Fragen angemessene Ansatz (wann kommt es schon einmal vor, dass eine Untersuchung ohne jegliche Vorkenntnisse über das zu erhebende Merkmal begonnen wird?) kann Parameterschätzungen deutlich verbessern. Er bereitet dem unerfahrenen Forscher in seinem Bemühen, Vorkenntnisse in eine angemessene Priorverteilung zu transformieren, allerdings anfänglich Schwierigkeiten. Dennoch ist es ratsam, sich im Umgang mit dieser Methode zu üben, denn schließlich ist es nicht einzusehen, warum das in vergangenen Forschungen erarbeitete Wissen bei einem aktuellen Schätzproblem unberücksichtigt bleiben soll. Zudem gestattet es der Bayes'sche Ansatz, den Einfluss der subjektiven Vorinformationen auf die Parameterschätzung zu kontrollieren. Die Verwendung sog. »diffuser Priorverteilungen« macht die Bestandteile, die in die Parameterschätzung eingehen, transparent. So verstanden kann der Bayes'sche Ansatz die sozialwissenschaftliche Forschungspraxis erheblich bereichern.

Kombinierte Stichprobenpläne. Geschichtete Stichproben, Klumpenstichproben, 2- oder mehrstufige Stichproben und Stichproben mit wiederholten Messungen wurden bisher als alternative Stichprobenpläne behan-

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

delt. Die Elemente dieser Stichprobenpläne lassen sich jedoch beliebig zu neuen, komplexeren Stichprobenplänen kombinieren. Man könnte beispielsweise bei einer 2-stufigen Stichprobe aus den Klumpen keine Zufallsstichproben, sondern geschichtete Stichproben ziehen, in denen sich zusätzlich einige Untersuchungsobjekte befinden, die bereits zu einem früheren Zeitpunkt untersucht wurden. (Beispiel: In jeder ausgewählten Universität wird eine nach dem Merkmal »Studienfach« geschichtete Stichprobe gezogen und in jeder Stichprobe befinden sich einige wiederholt untersuchte Personen.) Oder man entnimmt den Teilpopulationen einer geschichteten Stichprobe keine Zufallsstichproben (einfache geschichtete Stichprobe), sondern einzelne zufällig ausgewählte Klumpen, die ihrerseits vollständig erhoben werden (Beispiel: Man zieht eine geschichtete Stichprobe von Hortkindern mit dem Schichtungsmerkmal »Art des Hortes«: staatlich oder kirchlich gefördert. Statt zweier Zufallsstichproben aus diesen Teilpopulationen untersucht man einige zufällig ausgewählte Horte beider Schichten vollständig.)

Diese Beispiele verdeutlichen Kombinationsmöglichkeiten, deren »Mathematik« allerdings nicht ganz einfach ist. Als konkrete Realisierung eines komplexeren Stichprobenplanes sei das »ADM-Mastersample« der »Arbeitsgemeinschaft Deutscher Marktforschungsinstitute« erwähnt, das in Box 7.9 kurz beschrieben wird (vgl. Jacob & Eirmbter, 2000, S. 102 ff.; Schnell et al. 1999, S. 268 f.) bzw. ausführlicher von der Arbeitsgemeinschaft ADM-Stichproben/Bureau Wendt (1994).

Das empirische Untersuchungsergebnis, das im Bayes'schen Ansatz mit dem Vorwissen kombiniert wird, basiert nach unseren bisherigen Ausführungen auf einer einfachen Zufallsstichprobe. Aber auch diese Einschränkung ist nicht zwingend. Weiß man nicht nur, wie der zu schätzende Parameter ungefähr »verteilt« ist, sondern zusätzlich, dass bestimmte, einfach zu erhebende Merkmale mit dem untersuchten Merkmal hoch korrelieren bzw. dass sich die Population aus vielen Klumpen zusammensetzt, kann auch im Bayes'schen Ansatz die einfache Zufallsstichprobe durch eine geschichtete Stichprobe, eine Klumpenstichprobe oder eine mehrstufige Stichprobe ersetzt werden. Für Schätzprobleme steht damit ein Instrumentarium zur Verfügung, das alle vorhandenen Vorkenntnisse optimal nutzt.

Quotenstichprobe. Abschließend sei auf eine weitere, in der Umfrageforschung nicht unumstrittene Stichprobentechnik eingegangen: das Quotenverfahren. Bei diesem Verfahren werden dem Interviewer lediglich die prozentualen Anteile (Quoten) für bestimmte Merkmalskategorien vorgegeben (z. B. 30% Jugendliche aus Arbeiterfamilien, 20% aus Unternehmerfamilien, 20% aus Beamtenfamilien sowie 30% aus Angestelltenfamilien; gleichzeitig sollen sich 50% der befragten Jugendlichen in Ausbildung befinden, während die anderen 50% erwerbstätig sind). Die Auswahl der innerhalb dieser Quoten zu befragenden Personen bleibt dem Interviewer überlassen.

Dieses Vorgehen ist äußerst problematisch. Es resultieren keine repräsentativen Stichproben, da die Quoten nur die prozentuale Aufteilung der Quotierungsmerkmale, aber nicht die ihrer Kombinationen wiedergeben. Auch wenn – wie bei kombinierten Quotenplänen – die vorgegebenen Quoten für einzelne Merkmalskombinationen den entsprechenden Populationsanteilen entsprechen, kann man keineswegs sicher sein, dass die Stichprobe der Population auch bezüglich nichtquotierter Merkmale einigermaßen entspricht. Dies ist bei einer nach dem (den) Quotierungsmerkmal(en) geschichteten Stichprobe anders, weil hier innerhalb der Schichten Zufallsstichproben gezogen werden.

Bei der Quotenstichprobe »erfüllt« der Interviewer seine Quoten jedoch nicht nach dem Zufallsprinzip, sondern nach eigenem Ermessen (er meidet beispielsweise Personen in höheren Stockwerken oder Personen in entlegenen Gegenden). Die Stichprobe kann deshalb ein falsches Abbild der eigentlich zu untersuchenden Population sein.

Parameterschätzungen, die aus Quotenstichproben abgeleitet werden, beziehen sich deshalb nicht auf die eigentliche Zielpopulation, sondern auf eine fiktive, selten genau zu beschreibende Population. Dennoch wird diese Stichprobentechnik von Fall zu Fall als Notbehelf akzeptiert, wenn die Ziehung einer probabilistischen Stichprobe zu hohe Kosten oder zu viel Zeit erfordert. (Ausführlicher behandelt wird das Quotenverfahren bei Hoag, 1986; van Koolwijk, 1974b, Kap. 3; Noelle, 1967; Noelle-Neumann & Petersen, 1996.)

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Box 7.9

Das ADM-Mastersample

Stichproben, die den Anspruch erheben, für die gesamte Bundesrepublik Deutschland repräsentativ zu sein, werden seit 1978 von allen demoskopischen und wissenschaftlichen Instituten fast ausschließlich auf der Basis des ADM-Mastersamples erhoben. Bei diesem vom »Arbeitskreis Deutscher Marktforschungsinstitute (ADM)« und dem »Bureau Wendt« eingerichteten Stichprobensystem handelt es sich um eine 3-stufige Zufallsauswahl mit Stimmbezirken als erster, Haushalten als zweiter und Personen als dritter Auswahlstufe. Grundgesamtheit ist die erwachsene, deutsche Wohnbevölkerung in Privathaushalten. Aufbau und Anwendung dieses Stichprobensystems seien im Folgenden kurz erläutert:

Primäreinheiten. Die Bundesrepublik ist mit einem Netz von Stimmbezirken für die Wahl zum Deutschen Bundestag überzogen. Hieraus wurden ca. 64.000 synthetische Wahlbezirke mit mindestens 400 Wahlberechtigten gebildet, wobei Wahlbezirke mit weniger als 400 Wählern mit benachbarten Wahlbezirken fusioniert wurden. Diese Wahlbezirke sind die Primäreinheiten.

Mastersample. Aus der Grundgesamtheit der Primäreinheiten wurden nach dem PPS-Design (► S. 442) ca. 50% ausgewählt. Diese Auswahl konstituiert das sog. Mastersample.

Netze. Aus dem Mastersample hat man ca. 120 Unterstichproben mit jeweils 258 Wahlbezirken (PPS-Auswahl) gebildet. Eine Unterstichprobe wird als Netz bezeichnet und dessen Wahlbezirke als »Sampling Points«. Jedes Netz ist zusätzlich nach einem regionalen Index, dem sog. BIK-Index, geschichtet (Einwohner- und Arbeitsplatzdichte, primäre Erwerbsstruktur etc., vgl. Behrens, 1994). Jedes Netz stellt damit eine geschichtete Zufallsstichprobe aus der Grundgesamtheit der Primäreinheiten dar.

Haushaltsstichprobe. Die Grundlage einer Haushaltsstichprobe sind ein oder mehrere Netze. Für die Auswahl der Haushalte setzt man üblicherweise das Random-Route-(Zufallsweg-)Verfahren ein. Hierbei wird pro Sample-Point zufällig eine Startadresse festgelegt, von der aus ein Adressenermittler eine Ortsbegehung beginnt, die strikt einer für alle Sample-Points einheitlichen Begehungsanweisung folgt. Ziel des Random-Route-Verfahrens ist eine Zufallsauswahl von Adressen pro Sample-Point, aus der üblicherweise ca. 8 Haushalte zufällig ausgewählt werden. Diese Haushalte werden entweder direkt vom Adressenermittler (»Standard Random«) oder von einem von dem Untersuchungsleiter ausgewählten Interviewer befragt (»Address Random«). Bei dieser Vorgehensweise erhält man also pro Netz eine Bruttostichprobe von ca. $258 \cdot 8 = 2064$ zufällig ausgewählten Haushalten.

Personenstichprobe. Für eine repräsentative Personenstichprobe muss pro Haushalt zufällig eine Person ausgewählt werden. Hierfür wird üblicherweise der sog. Schwedenschlüssel (oder die »Last-Birthday-Methode«, ► S. 242) eingesetzt (der Erfinder dieses Auswahlverfahrens, Leslie Kish, stammt aus Schweden). Ziel des Verfahrens ist Chancengleichheit für alle Haushaltsmitglieder. Besteht ein Haushalt z. B. aus 4 potenziellen Zielpersonen, entscheidet eine zufällig ausgewählte Zahl zwischen 1 und 4, welche Person befragt wird (Einzelheiten zur Technik s. Jacob & Eirmbter, 2000, S. 106).

Eine so gewonnene Personenstichprobe ist bezüglich des Merkmals »Haushaltsgröße« nicht repräsentativ. Wenn jeder Haushalt der Grundgesamtheit mit derselben Wahrscheinlichkeit in die Stichprobe aufgenommen wurde, haben Personen aus 1-Personen-Haushalten eine Wahrscheinlichkeit von 1, befragt zu werden, Personen aus 2-Personen-Haushalten eine Wahrscheinlichkeit von 1/2, aus 3-Personen-Haushalten eine Wahrscheinlichkeit von 1/3 etc. Dieser Sachverhalt kann bei der Berechnung von statistischen Kennwerten für die Stichprobe (z. B.

▼

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,

Durchschnitts- oder Anteilswerte) durch eine zur Auswahlwahrscheinlichkeit reziproke Gewichtung kompensiert werden: Personen aus 1-Personen-Haushalten gehen mit einfachem Gewicht in das Ergebnis ein, Personen aus 2-Personen-Haushalten mit doppeltem Gewicht, Personen aus 3-Personen-Haushalten mit dreifachem Gewicht etc. (vgl. Schumann, 1997, S. 101 f., zum Stichwort »Design-Gewichtung«).

Übungsaufgaben

7.1 Wie ist die »einfache Zufallsstichprobe« definiert?

7.2 Was versteht man unter »probabilistischen Stichproben«?

7.3 Worin unterscheidet sich die Klumpenstichprobe von der geschichteten Stichprobe?

7.4 In der Zeitung lesen Sie unter der Überschrift »Haschisch macht müde und faul« folgende Meldung: »Wie eine neue amerikanische Repräsentativstudie zeigt, haben 70% aller Haschisch-Konsumenten unterdurchschnittliche Schulleistungen. Gleichzeitig schlafen sie überdurchschnittlich lange. Diese Befunde belegen eindrücklich, wie gefährlich eine liberale Drogenpolitik ist.« In dieser Nachricht sind fünf Fehler versteckt. Welche?

7.5 Es soll eine repräsentative Stichprobe (realisiert als geschichtete Stichprobe) von Museumsbesuchern gezogen werden. Entwerfen Sie einen Stichprobenplan mit drei Ziehungsstufen!

7.6 Erläutern Sie die Gütekriterien für Punktschätzungen!

7.7 Erklären Sie das Grundprinzip des Bayes'schen Ansatzes!

7.8 Angenommen, Sie interessieren sich dafür, wieviele Studierende im Fach Psychologie mit der Zuteilung ihres Studienortes unzufrieden sind und den Studienplatz tauschen. Ist der Anteil groß genug, damit sich die Einrichtung einer professionell betriebenen landesweiten »Tauschbörse« lohnt? Die Befragung von 56 Kommilitonen am heimischen Institut ergibt, dass 8% den Studienplatz mindestens einmal getauscht haben und 2% auf der Suche nach einem Tauschpartner sind.
   a) Wie groß muss der Umfang einer Zufallsstichprobe aus der Grundgesamtheit aller Psychologiestudenten sein, um bei einer vermuteten Tauscherrate zwischen 5% und 20% den »wahren« Populationsanteil mit einer Genauigkeit von ± 2% schätzen zu können?
   b) Welche Stichprobenart wäre alternativ zur einfachen Zufallsauswahl in diesem Beispiel gut geeignet?

7.9 In welcher Weise lässt sich Vorwissen über eine Population einsetzen, um die Genauigkeit von Parameterschätzungen zu erhöhen? Erläutern Sie diese Thematik für alle Stichprobenarten, die Sie kennen!

7.10 Warum ist es erforderlich, Parameterschätzungen mit einem Konfidenzintervall zu versehen?

7.11 Welche Aussagen stimmen?
   a) Je höher der Konfidenzkoeffizient, umso breiter ist das Konfidenzintervall.
   b) Die Varianz einer einfachen Zufallsstichprobe ist ein erwartungstreuer Schätzer der Populationsvarianz.
   c) Die Untersuchungsobjekte innerhalb eines Klumpens sollten bei einer Klumpenstichprobe möglichst ähnlich sein (homogene Klumpen).
   d) Der Mittelwert einer einfachen Zufallsstichprobe ist ein erwartungstreuer Schätzer des Populationsmittelwertes.
   e) Die Untersuchungsobjekte innerhalb der Schichten einer geschichteten Stichprobe sollten möglichst ähnlich sein (homogene Schichten).

▼

J. Bortz, N. Döring, Forschungsmethoden und Evaluation, DOI 10.1007/978-3-540-33306-7,