Repräsentativität, convenience samples und Signifikanztests

Wie das Konzept "Repräsentativität" zu Fehlanwendungen der Statistik verleitet

Peter von der Lippe 9.04.2014

Repräsentativität ist ein ebenso beliebter wie vager und unreflektiert wertender Begriff. Wir zeigen, was man darunter verstehen könnte und dass nichts davon widerspruchsfrei, exakt und nutzbar ist (um z.B. aus einer Definition der Repräsentativität ein Maß für den Grad der Repräsentativität abzuleiten oder bestimmen zu können, wie eine Stichprobe gezogen (oder allgemein generiert) werden soll und wie groß sie sein soll, also z.B. welche und wie viele Personen befragt werden sollten). Besonders bedenklich ist jedoch, dass "Repräsentativität" den Blick verstellt für das Besondere einer Zufallsauswahl und dafür, welche Voraussetzungen erfüllt sein müssen, um Methoden der induktiven Statistik (wie z.B. Signifikanztests) sinnvoll anwenden zu können. Ein großer Teil dieser Arbeit ist der Frage gewidmet, warum solche Methoden nicht auf die Auswertung von Daten anwendbar sind, die nicht auf einer Zufallsauswahl beruhen.

1. Der Begriff "Repräsentativität" ist unbrauchbar und schädlich

Anders als der Stichprobenfehler ist "Repräsentativität" (oder "Repräsentanz") kein Fachausdruck der Statistik.[1] Es gibt zwei naheliegende Gründe hierfür

- Es gibt kein Maß R für die Repräsentativität, wonach etwa die Repräsentativität bei Stichprobe 1 $R_1 = 12{,}7$ Punkte oder R_1 85% (Prozent von was?) beträgt. Selbst Vergleiche, ob etwa $R_1 > R_2$ sind nicht möglich (was ja nicht verlangt, exakt zu sagen, um wie viel R_1 größer ist als R_2). Es gibt noch nicht einmal eine halbwegs exakte Definition der "Repräsentativität", sondern nur einige meist sehr vage Vorstellungen, die man mit dem Konzept verbindet (wir gehen darauf in Abschn. 2 ein).
- Man kann dem Anliegen, dem der Begriff "Repräsentativität" dienen soll, viel besser mit dem Begriff "Stichprobenfehler" (oder "Standardfehler") einer "Statistik" wie z.B. dem arithmetischen Mittel $\bar{x}$ Rechnung tragen. Er ist definiert mit der Formel

(1) $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ ohne Endlichkeitskorrektur, bzw.

(2) $\sigma^*_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{\sigma_x}{\sqrt{n}} \cdot f$

[1] So auch Schnell, S. 16. Der Begriff kommt in Statistik-Lehrbüchern so gut wie nicht vor. Es gibt auch kaum entsprechende Aufsätze in statistischen Fachzeitschriften. Eine aktuelle Ausnahme ist S. Gabler und A. Quatember (2013), mit einem allerdings ziemlich misslungenen Versuch "Repräsentativität" zu definieren, als Daten, die es erlauben etwas "(zumindest näherungsweise) unverzerrt" schätzen zu können.

mit Endlichkeitskorrektur f, wobei f mit $N \to \infty$ gegen 1 strebt, wenn n der Umfang der Stichprobe, N der der Grundgesamtheit und $\sigma_x = \sigma$ die Standardabweichung der Variable x in der Grundgesamtheit (im Folgenden GG) ist.

"Repräsentativität" ist auch oft ein gewohnheitsmäßig verliehenes und unverdientes Prädikat und umgekehrt ist der Vorwurf, dass eine Erhebung (Datenbasis) "nicht repräsentativ" sei ein gern und gefahrlos (weil meist nicht nach einer Begründung oder einem Grad der "Nichtrepräsentativität" gefragt wird) vorgebrachter Einwand gegen eine Anwendung der Statistik, die einem nicht gefällt.

Es ist auch bezeichnend, dass sich zwar oft der Einwand "nicht repräsentativ" darauf stützt, dass der Stichprobenumfang n nicht sehr groß ist, aber andererseits – wie wir im folgenden Abschnitt 2 zeigen – alle Versuche, den Begriff "Repräsentativität" zu definieren gerade dabei versagen, aus dem Begriff konkrete Hinweise für das erforderliche n abzuleiten, während dies sehr wohl aus $\sigma_{\bar{x}}$ abgeleitet wird. Wie wenig hier die Dinge (also n und die "Repräsentativität") zusammenhängen wird schon daran deutlich, dass bei der Berechnung von n auch danach differenziert wird, ob wir mit der Stichprobe eine Genauigkeit der Schätzung, z.B. des mittleren Einkommens μ in der GG von ± 1€ oder nur von ± 10€ anstreben (es ist unmittelbar einleuchtend, dass man im ersten Fall mehr Leute befragen muss). Aber hat man schon jemals den Einwand gehört: Ihre Stichprobe mag zwar repräsentativ sein, wenn es nur um ± 10€ geht, aber sie ist es nicht, wenn die Genauigkeit ± 1€ sein soll?

Es gibt also einige Indizien für die Unbrauchbarkeit von Kategorien wie "Repräsentativität" oder "Strukturidentität". Solche Kategorien sind aber nicht nur unbrauchbar, sie sind sogar schädlich, weil sie dazu verleiten, auch dort Methoden der statistischen Auswertungen (wie z.B. Signifikanztests) vorzunehmen, die über die "bloß deskriptive" Statistik hinausgehen, wo dies nicht zulässig ist, weil diese Methoden eine Zufallsauswahl[2] voraussetzen. Denn der Gedanke, entscheidend sei statt dessen (also statt Zufallsauswahl) bei einer Teilerhebung, dass sie "repräsentativ" sei, was meist verstanden wird im Sinne von "strukturidentisch" mit der Grundgesamtheit (GG), scheint für viele die Anwendbarkeit der statistischen Schätz- und Testtheorie zu rechtfertigen.[3] Wir wollen – insbesondere im folgenden Abschnitt 3 – zeigen, dass das eine bedenkliche Konsequenz des Konzepts "Repräsentativität" ist. Dabei sind wir in guter Gesellschaft. In dem, gerade auch von Anwendern viel zitierten Buch von J. Bortz (1993) werden die "sog. 'anfallenden' oder 'ad hoc' Stichproben (z.B. die 'zufällig' in einem Seminar anwesenden Teilnehmer)[4] in der Hoffnung, auch so zu aussagefähigen Resultaten zu gelangen" explizit erwähnt (S. 85). Ihnen wird nicht selten (nachträglich) "Repräsentativität" für eine ähnlich strukturierte GG zugesprochen um die Anwendung inferenzstatistischer Verfahren (Schätzen und Testen) zu rechtfertigen. Bortz kritisiert dies ausdrücklich und sehr deutlich:

> "Vor dieser Vorgehensweise sei nachdrücklich gewarnt. Zwar ist die Verwendung inferenzstatistischer Verfahren nicht daran gebunden, dass eine Stichprobe aus einer wirklich existierenden Population gezogen wird: letztlich lässt sich für jede 'Stichprobe' eine fiktive Population konstruieren, für die diese 'Stichprobe' repräsentativ erscheinen mag. Die Schlüsse, die aus derartigen Untersuchungen gezogen werden, beziehen sich jedoch nicht auf real existierende Populationen und können deshalb wertlos sein" (S. 85).

[2] Die Stichprobentheorie in der Statistik beschäftigt sich nur mit solchen im engeren Sinne "echten" Stichproben (random samples oder – synonym – probability samples).

[3] Es wird leider auch oft nicht mehr richtig verstanden, dass Konfidenzintervalle und Signifikanztests nur zwei Varianten sind, den gleichen Sachverhalt darzustellen. So meinen z.B. Göritz u. Moser, dass Repräsentativität "weniger wichtig" sei, "wenn es statt der Schätzung von Populationsparametern um die Prüfung von Zusammenhangshypothesen … geht" (S. 161). Wenn man kein Konfidenzintervall ausrechnen kann (oder sollte), macht es auch keinen Sinn, einen angeblich weniger anspruchsvollen Signifikanztest durchzuführen.

[4] Genau das ist es was oft, und so auch im Folgenden, "convenience sample" genannt wird.

Das ist zweifellos die üblicherweise von Statistikern und auch von uns hier vertretene Position. Man findet jedoch in der statistischen Literatur kaum über die zitierte Aussage von Bortz hinausgehende ausführliche Begründungen hierzu. Was Statistiker meist nur beklagen ist, dass die Anwender falsche Vorstellungen darüber haben, was "Signifikanz" eigentlich bedeutet (z.B. Krämer 2010). Detaillierte Darstellungen, an welche Voraussetzungen inferenzstatistische Verfahren gebunden sind und warum es so wichtig ist, dies zu beachten, sind in der Literatur kaum zu finden. Wir mussten deshalb in Abschn. 3 versuchen, eigene Wege zu gehen.

Die zentrale Voraussetzung für Tests, Konfidenzintervalle etc. ist die Existenz der Stichprobenverteilung (sampling distribution) einer Schätzfunktion, wie z.B. von $\bar{x}$, dem arithmetischen Mittel der Stichprobe.[5] Das ist die Verteilung, die angibt, welche Werte $\bar{x}$ mit welcher Wahrscheinlichkeit annimmt, wenn man *alle*[6] aus einer gegebenen GG zu ziehenden Stichproben des gleichen Umfangs n zöge, wobei es keine Rolle spielt, ob eine Stichprobe bezüglich einer Variable x eine ähnliche Struktur hat wie die GG oder nicht.

Der Stichprobenfehler $\sigma_{\bar{x}}$ oder $\sigma^*_{\bar{x}}$ ist die Standardabweichung eben dieser Stichprobenverteilung von $\bar{x}$. Im Falle einer nichtzufälligen Auswahl, wie z.B. bei der Quotenauswahl oder dem "convenience sample" sind inferenzstatistische Verfahren nicht anwendbar, weil es – wie wir in Abschn. 3.1f zu zeigen versuchen – in solchen Fällen keine Stichprobenverteilung gibt.

Man kann diese Begründung leicht als zu "formal" oder rein "technisch" empfinden, weshalb wir in den weiteren Abschnitten auch noch andere Begründungen zu geben versuchen, indem wir u. a. fragen, welchen Sinn "Hypothesen" (die ja immer Vermutungen über die GG sind) haben, wenn es mehr oder weniger unklar bleibt, was genau die GG ist, für die die Stichprobe "repräsentativ" sein soll. Anders als z.B. bei ad hoc Befragungen von Studenten, wo ja kaum eine "Ziehung" vorliegt, kann es bei Ziehung einer Stichprobe nach dem Zufallsprinzip aus einem "Auswahlrahmen" (sampling frame) keine Unklarheit darüber geben, was hier die GG ist, denn der Auswahlrahmen stellt ja eine Auflistung der Einheiten einer "wirklich existierenden Population" im Sinne von Bortz dar.

Die nur bei einem random sample gegebene Anwendbarkeit der Wahrscheinlichkeitsrechnung und damit auch der statistischen Tests ist auch der Grund dafür, dass in (Lehr-) Büchern zu Stichprobentheorie – wie z.B. dem Standardwerk von Cochran – meist kaum ein Wort zur "Repräsentativität" zu finden ist, sondern nur vom random samples die Rede ist.[7] Auch auf den 616 Seiten des Buchs von Levy und Lemeshow (2013) findet man nur im einleitenden Kapitel eine kurze Bemerkung zu non-probability samples, die mit den Worten schließen:

> "In this text we will consider only probability samples, since we feel very strongly that sample surveys should yield estimates that can be evaluated statistically with respect to their expected values and standard errors."

Es gibt durchaus praktische Gründe, die es geraten erscheinen lassen, auf Stichproben im Sinne der Stichprobentheorie zu verzichten andere, in der Durchführung wesentlich weniger anspruchsvollen Verfahren der Auswahl (wie Quotenauswahl oder auch das convenience sample) vorzunehmen. Wir kommen darauf in Abschnitt 4 zurück. Dagegen ist nichts einzuwenden,

---

[5] Das arithmetische Mittel der Stichprobe ist eine Stichproben- oder Schätzfunktion weil mit ihm der "wahre" Mittelwert μ in der GG geschätzt werden soll. Was im Folgenden jeweils am Beispiel des arithmetischen Mittels demonstriert wird gilt entsprechend für Mittelwertdifferenzen μ_1-μ_2, Korrelations- und Regressionskoeffizienten (ρ, β_i) und viele andere Parameter.

[6] Im Markfordvhung-Lehrbuch von A. Kuß findet man (S. 215 – 217) eine Auflistung einiger Stichproben einer Stichprobenverteilung mit n = 10 bei N = 100 (aber nicht alle Stichproben, denn das wären bei diesen Werten für N und n nach Gl. (4) 17 Billionen, 1,731 10^{13} Stichproben), aber leider keine Bemerkung dazu, warum diese Stichprobenverteilung der Dreh- und Angelpunkt für die Anwendbarkeit von Tests und Konfidenzintervallen ist.

[7] Es gibt praktisch keine statistische Theorie zu non-random samples. Unsere Überlegungen in Abschn. 3 bauen deshalb also auf keine uns bekannte Veröffentlichung auf. Wir mussten hier versuchen eigene Wege zu gehen.

wenn man sich bei der Auswertung auf Methoden der deskriptiven Statistik beschränkt. Problematisch wird es nur, wenn man sich hiermit nicht begnügen will, etwa im sehr verbreiteten aber gänzlich unbegründeten Glauben, die Wissenschaftlichkeit beginne erst mit der Verwendung inferenzstatistischer Verfahren.

2. Was könnte man mit "Repräsentativität" meinen?

2.1. Einführung und Übersicht

In diesem Abschnitt wollen wir fünf Repräsentativitätsbegriffe unterscheiden:

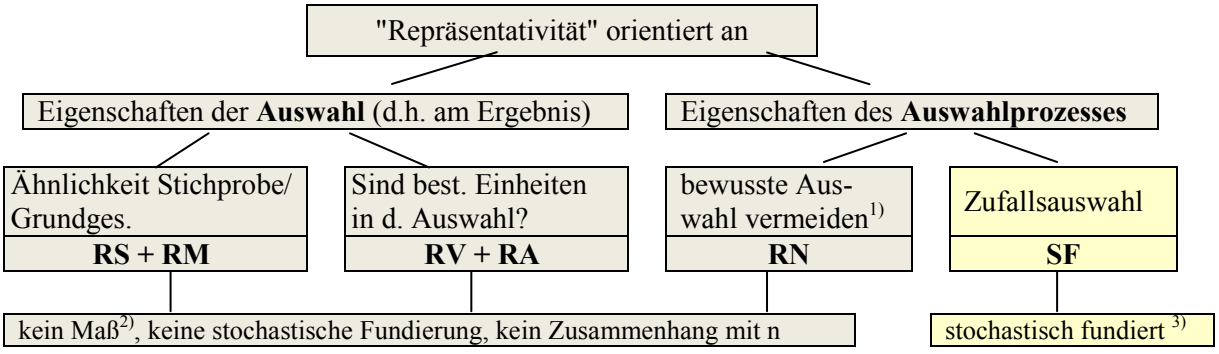
RS	Strukturkonzept*
RM	Miniaturkonzept
RV	Stellvertreter (Vize) Konzept
RA	Arche-Noah Konzept
RN	Nichtselektivitätskonzept

* es ist die mit Abstand am weitesten verbreitete Vorstellung von "Repräsentativität", weshalb wir im Folgenden hierauf auch am ausführlichsten eingehen werden.

und zeigen, dass *keines dieser Konzepte* bei genauerer Betrachtung

1. widerspruchsfrei und operational ist, d.h. etwas hergibt bei der Frage, wie eine Stichprobe konkret zu "ziehen" (oder allgemeiner: zu generieren) ist,[8] wie groß z.B. der Stichprobenumfang n sein soll und wie man nach unterschiedlichen Graden der "Repräsentativität" differenzieren sollte, weil keines der Konzepte
2. einen formelmäßigen Zusammenhang mit dem Schätzproblem erkennen lässt.

Bei einem brauchbaren Konzept der "Güte" einer Stichprobe i.w.S. (also einschließlich Verfahren der nichtzufälligen Auswahl) erwartet man auch Antworten auf praktische Fragen, wie z.B. wie ermittelt man "repräsentativ" den Ausschussanteil in einer Lieferung? Wie stellt man einen strukturgleichen Ausschnitt aus der Lieferung zusammen? Kein Repräsentativitätskonzept wird hier sonderlich konkret. Mit der folgenden Abbildung versuchen wir die fünf Konzepte unter Einschluss der Zufallsauswahl, also des Konzepts des Stichprobenfehlers (SF) als ein sechstes Konzept kurz einzuordnen:



1) nicht notwendig, dass stattdessen eine Zufallsauswahl durchgeführt werden sollte; das Konzept RN ist auch nicht an Auswahlwahrscheinlichkeiten orientiert
2) es gibt kein Maß für mehr oder weniger "Repräsentativität" oder Grade der Verallgemeinerungsfähigkeit von Daten
3) wahrscheinlichkeitstheoretisch fundiert; direkter Bezug zum Schätzproblem (SF ist ein Maß für die Güte einer Schätzung) und zum Stichprobenumfang n

Begriffe wie "repräsentativ" oder "typisch" sind nicht eindeutig. Man kann von der *Gesamtheit* der Ausgewählten verlangen, dass sie *als Auswahl typisch* ist oder man kann von den einzelnen Einheiten verlangen, dass sie *"repräsentative", "typische" Einheiten* sind. Wenn die getroffene Auswahl "typisch" sein soll, und für "typisch" die Strukturen maßgeblich sind,

[8] Das Konzept RN bietet bestenfalls eine Antwort auf die Frage, wie die Auswahl *nicht* durchzuführen ist.

läuft das auf RS hinaus. Sollen es die einzelnen *Einheiten* sein, die "typisch" sind läuft das auf RV hinaus und die Frage ist, wie typisch muss man sein, um "Stellvertreter" für alle sein zu können. [9]

Alle Konzepte, die an der Auswahl statt am Auswahlprozess ansetzen verlangen Kenntnisse über die GG, wie sie in der Regel nicht, oder nicht im geforderten Maße vorhanden sind. Verlangt man ein "getreues Abbild" oder eine exakte Miniatur der GG fragt es sich auch, ob dann überhaupt noch sinnvoll von einer "Schätzung" durch eine Stichprobe die Rede sein kann. Denn es bleibt ja nichts mehr zu schätzen übrig, wenn man schon weiß (woher eigentlich?), dass man eine[10] Miniaturausgabe der GG besitzt.

Bei Konzepten, die auf Ähnlichkeiten von Stichprobe und GG abstellen (also RS und RM) ist stets ein kritischer Punkt, wie genau (mit wie vielen und wie stark differenzierten Merkmalen) man die Ähnlichkeit beschreiben will. Nimmt man es in dieser Hinsicht sehr genau führt der Gedanke, dass die Stichprobe irgendwie ein "getreues Abbild" der GG sein soll (wie beim Konzept RS und RM) auch zu einer großen Mindestgröße der erforderlichen Stichprobe, während bei RV eher ein kleines n ausreichen dürfte. "Repräsentativität" erlaubt nicht nur viele Definitionsversuche ("Konzepte"), sondern diese führen auch zu keinen oder aber durchaus verschiedenen Folgerungen bezüglich des Stichprobenumfangs n oder der Technik, mit der eine repräsentative Auswahl im Sinne des entsprechenden Konzepts vorzunehmen ist. Keines der fünf genannten Konzepte liefert eine konkrete Zahl für den Stichprobenumfang n, der aber ganz offensichtlich bestimmend ist für die Schätzqualität[11] beim Schluss von der Stichprobe auf die GG.

Wir beschäftigen uns hier *nicht* mit *allen* Versuchen, "Repräsentativität" zu definieren, die man in der Literatur findet, so z.B. nicht mit solchen, bei denen keinerlei Begründungsversuch zu sehen ist, wie z.B. wenn auf die Erwartungstreue etc. einer Schätzung Bezug genommen wird[12] oder nur auf den Stichprobenumfang n, wie bei der so beliebten Aufforderung, man möge sich an der Befragung beteiligen, weil sonst die "Repräsentativität" gefährdet sei.

2.2. Das Strukturkonzept (RS)

Die häufigste Vorstellung ist wohl die, dass eine konkrete Stichprobe dann "repräsentativ" ist, wenn die Struktur der Stichprobe ähnlich der der GG ist. Ist z.B. die Aufteilung bezüglich des Geschlechts in der GG: $\pi = 0{,}5$ männlich und $1-\pi = 0{,}5$ weiblich (also 50% Männer und 50% Frauen) und erhält man mit einer Stichprobe $\hat{\pi} = 0{,}5$ oder $\approx 0{,}5$, so ist diese Stichprobe "repräsentativ" und auch repräsentativer als eine Stichprobe mit einer Aufteilung von 40:60.

Dagegen gibt es vier naheliegende und fünf mehr grundsätzliche, die statistische Theorie betreffende Einwände. Zunächst, was auf der Hand liegt:

1. nach dem RS-Konzept wäre eine Stichprobe von 3 Männern und 3 Frauen genauso gut wie eine Stichprobe von 30 Männer und 30 Frauen, aber besser als eine von 305 Männer und 295 Frauen.[13]

---

[9] Das Problem stellt sich bei Zufallsauswahl nicht, weil es dort auf die Wahrscheinlichkeit ankommt, mit der eine Einheit *ausgewählt* wird, nicht auf die Repräsentativität der ausgewählten *Einheit* (ob sie "typisch" ist).

[10] Kann man hier von "*einer*" Miniaturausgabe sprechen oder muss es heißen: "*die*" Miniaturausgabe?

[11] Viele kennen nur n als Bestimmungsfaktor für den Schätz- oder Stichprobenfehler SF und sie vergessen, dass es nach Gl. 1 auch auf σ ankommt. Das wird wie folgt unmittelbar einsichtig: Wären alle N Einheiten der GG gleich, dann wäre auch $\sigma^2 = 0$ und man würde mit $n = 1$ die komplette GG exakt (kein SF!) kennen.

[12] So Quatember 1996, oder Gabler u. Quatember 2013, oder auch Kutsch (2007), S. 110 bei denen "Repräsentativität" heißt, dass die Stichprobe eine erwartungstreue Schätzung eines Parameters erlaubt. Hier wird offensichtlich eine Eigenschaft einer Stichprobenfunktion mit einer Eigenschaft einer Stichprobe vermischt, obgleich zwischen beiden Dingen kein Zusammenhang zu erkennen ist.

[13] Befragt man 20 Männer und 22 Frauen könnte man nach der RS-Logik die Daten von zwei (egal welche!) Frauen aus der Stichprobe entfernen und man hätte dann bessere Daten. Man kann auch leicht Beispiele konstru-

2. Zweifel können einem auch kommen, wenn die Struktur nach einem trichotomen (drei Merkmalsausprägungen) statt dichotomen Merkmal bestimmt wird:

Familien-stand	Grundge-samtheit	Stich-probe 1	Stich-probe 2
ledig	36%	34%	38%
verheiratet	52%	54%	51%
sonstige	12%	12%	11%

Welche Stichprobe ist repräsentativer? Stichprobe 1 oder Stichprobe 2?

Wie man sieht, kann es ein "getreues Abbild" der GG gemäß dieser Tabelle auch nicht mit einer Stichprobe von n < 25 geben, weil es 9 Ledige, 13 Verheiratete und 3 Sonstige sein müssen bzw. ein Vielfaches dieser Zahlen damit die Proportionen stimmen (so dass n nur 25, 50, 75 usw. sein kann). Und wie sähe es bei einem quantitativen Merkmal X, etwa dem Einkommen aus? Soll man sich bei der Struktur in Gestalt von "Quoten" an einer Klasseneinteilung orientieren, wie 0 bis unter 500, 500 bis unter 1000 usw., oder muss es eine feinere Klasseneinteilung für die Quoten sein und wie fein muss sie sein?

3. Als nächstes kann man sich fragen, wie genau man es mit der "Struktur" nehmen muss: soll nur der Anteil der Frauen in der Stichprobe in etwa dem in der GG entsprechen oder ist dies auch zu fordern, für den Anteil der ledigen Ärztinnen zwischen 30 und 40. Und wenn mehrere Merkmale gleichzeitig betrachtet werden: wie sieht es aus mit der Korrelation? Reicht es zwei Merkmale (etwa Alter und Familienstand), jedes für sich zu betrachten, oder muss man beide gemeinsam betrachten? Anders gesagt: Reichen ähnliche "Strukturen" der Randverteilungen oder brauchen wir auch ein "getreues Abbild" der gemeinsamen Verteilung?

	J	A	Σ
L	0,08	0,28	0,36
V	0,17	0,35	0,52
S	0,07	0,05	0,12
Σ	0,32	0,68	1

	J	A	Σ
L	0,13	0,23	0,36
V	0,16	0,36	0,52
S	0,03	0,09	0,12
Σ	0,32	0,68	1

A = alt, J = jung, L = ledig, V = verheiratet, S = sonstige, Σ = Summe

Wie das Zahlenbeispiel zeigt, bedeuten gleiche Randverteilungen (die schattierten Felder) nicht auch eine gleiche gemeinsame (bivariate) Verteilung. Betrachtet man mehrere Merkmale bei der Definition der relevanten "Struktur" erreicht man schnell Dimensionen, die nicht mehr praktikabel sind. Bei 5 trichotomen Merkmalen müssten die Häufigkeiten von $3^5 = 243$ Tabellenfeldern in Stichprobe und in der GG proportional sein.

4. Da der Strukturbegriff auf *Merkmale* statt auf Einheiten (z.B. Personen) abstellt erscheint es nur konsequent, zu sagen, Repräsentativität sei keine absolute Eigenschaft, sondern "merkmalsgebunden".[14] Das führt zu der abwegigen Vorstellung, dass eine Zusammenstellung von n Personen repräsentativ sein kann, wenn es um das Merkmal x geht und dass man dann aber personell Änderungen an der Stichprobe vornehmen muss, wenn man von der Untersuchung von x zu der von y übergeht.[15] In der statistischen Stichprobentheorie kommt es dagegen allein auf die Art der Auswahl der *Einheiten* aus der GG an, nicht darauf, welche Verteilungen (Strukturen) sich daraus für bestimmte *Merkmale* ergeben.[16]

Nun fünf Punkte zu einer etwas grundsätzlicheren Kritik:

---

ieren, bei denen eine Zufallsauswahl (also eine echte Stichprobe, die nach statistischem Verständnis stets "repräsentativ ist) vorgenommen wird und überhaupt keine der Stichproben "repräsentativ" im Sinne von RS ist. Im Beispiel mit 50% Männer und 50% Frauen würde jedes ungerade n, etwa n = 193 eine Stichprobe liefern, die streng genommen nicht "repräsentativ" ist, denn es kann ja nicht 96,5 Männer und 96,5 Frauen geben.

[14] So K. Zerr in einem Buchbeitrag (2003) zur Online-Marktforschung, zitiert nach Kutsch 2007, S. 4, wobei Kutsch selbst solchen Überlegungen nahesteht und glaubt, Repräsentativität sei eine "merkmalsbezogene" Eigenschaft, die bei jedem Merkmal einzeln geprüft werden müsse (a.a.O., S. 128f.).

[15] Allerdings kommen wir unten in Abschn. 3.2 durchaus selbst auch in die Nähe einer solchen Argumentation.

[16] Zu erwähnen wäre auch noch bei RS, quasi als fünftes, allerdings wohl weniger bedeutsames Argument, dass es durchaus Sinn machen kann, eine Auswahl so vorzunehmen, dass die Proportionen ganz bewusst andere sind als in der GG, dass z.B. kleine Bundesländer stärker repräsentiert werden sollen als große um zu einer besseren Schätzung zu gelangen.

1. Dem RS-Konzept liegt die Vermutung zugrunde, dass eine Auswahl, die in der Lage ist, die Proportionen bezüglich bestimmter Merkmale "richtig" (wie zu messen?) wiederzugeben auch geeignet sein dürfte eine "gute" Schätzung z.B. eines Mittelwerts μ in der GG aufgrund des Mittelwerts $\bar{x}$ in der Stichprobe zu liefern.

   Das scheint in den Augen vieler ein plausibles Argument zu sein. Aber wenn es so plausibel ist sollte es auch möglich sein, für diese Behauptung einen Beweis zu liefern. Uns ist aber nicht bekannt, dass jemand den Versuch eines Beweises unternommen hätte. Aber ist es wirklich so plausibel? Aber es bleibt völlig offen, worauf sich hier die Erwartung einer besseren Schätzung stützen soll, wenn sich die Schätzung von μ und die Gütekriterien einer Schätzung, wie Erwartungstreue etc. aus der Wahrscheinlichkeitsrechnung ableiten, aber die "repräsentative" Auswahl (wenn überhaupt eine Auswahl vorliegt, was ja beim convenience sample zweifelhaft ist) gerade keine Zufallsauswahl ist.

   Man könnte argumentieren, dass eine Stichprobe, die die Struktur eines Merkmals y gut wiedergibt auch die des anderen Merkmals x gut wiedergibt, insbesondere wenn x und y hoch korrelieren.[17] Wie man aber an der Formel für $\sigma_{\bar{x}}$ in Gl. 1 sieht, kommt es bei $\sigma_{\bar{x}}$ allein auf die Varianz σ_x^2 von x an, nicht auf die von y oder die Kovarianz σ_{xy} und damit auch nicht auf die Korrelation $\rho_{xy} = \sigma_{xy}/\sigma_x\sigma_y$.

2. Genau das, wie gut nämlich μ durch $\bar{x}$ zu schätzen ist, ist aber eine Frage des Stichprobenfehlers $\sigma_{\bar{x}}$. Zwischen RS und $\sigma_{\bar{x}}$ besteht jedoch ein erheblicher Unterschied; denn $\sigma_{\bar{x}}$ bezieht sich nicht wie das Konzept RS auf eine einzelne konkrete Stichprobe, sondern auf alle Stichproben vom Umfang n, die man überhaupt aus einer GG vom Umfang N ziehen kann und ist somit eine *Wahrscheinlichkeitsaussage*, die sich auf eine Zufallsvariable d.h. unter gleichen Bedingungen beliebig oft wiederholbare Vorgänge bezieht.[18]

3. Wenn man sich den problematischen Gedanken zu eigen macht, "Repräsentativität" sei merkmalsbezogen kann man ein Defizit des RS-Konzepts darin sehen, dass es keine Maßstäbe dafür liefert, nach welchen Merkmalen die "Strukturidentität" definiert werden soll, welche besser und welche schlechter sind und ab wann sie als "erfüllt" gelten soll.

4. Dem RS Konzept liegt der legitime Gedanke zugrunde, dass vermieden werden sollte, dass wichtige Teilgesamtheiten in der Stichprobe gar nicht erscheinen oder unterrepräsentiert sind, während andere überrepräsentiert sind. Eine andere Möglichkeit als RS, so etwas zu vermeiden wäre eine geschichtete Stichprobe, bei der aus K Schichten[19] mit N_1, N_2, …, N_K Elementen in der GG Stichproben mit den Umfängen n_1, n_2, …, n_K zufällig gezogen werden ($N = \sum_k N_k$, $n = \sum_k n_k$, k = 1,…,K). Ein Schichtungsmerkmal ist umso besser (kleineres $\sigma_{\bar{x}}$ bei gleichem n) je kleiner die interne Varianz V_{int} von x in einer nach ihm strukturierten GG und je größere die externe Varianz V_{ext} ist,

   (3) $\sigma_x^2 = \sum_k f_k(\mu_k - \mu)^2 + \sum_k f_k\sigma_k^2 = V_{ext} + V_{int}$ mit $f_k = N_k/N$.

   Das Schichtungsmerkmal sollte also Schichten liefern, die in sich ("innerhalb" der Schichten) möglichst homogen sind (kleine Varianzen σ_k^2), im Vergleich untereinander ("zwischen" den Schichten) aber möglichst verschieden sind (ungleiche Mittelwerte μ_1, …, μ_K von x in den K Schichten).

---

[17] Man schließt damit aber auch wieder nur auf die *Struktur* von x nicht auch auf die *Schätzung* von μ_x!

[18] Wir gehen darauf in Abschn. 3.6 näher ein.

[19] Ein Beispiel wäre K = 2, Ein- und Mehrpersonenhaushalte wenn es z.B. um x als den in € gemessenen Wert der Konsumgüterkäufe geht.

Im Vergleich dazu haben wir (wie schon unter 3 gesagt) kaum operationale Kriterien bei der Frage, welche Eigenschaften Merkmale haben sollten, mit denen man die "Strukturidentität" von Stichprobe und GG beurteilen kann.

5. Bei der Frage, wie aus den K Schichten bei einer geschichteten Stichprobe auszuwählen ist (also bei der "Aufteilung" von n in $n_1, n_2, \ldots, n_K$), kann man sich nicht nur an den Größen $N_1, N_2, \ldots, N_K$ orientieren (also an Größen die bei einem Konzept wie "Struktur" im Vordergrund stehen), sondern z.B. auch an Größen wie $N_k\sigma_k$ (k = 1, 2, …, K), wie dies im Falle der sog. "optimalen Aufteilung " (die den Stichprobenfehler minimiert) geschieht. Dass Schicht*umfänge* N_k das Denken dominieren liegt nahe beim RS-Konzept. Darauf ist wohl auch das Missverständnis bei Quatember (1996; S. 238) und Kuß (S. 72) zurückzuführen, dass eine von der proportionalen Aufteilung $n_k/\Sigma n_k = N_k/\Sigma N_k$ abweichende Aufteilung nicht "repräsentativ" sei.[20]

Abschließend kann man festhalten:

Das RS Konzept ist so beliebt, weil es offenbar intuitiv verständlich ist. Es ist dagegen sehr viel schwieriger, plausibel zu machen, dass bei gleicher GG *jede* (Zufalls-) Stichprobe vom gleichen Umfang n gleich "repräsentativ" ist, wo doch die konkreten Stichproben (gerade wegen des Zufalls) sehr unterschiedlich ausfallen können.[21]

Andererseits ist aber RS ein widersprüchliches und inexaktes Konzept. Es gibt keinen Zusammenhang zwischen Strukturähnlichkeit (die vorausgesetzt wird) und Qualität der Induktion (die das Ziel ist). "Repräsentativität" gem. RS kann deshalb nicht die Rolle eines Qualitätsnachweises spielen. Das ist aber genau die Rolle, die es in der Praxis spielt.[22]

2.3. Das Miniaturkonzept (RM)

Nach diesem besonders vagen und daher ziemlich unbrauchbaren Konzept sollte die Stichprobe eine "getreue" Verkleinerung der Grundgesamtheit sein, ein Gedanke, der beim RS Konzept bereits angesprochen wurde, wenn mit "getreu" die gleichen Strukturen (Verteilungen) bezüglich der Merkmale gemeint sind. Aber bei RS geht es um die *Verteilung von Merkmalen*, bei RM dagegen (auch) um die *Anwesenheit oder Abwesenheit bestimmter Einheiten* in der Auswahl damit diese "repräsentativ" ist und die Vielfalt in der GG wiedergespiegelt (insofern besteht auch Ähnlichkeit mit dem RA Konzept).[23]

Auf Einheiten Bezug nehmende Konzepte haben Schwierigkeiten wenn es unterscheidbare "Einheiten" kaum gibt; z.B. bei Entnahme einer Probe aus einer gut durchmischten Flüssigkeit. In diesem Fall ist auch die GG in besonders hohem Maß homogen (und damit kaum strukturiert). Andererseits gilt aber: die Miniaturausgabe kann umso kleiner sein, je homogener die GG ist. Zum gleichen Schluss, dass eine homogenere GG einen geringeren Stichprobenumfang n erfordert als eine heterogene kommt man auch mit dem Konzept des Stichprobenfehlers (also SF).

---

[20] Dabei ist die Orientierung an $N_k\sigma_k$ statt an N_k intuitiv leicht einzusehen. Dass $n_k > n_j$ ist dürfte nicht nur geboten sein, wenn $N_k > N_j$, sondern auch dann, wenn die Schicht k weniger homogen ist als Schicht j und deshalb $\sigma_k > \sigma_j$ ist. Die optimale Aufteilung, die gerade *nicht* proportional ist (es sei denn $\sigma_1 = \ldots = \sigma_K$) ist keineswegs irgendwie schlechter als die proportionale Aufteilung (allocation) von n bei einer geschichteten Stichprobe; sie ist ganz im Gegenteil, gemessen am SF (der ja gerade minimiert ist), besser (genauer: "optimal", also am besten).

[21] Andererseits gilt: Orientiert man, sich am Ergebnis, nicht am Auswahlprozess kann trotz gleicher Umstände (gleicher Methode) der "Ziehung" eine Auswahl repräsentativ sein, eine aber andere nicht.

[22] "the concept of representativeness is used primarily as an assertive talisman, or as means of sounding more scientific" (Kruskal & Mosteller, S. 16).

[23] Der Gedanke einer Miniaturausgabe der GG spielt eine Rolle bei der Klumpenauswahl, (cluster sample) weil dort Klumpen zu 100% ausgezählt werden, also eine Auswahl nur auf der ersten Stufe (Auswahl der Klumpen) stattfindet, nicht auf der zweiten (Einheiten innerhalb des Klumpens). Von M Klumpen (z.B. ausgewählten Gemeinden) wird – ganz im Gegensatz zu den K Schichten – gefordert, dass sie in sich möglichst inhomogen und (zwischen den Klumpen) hinsichtlich der Mittelwerte $\mu_1, \ldots, \mu_M$ aber möglichst gleich sein sollen.

Warum ist das Konzept RM – wie gesagt –im besonderen Maße vage und unklar? Aus zwei Gründen

1. Es lässt offen, wie beurteilt werden soll, ab wann man von einer "Miniatur" sprechen kann: Ist bei N = 1000 eine Auswahl von n = 50 eine akzeptable Miniatur, oder reicht vielleicht schon n = 10? Und wenn 10 ausreicht, wie erkennt man, dass diese Miniatur nicht schlechter ist als die bei n = 50?

   Der Gedanke, dass es bei jeder Miniatur vielleicht eine noch kleinere Miniatur gibt, kann bei physischen Objekten zutreffend sein (das immer noch kleinere maßstabgetreue Spielzeugauto); ist aber wohl nicht immer auf das Stichprobenproblem übertragbar. Man findet ihn in der Literatur auch oft zitiert als den Homunkulus Gedanken. [24]

2. Das RM Konzept legt sich – anders als die nächsten beiden Konzepte, RV und RA – auch nicht fest, *welche Einheiten* konkret in die Auswahl gelangen sollten, und wie (mit welcher Technik) man sie auswählen soll; es kann erfüllt sein, sowohl wenn man gezielt (nicht zufällig) auswählt, als auch dann wenn zufällig ausgewählt wird.

Der erste Punkt führt zu einer Paradoxie: Ist die n-te (Sub-) Stichprobe ebenfalls "repräsentativ" im Sinne des RM-Konzepts, kann man einwenden, dass die (n-1)-te Stichprobe noch nicht die gewünschte Miniatur war, sondern noch "eine Nummer zu groß war". Nach der Logik des RM Konzepts müsste jeweils die Miniatur der Miniatur als noch repräsentativer gelten. Aber es dürfte generell schwer sein, zu entscheiden, ob etwas mit mehr oder weniger Berechtigung als (kleinste) "Miniatur" akzeptiert werden kann; denn

- der Begriff "Miniatur" liefert keine Anhaltspunkte, um zwischen mehr oder weniger "repräsentativ" zu unterscheiden, und
- selbst wenn man sich sicher sein könnte, dass eine Auswahl eher eine "Miniatur" darstellt als eine andere, weiß man deshalb noch nicht wie groß jeweils $\sigma_{\bar{x}}$ und damit um wie viel besser eine Schätzung von μ ist (was ja Kenntnis von $\sigma_{\bar{x}}$ voraussetzt).

Es ist auch bezeichnend, wie wenig man sich offenbar bei Berufung auf das RM-Konzept überhaupt Gedanken macht über die restriktiven Voraussetzungen, die notwendig wären, um tatsächlich eine "getreue" Miniaturisierung haben zu können.

Idealerweise müsste dann eine Gesamtheit nämlich aus N = λm Einheiten bestehen, also bei m = 3 Einheiten A, B, C etwa wie nebenstehend aussehen (nur λ = 1 wäre dann *die* Miniatur). Aber die Wirklichkeit einer GG wird natürlich in aller Regel nicht so aussehen.

λ	Einheiten der GG	N
1	A, B, C	3
2	A, A, B, B, C, C	6
3	A, A, A, B, B, B, C, C, C	9
usw.		

Wie schon beim RS Konzept ist die kritische Frage auch hier, wie genau man es mit der Ähnlichkeit zwischen der GG und ihrem Miniaturexemplar meint. Betrachtet man nur ein Merkmal mit wenigen Merkmalsausprägungen (z.B. Geschlecht mit den Ausprägungen männlich und weiblich) mögen bereits zwei Menschen, ein Mann und eine Frau, eine Miniatur darstellen. Verlangt man dagegen von einem "getreuen Abbild", auch dass es die Häufigkeit "getreu" wiederspiegelt, mit der es in der GG ältere verheiratete Frauen im Vergleich jüngeren ledigen Männern gibt wird man mit n = 2 nicht auskommen. Je mehr Aspekten Rechnung zu tragen ist, damit die Stichprobe im Sinne von RM "repräsentativ" ist, desto größer muss auch die Miniatur werden, bis sie schließlich nicht kleiner ist als die GG. Wenn man es nur streng genug nimmt ist die einzig korrekte "Miniatur" das Original selbst.

[24] Das ist die früher einmal für möglich gehaltene Vorstellung, wonach der Mann in seinen Spermien alle seine Nachkommen in Miniaturform (Homunkulus) in sich trägt. Wenn das richtig wäre, müsste er genau genommen auch noch die zweite, dritte, ... Generation in sich tragen, und zwar jeweils in noch kleinerer Miniaturform.

Bei dem im Folgenden behandelten Konzept RV ist es dagegen genau umgekehrt. Aus ihm folgt, dass eine im Sinne von RV "repräsentative" Stichprobe eher relativ klein sein müsste.

2.4. Das Stellvertreter (Vize) Konzept (RV)

Man kann unter "repräsentativ" – ausgehend von der wörtlichen Übersetzung – verstehen, dass die ausgewählten Einheiten A, B, … die nichtausgewählten Einheiten X, Y, … vertreten ("repräsentieren") können. Es gibt zwei Ideen dazu, wie das gewährleistet sein kann:

1. die Ausgewählten A, B, … sollten den Nichtausgewählten X, Y, … weitgehend gleich oder "ähnlich" sein, und
2. A, B, … gelangen mit der gleichen Wahrscheinlichkeit in die Auswahl wie X, Y, … (es ist nur Zufall, dass A, B, … drin, aber X, Y, … draußen sind).

Bei RV gilt 1 (*Repräsentation wegen Ähnlichkeit*), in der Stichprobentheorie dagegen 2 (*Repräsentation wegen gleicher Auswahlwahrscheinlichkeit* von Repräsentanten und Repräsentierten, wobei Repräsentanten und Repräsentierte dann sich auch durchaus unähnlich sein dürfen). Mit 1 gibt es aber Probleme, die man mit 2 nicht hat. Es dürfte schwer sein zu beurteilen, ob hinreichende Ähnlichkeit mit den Nichtausgewählten gegeben ist,

- weil man ja i.d.R. die Nichtausgewählten nicht kennt, und
- selbst wenn man sie kennen würde, es schwer wäre zu sagen, welches Maß (!) an Ähnlichkeit erforderlich ist, um jemand anderes gleichwertig "vertreten" zu können.

Hinsichtlich der Existenz von m "Stellvertretern" sind drei Fälle zu unterscheiden

1. es gibt nur eine solche Einheit, die "Stellvertreter" (vicarius) für alle sein kann,
2. es gibt m > 1 solche "typische" Einheiten, die aber alle gleich sind, und
3. die m > 1 Einheiten sind unterschiedlich, wobei es schwer vorstellbar ist, dass sie gleichwohl auch alle gleich typisch sind.

In beiden Fällen 1 und 2 genügt eine Stichprobe von n = 1. Typische Einheiten können tatsächliche oder entsprechend *konstruierte* Einheiten sein. Letzteres dürfte realistischer sein und es kann sich dann ohnehin nur um eine Einheit handeln. Streng genommen verlangt "Stellvertretung", dass man auf den Durchschnitt bezüglich *aller* relevanter Merkmale abstellt (die Einheit i ist nur "typisch" *für alle*, wenn sie auch in *jeder* Hinsicht dem Durchschnitt entspricht, wenn also $x_{1i} = \mu_1$, $x_{2i} = \mu_2$,… usw.). In jedem Fall haben wir das Problem, dass

- unsere Kenntnis der Verteilung (und damit auch der Durchschnitte) der interessierenden Merkmale in der GG unvollständig ist und man ja auch gerade deshalb eine Stichprobenuntersuchung und Schätzung vornimmt,[25]
- es eigentlich absurd ist, in einem solchen Fall noch von einer "Schätzung" von Mittelwerten (μ_1, μ_2, …) zu sprechen; denn es sind ja die als bekannt vorauszusetzenden Werte $x_{1i} = \mu_1$, $x_{2i} = \mu_2$,…, weshalb die Einheit i "repräsentativ" ist, und dass
- es streng genommen nicht nur problematisch wird, von einer "Schätzung" zu sprechen, es ist auch fraglich, ob man noch von "Auswahl" sprechen kann, wenn es um so sehr spezifische Einheiten (oder auch nur eine solche Einheit) geht, die alle andern vertreten können

[25] Dieser Punkt gilt – wie bereits gesagt – bei allen Konzepten die an der Auswahl statt am Auswahlprozess ansetzen (also bei RS, RM RV und RA). Sie alle setzen Kenntnisse über die GG voraus, wie sie in der Regel nicht, oder nicht im geforderten Maße vorhanden sind.

(bzw. kann). Je mehr Bedingungen erfüllt werden müssen, desto mehr ist ein "Finden", und nicht ein "Auswählen" gefragt.[26]

Nur im Fall ungleicher und nur *mehr oder weniger* typischer Einheiten (oben Fall 3, also im allein realistischen Fall ($\sigma^2 \neq 0$) haben wir ein Auswahlproblem. Aber auch hier versagt das RV Konzept, weil es uns nicht sagt, welche Einheit mehr "typisch" ist (also ausgewählt werden sollten) und welche weniger und wie man dies feststellen soll. Es versagt bei der Frage

1. *welche*, und
2. *wie viele* Einheiten, sowie auch
3. *wie* man die Einheit(en) auswählen soll.

Aber das Prinzip der Zufallsauswahl gibt mit seinem Fokus auf Frage 3 zugleich auch die Antwort auf die beiden ersten Fragen, die sonst, ohne Zufallsauswahl schwer (oder gar nicht) zu beantworten sind.

2.5. Das Arche-Noah Konzept (RA)

Während dem RV Konzept "implizit" das Ideal einer Stichprobe von n = 1 zugrundeliegt wird beim RA Konzept eher Repräsentativität im Sinne einer korrekten und vollständigen Wiedergabe der in der GG vorhandenen *Vielfalt* angestrebt (verbunden eher mit einen größeren Stichprobenumfang). Statt nach Einheiten (oder *der* Einheit) zu suchen, die Stellvertreter für alle (auch die nichtausgewählten) Einheiten sind, bzw. ist, geht es bei RA um die Abdeckung ("coverage") aller in der GG vorhandener Arten von Einheiten, so wie in der Arche von jeder Tierart wenigstens ein Exemplar vorhanden war.[27] Das setzt Kenntnis der in der GG vorhanden Vielfalt voraus: wer die Arche mit den entsprechenden Tieren bestückt muss bereits das über die GG wissen, wozu ihm eigentlich erst die Stichprobe verhelfen soll. Und auch hier kommt es wieder darauf an, wie genau man die Einheiten beschreibt: reicht ein Pferd in der Arche, oder muss man auch von jeder Rasse (Hannoveraner, Lipizzaner, Trakehner usw.) ein Exemplar haben?

Dem RA Konzept würde es entsprechen, wenn bei einer geschichteten Stichprobe jede der K Schichten mit einer und nur einer ausgewählten Einheit repräsentiert wird, so dass n = K ist und $n_1 = n_2 = ... = n_K = 1$. Das gilt, *wenn* – und hier liegt der Haken – die Schichten so gebildet sind, dass sie die Heterogenität der GG genau wiedergeben. Aber abgesehen davon, dass dieser Gedanke ($n_k = 1$ für alle k = 1, …, K Schichten) der geschichteten Stichprobe völlig fremd ist, bleibt offen, wie K Schichten, die diesen Anforderungen genügen, zu bilden sind, oder – um auf das Bild der Arche zurückzukommen – wie zwischen den Tieren zu differenzieren ist damit *alle* (!) Arten in der Arche vertreten sind.

Konzepte mit dem Fokus auf "welche Einheiten auswählen?" wie RV und RA, verlangen eine (bewusste) Selektion, aber ohne sagen zu können, wie konkret zu selektieren ist. Nach dem folgenden Konzept verlangt "Repräsentativität" aber gerade, dass *nicht* selektiert wird.

2.6. Das Nichtselektivitätskonzept (RN)

Dieses Konzept stellt darauf ab, dass Einwirkungen mit dem Ziel, die Auswahl einer bestimmten Einheit zu erzwingen oder zu verhindern unterbleiben. Abgesehen davon, dass diese

[26] Auch das Konzept "Repräsentation" selber kann absurd werden. Das gilt insbesondere im Grenzfall lauter gleicher Einheiten (oben Fall 2). Ein Klon ist kein "Stellvertreter", man "vertritt" immer nur jemand *anderes*.

[27] Das Prinzip ist hier "Selektivität" (bewusste Auswahl) und das scheint eher quasi das Gegenteil der Zufallsauswahl zu sein (und auch das Gegenteil des RN Konzept). Dieser Gedanke steht auch Pate bei der "repräsentativen" Demokratie, bei der im Parlament alle Schichten der Bevölkerung vertreten sein sollten.

Vermeidung einer "selection bias" eigentlich genau der Grund für eine Zufallsauswahl ist, hat man mit RN das Problem, dass

- das RN Konzept nur besagt, wie *nicht* ausgewählt werden sollte, nämlich gezielt (keine "bewusste" Auswahl), nicht aber wie statt dessen ausgewählt werden sollte (RN ist zwar mit Zufallsauswahl verträglich aber nicht mit ihr gleichzusetzen), und dass
- auch nicht gesagt wird, wie man die Nichtselektivität sicherstellen soll und beweisen kann dass jede Art von Selektivität und Einseitigkeit ausgeschlossen ist?

Letzteres ist aber genau der Punkt, der zur Zufallsauswahl führt, denn man kann ohne sie nie sicher sein, ob wirklich vollkommene (wie zu messen?) "Blindheit" gegenüber der konkreten Einheit herrscht. Und wie beweist man seine Unparteilichkeit? Zufallsauswahl und Selektivität schließen sich auch nicht gegenseitig aus. Sie lassen sich sogar vorteilhaft miteinander verbinden, z.B. im Falle der bereits genannten geschichteten Stichprobe, wo Kenntnisse über die GG zur Verbesserung der Schätzung genutzt werden und deshalb bewusst nicht alle Einheiten in einen Topf geworfen werden und unterschiedslos behandelt werden. Außerdem

- kann z.B. im Fall von Ausreißern bewusstes Nichtziehen oder Entfernen von Elementen aus einer bereits getroffenen Auswahl sinnvoll sein und
- eine Auswahl kann trotz Bemühung um Nichtselektivität (also um Repräsentativität im Sinne von RN) selektiv sein, wenn nämlich der Auswahlvorgang eine Selbstselektion ermöglicht (z.B. selbstbestimmter Austritt in Form von non-response).

Nichtselektivität kann nach alle dem nicht das alleinige Kriterium der "Repräsentativität" sein. Aber auch als ein Kriterium neben anderen ist sie nicht brauchbar, schon weil sie schwer operational zu definieren ist.

2.7. Warum überhaupt Zufall?

Ein Problem mit dem RN Konzept ist das Beweisprobleme bezüglich "Nichtbeeinflussung", "Unverzerrtheit" der Auswahl und ein Argument dafür, bei der Auswahl den Zufall sprechen zu lassen ist, dass einem genau dann solche Probleme erspart bleiben. Aus drei Gründen lässt man gern bei einer Auswahl den Zufall entscheiden

1. um eine Verzerrung zu vermeiden, die entstünde, wenn man systematisch (also gerade *nicht* zufallsgesteuert) nur ganz bestimmte Einheiten auswählt und
2. um die Wahrscheinlichkeitsrechnung (die Zufälligkeit voraussetzt[28]) anwenden zu können bei der Quantifizierung des Auswahlfehlers $\sigma_{\bar{x}}$, einer Größe die auch für Konfidenzintervalle und statistischen Tests benötigt wird, und
3. um sagen zu können, dass z.B. die ausgewählten Personen A, B und C "repräsentativ" für alle sind, die Ergebnisse also *verallgemeinerungsfähig* sind (was letztlich auf das Gleiche wie Nr. 2 hinausläuft, denn bei Tests geht es ja darum, ob Stichprobenbefunde für die GG zu verallgemeinern sind).[29]

Also "Zufall", um eine Bevorzugung oder Benachteiligung von Einheiten der GG zu vermeiden (bzw. Abwesenheit von Selektivität, Einseitigkeit und Parteilichkeit sicherzustellen) und

[28] Bei einem reinen Glücksspiel (wo es nur oder entscheidend auf den Zufall ankommt) kann man fragen, wie wahrscheinlich es ist, dass A gegen B gewinnen wird, nicht aber bei einem Geschicklichkeitsspiel (oder generell, wo Wissen und Können entscheidet, was bei A und B im unterschiedlichem Maße vorhanden sein kann).

[29] Das Argument dabei ist: hätte man A, B und C bewusst ausgewählt, dann würden die Ergebnisse nur für A, B und C gelten, hat man sie dagegen zufällig ausgewählt, dann gelten sie auch für andere Personen, wie X, Y und Z, weil der Zufall ja gerade darin besteht, dass es genauso gut sein konnte, dass X, Y und Z, statt A, B und C ausgewählt wurden (und das, weil es genauso wahrscheinlich war).

um Methoden, die auf der Wahrscheinlichkeitsrechnung beruhen, beim Schluss von der Stichprobe auf die GG anwenden zu können.

Die Frage die wir in Abschn. 3 behandeln wollen ist: gilt auch der Umkehrschluss, d.h. wenn kein Zufall, dann auch keine Methoden, die auf der Wahrscheinlichkeitsrechnung beruhen?

Unsere Antwort wird ganz klar "ja" sein. Dazu aber die Vorfrage: wann haben wir eine "Zufallsauswahl" und wann nicht? Zufallsauswahl verlangt – zumindest bei einer "einfachen" (nicht geschichteten) Stichprobe – dass *alle* Einheiten der GG eine gleich große Auswahlwahrscheinlichkeit haben. Kein Zufall hat man bei einer *"willkürlichen"* Auswahl (accidental sample) oder einem convenience sample; denn wenn die Erreichbarkeit (und Auskunftsbereitschaft[30]) entscheidet ist die Auswahlwahrscheinlichkeit gerade *nicht* gleich, sie ist groß bei den Erreichbaren und null bei den Nichterreichbaren. Keinen Zufall haben wir auch bei einer Totalerhebung, bei der jede Einheit der GG mit einer Wahrscheinlichkeit von 1 in die "Stichprobe" gelangt. Bei einem sicheren Ereignis (dass z.B. Dienstag auf Montag folgt), macht es keinen Sinn, mit Wahrscheinlichkeiten rechnen zu wollen.

3. Anwendbarkeit von Signifikanztests

3.1. Einführung, Voraussetzungen einer Stichprobenverteilung

Das Hauptproblem mit der "Repräsentativität" ist, wie schon in Abschn. 1 gesagt, dass der Begriff dazu verleitet, zu meinen, es sei vollkommen in Ordnung, auch dann Konfidenzintervalle zu bestimmen oder Signifikanztests durchzuführen, wenn gar keine Zufallsauswahl vorliegt. Wie so oft wird etwas klarer wenn man extreme Fälle betrachtet. Wir versuchen deshalb zu zeigen, was gegen die Praxis spricht, zu testen auch dann wenn

- die "Stichprobe" faktisch die GG ist (Abschn. 3.3), und wenn
- man eine "Stichprobe" "vorgefunden" hat oder sich nach einem Repräsentativitätskonzept zusammengestellt hat, und dann nachträglich die GG, für die diese Stichprobe repräsentativ sein soll aufgrund der Struktur der Stichprobe konstruiert (Abschn. 3.4).

Es ist gar nicht so selten, dass in solchen Fällen gleichwohl inferenzstatistische Methoden angewendet werden. Mc Closkey and Ziliak (1996) haben 182 in den 80er Jahren in der American Economic Review (AER) erschienene Arbeiten mit empirischen Regressionsanalysen auf die Frage hin untersucht, ob hier (bei 107 Arbeiten ging es um Querschnittsdaten, beim Rest um Zeitreihenanalysen[31]) korrekt mit Signifikanztests operiert wurde:

> "It requires thinking more rigorously about data – for example, asking what universe they are a 'sample' from. (Carelessness in such matters is more common than one might have expected. Of the 107 papers using cross-sectional data, for example 20 percent used tests of statistical significance on the entire population or on a sample of convenience. Only two of these offered some justification for the usage)" (S. 112).

Dass so etwas oft vorkommt, liegt auch daran, dass man von Signifikanztests oft den Beweis für die Richtigkeit eines Modells erwartet bzw. dafür, die richtige Auswahl unter alternativen Modellen getroffen zu haben, egal von welcher Natur die für den Test verwendeten Daten

[30] Es ist bezeichnend, dass non-response bei einer echten Stichprobe (Zufallsauswahl) ein Thema ist und Anlass ist, korrigierende Gewichtungen vorzunehmen oder fehlende Daten hinzu zu schätzen (zu "imputieren"), während es bei einem convenience sample kein Thema ist (es gibt dort keinen Vergleich zwischen dem Stichprobenumfang, wie er tatsächlich ist und wie er eigentlich sein sollte).

[31] Aus Platzgründen können wir hier nicht darauf eingehen, wie bei den in der Ökonometrie üblichen Zeitreihen als Daten argumentiert wird um die Anwendbarkeit von Tests und Konfidenzintervallen zu rechtfertigen. Es ist nicht einfach so, dass man behauptet, die Zeitreihe stelle Daten dar, die zufällig aus einem sehr viel längeren Zeitraum gezogen wurden (womit dann die Zeit eine Zufallsvariable wäre), denn dann könnte man ja fragen: warum hat man statt 2000, 2001, 2002, … nicht zufällig 1452, 1789, 1815, … gezogen?

sind. In diesem Sinne zitierten die Autoren dann auch eine Kritik von Zvi Griliches, zu einer Studie, "the 'sample' analyzed comes close to covering completely the relevant population. Tests of significance are used here as a metric for discussing the relative fit of different versions of the model." (p. 106) und stellten zusammenfassend fest: "that populations should not be treated as samples, and that statistical significance is not a substitute for economic significance …. But most authors in the AER do not understand these points" (p. 106).

Von Statistikern wird gegen diese Praxis üblicherweise eingewandt, dass $\sigma_{\bar{x}}$, die bestimmende Größe sowohl für Konfidenzintervalle, als auch für Signifikanztests, die den "wahren" Mittelwert μ der GG betreffen, die Existenz der Stichprobenverteilung von $\bar{x}$ voraussetzt, was aber wiederum voraussetzt

1. eine gegebene (bekannte oder hypothetisch angenommene) GG,
2. aus der *Stichproben* (und zwar *mehr als nur eine einzige Stichprobe*) gezogen werden könnten[32] (und i.d.R. auch eine nach dem Zufallsprinzip tatsächlich gezogen wird), und
3. die Kenntnis der Wahrscheinlichkeit, mit der die verschiedenen Stichproben gezogen werden; denn wenn man nichts darüber weiß, wie wahrscheinlich bestimmte Stichproben sind, kann man auch nichts darüber sagen, mit welcher Wahrscheinlichkeit bestimmte $\bar{x}$ Werte bei einem gegebenem oder angenommenen Wert für μ auftreten.

Dies wird im folgenden Abschnitt 3.2 verdeutlicht, indem wir für ein Zahlenbeispiel konkret Stichprobenverteilungen von $\bar{x}$ herleiten werden. Man verstößt gegen Voraussetzung 2 (oder auch Voraussetzung 3 in Form der extremen Wahrscheinlichkeiten von 0 bzw. 1), wenn nur eine ganz bestimmte Stichprobe in Frage kommt, weil

- faktisch eine Totalerhebung vorliegt, man also bei einer realen, also endlichen GG vom Umfang N quasi eine "Stichprobe" vom Umfang n = N mit der extremen Wahrscheinlichkeit von 1 hat und nach Gl. (2) $\sigma^*_{\bar{x}} = 0$ ist weil N - n = 0 ist; und wenn
- es die Art, die (nichtzufällige) "Stichprobe" zu definieren nicht zulässt (oder geboten erscheinen lässt), dass es mehrere verschiedene Stichproben aus der gleichen GG gibt, die als gleichwertig bzw. gleich "repräsentativ" zu betrachten sind.

In diesen Fällen ist eine Stichprobenverteilung faktisch nicht gegeben, die Berechnung eines Konfidenzintervalls oder einer Prüfgröße (Teststatistik) sinnlos[33] und $\sigma_{\bar{x}} = 0$.[34] Die Schätz- und Testtheorie ist also nur anwendbar wenn nicht nur eine, sondern mehrere Stichproben aus einer GG nicht nur möglich, sondern auch wahrscheinlich sind. Die Stichproben dürfen nicht nur, sie *müssen* sogar verschieden sein damit es verschiedene Werte von $\bar{x}$ gibt und $\sigma_{\bar{x}} > 0$ ist.[35] Aber dies scheint sich mit dem Begriff "Repräsentativität" nicht gut zu vertragen. Es ist beispielsweise schwer vorstellbar, dass es mehrere *verschiedene* Miniaturausgaben der gleichen GG gibt. Gäbe es mehrere, müsste man ja auch eine Regel haben, wie aus ihnen auszuwählen ist. Wenn es aber nur eine Miniaturausgaben gibt, hat man auch nur ein $\bar{x}$ (was wir $\bar{x}_1$ nennen wollen) auf Basis der n Einheiten der einen "repräsentativen" Stichprobe. Andere Werte $\bar{x} \neq \bar{x}_1$ und damit eine *Verteilung* von $\bar{x}$ hätten wir nur, wenn wir auch Einheiten aus den bisher nicht berücksichtigten N-n Einheiten betrachten würden. Aber warum sollte man das tun, wenn dies doch auf die Betrachtung auch "nicht repräsentativer" Stichproben hinaus-

[32] Das "werden *könnten*" ist nach dem oben zitierten Satz von Bortz eine Voraussetzung, dass eine tatsächlich gezogen *wurde* aber nicht.
[33] Wir zeigen in Abschn. 3.3 ff, dass in beiden Fällen schon die Fragestellung eines statistischen Tests sinnlos ist.
[34] Damit verbürgt auch ein verschwindender Stichprobenfehler nicht in jedem Fall eine gute Schätzqualität.
[35] Die Existenz mehrerer gleichwertiger Stichproben aus einer GG wird auch bei statistischen Tests nicht selten vorausgesetzt; wir kommen darauf in Abschn. 3.5 zurück.

liefe? Im Grunde sind es zwei Punkte, Zufall und Ähnlichkeit, bei denen sich die Welt der Statistik und die der "Repräsentativität" trennen; in der Statistik ist

1. der Zufall das Kriterium für eine Entscheidung: bei Tests geht es darum, ob eine Abweichung so gering ist, dass sie *noch zufällig* sein könnte (dann wird H_0 angenommen) oder so groß ist, dass sie *nicht mehr zufällig* (also "signifikant") sein dürfte, und
2. in der Statistik ist Ähnlichkeit von Stichprobe(n) und GG kein Kriterium; sie dürfen nicht nur unähnlich sein, sie sollen es auch;

aber bei der "Repräsentativität" ist gerade die Ähnlichkeit das Kriterium; das Abbild sollte "getreu" sein und wenn es das nicht ist spielt es keine Rolle inwiefern dies zufällig nicht so ist.

3.2. Ein Modell für eine "strukturidentische" Auswahl

In diesem Abschnitt versuchen wir zu zeigen dass

a) sich mit zunehmendem N wohl i.d.R. der Anteil "repräsentativer" (im Sinne von RS) Stichproben an der Gesamtzahl

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

der Stichproben vom Umfang n verringern dürfte,

b) sich widersprüchliche Folgerungen für die Verteilung der $\bar{x}$ bei Betrachtung nur repräsentativer und nur nicht-repräsentativer Stichproben ergeben, wenn man das Verhältnis der internen zur externen Varianz systematisch variiert,

c) man eine nichtzufällige Auswahl als Fall extremer Wahrscheinlichkeiten (von 0 oder 1) von Stichproben auffassen kann,

d) dass eine von Null verschiedene Wahrscheinlichkeit von *allen* möglichen Stichproben fundamental ist und dass bei einer nicht-zufälligen Teilerhebung keine der bei Herleitung der Stichprobenverteilung zu machenden Voraussetzungen erfüllt ist.

Zu a: Um die folgenden kombinatorischen Überlegungen anschaulicher zu machen sei eine GG mit N = 8 Einheiten angenommen, die sich nach einem dichotomen (binären) Merkmal in G = 2 gleich große Gruppen zu jeweils N/G = 8/2 = 4 Einheiten wie folgt aufteilt:[36]

	Gruppe 1				Gruppe 2			
Einheit	a	b	c	d	α	β	γ	δ
x-Wert	12	24	36	48	6	12	18	24

Wir betrachten im Folgenden als "strukturgleiche" Stichproben solche mit jeweils n = G Einheiten, eine aus jeder Gruppe wie a, α oder a, β usw. Die Anzahl repräsentativer Stichproben ist $R = (N/G)^n$. Im Fall von N = 8 und G = 2 sind $(N/2)^2 = 16$ Stichproben "repräsentativ" ("strukturgleich" bezüglich des dichotomen strukturbestimmenden Merkmals) wie die Stichproben aα, aβ,… und die restlichen N(N-2)/4 = 12 "nichtrepräsentativ", wie ab, ac usw. (nur Einheiten von Gruppe 1) oder auch αβ, αγ usw. (nur Gruppe 2).

	G = n = 2		
N	8 [a)]	16	32
R	16	64	256
S	28	120	496
R/S	0,5714 [b)]	0,5333	0,5161

G = n = 4		
8	16	32
16	256	4096
70	1820	35960
0,2286 [c)]	0,1407	0,1140

a) die sich für diesen Fall ergebene Stichprobenverteilung wird unten (unter c) dargestellt
b) die Relation ist $(2 - 1/N)^{-1}$ und strebt mit $N \to \infty$ gegen ½
c) die Relation ist $3N^3/32(N-1)(N-2)(N-3)$ und strebt mit $N \to \infty$ gegen 3/32 = 0,093785

[36] Wir wählen G = 2, weil man dann eine besonders einfache, durch einen Anteil p bestimmte Struktur erhält.

Ist G = 2 und n = 4 dann ist $R = \binom{N/2}{2}^2$ und $S = \binom{N}{4}$. Dabei enthält die "repräsentative" Stichprobe je zwei Elemente aus Gruppe 1 und aus Gruppe 2 und es gilt dann

N	8	16	32
R	36	784	14400
S	70	1820	35960
R/S	0,5143	0,4308	0,4004

Mit $M = N/2 \to \infty$ strebt die Relation R/S gegen 3/8 = 0,375.

Wie man sieht, klammert man einen erheblichen Teil aller S möglichen Stichproben (die aber *alle mit gleicher Wahrscheinlichkeit* in die Bestimmung der Stichprobenverteilung eingehen) aus der Betrachtung aus, wenn man sich nur mit "repräsentativen" Stichproben beschäftigt.

Zu b: Im eingangs präsentierten Beispiel ist $x_{1i} = fx_{2i}$ mit f = 2 und $\mu = ½(\mu_1 + \mu_2) = 22{,}5$ wobei $\mu_1 = 30 = f\mu_2 = 2\mu_2$; für die Varianzen innerhalb der Gruppen gilt damit $\sigma_1^2 = f^2\sigma_2^2$, im Beispiel also $\sigma_1^2 = 4 \cdot 45 = 180$ und $\sigma_2^2 = 45$. Nach Gl. 3 lässt sich danach die Gesamtvarianz bei $f_k = ½$ (bei k = 1, 2) zerlegen in die interne Varianz $V_{int} = \frac{1}{2}\sum_k \sigma_k^2 = \frac{1}{2}\sigma_2^2\left(1 + f^2\right)$ und die externe Varianz $V_{ext} = \frac{1}{2}\sum_k \left(\mu_k - \mu\right)^2 = \left[\frac{\mu_2}{2}\left(f - 1\right)\right]^2$ im Beispiel mit f = 2 ist also $V_{int} = 112{,}5$ und $V_{ext} = (15/2)^2 = 56{,}25$ und $\sigma^2 = V_{int} + V_{ext} = 168{,}75$. Der Anteil $V_{int}/\sigma^2 = [1+2{,}5(1 - 2f/(f^2+1)]^{-1}$ wird mit wachsenden f kleiner und strebt gegen 1/3,5 = 0,2857.

f	V_{ext}	V_{int}	σ^2	V_{int}/σ^2
1	0	45	45	1
2	56,25	112,5	168,75	2/3
3	225	225	450	0,5

f	V_{ext}	V_{int}	σ^2	V_{int}/σ^2
4	506,25	382,5	888,75	0,43
5	900	585	1485	0,3939
6	1406,25	832,5	2238,75	0,3719

Sucht man nach einer Begründung für eine Bevorzugung "strukturgleicher" Stichproben, so könnte man daran denken, dass man vielleicht ein geringeres $\sigma_{\bar{x}}$ erhält, wenn man eine spezielle Stichprobenverteilung nur für diese strukturgleichen ("repräsentativen") Stichproben berechnet. Dabei ist jedoch zu bedenken, dass

- man zwar – wie das hier zu Demonstrationszwecken geschieht – solche partielle Stichprobenverteilungen für die S_r "repräsentativen" und die $S_{nr} = S - S_r$ "nicht repräsentativen" Stichproben untersuchen kann, dass aber für das Schätzen und Testen immer nur allein die Stichprobenverteilung auf Basis *aller* S Stichproben relevant ist, und dass
- wir auch im Folgenden bei Stichproben, die keine "random samples" sind mit gleichen Wahrscheinlichkeiten für alle S_r bzw. S_{nr} Stichproben rechnen, was aber nicht der Praxis bei solchen "Stichproben" (wie z.B. convenience samples) entspricht.

Diese Einschränkungen relativieren unser folgendes, zudem auch noch etwas widersprüchliches Ergebnis: mit steigendem f nimmt die Unterschiedlichkeit von μ_1 und μ_2 und damit das Gewicht der externen Varianzkomponente an der Gesamtvarianz σ^2 zu und es werden repräsentative Stichproben (gemessen am relativ geringen "partiellen" $\sigma_{\bar{x}}$) vorteilhafter.[37] Andererseits nimmt aber auch σ^2 zu[38] und damit der Stichprobenfehler $\sigma_{\bar{x}}$ insgesamt, der ja korrekt bestimmt, stets die Berücksichtigung *aller* Stichproben verlangt. Außerdem kann das "partielle" $\sigma_{\bar{x}}$ (nur der repräsentativen Stichproben) nicht nur kleiner, sondern auch größer sein als

[37] Danach wären solche strukturbestimmenden Merkmale vorteilhaft, die Gruppen erzeugen, die in sich homogen (geringe V_{int}) und im Vergleich miteinander sehr unterschiedlich (V_{ext} groß) sind, also ähnlichen Kriterien genügen wie Schichtungsmerkmale bei einer geschichteten Stichprobe.

[38] wie die zuletzt präsentierte Tabelle zeigt: von 45 zu 2238,75, wenn f zunimmt von 1 bis 6.

das der nichtrepräsentativen Stichproben. Das gilt bei f =1 (V_{ext} = 0), wenn also die Struktur aus zwei Gruppen besteht, die nicht nur gleich groß sind ($N_1 = N_2 = N/2$), sondern sich auch wegen gleicher x-Werte ($x_{i1} = x_{i2}$) nicht wirklich unterscheiden. Eine Begründung für eine Bevorzugung "repräsentativer" Teilgesamtheiten dürfte also danach kaum beizubringen sein.

Im eingangs präsentierten Zahlenbeispiel (mit N = 8, f = 2) erhält man die folgenden Stichprobenverteilungen von $\bar{x}$

nur repräsentative Stichproben	
Anzahl	16
$E(\bar{X})$	360/16 = 22,5
$E(\bar{X}^2)$	9000/16 = 562,5
$\sigma_{\bar{x}}^2$ ($\sigma_{\bar{x}}$)	56,25 (7,5)
$\bar{x}$ von…bis	9 bis 36

nur nichtrepräsentative Stichpr.	
Anzahl	12
$E(\bar{X})$	270/12 = 22,5
$E(\bar{X}^2)$	7200/12= 600
$\sigma_{\bar{x}}^2$ ($\sigma_{\bar{x}}$)	93,75 (9,7)
$\bar{x}$ von…bis	9 bis 42

mit allen möglichen Stichproben	
Anzahl	28
$E(\bar{X})$	630/28 = 22,5
$E(\bar{X}^2)$	16200/28 = 578,6
$\sigma_{\bar{x}}^2$ ($\sigma_{\bar{x}}$)	72,32 (8,5)
$\bar{x}$ von…bis	9 bis 42

Die vollständige Stichprobenverteilung von $\bar{x}$ mit allen 28 möglichen Stichproben ist:

$\bar{X}$	9	12	15	18	21	24	27	30	33	36	42
Stichproben	1*	2*	4	4	3	3	2	4	1	2	1

*) Die Wahrscheinlichkeit für eine Stichprobe mit $\bar{x}$ =9 ist somit 1/28, mit $\bar{x}$ = 12 ist sie 2/28 usw.

Es gilt $\sigma_{\bar{x}}^2 = E(\bar{X}^2) - \left[E(\bar{X})\right]^2$. Alle Schätzungen sind erwartungstreu. d.h. es ist stets $E(\bar{X}) = \mu$. Man kann die hier berechneten Varianz $\sigma_{\bar{x}}^2$ (bei allen Stichproben nicht die partiellen" $\sigma_{\bar{x}}^2$) auch mit den oben berechneten Werten für σ^2 bestimmen und erhält gem. Gl. 2 mit $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = 72{,}32$ bei σ^2 = 168,75, n = 2, und N = 8. Das gilt für jedes für f, nicht nur für f = 2. Die Relation 7,5 < 9,7 für "partielle" $\sigma_{\bar{x}}$ scheint zugunsten "repräsentativer" Stichproben zu sprechen. Das muss aber nicht grundsätzlich so sein. Im Fall f = 1 gilt – wie erwähnt – das Gegenteil, hier ist die Relation 4,47 > 3,87.[39]

nur repräsentative Stichproben	
$\sigma_{\bar{x}}^2$ ($\sigma_{\bar{x}}$)	22,5 (4,47)
$\bar{x}$ von…bis	6 bis 24[a)]

nur nichtrepräsentative Stichpr.	
$\sigma_{\bar{x}}^2$ ($\sigma_{\bar{x}}$)	15 (3,87)
$\bar{x}$ von…bis	9 bis 21[b)]

mit allen möglichen Stichproben	
$\sigma_{\bar{x}}^2$ ($\sigma_{\bar{x}}$)	19,28 (3,39)
$\bar{x}$ von…bis	6 bis 24

[a)] man erhält 24 als Mittel von x_δ =24 und x_d = 24

[b)] maximal möglich ist jetzt 21, das Mittel aus x_γ = 18 und x_δ =24 oder aus x_c = 18 und x_d = 24.

Die Schätzung von μ war bei der bisher angenommenen Gleichwahrscheinlichkeit aller Stichproben stets erwartungstreu. Auch das muss nicht grundsätzlich so sein und wird sich gleich zeigen, wenn wir von der zentralen Annahme der Gleichwahrscheinlichkeit der möglichen Stichproben abrücken. Mit der Annahme ungleicher Wahrscheinlichkeiten der Stichproben (bis zu extremen Wahrscheinlichkeiten von 0 und 1) können wir eine Brücke bauen zwischen "probability samples" und nicht-zufälligen Teilerhebungen wie z.B. convenience samples.

Zu c: Nehmen wir für unser Ausgangsbeispiel (mit f = 2 und μ = 22,5) an, dass jede der 13 Stichproben[40] in denen a und/oder α vorkommt doppelt so wahrscheinlich ist, wie jede der übrigen 15 Stichproben, in denen weder a noch α vorkommt. Dann gilt

[39] Bei f = 3 ist die Relation 10,6:17,32 und auch bei f = 1,5 ist die Relation mit 6,05:6,20 noch zu Gunsten der "repräsentativen Stichproben". Der Fall f = 1 wäre vom Standpunkt einer Analogie zum cluster sample besonders günstig, aber es gibt viele Gründe weshalb es nicht zielführend ist, zu versuchen, eine mathematische Theorie der nichtzufälligen Auswahl per Analogien zu Stichprobendesigns (wie geschichtete Stichprobe oder Klumpenstichprobe) zu entwickeln. Aus Platzgründen können wir hier auf unsere Versuche nicht weiter eingehen.

[40] 7 von 16 repräsentativen und 6 von 12 nichtrepräsentativen Stichproben.

nur repräsentative Stichproben	
$E(\overline{X})$	20,74
$\sigma^2_{\bar{x}}$ ($\sigma_{\bar{x}}$)	55,49 (7,45)

nur nichtrepräsentative Stichpr.	
$E(\overline{X})$	21
$\sigma^2_{\bar{x}}$ ($\sigma_{\bar{x}}$)	84 (9,17)

mit allen möglichen Stichproben	
$E(\overline{X})$	20,85
$\sigma^2_{\bar{x}}$ ($\sigma_{\bar{x}}$)	68,03 (8,25)

Es gibt jetzt bei ungleichen Wahrscheinlichkeiten keine Erwartungstreue mehr ($E(\overline{X}) < \mu = 22{,}5$, was aber auch nicht überrascht wo doch x bei a und α jeweils unterdurchschnittlich (unter 30 bzw. 15) war. Noch deutlicher wird die Deformation oder Degeneration der Stichprobenverteilung wenn man Situationen betrachtet mit nur zwei gleichwahrscheinlichen Stichproben, nämlich mit den Elementen b, γ und c, δ oder gar ein convenience sample, z.B. nur mit den Einheiten b und γ als Grenzfall *einer* einzigen sich in einer Situation anbietenden Stichprobe, die deshalb auch mit Wahrscheinlichkeit 1 gezogen wird (womit dann auch $\overline{x}$ keine Zufallsvariable mehr ist). Dann ist die Stichprobenverteilung zu einer Zwei- bzw. Einpunktverteilung degeneriert, während wir es oben mit 28 Stichroben zu tun hatten

	$\overline{x} = 21$	$\overline{x} = 30$
Wahrscheinlichkeit bei zwei Stichproben b, γ und c, δ	1/2	1/2
bei nur einer Stichprobe b, γ	1	0

$E(\overline{X})$	$\sigma_{\bar{x}}$
25,5	20,25
21	0

In solchen Situationen kann man natürlich nicht mit dem $\sigma_{\bar{x}} = 8{,}5$ von allen 28 möglichen Stichproben rechnen. Allerdings könnte man selbst bei nur einer einzigen (vorgefundenen oder mit dem RS-Kriterium gebildeten) "Stichprobe" mit den Elementen b und γ einen Schätzwert $\hat{\sigma}_{\bar{x}} = \hat{\sigma}_x / \sqrt{n}$ und damit dann rein rechnerisch ein Konfidenzintervall bestimmen.[41] Analog zur üblichen Vorgehensweise würde man jetzt im Fall mit $x_b = 24$ und $x_\gamma = 18$ ($n = 2$) wie folgt rechnen: Mit $\hat{\sigma}^2_x = \hat{\sigma}^2 = \frac{(x_b - \overline{x})^2 + (x_\gamma - \overline{x})^2}{n-1} = 18$ erhält man einen Schätzwert für das unbekannte σ^2 der GG und dann $\hat{\sigma}_{\bar{x}} = \sqrt{18}/\sqrt{2} = 3$ mit Gl. 1, bzw. $3 \cdot \sqrt{6/7} = 2{,}778$ mit Gl. 2 einen Schätzwert für den Stichprobenfehler $\sigma_{\bar{x}}$, der aber eigentlich in dieser Situation – wie die obige Tabelle zeigt – $\sigma_{\bar{x}} = 0$ ist. Zwar ist das "wahre" σ^2 der GG von $\sigma^2 = 168{,}75$ mit $\hat{\sigma}^2 = 18$, und damit auch $\hat{\sigma}_{\bar{x}}$ mit 3 statt 8,5 weit verfehlt, aber im Prinzip würde man auch genau so rechnen,[42] wenn das Wertetupel x_b, x_γ nur eine von 28 Zufallsstichproben vom Umfang $n = 2$ ist und wie die anderen 27 Stichproben mit einer Wahrscheinlichkeit von 1/28 gezogen worden wäre, die Berechnung also nicht zu beanstanden wäre.

Zu d: Warum darf man in einem Fall (x_b, x_γ eine Zufallsstichprobe) so rechnen, im anderen (x_b, x_γ ein convenience sample) aber nicht? Es ist wichtig, sich klar zu machen, dass "rein äußerlich", beim Rechengang kein Unterschied besteht zwischen der richtigen und der falschen Rechnung. Der Unterschied liegt allein in den Voraussetzungen für die Zulässigkeit der Berechnung von $\hat{\sigma}^2_x = \hat{\sigma}^2$ und damit auch von $\hat{\sigma}_{\bar{x}}$.

Ohne Stichproben nach dem Zufallsprinzip hat man keine Stichprobenverteilung von $\overline{x}$ und damit auch keinen Stichprobenfehler $\sigma_{\bar{x}}$ und beim convenience sample ist keine der folgenden drei Voraussetzungen für die Existenz einer Stichprobenverteilung gegeben (wir haben die Voraussetzungen bereits in Abschn. 3.1 genannt und von ihnen in diesem Abschnitt jeweils bei der Herleitung einer Stichprobenverteilung Gebrauch gemacht), nämlich

[41] Wir sehen davon ab, dass n = 2 natürlich unrealistisch klein ist und dass der zentrale Grenzwertsatz, wonach mit der Normalverteilung als Stichprobenverteilung zu rechnen ist, $n \to \infty$ (faktisch eher nur $n > 30$) voraussetzt.
[42] Natürlich wieder abgesehen von den unrealistisch kleinen Werten n und N.

1. dass es eine klar definierte GG gibt, die unabhängig von den Stichproben existiert, von diesen durchaus verschieden sein kann (nicht nur wegen $n \neq N$) und Parameter hat (wie μ und σ^2), über die Kenntnis besteht, bzw. über die Hypothesen gebildet werden;[43]
2. dass mehr als nur eine Stichprobe möglich ist, so dass $\sigma_{\bar{x}} > 0$ ist weil verschiedene Werte für $\bar{x}$ auftreten können;
3. dass *jede* der N Einheiten der GG a priori (vor Ziehung einer konkreten Stichprobe) die gleiche positive (von null und eins verschiedene) Auswahlwahrscheinlichkeit hat, so dass dann jede der S Stichproben mit der gleich großen Wahrscheinlichkeit von $1/S > 0$ auftreten kann, wobei es irrelevant ist, welche und wie viele der S Stichproben "repräsentativ" oder "nicht repräsentativ" sind.[44]

Viele (oder eher die meisten) der Einheiten der GG haben demgegenüber beim convenience sample eine Auswahlwahrscheinlichkeit von null, andere (die Anwesenden) eine von eins, so dass 3 verletzt ist, aber auch Voraussetzung 1 weil es meist im Dunklen bleibt, was jeweils die relevante GG sein soll. Mehr noch: ob man in das Sample gelangt hängt von systematischen, nicht zufälligen Einflüssen ab. Anwesenheit ist i.d.R. nicht eine Sache des Zufalls. Es sind also systematische Einflüsse, die sich in der großen Zahl nicht ausgleichen, so wie dies von zufälligen Einflüssen zu erwarten ist und sie sind es, die eine "Verzerrung" (bias) von unbekannter Größe erzeugen.

Ohne die genannten drei Voraussetzungen ist $\hat{\sigma}^2$, bzw. $\hat{\sigma}_{\bar{x}}$ kein erwartungstreuer (unverzerrter) Schätzer für σ^2 bzw. für $\sigma_{\bar{x}}$, darf man nicht in der üblichen Art schätzen und testen und kann man auch noch viel Grundsätzlicheres nicht:

- es gibt jetzt auch keine Basis, zu folgern, dass sich die Schätzung verbessert (und um wie viel sie sich verbessert) wenn man den Stichprobenumfang n vergrößert (ein Ergebnis, das dagegen leicht aus Gl. 1 bzw. 2 herzuleiten ist), und
- es ist dann dubios, welche Erkenntnis man mit einem Hypothesentest überhaupt gewinnt; man kann noch nicht einmal sinnvoll Hypothesen formulieren, denn das sind ja Vermutungen über die GG und es muss klar sein, was bei einem Test die GG sein soll.

Das wird besonders deutlich bei der wohl krassesten Fehlanwendungen inferenzstatistischer Methoden, wenn nämlich mit einer Stichprobe, die faktisch die GG darstellt, ein Signifikanztest durchgeführt wird (was im Übrigen gar nicht so selten ist und in Zeiten von "big data" wohl noch viel häufiger passieren wird als es jetzt schon der Fall ist).

3.3. Die vermeintliche Stichprobe ist eigentlich die Grundgesamtheit

Ein Signifikanztest ist leider oft nur noch ein unverstandenes Ritual und wird besonders bedenklich, wenn man gar nicht bemerkt, dass die Daten eigentlich selbst schon die GG darstellen. Wir bringen hierzu nur zwei Beispiele, beide aus mehr volks- als betriebswirtschaftlichen Schriften. Man könnte natürlich auch viele andere ähnliche Fälle anführen, z.B. aus der Medizin, wo dies auch recht üblich ist.

---

[43] Bei den Betrachtungen in diesem Abschnitt haben wir eine vollständig (mit allen x-Werten der N Einheiten) bekannte GG angenommen und aus ihr n = 2 Einheiten entnommen. Wegen der Grenzwertsätze, ist sonst eine so detaillierte Kenntnis der GG natürlich nicht nötig ist. Aber im Falle eines convenience samples haben wir weit weniger als das: oft wird die GG noch nicht einmal genannt, von "Entnahme" ganz zu schweigen.

[44] Selbst dann, wenn man mehrere *verschiedene* Stichproben als "repräsentativ" zuließe, aber nichts darüber wüsste, mit welcher Wahrscheinlichkeit welche Stichprobe "gezogen" wird, hätte man noch keine Stichprobenverteilung), die sich im Übrigen *alle* aus einer GG zu ziehenden Stichproben des gleichen Umfangs n einbezieht. Wir haben – wie bereits gesagt – gegen dieses Prinzip verstoßen und spezielle ("partielle") Stichprobenverteilungen für "strukturgleiche" ("repräsentative") und "nicht-repräsentative" Stichproben bestimmt, aber dies geschah nur zu Demonstrationszwecken und wir haben sie auch nicht für Konfidenzintervalle oder Tests benutzt.

Beispiel 1: Wagschal et al.(2012), S. 214 stellten fest, dass "Sozialdemokratische Wohlfahrtsstaaten" (SW), worunter sie die fünf Länder Schweden, Dänemark, Norwegen, Finnland und die Niederlande verstanden, bei bestimmten Maßen der sozialen Gerechtigkeit (wie Verteilungs-, Bedarfsgerechtigkeit usw.) *signifikant* bessere (gerechtere) Werte aufweisen konnten als andere Ländergruppen, wie z.B. sechs "Liberale Wohlfahrtsstaaten" LW (Australien, Kanada, Irland, Neuseeland, UK und USA). Der Vergleich bezog sich primär auf Quoten (Anteile p), wie etwa der Anteil der Frauen im Parlament mit $p_{LW} = 0{,}210$, $p_{SW} = 0{,}406$, aber auch auf Gini-Koeffizienten und andere Statistiken.

Man fand (meist mit dem U Test von Mann und Whitney) heraus, dass 10 Unterschiede beim Vergleich zwischen sozialdemokratischen und liberalen Ländern "statistisch auf den 5%-Niveau signifikant" waren (S. 220). Es ist aber ziemlich offensichtlich, dass hier die fünf SW und die die sechs LM Länder keine Zufallsauswahl aus der GG aller "Sozialdemokratischer" oder "Liberaler" Staaten waren, sondern die GG selber darstellen. Jedenfalls wurde kein Versuch unternommen, weitere SW- bzw. LM-Länder zu benennen, die zu einer entsprechend größeren GG gehören würden.

Offenbar glaubte man hier Anteile p_{LW} und p_{SW} aus zwei unabhängigen Stichproben mit $n_{LW} = 6$ und $n_{SW} = 5$ (Anteile p oder $\hat{\pi}$ als Schätzwerte für die Anteile π in der GG) ermittelt zu haben und testete entsprechend die Hypothese H_0: $\pi_{SW} = \pi_{LW}$ (oder $\pi_{SW} - \pi_{LW} = 0$, daher der Ausdruck "Nullhypothese") gegen H_1: $\pi_{SW} \neq \pi_{LW}$.

Tatsächlich war aber offenbar $n_{LW} = N_{LW} = 6$ und $p_{LW} = \pi_{LW}$ und entsprechend $n_{SW} = N_{SW} = 5$, $p_{SW} = \pi_{SW}$ und auch $\pi_{SW} - \pi_{LW} = 0{,}406 - 0{,}210 = 0{,}196$. Was für einen Sinn macht es nun, die (bekanntermaßen falsche) Hypothese H_0: $\pi_{SW} - \pi_{LW} = 0$ zu testen? Das läuft darauf hinaus, dass "geprüft" wird, ob wir es mit *einer* GG gem. H_0 zu tun haben *könnten*, wo doch schon *feststeht*, dass dies nicht der Fall ist und H_1 ($\pi_{SW} - \pi_{LW} = 0{,}406 - 0{,}210 = 0{,}196 \neq 0$) gilt.

Natürlich kann man *ausrechnen*, wie wahrscheinlich eine Differenz $p_{SW} - p_{LW} = 0{,}196$ wäre *wenn* H_0 gelten würde, aber das Ergebnis ist aus mehreren Gründen von vornherein wertlos:

1. Wir haben es nicht mit der Zufallsvariable $p_{SW} - p_{LW}$ zu tun, sondern mit der per Totalerhebung festgestellte Differenz $\pi_{SW} - \pi_{LW} = 0{,}196$, die keine Zufallsvariable ist.
2. Wenn man schon Daten über alle Einheiten der GG hat gibt es eigentlich nichts mehr zu schätzen oder zu testen (schon gar nicht eine Hypothese, die im Gegensatz zu dem steht, was die vorhandenen Daten über die GG zeigen).
3. Man hat mit dem Ergebnis "signifikant" im Grunde festgestellt, dass die GG nur mit einer sehr geringen Wahrscheinlichkeit (p-Wert < 5% oder < 1%) ganz anders sein dürfte (nämlich mit $\pi_{SW} - \pi_{LW} = 0$) als sie es tatsächlich ist (nämlich $\pi_{SW} - \pi_{LW} = 0{,}196$). Obgleich man die GG im Einzelnen kennt, also *weiß*, dass H_0 falsch ist, hat man jetzt mit dem Test aufgrund einer Wahrscheinlichkeit nur Anlass zu *vermuten*, dass H_0 falsch ist. Was ist das für eine Erkenntnis: man weiß schon, dass H_0 definitiv *falsch ist* und stellt dann fest, dass die Wahrscheinlichkeit dafür spricht, dass H_0 *falsch sein könnte*?[45]
4. Der Fehler 1. Art (oder α-Fehler) besteht darin, die richtige H_0 nicht anzunehmen. In unserem Fall ist aber H_0 eine falsche Hypothese und eine falsche Hypothese nicht anzunehmen (zu verwerfen) ist kein Fehler.[46]

[45] Nach einem verbreiteten Missverständnis beweist "Signifikanz" dass die H_0 falsch ist. Das ist nicht richtig. "Signifikanz" zeigt normalerweise (außer im obigen Grenzfall einer Totalerhebung) nur, dass es Gründe gibt, H_0 für falsch zu *halten*.

[46] Mit anderen Worten, wir haben hier kein Entscheidungsproblem, bei dem man mit *zwei* Fehlern, dem α- und β-Fehler konfrontiert wird und es ist auch nicht angebracht, ein kleines Signifikanzniveau α zu wählen (z.B. 1% "hochsignifikant" statt nur 5%) und damit ein großes β und eine geringe "power" $1-\beta$ in Kauf zu nehmen.

5. Der Fehler 2. Art (β Fehler), H_0 annehmen (bei einem nichtsignifikanten Testergebnis) obgleich H_0 falsch ist oder gleichbedeutend H_1 ablehnen obgleich H_1 richtig ist) ist zwar ein Fehler, aber er sollte eigentlich mit einer Wahrscheinlichkeit von null auftreten.[47]
6. Die ganze Absurdität der unzulässigen Anwendung eines Tests zeigt sich aber, wenn das Ergebnis "nicht signifikant" ist und H_0 angenommen wird, was bei einem entsprechend kleinem n gar nicht so ausgeschlossen ist: man *weiß* dass H_0 nicht stimmt, weil ja gilt $\pi_{SW} - \pi_{LW} = 0{,}196$, und man erfährt nun, dass H_0 vielleicht doch stimmen *könnte*.

Fazit: Es gibt keinen Grund, Hypothesen über Parameter der GG, wie μ, ρ oder π (bzw. wie hier $\pi_1 - \pi_2$) aufzustellen und aufgrund von Wahrscheinlichkeiten über solche Hypothesen zu entscheiden, wenn die Daten keine Teilgesamtheit darstellen, sondern die GG selber vollständig beschreiben. Man stellt keine Vermutungen an über etwas, was einem bereits bekannt ist und man gewinnt nichts, wenn man erfährt, dass das, was definitiv als richtig *bekannt* ist, vielleicht auch mit einer gewissen Wahrscheinlichkeit nicht richtig sein *könnte*.

Beispiel 2: Jerger (2013) berechnete Korrelationen zwischen Indikatoren der wirtschaftlichen (x_t, BIP pro Kopf) und solchen der institutionellen (y_t, sog. transition indicators für Demokratisierung, Privatisierung etc.) Entwicklung. Es wurden "Korrelationskoeffizienten über die Zeit für jedes einzelne Land" berechnet (S. 132f.) also wohl mit Wertetupeln x_t,y_t $t = 1,\ldots,T$ für die $T = 21$ Jahre von 1990 bis 2010 gerechnet.[48] Andererseits wurden offenbar aber auch Korrelationen über n Länder bezogen jeweils auf ein Jahr t bestimmt, heißt es doch: "Die Korrelationskoeffizienten sind für jedes Jahr positiv und recht hoch, außer dem Wert für 1990 sind alle gezeigten Koeffizienten auch hochsignifikant. Die Insignifikanz des Werts für 1990" (als r seinerzeit etwa 0,4 betrug) sei u.a. der Tatsache geschuldet, dass "die Variabilität der transition indicators über die Länder (noch) recht gering ist" also σ_y klein ist (S. 131). Es wurde also "über die Länder in der Stichprobe" (S. 131) mit Wertetupeln x_i,y_i ($i = 1,\ldots, n$) gerechnet. Tatsächlich sind aber die $n = 26$ Länder von Albanien bis Weißrussland bereits die GG aller N osteuropäischen Länder, die hier Gegenstand der Untersuchung waren. Es gibt darüber hinaus keine weiteren Länder, die als "osteuropäische Länder" in Frage kämen.

Was besagt hier also ein Test der Hypothese $\rho = 0$ mit vermeintlichen Stichproben-Korrelationskoeffizienten r (die im Schnitt um etwa 0,7 lagen) wenn diese, in Wahrheit schon die GG-Korrelationskoeffizienten ρ waren?

Auch hier ist nicht erst die Berechnung einer Prüfgröße, sondern schon die Formulierung der Hypothese H_0 schlicht Unsinn. Die "Hypothese" kann nämlich nicht lauten H_0: $\rho = r = 0{,}7$; denn eine solche H_0 könnte nie als "signifikant" verworfen werden (die Teststatistik t wäre z.B. auch beim Test H_0: $\mu = \bar{x}$ Null, weil der Zähler $\bar{x} - \mu = 0$ ist), und man stellt ja keine Hypothese auf, die man nie verwerfen kann. Man prüft also H_0: $\rho \neq r$ oder speziell H_0: $\rho = 0$. Diese Hypothese ist aber nicht nur faktisch falsch, das Ergebnis "signifikant" ist auch nicht überraschend. Es besteht kein Grund, sich zu wundern (oder zu freuen), dass H_0 als "signifikant" verworfen wird, wenn die Korrelationen um 0,7 liegen, also hinreichend verschieden von Null sind und wenn nur $n = N$ hinreichend groß ist.

Es erscheint auch paradox, dass es bei einem solchen Test darauf ankommt, wie groß n ist, ob man die definitiv falsche H_0 $\rho = 0$ ablehnt und dass man bei kleinen Werten für $n = N$ die H_0 ($\rho = 0$) annehmen würde obgleich uns eigentlich doch schon bekannt ist, dass H_0 falsch ist,

[47] Bei der Abnahmeprüfung wäre dies das Käuferrisiko (eine schlechte Ware zu kaufen). Hier wird besonders gut deutlich, wie unsinnig im Grunde Tests bei Daten sind, die schon die komplette GG darstellen. Kennt der Kunde die GG (also alle Stücke einer Lieferung), dann gibt es kein Käuferrisiko (β wäre also in der Tat null); denn der Käufer könnte einfach die schlechten Stücke aussondern und nur die einwandfreien Waren annehmen.

[48] Dass "alle … Koeffizienten auch hochsignifikant" waren (S. 131) ist wenig überraschend; denn es gibt hier genug Gründe für einen gemeinsamen Trend von x und y (und damit für eine Scheinkorrelation).

also ρ nicht null ist.[49] Der t-Wert beim Test der Hypothese $\rho = 0$ berechnet sich mit $t = r\sqrt{n-2}/\sqrt{1-r^2}$ (Bortz, S. 199), d.h. man erreicht z.B. für $r = \pm 0{,}5$ den Wert $t = 2$ nicht, kann also die erwiesenermaßen falsche Hypothese $\rho = 0$ nicht verwerfen, solange $n < 14$ ist.[50]

Fazit: Jerger sah dass Korrelationen ρ, die er für r hielt, um 0,7 lagen und er stellte dann mit Signifikanztests fest, dass die Wahrscheinlichkeit nicht dafür sprach, dass sie $\rho = 0$ sind; ein zweifelhafter Erkenntnisgewinn, zugleich aber auch ein Zeichen, dass Methoden der "Induktiven Statistik" bei einer Totalerhebung fehl am Platz sind. Von einer GG kann man nicht mit der Wahrscheinlichkeitsrechnung auf eine GG schließen.

3.4. Die Grundgesamtheit wird nachgeliefert

Bei der "Repräsentativität" geht es um das Problem, eine Teilgesamtheit ("Stichprobe") zu definieren, die der GG möglichst ähnlich ist. Jetzt geht es umgekehrt darum, für eine vorgefundene (oder "anfallende" [Bortz]) Stichprobe die passende (möglichst ähnliche) GG zu definieren, wenn man überhaupt das Bedürfnis hat, ein paar Worte zu verlieren über die GG, für die die Stichprobe "Repräsentativität" beansprucht.

Das "Nachliefern" der GG (oder die "after the fact" Repräsentativität einer Stichprobe) ist eine Praxis, die offenbar in manchen Lehrbüchern aus den USA ernsthaft diskutiert wird.[51] Man hat Studenten *vorgefunden* (nicht zufällig gezogen) und befragt und sagt dann, dies sei repräsentativ für die Bevölkerung im Alter zwischen 20 und 25, vorwiegend aus der Mittel- und Oberschicht usw. Es wird also im Nachhinein die zur Stichprobe passende GG konstruiert und zwar so, dass die GG der Stichprobe ähnlich ist, weil ja andernfalls die Stichprobe nicht repräsentativ wäre.[52] Es ist klar, dass das "Nachliefern" der GG problematisch ist; denn

- es ist schwer vorstellbar, wie eine gegebene Stichprobe unter solchen Umständen jemals nicht "repräsentativ" sein könnte, wird doch die GG nachträglich so konstruiert, dass sie der Stichprobe ähnlich ist (womit dann auch Stichprobe der GG ähnlich also "repräsentativ" ist), anders gesagt: mit dem Nachliefern ist sichergestellt "dass jede Stichprobe für 'irgendeine' Grundgesamtheit repräsentativ" ist (Kutsch, S. 61)[53];
- so etwas kann es bei einer Zufallsauswahl nicht geben: man kann nicht aus den gezogenen Kugeln im Nachhinein wieder die Urne mit den gezogenen und den nichtgezogenen Kugeln machen;
- mit der Nachlieferung kommt man schnell in große Schwierigkeiten, wenn es gilt, konkret zu werden, in welchen Punkten die GG anders als die Stichprobe sein kann (man hat z.B. schon ein Problem, zu sagen, wie groß N ist, außer dass $N > n$ sein sollte); mehr noch: dann besteht – will man konsistent mit der "repräsentativen" Stichprobe sein – auch kein Grund eine Hypothese, nach der Stichprobe und GG verschieden sind zu formulieren und abzulehnen (aber auch H_0: $\mu = \bar{x}$ zu testen macht wenig Sinn);
- so manövriert man sich mit dem Nachliefern der GG unmerklich in die beschriebene Absurdität von Signifikanztest bei Totalerhebungen hinein, und schließlich

[49] Der Stichprobenumfang n spielt grundsätzlich eine Rolle für die Testentscheidung. Was speziell hier die Sache paradox macht, ist nur dass n = N ist und dass über eine bekanntermaßen falsche H_0 entschieden wird.

[50] Wir nehmen $t = 2$ als Faustregel für ein Signifikanzniveau von 5% und aus der Umformung der obigen Gleichung folgt $n = 4/r^2 - 2$. Danach muss z.B. auch bei $r = \pm 0{,}25$ $n \geq 62$ sein um H_0 verwerfen zu können.

[51] So bei Rebecca Warner und auch Timothy Urdan, die dem "representative sample" (was wir auch "Quotenverfahren" nennen) ähnliche Qualitäten zusprechen wie dem "random sample" (also der Zufallsauswahl).

[52] Das Nachliefern einer GG ist ein Vorgehen, das im Sinne von RS durchaus legitim ist. Es dürfte klar sein, dass das Denken in Kategorien wie "Repräsentativität" und "Strukturidentität" diesem Vorgehen Vorschub leistet.

[53] Nach Kutsch (S.61) wird die GG passend "zurecht gestutzt" wobei der Maßstab die "Struktur" im Sinne des RS-Konzepts ist. Es ist eigenartig, dass Kutsch dies "rein formal" für "nicht angreifbar", sondern nur für "nicht zum Ziel führend" hält.

- hängt auch hier wieder alles davon ab, wie genau man die "Struktur" von Stichprobe und GG beschreiben will, womit allerdings auch ein Dilemma verbunden ist.

Definiert man nämlich – wie oben – die GG als Bevölkerung im Alter zwischen 20 und 25 Jahren, vorwiegend aus der Mittel- und Oberschicht (weil dies für Studenten im Allgemeinen so zutrifft), dann fallen auch Nicht-Studenten darunter. Wird man aber so konkret, dass man die GG als "Studierende an deutschen Universitäten" definiert (wovor viele zurückschrecken dürften, denn was ist das schon für eine "Verallgemeinerung") dann sind die Hörer einer Anatomie-Vorlesung genauso "repräsentativ" wie die Hörer einer Marketing-Vorlesung und zwar egal um was es inhaltlich bei der empirischen Untersuchung geht. Weil das aber nicht akzeptabel ist, müsste man bei der Beschreibung der "nachgelieferten" GG noch detaillierter werden und würde schließlich bei der Gleichheit von Stichprobe und GG landen, was uns dann wieder auf die Situation von Abschn. 3.3 zurückwirft. Wir haben also ein Dilemma

- In den Abschnitten 3.1 und 3.2 wurde deutlich, dass die Stichprobe nicht nur anders als die GG sein darf, sie muss auch anders sein können. Es muss verschiedene Werte von $\bar{x}$ in verschiedenen Stichproben geben; denn sonst hätten wir keine Stichprobenverteilung mit $\sigma_{\bar{x}} > 0$ und es gäbe auch keine Basis für die Hypothese $\mu \neq \bar{x}$.
- Beim Nachliefern der GG geht es aber ganz im Gegensatz dazu darum, eine GG zu konstruieren, die möglichst wenig anders als die Stichprobe ist, um so die Stichprobe als "repräsentativ" deklarieren zu können.

Dieses Dilemma, in das man mit dem Denken in Kategorien, wie "Repräsentativität" und "Strukturidentität" gerät, macht klar, dass in Abwesenheit einer Zufallsstichprobe und einer klar zu benennenden GG, aus der die Stichprobe gezogen wurde, lediglich Methoden der "Deskriptiven Statistik" für die Auswertung der Daten angebracht sind (so wie dies auch bei Daten einer Totalerhebung der Fall ist) und inferenzstatistische Methoden fehl am Platz sind.

3.5. Für eine Grundgesamtheit ist nur eine Stichprobe repräsentativ

Es gibt noch weitere Unstimmigkeiten mit dem Konzept der "Repräsentativität", wenn man an Signifikanztests denkt, nämlich

- im *Ein-Stichprobentest* ist die Möglichkeit vorgesehen, dass ein und die gleiche Stichprobe $x_1, x_2, \ldots, x_n$ aus einer GG mit μ_0 stammen könnte (wenn also H_0: $\mu = \mu_0$ gilt), oder aber auch aus einer ganz anderen GG, gem. der Alternativhypothese H_1: $\mu \neq \mu_0$,
- beim Tests der Mittelwertdifferenz[54] mit *zwei unabhängigen Stichproben* können die unterschiedlichen Mittelwerte $\bar{x}_1$ und $\bar{x}_2$ nach H_0 aus der gleichen GG ($\mu_1 = \mu_2$), oder aber nach H_1 ($\mu_1 \neq \mu_2$) aus zwei verschiedenen Grundgesamtheiten (GGs) stammen (es kann ja auch dann wenn $\mu_1 = \mu_2$ ist eine Differenz $\bar{x}_1 - \bar{x}_2 \neq 0$ existieren).

Das sind deutlich zwei verschiedene Situationen. Man kann den Zwei-Stichprobentest nicht auf den Ein-Stichprobenfall reduzieren und etwa bei $\bar{x}_1 = 70$ und $\bar{x}_2 = 80$ die H_0: $\mu_2 = 70$ testen mit der Prüfgröße $\frac{\bar{x}_2 - \mu_2}{\hat{\sigma}_2}\sqrt{n_2} = \frac{10}{\hat{\sigma}_2}\sqrt{n_2}$, statt H_0: $\mu_1 - \mu_2 = 0$ zu testen mit der Prüfgröße $(\bar{x}_1 - \bar{x}_2) \Big/ \sqrt{\frac{\hat{\sigma}_1}{n_1} + \frac{\hat{\sigma}_2}{n_2}}$ weil nicht sicher ist, dass $\mu_1 = 70$ ist.

Im Falle von random samples hat man mit solchen Situationen konzeptionell keine Schwierigkeiten, weil sich GG und Stichprobe durchaus sehr unähnlich sein können, aber "Strukturidentität", wie immer man diese definiert, verträgt sich nicht gut damit, dass eine Stichprobe

[54] Oder einer Differenz von Anteilen π, wie oben im Beispiel der Arbeit von Wagschal et al. mit H_0: $\pi_1 - \pi_2 = 0$.

gleichermaßen repräsentativ für verschiedene Grundgesamtheiten (GGs) sein kann oder unterschiedlich strukturierte Stichproben gleich "repräsentativ" für ein und die gleiche GG.

Es ist – wie in Abschn. 3.1 schon gesagt – schwer vorstellbar, dass es verschiedene aber trotzdem gleich brauchbare Miniaturausgaben der gleichen GG gibt und wenn man eine "repräsentative Stichprobe" so konstruiert hat, dass sie "repräsentativ" ist für eine GG macht es eigentlich wenig Sinn, sich zu fragen, ob sie nicht auch repräsentativ für eine ganz anders strukturierte GG sein könnte. Umgekehrt: hat man eine Stichprobe und konstruiert man die dazu passende GG wird man *nur eine GG für diese eine Stichprobe* "nachliefern".

3.6. Wahrscheinlichkeit und einzelne Beobachtungen

Mit Vorstellungen wie "Repräsentativität" und "Strukturidentität" wird der Fokus geradezu abgelenkt von dem, was eine Zufallsauswahl ausmacht (und bei der Stichprobenverteilung vorausgesetzt wird), dass man nämlich statt der gezogenen Stichprobe genauso gut auch eine ganz andere Stichprobe aus der gleichen GG hätte ziehen können und diese – gerade wegen des Zufalls – nicht nur anders als andere Stichproben, sondern auch anders als die GG strukturiert sein darf.

In letzter Konsequenz beruht das Konzept der "Repräsentativität" von *jeweils einer konkreten gezogenen Stichprobe* auf einem Missverständnis: Die Wahrscheinlichkeitsrechnung macht nicht Aussagen über eine konkrete Stichprobe, sondern über *die Gesamtheit aller Stichproben*.[55] Das "klassische" Beispiel für eine Verkennung des Unterschieds dieser beiden Aussagen ist die "gamblers' fallacy", der Fehlschluss, dass eine Roulettekugel mit einer größeren Wahrscheinlichkeit auf rot (R) fällt, wenn sie in einer entsprechend langen Reihe (etwa 10-mal) hintereinander immer auf schwarz (S) gefallen ist.[56] Man schließt von der Struktur der GG (R und S beim Roulette) auf die Stichprobe von Würfen der Roulettekugel ("strukturidentisch" wären dann ganz streng genommen die Folgen SR oder RS, nicht aber SS oder RR).

Das hinter der "gamblers' fallacy" stehende Problem ist die *unzulässige Vorhersage des Eintretens eines (einzelnen) zufälligen* Ereignisses aufgrund eines vorangegangen Ablaufs. Denn es ist ja geradezu das Kennzeichen eines Zufallsvorgangs, dass man sein Ergebnis im Einzelfall nicht voraussagen kann, nicht angesichts einer Struktur der GG und auch nicht im Lichte einer gerade gemachten Erfahrung. Das ist ein Phänomen, das früher einige Philosophen als ein Paradoxon faszinierte: Zufall bedeutet, dass man gerade *nicht* etwas vorhersagen und berechnen kann, und trotzdem gibt es hier aufgrund der Wahrscheinlichkeits*rechnung* etwas zu *berechnen*. Es ist aber dann nicht mehr paradox, wenn man sieht, dass es nicht das Gleiche ist, um was es hier geht, *ein einzelne Ereignis* (wie die "Struktur" einer Stichprobe sein muss oder

[55] Wahrscheinlichkeitsaussagen betreffen nur Zufallsvariablen. Es ist deshalb auch nicht richtig zu sagen: μ liegt mit 95% Wahrscheinlichkeit zwischen … und …. (denn μ ist ja eine Konstante). Richtig wäre nur die Aussage: mit 95% Wahrscheinlichkeit sind die Grenzen des Konfidenzintervalls so, dass μ im Intervall zwischen diesen Grenzen liegt. Denn nur die Grenzen sind Zufallsvariablen, nicht aber μ.

[56] Auf der gleichen Ebene liegt auch das verbreitete Missverständnis, dass die Ablehnung einer Hypothese bei einem Signifikanztest so etwas sei, wie die von Karl Popper geforderte Falsifikation einer Hypothese. Eine Wahrscheinlichkeitsaussage kann nicht durch ein einziges entgegenstehendes Ereignis "falsifiziert" werden. Solche Missverständnisse erklären vielleicht auch das völlig überzogene Prestige, das statistische Tests genießen und die ebenso ungerechte Abwertung von "nur" deskriptiver Statistik. Besonders deutlich werden derartige Fehlvorstellungen über statistische Test bei T. Hillig 2006, der wie folgt zu einem Lob des convenience sample gelangt: "Der Wissenschaftstheorie Poppers folgend können Theorien nicht bewiesen, sondern lediglich falsifiziert werden. Demnach wäre die Verwendung von convenience samples durchaus wünschenswert, da eine Theorie, die für Konsumenten im Allgemeinen Gültigkeit haben soll, abgelehnt werden kann, falls sie für eine Untergruppe verworfen werden muss" (S. 127). Abgesehen vom Missverständnis, was das "Falsifizieren" betrifft, wird hier ein convenience sample ausgerechnet wegen der *Nicht*-Repräsentativität gelobt und die Wissenschaftstheorie geradezu pervertiert: man gewinnt Erkenntnisse durch Ablehnung von Hypothesen in statistischen Test. Das dürfte leicht sein, wenn nur die Hypothesen und die befragten "Untergruppen" exotisch genug sind.

wohin die Roulettekugel beim nächsten Wurf fallen wird), wo es nichts zu rechnen gibt, und Wahrscheinlichkeitsaussagen, die stets Aussagen über die *Gesamtheit aller möglicher Beobachtungen unter den gleichen Bedingungen* sind. Es ist sogar so, dass man *nur deshalb* mit Wahrscheinlichkeiten rechnen kann, *weil* der Vorgang (beim Roulette) *allein* vom Zufall bestimmt wird.[57] Es ist also der gleiche Grund, der dahinter steht, wenn man

- das Eintreten eines Ereignis E nicht vorhersagen kann,
- andererseits aber die Wahrscheinlichkeit P(E) von E sehr wohl berechnen kann.

Danach ist auch zu akzeptieren, was nach dem RS-Konzept befremdlich sein mag, dass nämlich wegen der *Zufalls*auswahl Stichproben aus der gleichen GG ganz unterschiedlich ausfallen können und damit auch deren mit Quoten beschriebenen Strukturen unterschiedlich sein können, ohne dass deshalb eine Stichprobe "repräsentativer" wäre als eine andere.[58] Und so gesehen ist es auch nicht überraschend, dass ein gleich großer Unterschied zwischen einer Stichprobengröße, wie $\bar{x}$ und dem entsprechenden Parameter μ der GG bei einem geringen Stichprobenumfang n nichtsignifikant, bei einem größeren Stichprobenumfang aber durchaus signifikant sein kann, was ja schon an der Prüfgröße

(5) $$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{x} - \mu}{\hat{\sigma}} \sqrt{n}$$ erkennbar ist.[59]

4. Praktische Fragen und Konsequenzen

Welche Folgerungen kann man aus alle dem für empirische Arbeiten ziehen? Zur Beantwortung dieser Frage ist es nützlich, einen Blick auf typische Themen und Ziele solcher Arbeiten zu werfen, denn davon hängt es ab, welche Anforderungen an die Befragten zu stellen sind und damit auch ob dann eine Stichprobenziehung überhaupt praktisch realisierbar ist. Wir betrachten hierzu (nur beispielhaft) zwei Arbeiten aus dem Marketing.

In Müller et al. (2010) ging es um die Ermittlung der Zahlungsbereitschaft für bestimmte Produkte (Schokoladenriegel) auf der Basis von hypothetischen Kauf- und Preisabfragen. Dafür war es wichtig, dass den Befragten das Produkt vertraut war und sie auch kauferfahren waren. Da es aber kein aktuelles und vollständiges Verzeichnis von Menschen mit einer solchen Qualifikation gibt, das als Auswahlrahmen (sampling frame) für eine Stichprobe hätte dienen können, war hier – wie auch sonst aus praktischen Gründen recht oft – eine Zufallsstichprobe nicht durchführbar. Die Datenerhebung bestand deshalb in einer Befragung "unter mehr als 500 Studenten auf dem Campus der Otto-von-Guerike-Universität Magdeburg" (S. 122). Weil Studenten bei den fraglichen Produkten zu den "typischen Konsumenten" rechnen dürften war "von einer hinreichenden Aussagekraft der erhobenen Zahlungsbereitschaften auszugehen" (S. 122). Zur Validierung fand "im Anschluss an hinreichend frequentierte Lehrveranstaltungen" eine weitere Befragung unter "mehr als 200 produktgruppenaffinen Studenten" statt (S. 123).

Gegenstand einer Arbeit von Strebinger et al. (2000) waren die meist sehr kompliziert zu operationalisierenden Bestimmungsgründe für den Erfolg von verschiedenen Arten der Produkt-

[57] Es hängt vom Zufall ab, was man *bei einer Stichprobe* für das arithmetische Mittel erhält, nicht aber was man hierfür *bei einer Totalerhebung* erhält (nämlich μ). So viel noch einmal zum Testen mit Daten einer Totalerhebung (es gibt dort nichts, was mit mehr oder weniger großer Wahrscheinlichkeit nur zu *erwarten* ist).

[58] Es ist bemerkenswert, dass die *Struktur* einer Stichprobe betreffende Forderungen in Barth et al. (2008) fehlen.

[59] Die Abhängigkeit einer Testentscheidung von n mag aus der Sicht anderer Entscheidungen ungewöhnlich und schwer verständlich sein, ist aber eine Konsequenz der Wahrscheinlichkeitsrechnung. Man begegnet diesem Problem z.B. oft bei Mietspiegeln, wenn sich herausstellt, dass ein Wohnungsmerkmal nicht signifikant ist und deshalb keinen Zu- oder Abschlag der Miete rechtfertigt, nur weil n bzw. die Anzahl der betroffenen Wohnungen zu klein ist und die Situation bei größerem n ganz anders wäre, obgleich sich dabei in der Sache nichts ändert. Da der Einfluss von Mietspiegeln zunehmen dürfte ist es nicht unwichtig zu erkennen, dass die Statistik hier schnell von einem Beschreibungs- zu einem Gestaltungsinstrument werden kann.

präsentationen. Zur Datenbasis heißt es: "Das Experiment fand in Form einer Shopping-Center-Befragung in einem großen Einkaufszentrum in der Nähe von Salzburg statt. Ziel war die Gewinnung eines allgemeinen Konsumentensamples, da ein zu hohes Bildungsniveau (z.B. durch ein Studentensample) die Aussagekraft der Ergebnisse für praktische Anwendungen der Conjointanalyse beeinträchtigt hätte" (S. 59).

Gemeinsam war den beiden empirischen Arbeiten, dass

- die Befragung sehr kompliziert war, weshalb auch nur bestimmte Versuchspersonen in Frage kamen und eine "echte" Zufalls-Stichprobe kaum möglich war, aber
- Überlegungen angestellt wurden, inwieweit die Befragten "typisch" sind (für wen?) und ob, bzw. unter welchen Voraussetzungen ihre Auskünfte "aussagefähig" sind,
- die GG, für die das Sample repräsentativ sein soll nicht explizit erwähnt wurde,[60] aber
- gleichwohl Gebrauch gemacht wurde von Signifikanztests.

Man kann sich nachgerade fragen: warum glaubt man trotz nicht erfüllter Voraussetzungen für inferenzstatistische Methoden, auf diese nicht verzichten zu können? Das ist umso erstaunlicher, als das Ziel dieser Methoden ja ist, zu prüfen, ob ein Stichprobenbefund auch für die GG, aus der die Stichprobe gezogen wurde zu verallgemeinern ist, hier aber die GG, auf die sich die Verallgemeinerung bezieht, mehr oder weniger im Dunklen bleibt.

Als Konsequenz aus unseren Überlegungen könnte man ein Umdenken weniger bei der Datengewinnung und mehr bei der Analyse der gewonnen Daten wünschen:

1. Die sehr anspruchsvollen Fragestellungen einschlägiger empirischer Studien erlauben oft keine Zufalls-Stichproben und der hierfür erforderliche Aufwand dürfte auch oft kaum zu rechtfertigen sein. Andere Erhebungsformen haben also durchaus auch weiter eine Existenzberechtigung. Das dürfte wohl für das in der Markt- und Meinungsforschung so beliebte Quotenverfahren[61] gelten, was allerdings genau nach dem – von uns kritisierten – RS-Konzept konzipiert ist, deutlich weniger aber wohl für das "convenience sample" (oder "accidental sample"), bei dem im besonderen Maße die bequeme Erreichbarkeit über die Zusammensetzung (von "Auswahl" kann man wohl nicht sprechen) entscheidet.
2. Es sollte aber mehr bedacht werden, dass immer dann, wenn von Zufallsauswahl keine Rede sein kann die Wahrscheinlichkeitsrechnung grundsätzlich nicht anwendbar ist und dass deshalb die Aussagefähigkeit von Konfidenzintervallen und Signifikanztests zumindest problematisiert werden sollte, wenn man schon meint, auf so etwas nicht verzichten zu können. Das bleibt auch dann richtig wenn entsprechende dubiose Rechnungen oft durchgeführt werden, so dass man sich stets bequem mit dem Argument exkulpieren kann, dass dies gängige Praxis sei.
3. Man sollte zumindest dann auf solche Berechnungen verzichten, wenn die "Stichprobe" in Wahrheit schon die interessierende GG ist. Obgleich dies eigentlich eine sehr leicht erkennbare Fehlanwendung der Wahrscheinlichkeitsrechnung mit geradezu absurden Implikationen ist, kommt so etwas gleichwohl erstaunlich oft vor.
4. Es ist die Frage, ob die Methode der "nachgelieferten" GG weniger bedenklich ist.

[60] Man könnte aus dem Ziel, ein "allgemeines" Konsumentensample zu realisieren folgern, dass die Zielgesamtheit die Menge aller Konsumenten war, nicht nur die Besucher eines Einkaufszentrums in Salzburg. Aber was heißt "alle"? Die Konsumenten in Österreich, in Europa oder weltweit?

[61] Man findet leider oft Darstellungen in denen der Unterschied zwischen der Quotenauswahl ("representative sample" bei Warner und Urdan, die diesem ähnliche Qualitäten zusprechen wie dem "random sample") und der geschichteten Stichprobe nicht gesehen wird. Wenn der Interviewer drei Beamte befragen muss, weil dies die Einhaltung der Quote erfordert, kann er der Einfachheit halber drei gute Bekannte befragen, die Beamte sind. Das kann eine Verzerrung, also einen systematischen Fehler unbekannter Größe bewirken Bei einer geschichteten Stichprobe müssen dagegen die zufällig gezogenen Personen interviewt werden.

5. Man sollte aufhören, eine Auswertung, die sich auf beschreibende Statistiken beschränkt als minderwertig zu empfinden. So etwas ist in jedem Fall besser als zu testen, ohne sich dabei zu fragen, was die Hypothesen und das Testergebnis eigentlich besagen.
4. Anwender der Statistik sollten aufhören, Worte wie "repräsentativ" und "signifikant" unüberlegt als Worthülsen zu gebrauchen und Statistiker sollten mehr mit den Anwendern kommunizieren und die Besonderheit der Zufallsauswahl besser herausarbeiten, die einen großen Vorteil bietet, der aber auch seinen Preis hat.

Der Vorteil einer Zufallsauswahl ist, dass nur bei ihr der Auswahlfehler, der bei jeder Teilerhebung (natürlich auch beim convenience sample) auftritt mit der Wahrscheinlichkeitsrechnung als *Stichproben*fehler (SF) bestimmt werden kann. Der Preis dafür ist, dass eine echte Zufallsauswahl oft schwer zu realisieren ist. Aber man kann diesen Vorteil einer Zufallsauswahl nicht ohne den Preis haben und der Vorteil wird nicht dadurch geringer, dass es auch andere Fehler, die "Nichtstichprobenfehler" (NSF) gibt, z.B. in Gestalt von non-response (eine Art von Selbstselektion), die übrigens nicht – wie die SF – mit größerem n abnehmen. Ein oft gebrachtes Argument ist, dass die Möglichkeit, den SF zu quantifizieren angesichts der NSF zu relativieren sei. Es wird auch oft gesagt, diese Fehler können quantitativ sogar bedeutsamer sein als der SF.[62] Dabei wird gerne übersehen, dass es selten gelingt, Zahlen zum NSF zu präsentieren und, dass es in jedem Fall besser ist nur den NSF nicht quantifizieren zu können als NSF *und* SF nicht quantifizieren zu können.

Wir können darauf hier nicht weiter eingehen. Das gilt auch für Versuche, den Mangel einer nichtzufälligen Auswahl (und damit den Mangel der Nichtanwendbarkeit der Wahrscheinlichkeitsrechnung) nachträglich durch Selektion[63] oder Gewichtung (redressment)[64] heilen zu wollen, damit dann Konfidenzintervalle und statistische Tests gerechtfertigt werden können. So etwas könnte nur erfolgreich sein, wenn man es der Auswahl ansähe, ob sie eine Zufallsauswahl ist oder nicht. Das ist aber nicht der Fall. Was eine Auswahl zu einer Zufallsauswahl macht, ist wie sie zustande gekommen ist.[65]

Literatur

Assaei Henry, John Keon, Nonsampling vs. Sampling Errors in Survey Research, Journal of Marketing, Vol. 62, No.2 (1982), pp. 114 -123

Barth Wolfgang, Eckart Bomsdorf, Uwe Kaletsch u. Angela Knickel, Die systematische Bestimmung von Mindeststichprobenumfängen bei quantitativen Werbewirkungstests im Direktmarketing. Marketing ZFP 30, H. 2 (2008), S. 69 - 76

Bortz Jürgen, Statistik für Sozialwissenschaftler, 4. Aufl. , Berlin etc. (Springer), 1993

Cochran William G., Sampling Techniques, 3rd. ed., New York (Wiley) 1977,

Gabler Siegfried und Andreas Quatember, Repräsentativität von Subgruppen bei geschichteten Zufallsstichproben, in: AStA, Wirtschafts- und Sozialstatisches Archiv, Bd. 7, H. 3-4, 2013, S. 105 – 119 (107).

Göritz Anja, Klaus Moser, Repräsentativität im Online-Panel, Der Markt

Hillig Thomas, Verfahrensvarianten der Conjoint-Analyse zur Prognose von Kaufentscheidungen, Eine Monte Carlo Simulation (Diss. TU Berlin 2004), Wiesbaden 2006

[62] Sehr ausführlich in Kutsch 2007, S. 248. Solche Betrachtungen sind selten empirisch. Nur unter sehr speziellen Bedingungen kann man Nichtstichprobenfehler quantifizieren und mit dem Stichprobenfehler vergleichen und wenn so etwas geschieht (z.B. bei Assaei und Keon) ist es sehr fraglich ob dies verallgemeinert werden darf.
[63] Kutsch 2007 stellt mir Recht fest "Es ist nicht möglich aus einer verzerrten Untermenge der interessierenden GG mit Hilfe der Zufallsauswahl eine bezüglich der GG unverzerrte Stichprobe zu gewinnen" (S. 74).
[64] Zu kritischen Einwänden gegen diese beliebte Methode vgl. Kutsch und vor allem Schnell.
[65] Das ist übrigens auch etwas, was bei allen Versuchen "Repräsentativität" zu definieren, immer vergessen wird.

Jerger Jürgen, Zur Akzeptanz politischer und marktwirtschaftlicher Reformen in Osteuropa: Empirische Befunde und Erklärungsansätze, in: Theresia Theurl (Hrsg.), Akzeptanzprobleme der Marktwirtschaften: Ursachen und wirtschaftspolitische Konsequenzen, Berlin (Duncker & Humblot) 2013, S. 121 – 145

Krämer Walter, The cult of Statistical significance, SFB Discussion Paper Nr. 44/2010

Kruskal William, Frederick Mosteller, Representative Sampling, I: Non-scientific Literature, part II: Scientific Literature Excluding Statistics, part III: The Current Statistical Literature, in: International Statistical Review, 47 (1979), 13 - 24, 111 - 127, 245 -265.

Kuß, Alfred, Marktforschung, Grundlagen der Datenerhebung und Datenansanalyse, 4. Aufl., Wiesbaden 2012

Kutsch Horst B., Repräsentativität in der Online-Marktforschung, Köln 2007

Levy Paul S., and Stanley Lemeshow, Sampling of Populations, Methods and Applications, New York (Wiley), 4th. ed. 2013

Mc Closkey Deirde N. and Stephen T. Ziliak.T., The Standard Error of Regression, Journal of Economic Literature vol. 34 (1996), No. 1, pp. 97 – 114

Müller Holger, Steffen Voigt und Bernd Erichson, Ermittlung von Zahlungsbereitschaften mittels monadischer Preis- und Kaufabfragen. Neue empirische Erkenntnisse, Marketing ZFP 32, H. 2 (2010), S. 117 - 127

Quatember Andreas, Das Problem mit dem Begriff Repräsentativität, Allgemeines Statistisches Archiv, Bd. 80 (1996), S. 236 – 241

Schnell Rainer, Die Homogenität sozialer Kategorien als Voraussetzung für "Repräsentativität" und Gewichtungsverfahren, Zeitschrift für Soziologie, 22/1 (1993), S. 16 - 32

Strebinger Andreas, Sabine Hoffmann, Günter Schweiger u. Thomas Otter, Zur Realitätsnähe der Conjointanalyse. Der Effekt von Präsentationsformat, Involvement und Hemisphärizität auf die subjektive Beurteilung der Aufgabe durch die Auskunftspersonen und die Vorhersagevalidität, Marketing ZFP 22, H. 1 (2000), S. 55 - 74

Urdan Timothy, Statistics in Plain English, 3rd. ed. New York 2012

von der Lippe Peter und Andreas Kladroba, Repräsentativität von Stichproben, in Marketing 24 (2002), S. 227 – 238

von der Lippe Peter: Wie groß muss meine Stichprobe sein, damit sie repräsentativ ist? Diskussionsbeiträge aus dem FB Wirtschaftswissenschaften Univ. Duisburg-Essen (Campus Essen), Heft 187, Februar 2011.

Wagschal Uwe, Frieder Neumann u. Sebastian Jäckle, Gerechtigkeit und Marktwirtschaft in der OECD – ein Benchmarkvergleich, in: Viktor Vanberg (Hrsg.), Marktwirtschaft und soziale Gerechtigkeit, Tübingen (Mohr/Siebeck) 2012, S. 195 - 229

Warner Rebecca, Applied Statistics. From Bivariate Through Multivariate Techniques, Los Angeles (Sage Publications) 2012