UNIVERSITÄT POTSDAM

Wirtschafts- und Sozialwissenschaftliche Fakultät

Hans Gerhard Strohe (Hrsg.)

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 30

Monique Newiak

Prüfungsurteile mit Dollar Unit Sampling

Ein Vergleich von Fehlerschätzmethoden für Zwecke der Wirtschaftsprüfung: Praxis, Theorie, Simulation

Potsdam 2009

ISSN 0949-068X

STATISTISCHE DISKUSSIONSBEITRÄGE

Nr. 30

Monique Newiak

Prüfungsurteile mit Dollar Unit Sampling

Ein Vergleich von Fehlerschätzmethoden für Zwecke der Wirtschaftsprüfung: Praxis, Theorie, Simulation

Herausgeber: Prof. Dr. Hans Gerhard Strohe, Lehrstuhl für Statistik und Ökonometrie
Wirtschafts- und Sozialwissenschaftliche Fakultät der Universität Potsdam
Postfach 90 03 27, D-14439 Potsdam
Tel. +49 (0) 331 977-3225
Fax. +49 (0) 331 977-3210
Email : strohe@uni-potsdam.de
2009, ISSN 0949-068X

Inhaltsverzeichnis

Symbolverzeichnis

α	Irrtumswahrscheinlichkeit
α_i	Anteil an Fehleranteilen der Größe i
bw_i	Buchwert des i-ten Objekts
BW_i	Kumulierter Buchwert
C	Obere Fehlergrenze
G	Maß für die Genauigkeit in der Simulation
I_i	Stichprobenintervall
$\overline{I}$	Durchschnittliches Stichprobenintervall
$\lambda = np$	Parameter der Poisson-Verteilung, erwartete Fehleranzahl, „Fehlerintensität“
m	Mittlerer Fehleranteil in der Grundgesamtheit
n	Umfang der Stichprobe
N	Umfang der Grundgesamtheit (Anzahl der potentiell prüfbaren Objekte)
$p = \frac{X}{N}$	Fehlerrate
π_i	Wahrscheinlichkeit, dass die Zufallsvariable den j-ten Zustand annimmt
$\overline{r}$	Durchschnittlicher Fehleranteil in der Stichprobe, Schätzer für m
Q	Anzahl gezogener Stichproben in einem Simulationsdurchgang
RN_j	j-tes nicht zentriertes Moment der Verteilung der Fehlerrate
S	Standardfehler
T	Relative Trefferhäufigkeit in der Simulation
θ	Parameter für den gesamten monetären Fehler im Quasi-Bayesian-Bound-Ansatz
TN_j	j-tes nicht zentriertes Moment der Fehleranteilsverteilung

UC_j	j-tes zentrales Moment der Stichprobenverteilung des mittleren Fehlers
UN_j	j-tes nicht zentriertes Moment der Stichprobenverteilung des mittleren Fehlers
V	Maß für die Vorteilhaftigkeit einer Hochrechnungsmethode gegenüber einer anderen in der Simulation
x	Anzahl der in der Stichprobe gefundenen Fehler
x_i	Im Quasi-Bayesian-Bound-Ansatz: Anzahl der monetären Einheiten, die durch den Fehleranteil t_i charakterisiert sind
X	Anzahl nicht korrekt ausgewiesener Elemente in der Grundgesamtheit
Z_α	α-Quantil der Normalverteilung
z_i	i-te Ausprägung des Sets S
ZF	Zufallszahl

Abkürzungsverzeichnis

ACL	Audit Control List, eine Prüfungssoftware
AICPA	American Institute of Certified Public Accountants
BP	Basic Precision
BW	Buchwert
CAV	Combined Attributes Variables
CMA	Cumulative Monetary Amount
CIPFA	The Chartered Institute of Public Finance and Accountancy
CPA	Certified Public Accountant
DUS	Dollar Unit Sampling
HGB	Handelsgesetzbuch
IDEA	Integrated Data Evaluation and Analysis System, eine Prüfungssoftware
IDW	Institut der Wirtschaftsprüfer
ISA	International Standards on Auditing
LS	Load and Spread
MLE	Most Likely Error
MUS	Monetary Unit Sampling
PGW	Precision Gap Widening
PPS	Probability Proportional to Size
PS	Prüfungsstandard
SAS	Statements on Auditing Standards
SS	Single Spread
UEL	Upper Error Limit

1 Einleitung

Seit dem Ende der 1980er Jahre ist es in Bezug auf die Beschreibung und Diskussion statistischer Verfahren für den Bereich der Wirtschaftsprüfung in der Literatur relativ still geworden. Diese Stille wiegt uns in dem angenehmen Glauben, dass die Verfahren, die durch ihre Anwendung in Prüfungsprozessen letztlich dem Schutz unserer wirtschaftlichen Ordnung dienen sollen, vollkommen ausgereift sind, verlässliche Projektionen von in Stichproben gefundenen Fehlern bieten und von den Trägern der „Vorbehaltsaufgabe Jahresabschlussprüfung“[1] sinnvoll eingesetzt werden.

In der Tat beginnt die Geschichte des *Dollar Unit Samplings*[2] – des Verfahrens, welchem bei der Wahl eines statistischen Stichprobenverfahrens heute in vier von fünf Fällen der Vorzug gegeben wird – sehr vielversprechend. Nachdem 1961 der erste Aufsatz zum *Dollar Unit Sampling* veröffentlicht wurde,[3] stieß dieses auf einer größenproportionalen Zufallsauswahl basierende Verfahren in Prüfungen schnell auf Akzeptanz.[4] Nicht ohne Grund: Für die Hochrechnung des Ausmaßes des monetären Fehlers in Grundgesamtheiten aus einer geprüften Stichprobe führen traditionelle Hochrechnungsverfahren wie die Mittelwert-, Differenzen- oder Verhältnisschätzung aufgrund zu hoher Konfidenzintervalle,[5] für Prüfungszwecke ungeeigneter Schlussfolgerungen im Fall der Nullfehlerstichprobe[6] und hoher Stichprobenumfänge nur zu unbefriedigenden Ergebnissen.[7] Das *Dollar Unit Sampling* wird dagegen als eine elegante Alternative gesehen, da es die homograde Fragestellung (die Frage nach der Fehlerrate oder Anzahl) und die heterograde Fragestellung (die Frage nach dem Wert des Fehlers) miteinander verbindet,[8] durch eine maximale Schichtung der Grundgesamtheit zu mehr Prüfungseffizienz

[1] Nach §319 Abs. 1 HGB (Handelsgesetzbuch) ist die Prüfung des handelsrechtlichen Jahresabschlusses bestimmten Berufsgruppen vorbehalten.

[2] Auch: *Monetary Unit Sampling, Cumulative Amount Sampling, Probability Proportional to size Sampling.*

[3] Vgl. van Heerden (1961).

[4] Vgl. Kaplan (1975), S. 126.

[5] Vgl. Lillestol (1981), S. 263.

[6] Vgl. Neter (1978), S. 77.

[7] Vgl. Reneau (1978), S. 670.

[8] Vgl. Eggenberger (1986), S.84.

führt[9], besonders für kleine Fehlerraten und große Fehlerbeträge geeignet ist[10] und ohne Verteilungsannahmen auskommt.[11]

Jedoch wurde an der von Stringer (1963) und Stephan (1963) für das *Dollar Unit Sampling* ursprünglich vorgestellten Fehlerhochrechnungsmethode kritisiert, dass die durch sie ausgegebenen Konfidenzgrenzen zu konservativ sind.[12] Dies motivierte in den 1970er und 1980er Jahren eine Reihe von Veröffentlichungen, die alternative Möglichkeiten der Hochrechnung anpriesen,[13] von denen sich einige wenige in der Praxis durchgesetzt haben.

In der vorliegenden Arbeit soll die Betrachtung des *Dollar Unit Samplings* aufgrund seiner Relevanz für die Jahresabschlussprüfung wieder aufgenommen werden. Neben einer Beschreibung des Verfahrens möchte ich insbesondere auf die Schwachstellen der angewandten Hochrechnungsmethoden eingehen, und zeigen, mit welchen Konsequenzen die Verwendung einzelner Methoden verbunden ist.

Um das zu beschreibende Verfahren in seinen Kontext zu stellen, gebe ich zuerst einen Überblick über die Grundzüge der Stichprobenprüfung für Zwecke der Wirtschaftsprüfung (Abschnitt 2), wobei mein Fokus hier auf den rechtlichen Grundlagen und der aktuellen Prüfungspraxis liegen soll. Nach einigen notwendigen Begriffsbestimmungen wird das *Dollar Unit Sampling* der chronologischen Abfolge der Stichprobenziehung entsprechend vorgestellt. Zuerst werden die einzelnen Möglichkeiten der Stichprobenziehung im *Dollar Unit Sampling* erläutert (Abschnitt 3). Darauf folgt eine Beschreibung der Fehlerhochrechnung unterteilt in Schätzung der Fehlerrate (Abschnitt 4) und des Fehleranteils (Abschnitt 5). Vor dem Hintergrund dessen, dass jede der dargestellten Methoden letztlich nur eine Näherungslösung für die Hochrechnung des Fehlers bietet und sich die vorgestellten Methoden stark in Bezug auf die ausgegebenen oberen Fehlergrenzen unterscheiden, werden in Abschnitt 6 drei praktisch relevante Methoden jede für sich und im Vergleich zueinander einer Untersuchung auf

[9] Vgl. Steele (1992), S. 152, Apostpolou/Allemann (1991), S. 67.

[10] Vgl. Guy (2002), S. 195.

[11] Vgl. Biaggio (1987), S. 221.

[12] Vgl. zum Beispiel Leitch/Neter/Plante/Sinha (1982), S. 384, vgl. Rohrbach (1986), S. 128.

[13] Vgl. zum Beispiel Neter (1978), McCray (1984), Dworin (1984).

Verlässlichkeit und Genauigkeit anhand von 60.000 simulierten Stichproben unterzogen. Ich schließe meine Arbeit mit einer Zusammenfassung der Ergebnisse und einem Ausblick ab.

2 Grundsätzliches zu Stichprobenprüfungen für Zwecke der Jahresabschlussprüfung

2.1 Notwendigkeit von Stichprobenprüfungen für die Jahresabschlussprüfung, rechtliche Grundlagen

Bei der Prüfung des handelsrechtlichen Jahresabschlusses kann aus Effizienzgründen grundsätzlich keine Vollprüfung erfolgen.[14] Die Aufgabe des Abschlussprüfers besteht nicht darin, jeden einzelnen Geschäftsvorfall des vorliegenden Geschäftsjahres auf richtige Erfassung zu überprüfen, sondern vielmehr sicherzustellen, dass in dem veröffentlichten Jahresabschluss mit großer Wahrscheinlichkeit keine wesentlichen Fehler enthalten sind.[15] Um diese Freiheit von wesentlichen Fehlern mit hinreichender Sicherheit bescheinigen zu können, bedient sich der Prüfer unterschiedlicher Prüfungstechniken. Dabei lassen sich seine Prüfungshandlungen im Rahmen der Ergebnisprüfung grundsätzlich in analytische Prüfungshandlungen und in die Prüfung von Einzelnachweisen aufteilen.[16] Bei den Einzelprüfungshandlungen nimmt das Ziehen von Stichproben eine besondere Rolle ein.[17]

Nach SAS 39[18] kann der Prüfer für Zwecke der Stichprobenprüfung sowohl auf mathematisch-statistische als auch auf Urteilsstichproben[19] zurückgreifen, um durch sie entweder ein repräsentatives Bild von der Grundgesamtheit zu erhalten, oder – durch die

[14] Vgl. Marten (2003), S. 444.

[15] §264 Abs. 2 i. V. m. §317 Abs. 1 HGB, oder auch IDW PS 200.8-10 (Prüfungsstandard des Instituts der Wirtschaftsprüfer).

[16] Vgl. Marten/Quick/Ruhnke (2003), S. 240.

[17] Vgl. Barron/Groomer/Swink (1998), S. 1008.

[18] SAS steht für *Statements on Auditing Standards*. Es handelt sich um eine Prüfungsnorm des Amerikanischen Instituts der Wirtschaftsprüfer.

[19] Im statistischen Sinne handelt es sich bei der Urteilsstichprobe um keine echte Stichprobe, da dem Gedanken der repräsentativen Auswahl bei ihren Anwendung nicht Rechnung getragen wird.

Urteilsstichprobe – die Aussage zu treffen, dass neben den gefundenen Fehlern wahrscheinlich keine weiteren Fehler in der Grundgesamtheit vorliegen.[20]

Auch nach ISA 530[21] ist sowohl die statistische als auch die nicht-statistische Stichprobenziehung zulässig (ISA 530, 3). Dabei gilt als statistisches Verfahren ein solches, welches über die Merkmale der zufälligen Auswahl der Elemente in die Stichprobe und die Verwendung der Wahrscheinlichkeitstheorie für die Stichprobenauswertung verfügt (ISA 530, 10). Die Entscheidung, welches Verfahren anzuwenden ist, liegt im Ermessen des Abschlussprüfers (ISA 530, 28). Entscheidet sich der Prüfer für ein statistisches Verfahren, ist der Stichprobenumfang wiederum eine Ermessensfrage (ISA 530, 29). In dem internationalen Standard werden die Zufallsauswahl, die geschichtete Auswahl sowie die „nach Wert gewichtete Auswahl“ (*value weighted selection*) vorgestellt. Damit eröffnet ISA 530, 39 die Möglichkeit der Anwendung des *Dollar Unit Samplings* für die Zwecke der Erlangung von Prüfungsnachweisen. Es wird durch den Standard dabei ausdrücklich darauf verwiesen, dass sich dieses Verfahren als vorteilhaft erweist, wenn in dem entsprechenden Prüffeld vorrangig nach Überbewertungen gesucht wird.

2.2 Ein Überblick über Stichprobenprüfungen in der Praxis der Jahresabschlussprüfung

Obwohl eine Aussage über die Grundgesamtheit aufgrund einer Stichprobe nur möglich ist, wenn die Stichprobe nach mathematisch-statistischen Methoden gezogen wird, willkürliche Stichprobenziehungen regelmäßig zu verzerrten Schlussfolgerungen führen,[22] und für die Ziehung von Urteilsstichproben das prüfende Personal über einen sehr hohen Erfahrungsschatz verfügen muss,[23] spielen die mathematisch-statistischen Verfahren in der Praxis nur eine untergeordnete Rolle. In der Studie von Hall/Hunton/Pierce (2002) geben nur 15 % der befragten Prüfer an, überhaupt mathematisch-statistische Verfahren zur Bildung ihrer Stichprobe zu verwenden.[24] Als Begründung für das Nichtanwen-

[20] Vgl. Hitzig (2004b), S. 30.

[21] ISA steht für *International Standards on Auditing*.

[22] Vgl. Hall/Herron/Pierce (2006), S. 26 f. sowie Hall/Herran/Pierce/Witt (2001), S. 10.

[23] Hauptfachausschuss des Instituts der Wirtschaftsprüfer (1988), S. 7.

[24] Weiterhin wird von Matthews (2006), S. 86, festgestellt, dass viele Prüfer angeben, mathematisch-statistische Verfahren zu verwenden, wenn sie eigentlich nur die Auswahl der Elemente nach

den von mathematisch-statistischen Verfahren werden folgende Punkte angebracht:

- Es besteht die Auffassung, dass durch willkürliche Ziehung von Elementen für die Richtigkeit des Jahresabschlusses *wesentliche* Objekte näher untersucht werden können und durch mathematisch-statistische Verfahren die prüferische Urteilskraft untergraben wird.[25] In Gesprächen mit Vertretern zweier Wirtschaftsprüfungsunternehmen wurde mir bestätigt, dass willkürlichen Stichproben und Urteilsstichproben in Prüfungen meist Vorrang gegeben wird. Von vier von mir angesprochenen Wirtschaftsprüfungsunternehmen gab ledigliches eines an, mathematisch-statistische Verfahren grundsätzlich bei jeder Prüfung zu verwenden.

- Der mit Hilfe mathematisch-statistischer Verfahren ausgegebene Stichprobenumfang wird als zu hoch erachtet.[26] Da im Rahmen einer Urteilsstichprobe (im Gegensatz zu einer repräsentativen Auswahl) nur die Elemente geprüft werden müssen, die als risikobehaftet eingestuft wurden, sind die Stichprobenumfänge dort meist geringer. Die Irrtumswahrscheinlichkeit ist für die Urteilsstichprobe natürlich nicht objektiv quantifizierbar.
  Einer der Grundsätze für Abschlussprüfungen ist der der Wirtschaftlichkeit.[27] So soll mit minimalem Aufwand eine Aussage zur Richtigkeit der Angaben im Jahresabschluss gemacht werden. Zusätzlich sehen sich die Wirtschaftsprüfungsunternehmen immer mehr einem Budgetdruck[28] ausgesetzt, der zu Programmen mit Bezeichnungen wie *„Lean Audit“*[29] führt. Hier geraten Einzelfallprüfungshandlungen im Allgemeinen und mathematisch-statistische Verfahren im Besonderen in den Hintergrund.

- Es fehlt die Kenntnis über die Anwendung von mathematisch-statischen Verfahren im Prüfungsbereich.[30] Die Methoden erscheinen vielen Prüfern undurchsichtig

---

mathematisch-statistischen Prinzipien durchführen.

[25] Vgl. Guy/Carmichael/Whittington (2002), S. 6 oder auch Hitzig (2004), S. 35, der diesen Bedenken mit dem Argument der Beweiskraft von mathematisch-statistischen Verfahren in Haftungsfällen begegnet.

[26] Vgl. Hitzig (2004b), S. 31.

[27] Vgl. Lück/Lexer (2004), S. 270.

[28] Vgl. Lindgens (1999), S. 169, Braun (1996), S. 999.

[29] „Die schlanke Prüfung“.

[30] Vgl. Matthews (2006), S. 86.

und zu schwer zu handhaben. So wurde mir in einer Wirtschaftsprüfungsgesellschaft von der für die statistischen Schulungen verantwortlichen Mitarbeiterin erklärt, dass die statistischen Verfahren von den Mitarbeitern als *„Blackbox“* – d.h. ohne Kenntnis über die theoretischen Grundlagen – angewandt werden.

Wird trotz der beschriebenen Probleme ein mathematisch-statistisches Verfahren für die Prüfung eingesetzt, so ist es in 80 % der Fälle das *Dollar Unit Sampling*.[31]

2.3 Der Begriff des Fehlers

Laut IDW PS[32] 210.7 und ISA 240.3 stellen Fehler oder Unrichtigkeiten unbeabsichtigte Falschaussagen im Abschluss oder Lagebericht eines Unternehmens dar. Als Beispiele werden Schreib- und Rechenfehler, die irrtümlich falsche Anwendung von Rechnungslegungsnormen und unbeabsichtigte Fehlbuchungen genannt.

Im Rahmen des messtheoretischen Ansatzes[33] werden die Merkmalsausprägungen eines Prüfungsgegenstandes (etwa: eines Beleges oder eines Vertrages) für einen Sollzustand aufgrund von Rechnungslegungsvorschriften und Prüfungsnormen abgeleitet und der Ist-Zustand des Objektes festgestellt. Letztendlich werden Ist- und Soll-Zustand aufgrund der relevanten Merkmalsausprägungen miteinander verglichen. Geht es um den Vergleich der wertmäßigen Erfassung im Rahmen des Soll-Ist-Vergleiches, wird von dem Messwert des Prüfungsgegenstandes gesprochen.

Während die rechtlichen Grundlagen die Ungewolltheit der Fehler betonen, geht es in dem messtheoretischen Ansatz lediglich um die Abweichung von Soll- und Ist-Zustand, unabhängig von der Entstehungsursache, was für die Definition für den Fehler in der vorliegenden Arbeit übernommen wird. Ich schränke die Abweichung weiter auf monetäre Fehler, die Einfluss auf die Vermögens- oder Einkommensdarstellung des Unternehmens haben, ein. Das bedeutet, dass reine Zuordnungsfehler, die zwar Auswirkungen auf die einzelne Buchung oder das einzelne Konto haben können, außer Acht bleiben, wenn Sie letztendlich das Vermögen oder das Einkommen unverändert lassen

[31] Vgl. Hall/Hunton/Pierce (2002), S. 129.

[32] Prüfungsstandard des Instituts der Wirtschaftprüfer.

[33] Vgl. von Wysocki (2002), Sp. 1886-1889.

(bspw. Buchung eines Umsatzerlöses auf das Konto „sonstige Erlöse“), wogegen Fehler in der Rechnungsabgrenzung in die Definition des Fehlers fallen würden, da sie das Einkommen des betrachteten Geschäftsjahres beeinflussen.[34]

Somit definiere ich den Fehler als den Wert der Über- oder Unterbewertung von Einkommen oder Vermögen, ungeachtet des Entstehungsgrundes.[35]

Weitere Begrifflichkeiten, die für die Arbeit definiert werden müssen, betreffen die Klassifizierungen von Fehlern nach der Kenntnis des Prüfers: Es lassen sich bekannte (*known*), wahrscheinliche (*most likely*) und mögliche (*possible*) Fehler unterscheiden.[36] Die bekannten Fehler sind dabei Abweichungen eines monetären Betrages im Jahresabschluss, die dem Prüfer tatsächlich bekannt sind, d. h. von diesem entdeckt werden. Als *most likely error* (MLE) wird der durch die Stichprobe ermittelte Punktschätzer des wahren Fehlers in der Grundgesamtheit bei repräsentativer Auswahl verstanden. Mögliche Fehler sind dagegen Fehler, die, auch wenn sie nicht beobachtet wurden und unwahrscheinlich sind, nicht vollständig aufgrund der Prüfungshandlungen ausgeschlossen werden können.

Als tolerierbarer Fehler wird der maximale Gesamtfehler verstanden, bei dem ein den wirtschaftlichen Verhältnissen entsprechender Jahresabschluss noch testiert werden kann. Er entspricht der Wesentlichkeitsgrenze (*Materiality*), die vom Prüfer bei jeder Prüfung festgelegt wird.[37]

Speziell für die Beschreibung des *Dollar Unit Samplings* ist die Abgrenzung zweier weiterer Begriffe notwendig. Wird von der Fehlerrate[38] p gesprochen, so ist das Verhältnis der Anzahl fehlerhafter Elemente zur Anzahl sich in der Grundgesamtheit befindender Elemente gemeint:

$$p = \frac{X}{N}.$$

[34]Zu Bilanzierungsvorschriften vgl. zum Beispiel Coenenberg (2003).

[35]In Anlehnung an Leslie/Teitlebaum/Anderson (1979), S. 163.

[36]Vgl. Wolz (2003), S. 172.

[37]Zur Festlegung der Wesentlichkeitsgrenze vgl. zum Beispiel Barron/Groomer/Swink (1998).

[38]Der Begriff *error rate* ist der Literatur zur Statistik in der Wirtschaftsprüfung üblich und wird deshalb anstelle des statistisch üblichen Begriffs von Anteil oder Quote verwendet.

Hier ist X die Anzahl fehlerhafter Elemente in der Grundgesamtheit und N der Umfang der Grundgesamtheit. Der Fehleranteil (*„Error Tainting“*[39]) gibt dagegen das Verhältnis von der Differenz aus Ist- und Soll-Buchwert zum Ist-Buchwert[40] an:

$$r = \frac{bw_{ist} - bw_{soll}}{bw_{ist}}.$$

2.4 Das Verständnis des Risikos

Der Schluss über den Zustand der Grundgesamtheit aufgrund einer Stichprobenprüfung ist mit zwei Arten von Risiken der Fehlbeurteilung verbunden. Auf der einen Seite steht das α-Risiko bzw. der Fehler erster Art, welcher das Risiko angibt, dass ein in Wirklichkeit ordnungsmäßiges Prüffeld als nicht ordnungsgemäß eingeschätzt wird. Dieses wird auch als das Risiko des Auftraggebers bezeichnet. Auf der anderen Seite steht das β-Risiko (Fehler 2. Art). Hierunter wird das Risiko des Prüfers verstanden, einem Prüffeld Ordnungsmäßigkeit zu bescheinigen, obwohl es in Wirklichkeit nicht frei von wesentlichen Fehlern ist.[41]

Da in der Statistik das Stichprobenrisiko bzw. die Irrtumswahrscheinlichkeit üblicherweise mit dem Buchstaben α verbunden wird, möchte ich mich von den oben genannten Bezeichnungen für diese Arbeit jedoch lösen, und, wie es in der Literatur zu statistischen Prüfverfahren allgemein üblich ist, das Risiko der fehlerhaften Annahme eines nicht ordnungsgemäßen Prüffeldes mit α bezeichnen.

---

[39] Vgl. Martel-Escobar/Vazquez-Polo/Hernandez-Bastida (2005), S. 795.

[40] Wir beobachten in der Grundgesamtheit den Ist-Buchwert, welcher Fehler enthalten kann. Der Soll-Buchwert ist dagegen der Wert den wir als Idealwert für das entsprechende Objekt annehmen.

[41] Vgl. *International Auditing and Assurance Standards Board* (2007), S. 8 f., oder auch Hitzig (2004b), S. 31.

3 Die Stichprobenbildung im Dollar-Unit-Sampling-Verfahren

3.1 Der Grundgedanke der Auswahl im Dollar-Unit-Sampling-Verfahren

In Prüfungen mit Hilfe traditioneller Stichprobenverfahren[42] wird die Grundgesamtheit als Gesamtheit von Daten, aus der eine Stichprobe zum Zwecke der Erlangung von Schlussfolgerungen gezogen wird, verstanden.[43] Im Einzelfall kann es sich bei der Grundgesamtheit somit je nach Festlegung des Prüfungsobjektes um alle zu prüfenden Konten oder die Gesamtheit von Buchungen auf einem Konto handeln. Der Umfang der Grundgesamtheit entspricht dann der Anzahl der Konten in einem Jahresabschluss oder der Anzahl der Buchungen auf einem Konto.

Das *Dollar Unit Sampling*[44] löst sich von dieser Vorstellung des Umfangs der Grundgesamtheit als Anzahl von Prüfungsobjekten und definiert ihn als Gesamtbuchwert.[45] Der Grundgedanke der repräsentativen Auswahl – jedes Element hat die gleiche, von Null verschiedene Wahrscheinlichkeit, in die Stichprobe zu gelangen – wird umdefiniert zu: „Jede Geldeinheit hat die gleiche, von Null verschiedene Wahrscheinlichkeit, in die Stichprobe gezogen zu werden." Letztendlich wird damit die Wahrscheinlichkeit eines Prüfungsobjektes davon abhängig gemacht, mit welchem monetären Wert es ausgewiesen ist. Diese Eigenschaft des Verfahrens wird auch als maximale Schichtung der Grundgesamtheit interpretiert.[46]

Konkret bedeutet das: Besteht die Grundgesamtheit aus N Prüfungsobjekten, die in Summe mit einem Wert BW ausgewiesen sind, ist der Umfang der Grundgesamtheit

[42] Unter traditionelle Stichprobenverfahren werden Mittelwert-, Differenzen- und Verhältnisschätzungen subsumiert.

[43] Vgl. ISA 530, 6.

[44] DUS, auch: *Probability-Proportional-to-Size* (PPS), *Combined Attributes Variables* (CAV), *Cumulative Monetary Amount* (CMA) oder *Monetary Unit Sampling* (MUS).

[45] Vgl. Kaplan (1975), S. 126; ISA 530, 42.

[46] Vgl. Steele (1992), S. 152.; Hall/Pierce/Ross (1989), S. 65.

beim *Dollar Unit Sampling* definiert als BW (Anzahl der monetären Einheiten, der „Dollar“, die den ausgewiesenen Gesamtwert aller Prüfungsobjekte ausmachen, kurz: Gesamtbuchwert).[47] Die Wahrscheinlichkeit eines Prüfungsobjekts, in die Stichprobe zu gelangen, ist proportional zu dessen Buchwert.

Der Stichprobenumfang n ist die Anzahl der in die Stichprobe ausgewählten Geldeinheiten. Jede dieser Geldeinheiten gehört jedoch zu einem der N Prüfungsobjekte, sodass zu jeder ausgewählten monetären Einheit ein (reales) Prüfungsobjekt ausgewählt wird. Der geplante Stichprobenumfang ist damit erst einmal unabhängig davon, ob wir den Dollar oder die Buchung als Element der Stichprobe definieren.

In der Literatur wird häufig das Bild verwendet, dass die Dollar als Haken an den eigentlichen Prüfungsobjekten fungieren, an denen sie, durch Auswahl eines Dollars, in die Stichprobe buchstäblich „gezogen“ werden.[48] Folgt man dieser Vorstellung, ist es schon intuitiv verständlich, dass Prüfungsobjekte mit höherem Wert mit höherer Wahrscheinlichkeit in die Stichprobe gelangen, haben sie doch mehr „Haken“, an denen man sie in diese ziehen kann. In Bezug auf Jahresabschlussprüfungen wird durch diesen Gedanken der Annahme Rechnung getragen, dass sich hinter großen Beträgen in der Regel auch größere Fehler im Sinne von Überbewertungen verstecken.[49] Ein Nachteil ergibt sich dagegen für die Prüfung von Unterbewertungen, für die angenommen wird, dass diese eher mit kleinen Beträgen in Verbindung stehen. Diese gelangen nur mit einer geringeren Wahrscheinlichkeit in die Stichprobe. Für Nullsalden (etwa ausgeglichene Debitoren- oder Kreditorenkonten) gilt sogar eine Auswahlwahrscheinlichkeit von Null.[50]

Es bleibt festzuhalten, dass die Definition von Grundgesamtheit und Stichprobenelement von der ursprünglichen Definition abweicht. Dagegen bleibt der Begriff des Prüfungsobjekts beim *Dollar Unit Sampling* unangetastet, schließlich soll nicht ein

[47] Der Gedanke, eine monetäre Einheit für Prüfungszwecke als Stichprobenelement zu definieren, findet sich erstmals in van Heerdens 1961 veröffentlichtem Aufsatz. In diesem definiert er den Gulden als Stichprobenelement.

[48] Vgl. Guy/Carmichael/Whittington (2002), S. 196, Steele (1992), S. 152.

[49] Vgl. Johnson/Leitch/Neter (1981), S. 291 sowie Neter/Johnson/Leitch (1985), S. 489.

[50] Vgl. Guy (2002), S. 195.

einzelner Dollar auf richtige Merkmale geprüft, sondern die Richtigkeit des anhand eines Dollars ausgewählten Prüfungsobjekts aufgrund bestimmter Merkmale und auf diese gerichtete Prüfungshandlungen festgestellt werden.[51] Die Definition der einzelnen monetären Einheit als Stichprobenelement für die Zwecke der Prüfung ist damit formeller Natur.[52]

3.2 Methoden der Stichprobenziehung im Dollar-Unit-Sampling-Verfahren

3.2.1 Reine Zufallsauswahl

Bei der reinen Zufallsauswahl im *Dollar Unit Sampling* werden die einzelnen Buchungen oder Salden aufgereiht und die Buchwerte bw_i der einzelnen Positionen aufkumuliert. Im Ergebnis wird jeder Position i der bis zu ihr erreichte kumulierte Buchwert $BW_i = \sum_{j=1}^{i} bw_j$ zugeordnet. Der Gesamtbetrag aus den einzelnen Buchwerten ($BW_N \equiv BW$) bestimmt das Intervall, in welchem die Zufallszahlen zu ermitteln sind. Es werden n Zufallszahlen zwischen 1 und dem Gesamtbetrag gezogen. Das i-te Element gelangt in die Stichprobe, wenn die ermittelte Zufallszahl ZF folgende Voraussetzung erfüllt:

$$BW_{i-1} < ZF \leq BW_i.$$

Ein Nachteil der beschriebenen Methode ist, dass der geplante Stichprobenumfang häufig größer ist als die Anzahl unterschiedlicher Elemente in der Stichprobe. Dies ist der Fall, da jedes Element durch das Bilden der Zufallszahlen jedes Mal in die Stichprobe gelangen kann.

In der Literatur (und auch in der Software, die von Prüfungsgesellschaften verwendet wird) wird für den Fall der Mehrfachziehung meist die Mehrfachberücksichtigung des Elementes empfohlen.[53] Insofern handelt es sich bei der Methode nicht mehr um eine

[51] So könnte die Richtigkeit einer Buchung bspw. mit Hilfe eines vorhandenen Beleges ermittelt werden, indem für ihn charakteristische Merkmale wie Datum, ausgewiesener Betrag und Sachverhalt untersucht werden.

[52] Vgl. Hitzig (2004a), S. 31.

[53] Vgl. zum Beispiel die Benutzungshinweise der Prüfungssoftware ACL Version 8.0 (*Audit Conrol List*).

Ziehung der Elemente „ohne Zurücklegen“.
Ferner wird von Wurst (1989) und Horgan (1996) kritisiert, dass sich durch das notwendige Zurückverfolgen des Dollars zu seiner Buchung Implementierungsprobleme ergeben können.[54] In der aktuellen Prüfungssoftware wird das Verfahren der reinen Zufallsauswahl angeboten. So bietet ACL[55] Version 8.0 die Option, die Elemente durch zufällige Auswahl der Geldeinheiten in die Stichprobe zu ziehen.

3.2.2 Cell-Sampling

Im Rahmen des *Cell-Samplings* wird der Gesamtbuchwert des betrachteten Prüffeldes durch den Stichprobenumfang geteilt, wodurch sich das durchschnittliche Stichprobenintervall $\overline{I}$ ergibt:

$$\overline{I} = \frac{BW}{n}.$$

Wie bei der reinen Zufallsauswahl werden die einzelnen Buchungen hintereinander angeordnet, das gesamte Prüffeld dann jedoch in n „Zellen“ der Größe $\overline{I}$ geteilt. Für jede dieser n Zellen wird eine Zufallszahl zwischen 1 und $\overline{I}$ gebildet. Somit wird aus jeder Zelle ein Element in die Stichprobe gezogen.[56] Dieses Vorgehen entspricht einer zusätzlichen Schichtung des Prüffeldes. Durch die Auswahltechnik wird der Fall der Mehrfachziehung natürlich nicht ausgeschlossen, da ein Element der Grundgesamtheit auch zwei Zellen füllen kann, er tritt jedoch seltener auf als bei der reinen Zufallsauswahl, da extreme, bei der reinen Zufallsauswahl mögliche Stichprobenzusammensetzungen vermieden werden.
Es wird in der Literatur empfohlen und hat sich auch in den praktischen Richtlinien der Wirtschaftsprüfungsgesellschaften durchgesetzt, vor dem Ziehen der Stichprobe diejenigen Elemente unabhängig von der Stichprobe zu prüfen, deren Buchwert über dem durchschnittlichen Stichprobenintervall oder der Wesentlichkeitsgrenze liegt.[57] Dies wird zum einen dadurch begründet, dass diese Elemente mit großer Wahrscheinlich-

[54] Vgl. Horgan (1996), S. 216 über Wurst (1989).
[55] ACL steht für *Audit Control List*.
[56] Vgl. Eggenberger (1986), S. 85.
[57] Vgl. Boockholdt/Chang/Finley (1992) S. 63, CIPFA (1991), S. 47.

keit sowieso in die Stichprobe gelangen würden. Zum anderen sollen somit wertmäßig wesentliche Sachverhalte einer Untersuchung auf keinen Fall entfliehen können. Wie bei der reinen Zufallsauswahl hat auch beim *Cell-Sampling* eine Zurückverfolgung der Geldeinheit zum zu prüfenden Element zu erfolgen.

Auch das *Cell-Sampling* ist in aktuelle Prüfungssoftware implementiert. Sowohl ACL Version 8.0 als auch IDEA[58] Version 7 bieten das *Cell-Sampling* als Auswahlmethode an. Bei der Prüfungssoftware IDEA *muss* die Auswahl der Elemente aufgrund des *Cell-Samplings* erfolgen, da später die Bewertung auch durch die *Cell-Evaluation* durchgeführt wird.[59]

3.2.3 Sieve-Sampling

Einen leicht abgewandelten Ansatz der Stichprobenziehung bietet das *Sieve-Sampling*. Es geht nicht von einer Grundgesamtheit von monetären Einheiten aus, sondern bleibt bei dem Verständnis der Grundgesamtheit als Menge von Prüfungselementen.[60] Wie beim *Cell-Sampling* wird als Erstes das durchschnittliche Stichprobenintervall $\overline{I}$ als Quotient aus Gesamtbuchwert und Stichprobenumfang gebildet. Für jedes Element der Grundgesamtheit wird eine Zufallszahl zwischen 1 und $\overline{I}$ ermittelt. Ein Element gelangt in die Stichprobe, wenn sein Buchwert gleich oder größer der für dieses Element gebildeten Zufallszahl ZF_i ist:[61]

$$bw_i \geq ZF_i.$$

Somit ist die Wahrscheinlichkeit, dass ein Element in die Stichprobe gelangt, wie beim *Monetary Unit Sampling* gewollt, abhängig von seiner relativen Größe in dem Prüffeld. Zugleich ist es jedoch – im Gegensatz zu den zuvor beschriebenen Verfahren – nicht notwendig, eine bestimmte Geldeinheit einem Prüfelement zuzuordnen, da jedes Element für sich allein betrachtet wird.

[58] IDEA steht für *Integrated Data Evaluation and Analysis System*.

[59] Vgl. CaseWare IDEA (2003), S. 17.

[60] Vgl. Gill (1983), S. 1.

[61] Vgl. Horgan (1996), S. 216.

Das Problem der Verfahrensweise im *Sieve-Sampling* ist, dass nicht immer der gewünschte Stichprobenumfang n erreicht wird. Dieser ist jeweils von den gerade ermittelten Zufallszahlen abhängig. Die erreichte Stichprobengröße ist jedoch ein unverzerrter Schätzer der gewollten Stichprobengröße n, mit einer Standardabweichung von höchstens $\sqrt{n}$, wenn alle Buchwerte kleiner oder gleich dem Stichprobenintervall sind.[62]

Durch Kombination mit dem im nächsten Abschnitt vorgestellten *Lahiri-Sampling* kann das Problem des schwankenden Stichprobenumfangs behoben werden.[63] Ist die erzielte Stichprobe geringer als die gewünschte, wird das *Lahiri-Sampling* ausgeführt, bis der gewünschte Stichprobenumfang erreicht ist. Die Auswahl erfolgt dabei jedes Mal aus der Ausgangsgrundgesamtheit (inklusive der Elemente, die schon in der Stichprobe sind, und mit Zurücklegen). Ist die Stichprobenanzahl zu groß, wird aus ihr durch reine Zufallsauswahl eine weitere Stichprobe in der gewünschten Größe ohne Zurücklegen gezogen. Ein Element gelangt in die Stichprobe, wenn es in beiden Schritten ausgewählt wird.[64]

3.2.4 Lahiri-Sampling

Das von Lahiri (1951) vorgeschlagene und von Horgan (1997, 1998 und 1999) auf den Prüfungsbereich übertragene Stichprobenziehungsverfahren wird in zwei Schritten durchgeführt:[65]

1. Es wird eine Zufallszahl zwischen 1 und N (Umfang der Grundgesamtheit, etwa: Anzahl der Buchungen auf einem Konto) ermittelt, um zu entscheiden, welches Element für eine weitere Betrachtung in Frage kommt.

2. Eine weitere Zufallszahl (zum Beispiel zwischen 1 und $\overline{I}$) ZF_i wird ermittelt, um zu entscheiden, ob das im ersten Schritt ausgewählte Element i in die Stichprobe gelangt. Dieses wird in die Stichprobe gezogen, wenn für dessen Buchwert $bw_i \geq ZF_i$ gilt.

[62] Vgl. Horgan (1996), S. 216.

[63] Vgl. Horgan (1997), S. 41 über Lahiri (1951).

[64] Vgl. Horgan (1998), S. 48.

[65] Vgl. Horgan (1999), S. 20.

Die zwei Schritte werden solange ausgeführt, bis der gewünscht Stichprobenumfang erreicht ist.

3.2.5 Intervallziehungen

– **Die fixe Intervallziehung**

Im Rahmen der fixen Intervallziehung werden die Buchwerte der einzelnen Prüfungsobjekte bw_i aufkumuliert und wieder ein durchschnittliches Stichprobenintervalle $\overline{I}$ durch Teilung des Gesamtbuchwertes durch den Stichprobenumfang berechnet. Es wird eine Zufallszahl ZF bestimmt, die zwischen 1 und $\overline{I}$ liegt. Die Elemente, welche die Geldeinheiten

$$ZF, \quad ZF + \overline{I}, \quad ZF + 2 \cdot \overline{I}, \quad \ldots, \quad ZF + (n-1)\overline{I}$$

beinhalten, gelangen in die Stichprobe. Die Auswahl des ersten Elementes ist zufällig, während die Auswahl aller weiteren Elemente von der ersten Auswahl abhängt.[66] Stehen andere Auswahlmethoden zur Verfügung, sollte von der fixen Intervallziehung abgesehen werden, da sie Risiken bei eigentümlichen Fehlerverteilungen in der Prüfungsgesamtheit bürgt. Fehler können dann in der Stichprobe über- oder unterrepräsentiert sein.[67] Trotz dieses Nachteils bietet die Prüfungssoftware ACL neben anderen Auswahlmethoden auch die fixe Intervallsziehung als Alternative für den Prüfer. Die Methode ist mir gegebenen Informationen zufolge auch in den Prüfungsgrundsätzen von mindestens einem von mir angesprochenen Wirtschaftsprüfungsunternehmen verankert.

– **Die variable Intervallziehung**

Durch die variable Intervallziehung soll eine reine Zufallstichprobe simuliert werden, die jedoch einfacher zu ziehen ist und die Probleme der reinen Zufallsauswahl umgeht. Es wird ein durchschnittliches Stichprobenintervall $\overline{I}$ (siehe oben) aufgrund des gewünschten Stichprobenumfanges festgelegt. Die Zufallsstichprobe

[66] Vgl. Apostolou/Alleman (1991), S. 72.

[67] Vgl. Eggenberger (1986), S. 85. Als Beispiel können Lohnabrechnungen gesehen werden, wobei einem bestimmten Stichprobenintervall immer der gleiche Mitarbeiter in die Stichprobe gelangen und geprüft würde.

wird dann zum Beispiel aufgrund einer Reihe exponentiell variierender Intervalle generiert.[68] Die erzeugten Zufallszahlen ZF_i sind gleichverteilt zwischen 0 und 1. Die Gegenzahl ihres natürlichen Logarithmus wird mit dem durchschnittlichen Stichprobenintervall multipliziert:

$$I_i = -\overline{I} \cdot \ln ZF_i.$$

Die Stichprobenintervalle I_i werden nacheinander aufaddiert und das Element, welches die Geldeinheit $\sum_{i=1}^{j} I_i$ enthält, gelangt in die Stichprobe.

Ein Problem der beschriebenen Methode ist natürlich, dass, obwohl im Mittel zwar dem gewünschten Stichprobenumfang entsprechend, der Stichprobenumfang durch die Zufälligkeit der einzelnen Intervallsgrößen schwankt.

[68] Vgl. Leslie/Teiltlebaum/Anderson (1979), S. 101.

4 Die Schätzung der Fehlerrate – Relevante Verteilungen

In der Anwendung des *Dollar Unit Samplings* haben sich für den Teil der Schätzung der Fehlerrate einige Approximationen durchgesetzt. Im Folgenden werden die für die Schätzung der Fehlerrate relevanten Verteilungen inklusive der Voraussetzungen für die Zulässigkeit der Anwendungen der Verteilungsannahmen dargestellt.

4.1 Die hypergeometrische Verteilung

Nehmen wir eine Grundgesamtheit von N Elementen an, in der sich jedes Prüfungsobjekt nach Untersuchung in eine der zwei Kategorien „korrekt ausgewiesen“ oder „nicht korrekt ausgewiesen“, d. h. aufgrund eines zweipunktverteilten Merkmals charakterisieren lässt. Ist X die Anzahl der nicht korrekt ausgewiesenen Elemente in der Grundgesamtheit, sind $N - X$ Elemente korrekt ausgewiesen. In der Realität ist X natürlich unbekannt, und es wird versucht, auf der Grundlage von den x in der Stichprobe gefundenen Fehlern Aufschluss über die wahre Fehleranzahl zu erhalten. Der Stichprobenumfang ist gegeben durch n. Die Wahrscheinlichkeit $\mathcal{P}$, bei einer Grundgesamtheit mit Umfang N und X nicht korrekt ausgewiesenen Elementen in einer Stichprobe vom Umfang n genau x Fehler zu entdecken, ergibt sich dann aus:

$$\mathcal{P}\left(x, n, X, N\right) = \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}}.$$

Die Wahrscheinlichkeit, höchstens x Fehler zu finden, ist somit gegeben durch

$$\sum_{i=0}^{x} \frac{\binom{X}{i}\binom{N-X}{n-i}}{\binom{N}{n}} = \alpha.$$

Hierin steckt die Annahme, dass einmal in die Stichprobe gezogene Elemente nicht ein zweites Mal gezogen werden können. Dies macht für den Prüfungsbereich auch Sinn, schließlich soll ein einmal betrachtetes Prüfungsobjekt nicht zweimal zum Untersuchungsgegenstand werden.

Aufgrund der beschriebenen Verteilung wird somit das Risiko α quantifizierbar, dass eine Stichprobe höchstens x Fehler enthält, obwohl die Grundgesamtheit X Fehler beherbergt. Umgekehrt, und das ist die praktisch relevante Fragestellung, kann für ein gegebenes α (Irrtumswahrscheinlichkeit) ein Wert für die Fehlerrate $p = \frac{X}{N}$ in der Grundgesamtheit auf der Grundlage des Stichprobenergebnisses hergeleitet werden.

4.2 Die Binomialverteilung

Auch wenn die hypergeometrische Verteilung für Fälle, in denen der Umfang der Grundgesamtheit bekannt ist, durch die Annahme, dass Elemente nach einmaliger Ziehung nicht zurückgelegt werden, die Realität widerspiegelt, ist sie, werden die Berechnungen manuell durchgeführt, nur relativ aufwändig handhabbar. In den 1960er bis 1970er Jahren, in denen mathematisch-statistische Stichprobenverfahren im Prüfungsbereich an Bedeutung zunahmen und das *Dollar Unit Sampling* erstmals beschrieben und angewandt wurde,[69] waren zudem die verfügbaren Rechenkapazitäten nicht groß genug, um problemlos mit der hypergeometrischen Verteilung arbeiten zu können.[70] Eine Möglichkeit, den Rechenaufwand zu verringern, bietet die Approximation durch die Binomialverteilung für im Verhältnis zur Stichprobe große Grundgesamtheiten (als Faustregel gilt $\frac{n}{N} < 0,05$).[71] Ist die Fehlerrate p der Anteil fehlerhafter Elemente in der Grundgesamtheit und spiegelt $1 - p$ entsprechend den Anteil der korrekten Elemente in der Grundgesamtheit wider, ist die Wahrscheinlichkeit, genau x fehlerhafte Elemente in

[69] Vgl. Van Heerden (1961), Kaplan (1975).

[70] Vgl. Leslie/Teitlebaum/Anderson (1979), S. 70.

[71] Vgl. Bredeck (1993), S. 41

einer Stichprobe von n zu finden, gemäß der Binomialverteilung:[72]

$$\mathcal{P}(x, n, p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Eine Berechnung auf diesem Weg beinhaltet natürlich die Annahme, dass das in die Stichprobe gezogene Element nach der Ziehung zurück in die Grundgesamtheit gelangt, d. h. die Möglichkeit besteht, es ein weiteres Mal in die Stichprobe zu ziehen. Somit bleibt die Wahrscheinlichkeit, ein fehlerhaftes Element in die Stichprobe zu ziehen, bei jeder Ziehung gleich, und die Berechung enthält gegenüber der hypergeometrischen Verteilung einen Parameter weniger. Je größer die betrachtete Grundgesamtheit ist, desto weniger unterscheiden sich die Ergebnisse der hypergeometrischen und der Binomialverteilung.[73]

4.3 Die Poisson-Verteilung

In einem nächsten Schritt wird die Annahme getroffen, dass es sich bei dem Auftreten von Fehlern in der Grundgesamtheit um seltene Ereignisse handelt. Diese Annahme findet ihre Berechtigung in folgender Logik: Ginge man davon aus, dass in der Grundgesamtheit häufig Fehler auftreten, Fehler dementsprechend kein seltenes Ereignis darstellen, hieße das bereits, dass davon ausgegangen wird, dass die Grundgesamtheit als nicht unwesentlich fehlerhaft eingestuft wird. Ist die Grundgesamtheit hinreichend groß und die Fehlerrate hinreichend klein, kann die Binomialverteilung durch die Poisson-Verteilung approximiert werden.[74]

Ist p wieder die Fehlerrate der Grundgesamtheit und n der Stichprobenumfang, so stellt das Produkt $np = \lambda$ die erwartete Anzahl an Fehlern in der Stichprobe dar, welche in der Prüfungsliteratur häufig auch als Fehlerintensität, Verlässlichkeitsfaktor oder Faktor für die obere Fehlergrenze bezeichnet wird.[75] Die Wahrscheinlichkeit, genau x

[72] Wiederum der Logik im Abschnitt über die hypergeometrische Verteilung folgend, wird später eine Wahrscheinlichkeit α gesetzt und auf der Grundlage des Stichprobenergebnisses Rückschluss auf die maximale Fehlerrate in der Grundgesamtheit gezogen.

[73] Vgl. Bamberg/Baur (2002), S. 102.

[74] Vgl. Domschke/Drexl (2002), S. 196.

[75] Vgl. Wolz (2004), AICPA (1999), Leslie/Teitlebaum/Anderson (1979).

Fehler in einer Stichprobe mit $\lambda = np$ erwarteten Fehlern zu finden, berechnet sich dann nach:

$$\mathcal{P}(\lambda, x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}.$$

Auch wenn der idealtypische Fall, dass p infinitesimal klein und n infinitesmal groß ist, in der Prüfung nicht eintritt, bildet die Poisson-Verteilung doch eine recht genaue Approximation an die Binomialverteilung.[76] Ihr Vorteil liegt in ihrer leichten Handhabung. Sie ist auch ohne Prüfungssoftware und auch manuell sehr leicht anwendbar. In der Literatur werden die Berechnungen fast ausschließlich anhand der Poisson-Verteilung ausgeführt. Zum einen lassen sich dadurch Rechenbeispiele leichter darstellen, zum anderen stammt der Großteil der Literatur über das *Dollar Unit Sampling* aus den 1970er und 1980er Jahren, in denen noch größerer Wert auf die leichte manuelle Handhabung von Stichprobenverfahren geachtet wurde. Auch heute wird meist mit der Approximation durch die Poisson-Verteilung gearbeitet, und Leitlinien für Prüfer geben stets die Fehlerintensitäten gemäß der Poisson-Verteilung für die Berechnungen an.[77]

Tabelle 1 gibt ein Beispiel für eine solche Berechnungstabelle an, die von dem *American Institute of Certified Public Accountants (AICPA)* veröffentlicht wurde. Die Tabelle mag auf den ersten Blick mathematisch-statistisch ungewohnt erscheinen, legt aber folgende Logik zu Grunde: Im Rahmen der Stichprobenplanung legt der Prüfer ein Sicherheitsniveau $1-\alpha$ für sein Prüfungsurteil fest (zum Beispiel 95%, d. h. $\alpha = 0,05$). Es wird eine Stichprobe gezogen, und diese enthält x Fehler (zum Beispiel $x = 2$ Fehler). Aufgrund dieser zwei Parameter soll nun Rückschluss gezogen werden auf die Fehlerintensität $\lambda = pn$, sodass bei gegebenem Stichprobenumfang n (zum Beispiel $n = 100$), die Fehlerrate $p = \frac{\lambda}{n}$ der Grundgesamtheit schätzbar wird. Für die Beispielzahlen ergibt sich $\lambda_{2;\ 0,05} = 6,30$. Das entspricht einer geschätzten Fehlerrate von $\hat{p} = 0,063$. Die Aussage des Prüfers ist nun folgende: Läge die wahre Fehlerrate in der Grundge-

[76] Für den Beweis, dass die Poisson-Verteilung auch unter dem Aspekt der Sicherstellung konservativer Prüfungsergebnisse ihre Berechtigung besitzt, wird für den Nullfehlerstichprobenfall auf Leslie/Teitlebaum/Anderson (1979), S. 265, und für den allgemeinen Fall auf Anderson und Samuels (1967) verwiesen.

[77] Eigene Beobachtungen/Berichte aus Wirtschaftsprüfungsunternehmen. Veröffentlichung des Amerikanischen Instituts für Wirtschaftsprüfer (AICPA) im Jahr 1999.

samtheit bei über 6,3%, dann hätte ich mit 95% Sicherheit mehr als $x = 2$ Fehler in der Stichprobe vom Umfang $n = 100$ gefunden.

Die Fehlerintensität λ ist die Grundlage für die meisten Verfahren der Fehlerhochrechnung in Abschnitt 5, ihre Zusammensetzung wird deshalb in Abschnitt 5.1 ausführlich beschrieben.

Faktoren zur Berechnung der oberen Fehlergrenze ($\lambda_{x;\ \alpha}$)

Anzahl der Überbewertungen in der Stichprobe (x)	Risiko der inkorrekten Annahme des Prüffeldes (α)								
	1%	5 %	10 %	15 %	20 %	25 %	30 %	37 %	50 %
0	4,61	3,00	2,31	1,90	1,61	1,39	1,21	1,00	0,70
1	6,64	4,75	3,89	3,38	3,00	2,70	2,44	2,14	1,68
2	8,41	6,30	5,33	4,72	4,28	3,93	3,62	3,25	2,68
3	10,05	7,76	6,69	6,02	5,52	5,11	4,77	4,34	3,68
4	11,61	9,16	8,00	7,27	6,73	6,28	5,90	5,43	4,68
5	13,11	10,52	9,28	8,50	7,91	7,43	7,01	6,49	5,68
6	14,57	11,85	10,54	9,71	9,08	8,56	8,12	7,56	6,67
7	16,00	13,15	11,78	10,90	10,24	9,69	9,21	8,63	7,67
8	17,41	14,44	13,00	12,08	11,38	10,81	10,31	9,68	8,67
9	18,79	15,71	14,21	13,25	12,52	11,92	11,39	10,74	9,67
10	20,15	16,97	15,41	14,42	13,66	13,02	12,47	11,79	10,67
11	21,49	18,21	16,60	15,57	14,78	14,13	13,55	12,84	11,67
12	22,83	19,45	17,79	16,72	15,90	15,22	14,63	13,89	12,67
13	24,14	20,67	18,96	17,86	17,02	16,32	15,70	14,93	13,67
14	25,45	21,89	20,13	19,00	18,13	17,40	16,77	15,97	14,67
15	26,75	23,10	21,30	20,13	19,24	18,49	17,84	17,02	15,67
16	28,03	24,31	22,46	21,26	20,34	19,58	18,90	18,06	16,67
17	29,31	25,50	23,61	22,39	21,44	20,66	19,97	19,10	17,67
18	30,59	26,70	24,76	23,51	22,54	21,74	21,03	20,14	18,67
19	31,85	27,88	25,91	24,63	23,64	22,81	22,09	21,18	19,67
20	33,11	29,07	27,05	25,74	24,73	23,89	23,15	22,22	20,67

Tabelle 1: Fehlerintensitäten, Quelle: AICPA (1999), S. 109.

5 Die Schätzung des in der Grundgesamtheit vorhandenen Fehlerbetrages

5.1 Zerlegung der oberen Fehlergrenze

Abweichend von der üblichen statistischen Betrachtungsweise eines Konfidenzintervalls, das mit einer bestimmten Sicherheit den in der Grundgesamtheit enthaltenen unbekannten Parameter umschließt, geht es in der Prüfung auf wesentliche Falschangaben vorrangig um eine obere Fehlergrenze, die mit einer bestimmten Sicherheit den wahren Fehler überschreitet. Der Prüfer möchte ein Urteil in der Art abgeben: „Mit 95%-iger Sicherheit übersteigt der fehlerhafte Betrag in der Grundgesamtheit € y nicht." Der Betrag von € y in der Aussage ist die obere Fehlergrenze, die letztendlich mit dem maximalen Fehlerbetrag, bei dem der Prüfer eine Richtigkeit des Prüffeldes noch attestieren würde, verglichen wird.

Der in Tabellenform verfügbare Faktor ($\lambda_{x,\alpha}$),[78] mit dessen Hilfe die obere Fehlergrenze berechnet wird, kann für die Zwecke des Dollar-Unit-Sampling-Verfahrens in drei Teile zerlegt werden:[79]

1. Der Faktor für die *Basic Precision* (BP) ist der Faktor, der bei einer fehlerlosen Stichprobe Anwendung finden würde ($\lambda_{0,\alpha}$).

2. Der Faktor für den wahrscheinlichen Fehler (*Most Likely Error*, MLE) ist 1, da der MLE der Schätzer für den wahren Fehlers aufgrund der Stichprobe ist und sich als Produkt aus durchschnittlichem Stichprobenintervall und der Summe aus den gefundenen Fehleranteilen ergibt ($MLE = \overline{I} \cdot \sum_{i=1}^{x} r_i$), wobei $\overline{I}$ das durchschnittliche Stichprobenintervall ist und r_i den mit dem i-ten Fehler in der Stichprobe assoziierte Fehleranteil darstellt.

3. Die Zerlegung erfolgt ferner in den Faktor für das *Precision Gap Widening* (PGW), welcher dafür sorgt, dass sich die obere Fehlergrenze mit jedem gefunden Fehler

[78] Vgl. Tabelle 1. Zur Erinnerung: x ist die Anzahl der beobachteten Fehler in der Stichprobe und $1 - \alpha$ das Konfidenziniveau.

[79] In Anlehnung an CIPFA (1995), S. 37 f.

überproportional erhöht. Der PGW-Faktor ist die Differenz zwischen den Fehlerintensitäten für zwei aufeinander folgende Fehleranzahlen weniger dem MLE-Faktor auf einem bestimmten Konfidenzniveau:

$$PGW = \lambda_{i,\alpha} - \lambda_{i-1,\alpha} - 1.$$

Als Gleichung formuliert, ergibt sich folgender Ausdruck für den Faktor der oberen Fehlergrenze (*Upper Error Limit*, UEL):

$$UEL = \lambda_{x,\ \alpha} = \lambda_{0,\ \alpha} + \sum_{i=1}^{x} (\lambda_{i,\ \alpha} - \lambda_{i-1,\ \alpha}).$$

Abbildung 1 verdeutlicht die Zerlegung der Fehlerintensität (des Faktors für die obere Fehlergrenze) grafisch für das 95%-Sicherheitsniveau.

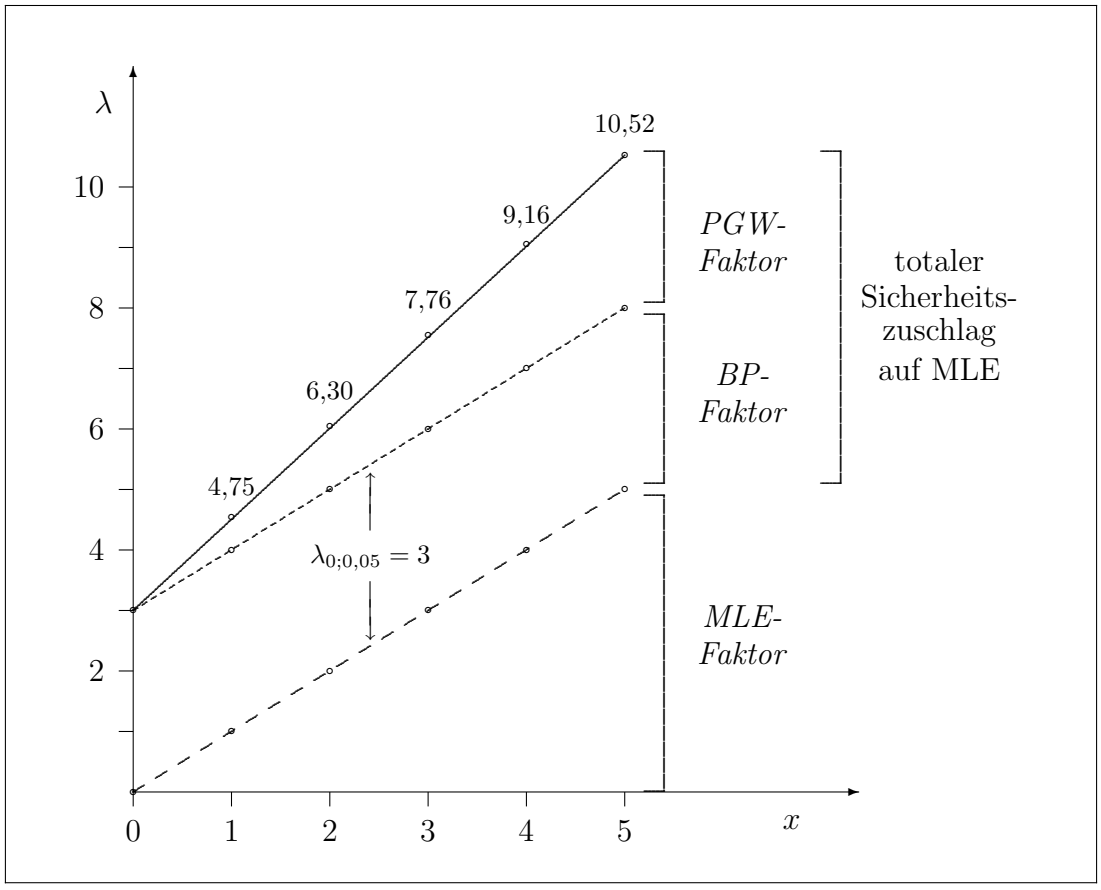


Abbildung 1: Zerlegung der Fehlerintensität λ für das 95%-Konfidenzniveau, in Anlehnung an Leslie et al. (1979), S. 126.

Werden in einer Stichprobe nur 100%-ige Fehler beobachtet (ein Betrag wurde gebucht, obwohl der Sachverhalt gar nicht hätte gebucht werden dürfen), könnte dieser Faktor zur realistischen Hochrechnung des Fehlers auf die Grundgesamtheit einfach mit dem durchschnittlichen Stichprobenintervall $\overline{I}$ multipliziert werden. Für den Fall partieller Überbewertungen gibt es jedoch keine eindeutige Lösung des Problems.[80] Die im Folgenden vorgestellten Verfahren stellen deswegen Näherungslösungen für die Fehlerhochrechnung für Fälle dar, in denen von 100%-Fehlern abweichende Fehleranteile in der Stichprobe entdeckt werden.

Die oberen Fehlergrenzen bezeichne ich im Folgenden C (für *„Combined Attributes and Variables Bounds“*). Nach einer grundsätzlichen Beschreibung der Verfahren für den Überbewertungsfall erfolgt eine Erläuterung dessen, wie in der Stichprobe gefundene Unterbewertungen in die Hochrechnung des Fehlers einbezogen werden.

5.2 Vorstellung der einzelnen Fehlerhochrechnungsmethoden

5.2.1 Die Maximalfehlermethode

Die Maximalfehlermethode ist die konservativste aller Fehlerprojektionsmethoden. Die Methode unterstellt, dass, wenn die monetäre Bewertung eines Prüfungsobjektes fehlerhaft ausgeführt wurde, eine hundertprozentige Überbewertung vorliegt. D. h. der mit einem positiven Betrag ausgewiesene Sachverhalt hätte eigentlich gar nicht gebucht werden dürfen. Der maximale Fehler, die obere Fehlergrenze im Prüffeld, wird dann als

$$C_{max} = \lambda_{x,\ \alpha} \cdot \overline{I}$$

berechnet, wobei λ die Fehlerintensität und $\overline{I}$ das Stichprobenintervall ist.[81]

Der Nachteil der beschriebenen Methode ist offensichtlich: Sind in einem Prüffeld eher viele kleine Fehler anstelle von einigen großen Fehlern vorhanden, neigt das Verfahren zu massiven Überschätzungen des wahren Fehlers. Zu einem realistischen Ergebnis führt

[80] Vgl. Biaggio (1987), S. 217.

[81] Vgl. Wolz (2003), S. 124, mit leicht abweichender Darstellung vgl. Reneau (1978), S. 672.

die Methode nur, wenn tatsächlich alle fehlerhaften Elemente 100 % falsch bewertet sind. Eine Zerlegung des UEL-Faktors für diese Methode erfolgt nicht.

Beispiel: Zur Veranschaulichung der Methode sei folgendes Beispiel gegeben, an dessen Ausgangsdaten auch die folgenden Fehlerhochrechnungsmethoden verdeutlicht werden. Gegeben sei ein Konto, auf welchem Umsatzerlöse in Höhe von rund 11,1 Millionen Euro verbucht wurden. Aufgrund der Festlegung des maximal tolerierbaren Fehlers auf € 300.000, der Erwartung, in der Stichprobe keine Fehler vorzufinden und einer angestrebten Sicherheit von 95 % wird ein Stichprobenintervall von € 100.000 festgelegt.[82] Dies entspricht einem Stichprobenumfang von 111 Elementen. Unter Verwendung der *Cell-Selection* werden die Elemente ausgewählt, und zwei Fehler – eine 50%-ige und eine 30%-ige Überbewertung – werden festgestellt.
Da das Verfahren jeden gefundenen Fehler als 100%-igen Fehler für die Hochrechnung betrachtet, erfolgt die Berechnung wie folgt:

$$C_{max} = \lambda_{2;\ 0,05} \cdot \overline{I} = 6,296 \cdot 100.000 = 629.600.$$

In diesem Fall liegt der hochgerechnete Fehler deutlich über dem maximal tolerierbaren Fehler. Dem Prüffeld könnte keine Ordnungsmäßigkeit bescheinigt werden. Als Alternative wird angeboten, den Stichprobenumfang auszuweiten,[83] obwohl von diesem Vorgehen in bestimmten Fällen abzuraten ist.[84] Auch sind in der Literatur Vorschläge, das Konfidenzniveau nach unten anzupassen oder den tolerierbaren Fehler zu erhöhen, zu finden.[85] Eine nachträgliche Korrektur des Konfidenzniveaus ist meines Erachtens fragwürdig. Wäre die korrigierte Sicherheit ausreichend, hätte die Berechnung von Anfang an auf ihrer Grundlage durchgeführt werden können.

Die Maximalfehlermethode wird in der Prüfungspraxis nicht angewandt, soll hier jedoch als Vergleichsmaßstab für die anderen Verfahren dienen.

5.2.2 Die Durchschnittsfehlermethode

Die Durchschnittsfehlermethode berichtigt den offensichtlichen Nachteil der Maximalfehlermethode. Werden Fehler in dem Prüffeld gefunden, die nicht sämtlich vollständige

[82] Zur Stichprobenplanung vgl. zum Beispiel Kaplan (1975), Grimlund (1988), Hall/Pierce/Ross (1989).

[83] Vgl. Hitzig (2004b), S. 35.

[84] Biaggio (1987), S. 219, rät von einer Vergrößerung des Stichprobenumfanges ab, da diese erst ab einer Verdoppelung Wirkung zeigt.

[85] Vgl. Wampler (2005), S. 37.

Falschbewertungen darstellen, wird ein einfaches arithmetisches Mittel $\overline{r}$ aus den beobachteten Fehleranteilen $r_i = \frac{bw_{i,ist} - bw_{i,soll}}{bw_{i,ist}}$ gebildet. Für den Nullfehlerfall gilt $\overline{r} = 1$.[86] Im Gegensatz zur Maximalfehlermethode wird somit dem Fakt Rechnung getragen, dass gegebenenfalls nur Teile des Buchungsbetrages fehlerhaft sind. Die Berechnung der oberen Fehlergrenze erfolgt folgendermaßen:

$$C_{D.} = \lambda_{x,\alpha} \cdot I \cdot \overline{r} \qquad \text{mit} \qquad \overline{r} = \begin{cases} \sum_{i=1}^{x} r_i & \text{für} \quad x > 0 \\ 1 & \text{für} \quad x = 0. \end{cases}$$

Dieses Verfahren führt zu logischen Inkonsistenzen, die bereits in der Literatur beschrieben, jedoch durch Konservativitätsüberlegungen gerechtfertigt wurden.[87]
Die Inkonsistenz des Verfahrens besteht darin, dass für eine fehlerfreie Stichprobe davon ausgegangen wird, dass noch ein Risiko für das Bestehen von 100%-ig fehlerhaften Elementen in der Grundgesamtheit gegeben ist, während für Stichproben mit gefundenen Fehlern die tatsächlichen Fehlerraten (die unter 100% liegen können) verwendet werden. So kann es vorkommen, dass in einer fehlerfreien Stichprobe die berechnete obere Fehlergrenze höher ist als in einer Stichprobe mit fehlerhaften Elementen, also der Fall, dass $\lambda_{0,\alpha} > \lambda_{x,\alpha}\overline{r}$ ist, zugelassen wird.
Meines Erachtens lässt sich die Inkonsistenz des Verfahrens *nicht* durch die Konservativität für unbeobachtete Fehler rechtfertigen. Die Annahme für eine fehlerfreie Stichprobe, dass noch 100%-ige Fehler in den restlichen Elementen vorhanden sein könnten, kann *nicht* durch das Finden von Fehlern „neutralisiert“ werden. Durch das Finden von Fehlern, auch wenn sie nur eine geringfügige Fehlerrate mit sich bringen, sinkt das Risiko, dass in der Grundgesamtheit ein vollständig falsch bewertetes Objekt vorliegt, gegenüber einer fehlerfreien Stichprobe *nicht*.[88]

Mit anderen Worten: Durch das Einbauen von kleinen Fehlern in die Stichprobe darf sich die maximale Fehlerrate *nie* verringern. Da dies bei dem vorgestellten Verfahren jedoch auftreten kann, ist dessen Anwendung meines Erachtens fraglich.

---

[86] Vgl. Reneau (1978), S. 672.

[87] Vgl. Wolz (2003), S. 126.

[88] Eine Diskussion von Abhängigkeiten von Fehlern in Prüfungssituationen befindet sich in Wheeler/Dusenbury/Reimers (1997).

Die Inkonsistenz des Verfahrens könnte theoretisch abgeschwächt werden, wenn auch für eine Nullfehlerstichprobe eine Fehlerrate von unter 100 % angenommen würde. Diese Festlegung müsste jedoch vor der Ziehung der Stichprobe aus Erfahrungswerten und logischen Überlegungen getroffen werden. Hält es der Prüfer für ausgeschlossen, dass vollständige Überbewertungen in der Grundgesamtheit vorliegen, kann der BP-Faktor mit einem Bruch multipliziert werden ($\lambda_{0,\alpha} \cdot x\%$), wie es von Leslie et al. (1979) für das Stringer-Verfahren vorgeschlagen wird. In Hinblick auf die Prüfungspraxis halte ich diese Anpassung wiederum für weniger geeignet, da sich dadurch der Toleranzbereich für den Stichprobenumfang erhöht.[89] Die irrtümliche Logik des Prüfers könnte darin bestehen, vorerst einen sehr geringen möglichen Fehleranteil anzunehmen, wodurch der notwendige Stichprobenumfang verkleinert würde. Befinden sich in der gezogenen Stichprobe dann keine Fehler oder nur der Annahme entsprechende Fehler, wird die ursprüngliche Annahme nicht verworfen (auch wenn in der Grundgesamtheit in Wirklichkeit auch größere Fehlerraten enthalten sein können). Der Stichprobenumfang wäre zu klein, die Prüfungssicherheit verzerrt.

Beispiel: Die gegebenen Daten sollen wieder denen im obigen Abschnitt entsprechen. Bei zwei gefundenen Fehlern mit Fehleranteilen von 50 % und 30 % und einem Konfidenzniveau von 95 % ergibt sich:

$$C_{D.} = \lambda_{2;\ 0,05} \cdot \overline{I} \cdot \overline{r} = 6,296 \cdot 100.000 \cdot 0,4 = 251.840.$$

Es zeigt sich, dass die mit der Durchschnittsfehlermethode berechnete obere Fehlergrenze deutlich unter der durch die Maximalfehlermethode ermittelten liegt. Sie umfasst nur 40 % des unter der Maximalfehlermethode ermittelten Betrages. Mit der aus dem letzten Beispiel gegebenen Wesentlichkeitsgrenze von € 300.000 würde dem Prüffeld somit aufgrund des Stichprobenergebnisses Ordnungsmäßigkeit bescheinigt werden.
Die Mangelhaftigkeit des Verfahrens zeigt sich jedoch, wenn man im Kontrast dazu die Fehlerhochrechnung für eine *fehlerfreie* Stichprobe durchführt. Da der durchschnittliche Fehleranteil im Nullfehlerfall nach Reneau (1979) 100 % gesetzt wird, würde hier die obere Fehlergrenze wesentlich höher liegen als im Zweifehlerfall:

$$C_{D.} = \lambda_{0;0,05} \cdot I = 2,996 \cdot 100.000 = 299.600.$$

[89] Ein Ansatz in der Stichprobenumfangsplanung geht von den erwarteten Fehlern in der Stichprobe aus, vgl. Grimlund (1988), S. 79.

Die Reihe an unlogischem Verhalten kann fortgesetzt werden, indem man sich zum Beispiel vorstellt, ein dritter Fehler mit einem 10%-Fehleranteil würde in der Stichprobe gefunden werden. Die obere Fehlergrenze würde dann im Vergleich zum Zweifehlerfall auf € 232.000 sinken.

5.2.3 Die Fehlerreihungsmethode – Stringer-Bound

Werden mehrere Fehler in einer Stichprobe beobachtet, die mit unterschiedlichen Fehleranteilen verbunden sind, liefert die Fehlerreihungsmethode, die von Springer (1963) und Stephan (1963) vorgeschlagen wird, ein konsistent konservatives Ergebnis bezüglich der Berechnung der oberen Fehlergrenze.[90] Die in der Stichprobe gefundenen Fehler werden bei der Verwendung der Fehlerreihungsmethode zuerst nach den Fehleranteilen r_i absteigend sortiert. Jedem Fehler wird dann der Anteil der Erhöhung des UEL-Faktors[91] zugeordnet, den er „verursacht“ hat. Die Berechnung ist wie folgt:

$$C_{Str.} = \overline{I} \cdot \left(\lambda_{0,\alpha} + \sum_{i=1}^{x} (\lambda_{i,\alpha} - \lambda_{i-1,\alpha}) \cdot r_i \right).$$

Wird das *Cell-Sampling* zur Ziehung der Elemente in die Stichprobe angewandt, so unterstellt die Fehlerreihungsmethode diesbezüglich die ungünstigste Verteilung in den Zellen.[92]

Das Stringer-Verfahren ist aber auch in anderer Hinsicht konservativ: Da λ mit jedem zusätzlichen Fehler nur degressiv wächst, wird höheren Fehleranteilen eine höhere Gewichtung für die Fehlerhochrechnung gegeben. Es wird argumentiert,[93] dass diese Methode regelmäßig zu einer Überbewertung der oberen Fehlergrenze im Verhältnis zum wahren Fehler führen muss.[94]

Trotz der oben genannten Mängel der Fehlerreihungsmethode erfreut sie sich in der Pra-

[90] Vgl. Biaggio (1987), S. 217, Reneau (1978), S. 673.

[91] Vgl. Abschnitt 5.1.

[92] Vgl. Stephan (1963), S. 404.

[93] Vgl. Dworin/Grimlund (1984), S. 218, Felix et al. (1990), S. 2.

[94] Diese Ansicht wird durch die in dieser Arbeit durchgeführte Simulationsstudie bestätigt, vgl. Abschnitt 6.

xis großer Beliebtheit. So wird ihre Anwendung in dem CPA[95] Journal im Jahr 2005 in Verbindung mit Excel empfohlen.[96] Gespräche und Einblicke in die Arbeitsdateien von Wirtschaftsprüfungsunternehmen bestätigen, dass dieses Verfahren angewandt wird.

Beispiel: Für unser Beispiel würde sich mit dem Stringer-Verfahren folgender hochgerechneter Fehler ergeben:

$$\begin{aligned} C_{Str.} &= 100.000 \cdot (\lambda_{0;\ 0,05} + 0,5 \cdot (\lambda_{1;\ 0,05} - \lambda_{0;\ 0,05}) + 0,3 \cdot (\lambda_{2;\ 0,05} - \lambda_{1;\ 0,05})) \\ &= 100.000 \cdot (2,996 + 0,5 \cdot 1,748 + 0,3 \cdot 1,552) = 433.560. \end{aligned}$$

Wiederum kippt das Prüfungsurteil im Vergleich zur vorherigen Methode. Bei einer Wesentlichkeitsgrenze von € 300.000 könnte dem Prüffeld keine Ordnungsmäßigkeit bescheinigt werden.

Im Gegensatz zum Durchschnittsfehlerverfahren wird bei der Springer-Methode immer angenommen, dass noch 100%-ige Fehler in der Grundgesamtheit vorliegen können. Der UEL-Faktor für die fehlerfreie Stichprobe erfährt deshalb keine Gewichtung. Es wird vorgeschlagen, dass für Grundgesamtheiten, bei denen der Prüfer sicher ist, dass keine vollständigen Fehlbewertungen vorliegen, den Faktor mit der höchsten anzunehmenden Fehlerrate zu gewichten. Finden sich in der Stichprobe dann jedoch Fehler, die der ursprünglichen Aussage im Sinne einer Abweichung der Fehlerrate nach oben widersprechen, muss die Gewichtung im Nachhinein angepasst werden (dies gilt natürlich nicht für eine Abweichung der Fehlerrate nach unten).[97]

5.2.4 Die Cell-Bewertung

Diese Art der Fehlerhochrechnung, die von Leslie/Teitlebaum/Anderson (1979), S. 135 ff., beschrieben wurde, kann nur in Verbindung mit der Cell-Auswahl angewandt werden, ist also abhängig von der Art der Auswahl der Elemente in die Stichprobe. Ihr Vorteil ist, dass sie für die Fehlerhochrechnung aus der zusätzlichen Schichtung des Cell-Sampling-Verfahrens Nutzen zieht.

---

[95] CPA steht für *Certified Public Accountant*.

[96] Vgl. Wampler/McEacharn (2005), S. 37 und das Excel-Template, welches dem Aufsatz in der Ausgabe des CPA Journals beigefügt ist.

[97] Wie in den Ausführungen zu der Durchschnittsfehlermethode angedeutet, halte ich dieses Vorgehen für weniger geeignet.

Da bei der Cell-Auswahl jeweils ein Dollar aus jeder Zelle ausgewählt wird, hängt das Risiko, einen Fehler nicht zu finden, von der Verteilung der Fehler in den Zellen ab. Da die wahre Fehlerverteilung jedoch unbekannt ist, wird in der Hochrechnung stets von der ungünstigsten Verteilung ausgegangen. Dies ist die Verteilung, für die das Stichprobenrisiko am größten ist. Wird eine Stichprobe gezogen, die keinen Fehler enthält, ist die Annahme, dass die in der Grundgesamtheit möglichen 100%-igen Fehler gleichmäßig in den Zellen verteilt sind, die konservativste. Bei gleichmäßiger Verteilung der Fehler in den Zellen ist die Wahrscheinlichkeit (im Prüfungssinn: das Risiko), eine fehlerfreie Stichprobe zu ziehen, am größten. Zur besseren Verdeutlichung sollte man sich zwei Extremfälle vorstellen: Bei Anwendung des *Cell-Samplings* würde ein fehlerhaftes Element, welches eine ganze Zelle füllt, auf jeden Fall in die Stichprobe gelangen, wogegen bei Gleichverteilung der Fehler in allen Zellen der Fall nicht ausgeschlossen ist, dass kein Fehler entdeckt wird.

Für den Nullfehlerfall ergibt sich damit dieselbe Hochrechnung wie bei der Fehlerreihungsmethode.

Für den Fall einer Einfehlerstichprobe, wobei das betroffene Element zu T % falsch bewertet ist, sind zwei Fälle als schlimmste Szenarien denkbar: Eine Zelle ist vollständig mit T%-Fehlern geladen, und 100%-Fehler sind gleichmäßig über alle anderen Zellen verteilt. Der zweite Fall ist, dass T%-Fehler gleichmäßig über die Zellen verteilt sind. Die Logik der Unterscheidung gilt für die Mehrfehlerstichprobe analog, wodurch schließlich die obere Fehlergrenze durch eine jeweilige Betrachtung des schlimmeren Falls aus den zwei genannten Alternativen für jeden zusätzlichen Fehler angestellt wird. Für die erste Möglichkeit – die vollständige Ladung einer Zelle mit T%-igen Fehlern und eine gleichmäßige Verteilung der 100%-igen Fehler im Rest der Grundgesamtheit (*„Load and Spread“*) – wird die obere Fehlergrenze wie folgt berechnet:

$$C_{LS} = \overline{I}(\lambda_{0,\,\alpha} + T/100).$$

Abbildung 2 visualisiert die Fehlerverteilung in den Zellen beim *„Load and Spread“*,

die Proportionen der Fehler in den Zellen wurden dabei überspitzt groß dargestellt.

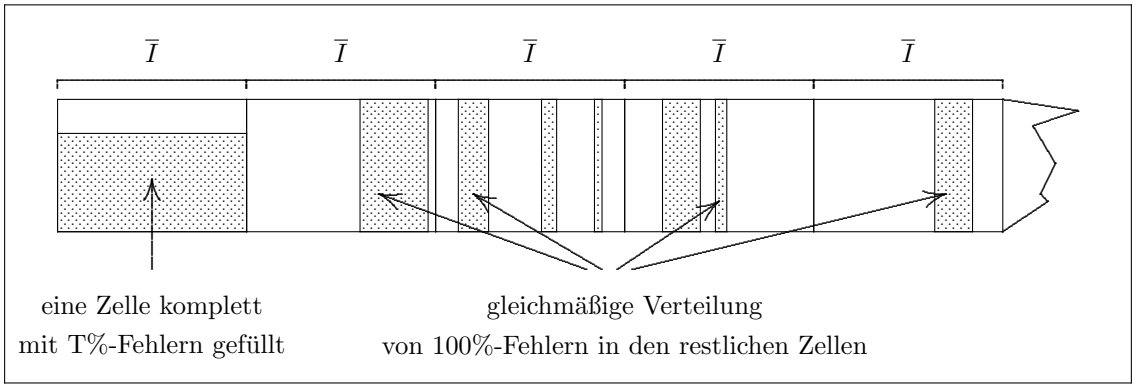


Abbildung 2: „*Load and Spread*“, in Anlehnung an Leslie et al. (1979), S. 140.

Für die gleichmäßige Verteilung von T%-igen Fehlern in der ganzen Grundgesamtheit („*Simple Spread*“) berechnet sich die obere Fehlergrenze analog zur Durchschnittsfehlermethode nach:

$$C_{SS} = \lambda_{1;\alpha} \cdot r_1 \cdot \overline{I}.$$

Die Verteilung der Fehler wird in Abbildung 3 dargestellt (die Fehler sind zu Verdeutlichungszwecken übertrieben groß gewählt).

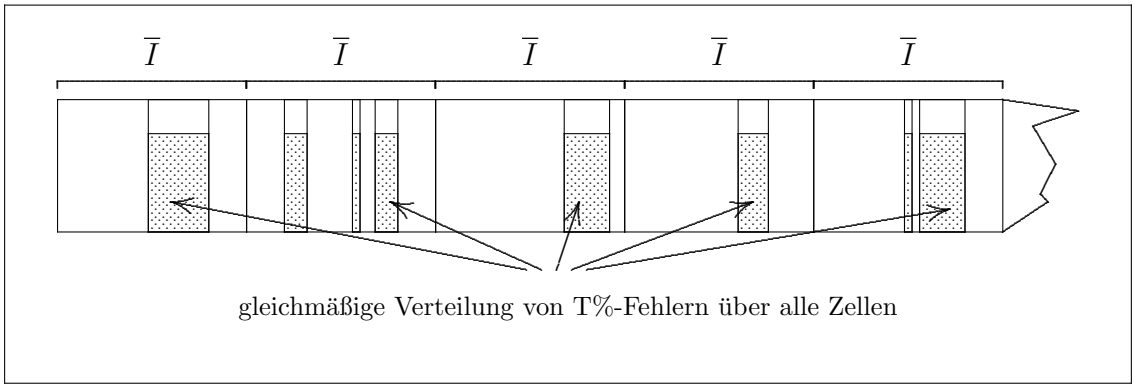


Abbildung 3: „*Simple Spread*“, in Anlehnung an Leslie et al. (1979), S. 140.

Für die Bestimmung der oberen Fehlergrenze muss dann der größere der Werte C_{LS} und C_{SS} ausgewählt werden. Es wird immer die höhere der Fehlergrenzen der vorhergehenden Stufe berechnet und dann der aktuelle Fehleranteil mit dem Stichprobenintervall multipliziert und hinzugerechnet. Dabei werden die Fehler wie bei der Fehlerreihungs-

methode vorher absteigend sortiert.

Zusammengefasst ergibt sich folgende iterative Formel, die jeweils das Maximum von *Load and Spread* und *Simple Spread* auf jeder Stufe ermittelt, zur Fehlerhochrechnung:

$$C_{Cell}(x) = \begin{cases} \overline{I} \cdot \lambda_{0;\alpha} \ , & \text{wenn} \quad x = 0 \\ \max\left\{\overline{I} \cdot \lambda_{x;\alpha} \cdot \overline{r}_x \ , \ C_{Cell}(x-1) + I \cdot r_x\right\}, & \text{wenn} \quad x > 0. \end{cases}$$

Hierbei stellt $\overline{r}_x$ das einfache arithmetische Mittel aus den x gefundenen Fehleranteilen dar.

Beispiel: Die Methode wird für die Beispielzahlen demonstriert:

$$\begin{aligned} C_{Cell}(0) &= \lambda_{0;0,05} \cdot 100.000 = 299.600 \\ C_{Cell}(1) &= \max\left\{100.000 \cdot 4,744 \cdot 0,5 \ , \ 299.600 + 0,5 \cdot 100.000\right\} \\ &= \max\{238.700 \ , \ 349.600\} = 349.600 \\ C - Cell(2) &= \max\{100.000 \cdot 6,296 \cdot 0,4 \ , \ 349.600 + 0,3 \cdot 100.000\} \\ &= \max\{251.840 \ , \ 379.600\} = 379.600. \end{aligned}$$

Für die Beispielzahlen und bei der gegebenen Wesentlichkeitsgrenze von € 300.000 könnte dem Prüffeld aufgrund des Stichprobenergebnisses bei Anwendung der Cell-Bewertung wiederum keine Ordnungsmäßigkeit bescheinigt werden. Der berechnete maximale Fehler liegt jedoch deutlich unter dem mit dem Stringer-Verfahren ermittelten.

Wird in Verbindung mit dem *Cell-Sampling* die Cell-Bewertung anstelle der Fehlerreihungsmethode angewandt, lässt sich der hochgerechnete Fehler reduzieren, ohne das Konfidenzniveau zu verringern.

5.2.5 Die Momenten-Methode – Moment-Bound

Die Idee der Momenten-Methode basiert auf der Schätzung einer Dichtefunktion $f_u(m)$ für den mittleren Fehleranteil m in der Grundgesamtheit. Dieser mittlere Fehleranteil gibt an, wie viel Prozent jede Geldeinheit im Durchschnitt falsch bewertet ist, und stellt demzufolge eine Verbindung aus Fehlerrate und Fehleranteil dar. Ist die Dichtefunktion ermittelt, so kann mit Hilfe ihrer Wahrscheinlichkeitsfunktion der Wert M gefunden

werden, für welchen ein bestimmtes Quantil $1-\alpha$ korrespondierend zu der gewünschten Irrtumswahrscheinlichkeit α erreicht wird:

$$\int_{-\infty}^{M} f_u(m)dm = 1 - \alpha.$$

Dabei geben Dworin und Grimlund (1984), S. 237, an, dass sich durch weit reichende Simulationsstudien ergab, dass die Form der Dichtefunktion für den mittleren Fehleranteil gut durch eine regularisierte Gammafunktion mit drei Parametern (A, B und D) approximiert werden kann:

$$f_u(m) \approx \frac{\left(\frac{m-D}{B}\right)^{A-1} \mathrm{e}^{-\frac{m-D}{B}}}{B\Gamma(A)}.$$

Es gilt, die Parameter A, B und D aus den durch die Stichprobenergebnisse ermittelten Schätzern für die ersten drei Momente des mittleren Fehleranteils UC_1, UC_2 und UC_3 zu bestimmen. Danach kann eine Approximation benutzt werden, um mit der geschätzten Funktion $f_u(m)$ den Wert M für ein gegebenes Konfidenzniveau zu berechnen.[98] Somit reduziert sich die Schätzung des Fehlers der Grundgesamtheit auf die Berechnung dreier Momente aus der Stichprobe.[99]

Die folgenden fünf Schritte zur Berechnung des *Moment-Bounds* könnten theoretisch auch durch eine Gleichung ausgedrückt werden. Zur besseren Veranschaulichung wird jedoch auch die Berechnung der Zwischenschritte gezeigt:

1. Es wird der durchschnittliche Fehleranteil $\overline{r}$ für die Stichprobe berechnet und ein hypothetischer Fehleranteil r^* konstruiert:

$$\begin{aligned} \overline{r} &= \sum_{i=1}^{x} \frac{r_i}{x} \\ r^* &= 0,81 \cdot (1 - 0,667 \tanh(10\overline{r})) \cdot \left(1 + 0,667 \tanh\left(\frac{x}{10}\right)\right). \end{aligned}$$

[98] Vgl. Dworin/Grimlund (1984), S. 237 f. über Kotz (1970), S. 186.

[99] Für eine Herleitung des Moment-Bounds wird auf Dworin/Grimlund (1984), S. 222-223 sowie 237-241, verwiesen.

In dieser Formel wurden die Koeffizienten (0,81; -0,667; 10; 0,667; 10) für den hypothetischen Fehleranteil numerisch von Dworin/Grimlund (1984) durch Tests an verschiedenen Grundgesamtheiten bestimmt. Der hypothetische Fehleranteil r^* wird wie ein weiterer gefundener Fehler in den folgenden Berechnungen verwendet. Wurden keine Fehler beobachtet, reduziert sich die Formel auf $r^* = 0,81$. Die Parameter wurden von den Begründern des Verfahrens so gewählt, dass die Verlässlichkeit des Verfahrens gewährleistet und die Fehlerhochrechnung für den Fall einer fehlerfreien Stichprobe mit dem *Multinomial Bound* und der Stringer-Methode vergleichbar ist.[100]

2. (a) Die nicht zentrierten Momente der Fehleranteilsverteilung (TN_j) werden berechnet (j steht für das j-te Moment):

$$TN_j = \frac{\left((r^*)^j + \sum_{i=1}^{x} r_i^j\right)}{x+1}, \quad j = 1, 2, 3.$$

(b) Die nicht zentrierten Momente der Verteilung der Fehlerrate RN_j werden berechnet:[101]

$$RN_1 = \frac{x+1}{n+2}, \quad RN_2 = \frac{x+2}{n+3} RN_1, \quad RN_3 = \frac{x+3}{n+4} RN_2.$$

3. Die Momente aus Schritt 2 werden zusammengeführt, um die nicht zentrierten Momente der Stichprobenverteilung des mittleren Fehlers UN_j zu berechnen:

$$\begin{aligned} UN_1 &= RN_1 \cdot TN_1 \\ UN_2 &= \frac{RN_1 \cdot TN_2 + (n-1) RN_2 \cdot TN_1^2}{n} \\ UN_3 &= \frac{RN_1 \cdot TN_3 + 3 \cdot (n-1) RN_2 \cdot TN_1 \cdot TN_2 + (n-1)(n-2) \cdot RN_3 \cdot TN_1^3}{n^2}. \end{aligned}$$

[100] Vgl. Dworin/Grimlund (1984), S. 221.

[101] Diese kommen aus der Beta-Verteilung:

$$f_R(p) = \frac{(n+1)!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

4. In diesem Schritt werden schließlich die in 3. ermittelten Momente in zentrale Momente (Momente über den Mittelwert) UC_j umgewandelt:

$$\begin{aligned} UC_2 &= UN_2 - UN_1^2 \\ UC_3 &= UN_3 - 3 \cdot UN_1 \cdot UN_2 + 2 \cdot UN_1^3. \end{aligned}$$

5. Die benötigten Parameter A, B und D der Gamma-Funktion werden dann wie folgt berechnet:[102]

$$A = 4 \cdot \frac{UC_2^3}{UC_3^2}, \quad B = \frac{0,5 \cdot UC_3}{UC_2}, \quad D = UN_1 - 2 \cdot \frac{UC_2^2}{UC_3}.$$

6. Letztendlich benutzen die Autoren die Wilson-Hilferty-Approximation,[103] um die obere Fehlergrenze auf einem Konfidenzniveau von $(1 - \alpha)$ zu berechnen:

$$M = D + A \cdot B \cdot \left(1 + \frac{Z_\alpha}{3\sqrt{A}} - \frac{1}{9A}\right)^3.$$

Z_α stellt dabei das α-Quantil der Normalverteilung dar.

Ein Vorteil des Verfahrens ist, dass es bei Vorliegen von sowohl Überbewertungen als auch Unterbewertungen die obere Fehlergrenze saldiert ausgibt. Diese muss in anderen Verfahren separat berechnet werden.[104]

Eine Abwandlung der Momentenmethode wurde von dem ehemaligen Wirtschaftsprüfungsunternehmen Arthur Andersen Ende der 1980er Jahre eingeführt.[105]

Beispiel: Die sechs Schritte werden für die bekannten Beispieldaten demonstriert:

1. $\overline{r} = \frac{1}{2} \cdot (0,3 + 0,5) = 0,4$

   $r^* = 0,81(1 - 0,667 \tanh(4))(1 + 0,667 \tanh(0,2)) = 0,3056$

2. (a) $TN_1 = 0,369, \quad TN_2 = 0,144, \quad TN_3 = 0,06$

   (b) $RN_1 = \frac{3}{113}, \quad RN_2 = \frac{2}{2147}, \quad RN_3 = \frac{2}{49381}$

[102] Dworin/Grimlund (1984), S. 237, über Kotz (1970), S. 186.

[103] Vgl. Johnson/Kotz (1970), S. 176.

[104] Vgl. Abschnitt 5.3.

[105] Vgl. Felix et al. (1990), S. 15.

3. $UN_1 = 0,00978, \quad UN_2 = 0,00016, \quad UN_3 = 0,000003$

4. $UC_2 = 0,000064, \quad UC_3 = 0,0000006$

5. $A = 2,848, \quad B = 0,0047, \quad D = -0,0037 \quad \Longrightarrow \quad M = 0,02501324.$

6. $C_{Mom.} = M \cdot 11.100.000 = 277.647.$

Für das gegebene Beispiel gibt der *Moment-Bound* die zweitkleinste obere Fehlergrenze $C_{Mom.}$ (ein kleinerer maximaler Fehler wurde nur durch die Durchschnittsfehlermethode erreicht). Würde das Prüfungsurteil auf der Grundlage der Momenten-Methode gefällt, könnte dem Prüffeld Ordnungsmäßigkeit bescheinigt werden. Dieses Ergebnis steht im Gegensatz zu den Erscheidungen, die auf der Grundlage der zwei vorhergehenden Methoden getroffen würden.

5.2.6 Weitere, häufiger diskutierte obere Fehlergrenzen

- **Der multinomiale Ansatz – Multinomial Bound**

  Wie der Name vermuten lässt, baut die Berechnung der oberen Fehlergrenze, des *Multinomial Bounds*, auf einer multinomialen Wahrscheinlichkeitsverteilung auf. Diese Verteilung beschreibt eine Zufallsvariable mit s verschiedenen Ausprägungen. Im Folgenden stellt π_i $(i = 1, \ldots, s)$ die Wahrscheinlichkeit dar, dass die Zufallvariable die i-te Ausprägung annimmt.

  Der Grundgedanke des Verfahrens basiert nun darauf, dass in einer Prüfung ein maximaler Fehler von zum Beispiel 100 Cent pro Euro angenommen wird, wodurch sich für den Fehleranteil 101 unterschiedliche Ausprägungen (kein Fehler und Fehler in Ein-Cent-Schritten) ergeben. Natürlich ist auch eine andere Definition von der Anzahl der Ausprägungen möglich (etwa in Zehn-Cent-Abschnitten oder in Schritten, die nur einen Bruchteil von einem Cent darstellen).[106] Die multinomiale Verteilung gibt die möglichen Fehlerausprägungen und die mit ihnen verbundenen Ziehungswahrscheinlichkeiten an. Der totale, sich aus Überbewertungen zusammensetzende Fehler $C_{multi.}$ für die Grundgesamtheit lässt sich damit wie folgt berechnen:

$$C_{multi.} = \frac{BW}{100} \sum_{i=1}^{100} i\pi_i,$$

[106] Vgl. Neter/Leitch/Fienberg (1978), S. 81.

wobei BW wieder den Gesamtbuchwert des Prüffelds darstellt. Für die Parameter π_i gilt für ein Konfidenzniveau von $1-\alpha$:

$$\sum_S \frac{n!}{z_0!z_1!\cdots z_{100}!}\prod_{i=0}^{100}\pi_i^{z_i} = \alpha.$$

In diesem Ausdruck stellen die z_i mögliche Ausprägungen des Sets S dar. Dieses ist die Menge aller möglichen Ausprägungen, die genauso extrem oder weniger extrem als die beobachtbaren Ergebnisse der Stichprobe sind. Es wird die folgende Maximierungsaufgabe für die Ermittlung der oberen Fehlergrenze gelöst (für ein gegebenes Konfidenzniveau sollen die Parameter π_i gefunden werden, welche den Fehler in der Grundgesamtheit maximieren):

$$C_{multi.} = \frac{BW}{100}\sum_{i=1}^{100} i\pi_i \quad \longrightarrow \quad Max!$$

unter den Nebenbedingungen, dass

$$\begin{aligned} \alpha &= \sum_S \frac{n!}{z_0!z_1!\cdots z_{100}!}\prod_{i=0}^{100}\pi_i^{z_i}, \\ \pi_i &\geq 0 \qquad (i=0,1,\ldots,100), \\ \sum_{i=0}^{100}\pi_i &= 1. \end{aligned}$$

Im Ergebnis wird so die obere Fehlergrenze $C^*_{multi.} = \frac{BW}{100}\sum_{i=1}^{100} i\pi_i^*$ bestimmt.

Das Set S, welches von Neter/Leitch/Fienberg (1978) vorgeschlagen wird,[107] führt dazu, dass, mit Ausnahme der mit den Fehleranteilen in der Stichprobe korrespondierenden π_i, π_0 und π_{100}, sämtliche π_i durch den Maximierungsprozess Null gesetzt werden.

[107] Die Autoren wählen in ihrem Aufsatz (S. 82 f.) das von ihnen als *„Step Down“* bezeichnete Set, welches (1) keine Ausprägungen mit mehr als der in der Stichprobe gegebenen Anzahl an Fehlern beinhaltet, (2) in dem die Summe der Fehleranteile in der Stichprobe nie überschritten wird und (3) keiner der Fehleranteile den korrespondierenden Fehleranteil in der Stichprobe überschreitet.

Ohne entsprechende Prüfungssoftware ist das Verfahren für den Prüfungsbereich zu kostenintensiv, wie auch von seinen Begründern später festgestellt wurde.[108] Die oberen Fehlergrenzen, die Leitch et al. in ihren Simulationen berechnen, übersteigen nie die Fehlergrenzen, die unter Zugrundelegung des *Stringer-Bounds* ermittelt wurden. In Fällen, in denen der durch die Fehlerreihungsmethode ermittelte Fehler die Wesentlichkeitsgrenze nicht übersteigt, führt die soeben beschriebene Methode zu keinem anderen Prüfungsergebnis.

In der Praxis hat sich das beschriebene Verfahren bislang nicht durchgesetzt.[109]

- **Quasi-Baysian-Bounds**

  Einen weiteren Ansatz zur Bestimmung von Fehlergrenzen, in welchem sowohl Über- als auch Unterbewertungen Berücksichtigung finden können, beschreibt McCray (1984) im Rahmen seines „Quasi-Bayesianischen Prüfungsrisikomodells für das *Dollar Unit Sampling*“.[110] Die Idee besteht darin, dass der mittlere Fehleranteil einer Buchungsgesamtheit durch eine Art Maximum-Likelihood-Funktion repräsentiert werden kann, welche die beobachtete Fehlerrate und die Fehleranteile in der Stichprobe widerspiegelt.[111] Für ein gegebenes Stichprobenergebnis wird der wahrscheinlichste Zustand der Grundgesamtheit ermittelt.[112]

  McCray berechnet für jeden vorher definierten möglichen Fehlerbetrag in einer Grundgesamtheit die Posterior-Wahrscheinlichkeiten auf der Grundlage des Konzepts von Bayes mit der Abwandlung, dass anstelle der Wahrscheinlichkeiten *Maximum-Likelihoods* benutzt werden. Die Posterior-Wahrscheinlichkeit für jeden Zustand ist durch

$$\mathcal{P}(\theta \Big| \cdot) \quad = \quad \frac{L(\theta_j)\mathcal{P}(\theta_j)}{\sum\limits_j L(\theta_i)\mathcal{P}(\theta_i)}$$

[108] Vgl. Leitch/Neter/Plante/Sinha (1982), S. 385.

[109] Der oben beschriebene Ansatz wurde von den Autoren Leitch/Neter/Plante/Sinha (1986) später für größere Fehleranzahlen modifiziert.

[110] McCray (1984): „A Quasi-Baysian Audit Risk Model for Dollar Unit Sampling“.

[111] Vgl. Dworin/Lowell (1986), S. 36.

[112] Dieser ist der Zustand, für den die Wahrscheinlichkeit für die gezogene Stichprobe maximiert wird.

gegeben. Hierbei ermittelt sich $L(\theta)$ aus der Maximierungsaufgabe:

$$L(\theta) = K \prod_{-100}^{100} (a_i + \pi_i)^{x_i} \longrightarrow \quad Max!$$

unter den Nebenbedingungen, dass

$$\begin{aligned} \sum_{i=-100}^{100} (a_i + \pi_i) &= 1, \\ \frac{BW}{100} \sum_{-100}^{100} i(a_i + \pi_i) &= \theta, \\ 0 \leq a_i, \quad \pi_i &\leq 1. \end{aligned}$$

In den Formeln ist θ der Gesamtfehler, $\mathcal{P}(\theta)$ die Prior-Wahrscheinlichkeit seines Auftretens, K eine Konstante, n der Stichprobenumfang, a_i der Anteil an Fehleranteilen der Größe t_i, welche bekannt sind oder für die Grundgesamtheit als existierend angenommen werden, BW der Ist-Buchwert, x_i die Anzahl der monetären Einheiten, die durch einen Fehleranteil von t_i charakterisiert sind, und π_i sind die Anteile an Fehleranteilen der Höhe t_i, welche die Wahrscheinlichkeit maximieren.[113]

Auf diese Weise können die Posterior-Wahrscheinlichkeiten berechnet werden, sobald die *Maximum-Likelihoods* für jeden möglichen Zustand berechnet wurden. Eine obere und untere Fehlergrenze auf einem bestimmten Konfidenzniveau kann dann ermittelt werden.[114]

5.3 Die Verrechnung von in der Stichprobe gefundenen Unterbewertungen

Die Anwendung des *Dollar Unit Samplings* unterstellt, dass sich Überbewertungen eher in größeren Positionen „verstecken“. Daraus leitet sich die Eignung des Verfahrens für

[113] Vgl. McCray (1984), S. 37.

[114] Vgl. McCray (1984), S. 38. So wie das Modell hier formuliert wurde, werden Fehler, die von 100%-iger Überbewertung bis hin zu 100%-iger Unterbewertung reichen können, angenommen.

die Prüfung von Überbewertungen ab.[115] Dies schützt das Verfahren natürlich in keiner Weise davor, dass in einer Grundgesamtheit auch Unterbewertungen vorliegen können und unterbewertete Positionen in die Stichprobe gelangen. Für den Fall, dass in einem Prüffeld verstärkt Unterbewertungen erwartet werden, sind bereits modifizierte Verfahren für die Auswahl der Elemente in die Stichprobe entwickelt worden,[116] die jedoch nicht Fokus dieses Abschnitts sein sollen und auch noch nicht in die Prüfungspraxis implementiert wurden.

Die Fragestellung dieses Absatzes ist vielmehr, wie gefundene Unterbewertungen auf die Grundgesamtheit projiziert werden sollen. Hierfür gibt es grundsätzlich zwei Ansätze, die jeweils auf den Annahmen des Prüfers aufbauen. Der erste Ansatz geht davon aus, dass die Fehlerrichtung in dem Prüffeld unbekannt ist. Der Prüfer hat keine klare Erwartung darüber, ob es sich bei dem Gesamtfehler in dem Prüffeld (dem Nettofehler, der sich aus der Saldierung von Über- und Unterbewertungen ergeben würde) um eine Netto-Unterbewertung oder Netto-Überbewertung handelt. Steht die Fehlerrichtung nicht fest, werden zwei separate Konfidenzgrenzen für Über- und Unterbewertungen gebildet. Diese Fehlergrenzen werden dann jeweils um den Punktschätzer für die gegenseitige Fehlerrichtung bereinigt, sodass sich zwei Netto-Fehlergrenzen ergeben.[117] Aufgrund dieser wird eine Aussage der Art „Mit 95 % Sicherheit liegt die obere Fehlergrenze von + € 1.000 Überbewertungen über dem wahren Fehler, und mit 95 % liegt die berechnete obere Fehlergrenze für Unterbewertungen von − € 700 unter dem tatsächlichen Unterbewertungsbetrag“ möglich. Es ist klar, dass eine Aussage, die jeglicher Information über die Richtung des Netto-Fehlers entbehrt, für Prüfungszwecke keinen Sinn macht, da die Auswirkungen der Fehler auf die Darstellung der Vermögens- und Ertragslage so nicht abgesehen werden können.

Der zweite Ansatz geht deswegen davon aus, dass die Richtung des Netto-Fehlers in der Grundgesamtheit bekannt ist (etwa: Der Prüfer geht davon aus, dass, wenn Unter- und Überbewertungen in der Grundgesamtheit saldiert würden, eine Netto-Überbewertung

[115] Guy (2002), S. 207. Guy warnt auch davor, das Dollar-Unit-Sampling-Verfahren zur Prüfung von Unterbewertungen zu verwenden.

[116] Vgl. Wolz (2004).

[117] Vgl. Leslie/Teitlebaum/Andrson (1979), S. 130.

ausgegeben würde). Dieser zweite Ansatz findet auch in der Prüfungspraxis Anwendung. Wird das *Dollar Unit Sampling* angewandt, wird von einer Netto-Überbewertung ausgegangen.[118] Für die Überbewertungen wird nach diesem Ansatz eine obere Fehlergrenze nach einer der oben genannten Methoden (zumeist der Fehlerreihungsmethode oder der Cell-Bewertung) berechnet und dann eine Punktschätzung für die Unterbewertung aus der Stichprobe gegengerechnet. Diese Punktschätzung (MLE_U[119]) ergibt sich dabei einfach als Summe der negativen Fehleranteile multipliziert mit dem durchschittlichen Stichprobenintervall:[120]

$$MLE_U = \overline{I} \cdot \sum_{i=1}^{u} r_i \qquad \forall \quad r_i < 0.$$

Für die Methoden des *Moment-Bounds* und des *Quasi-Bayesian-Bounds* werden die oberen Fehlergrenzen bereits als Netto-Fehlergrenzen ausgegeben.

6 Auswirkungen der Fehlerhochrechnungsmethoden auf das Prüfungsurteil – Eine Simulationsstudie

6.1 Ziele der Studie

Der vorherige Abschnitt macht deutlich, dass das Ergebnis der Fehlerhochrechnung beim *Dollar Unit Sampling* stark mit dem verwendeten Hochrechnungsverfahren variiert. Die Beispielrechnungen verdeutlichen auf simple Weise, dass ein Prüfungsurteil davon abhängig sein kann, welche Methode zur Fehlerhochrechnung verwendet wurde. Im Folgenden sollen deswegen einige ausgewählte Verfahren anhand einer Simulationsstudie einem Vergleich unterzogen werden. Konkret möchte ich auf der Grundlage von mit dem Sieve-Sampling-Verfahren gezogenen Stichproben die Durchschnittsfehlermethode, die Fehlerreihungsmethode (*„Stringer-Bound“*) und die Momenten-Methode (*„Moment-Bound“*) miteinander vergleichen. Die drei Methoden wähle ich aufgrund ihrer Relevanz für die Anwendung in der Wirtschaftsprüfung: Eine Variante des *Moment-*

[118] Vgl. ISA 530, 39.

[119] MLE steht für „Most Likely Error“.

[120] Vgl. CIPFA (1991), S. 71.

Bounds wurde von der Wirtschaftsprüfungsgesellschaft Arthur Andersen Ende der 1980er Jahre als Methode implementiert.[121] Der *Stringer-Bound* wird von dem AICPA als einzige Fehlerhochrechnungsmethode im Dollar-Unit-Sampling-Verfahren vorgestellt,[122] von dem CPA[123]-Journal mit beigelegter Excel-Kalkulations-datei angepriesen[124] und von Wirtschaftsprüfungsunternehmen in der Praxis entsprechend angewandt. Die Durchschnittsfehlermethode wird in der aktuellen deutschen Literatur von ihrer Konzeption her als das zu bevorzugende Verfahren dargestellt.[125]

Natürlich hat es für die Hochrechnungsverfahren im *Dollar Unit Sampling* schon einige Simulationsstudien gegeben.[126] Die vorliegende Arbeit unterscheidet sich jedoch von den meisten dieser Studien in folgenden Punkten:

1. Existierende Simulationsstudien zu Hochrechnungen im *Dollar Unit Sampling* gehen fast ausschließlich von simulierten Datensätzen für ihre Untersuchungen aus.[127] Der vorliegende Vergleich unterscheidet sich von diesen durch die Verwendung von Originalbuchungsgesamtheiten, sodass keine Verteilungsannahmen für die zu prüfenden Datensätze getroffen werden müssen.

2. Während in anderen Studien meist die reine Zufallsauswahl oder das *Cell-Sampling* im Vordergrund stehen, bediene ich mich der Methode des *Sieve-Samplings* für die Ziehung der Stichproben.

3. Aufgrund der Relevanz unterschiedlicher Konfidenzniveaus für die Praxis untersuche ich neben dem 95%-Niveau auch das 80%- und 50%-Niveau, während sich die mir bekannten Studien vorrangig auf das erstgenannte Konfidenzlevel beziehen.

4. Für Fehlerraten im Bereich von 1 % bis 10 % werden verschiedene Verhältnisse

[121] Vgl. Felix/Grimlund/Koster/Roussey (1990).

[122] Vgl. AICPA (1999), S. 72.

[123] CPA steht für *Certified Public Accountant*.

[124] Vgl. Wampler/McEacharn (2005), S. 36 ff.

[125] Vgl. Wolz (2003), S. 127.

[126] U. a. Neter (1978), Reneau (1978), Leitch/Neter/Plante/Sinha (1982), Dworin/Grimlund(1986), Horgan (1996), Wolz (2004).

[127] Vgl. aktuell Wolz (2004).

von Über- und Unterbewertungen betrachtet und somit die Anwendbarkeit der einzelnen Hochrechnungsverfahren insbesondere auch für den Unterbewertungsfall überprüft.

Durch die genannten Unterschiede und Abwandlungen soll die bestehende Literatur zu den Hochrechnungsverfahren ergänzt werden.

6.2 Verwandte Daten – Ist-Grundgesamtheiten

Die Daten für die vorliegende Studie wurden mir von zwei Wirtschaftsprüfungsunternehmen und einem Potsdamer Unternehmen zur Verfügung gestellt. Der Kontakt zu den Wirtschaftsprüfungsunternehmen wurde deshalb gesucht, um möglichst unterschiedliche Datensätze für die Untersuchungen zu erhalten. Die Mandanten der von mir angesprochenen Unternehmen unterscheiden sich in Bezug auf Unternehmensgröße, Vermögenslage und Betätigungsfeld. Für die Studie erwiesen sich jedoch nur 15 Datensätze nutzbar: Neun Umsatzerlöskonten, drei Debitorensaldenlisten, Sollbuchungen für jeweils ein Debitoren- und ein Kreditorenkonto sowie eine Bestandsliste für Roh-, Hilfs- und Betriebsstoffe. Andere Konten mussten aufgrund ihrer Größe oder der Verteilung der Buchungsbeträge aussortiert werden. So wurden Konten mit weniger als 200 Buchungen aufgrund des geplanten Stichprobenumfangs von $n = 100$ aussortiert, Jahreskonten mit mehr als 15.000 Buchungen mussten aufgrund von mangelnden Rechenkapazitäten aufgespalten werden. Konten, auf denen die 50 größten Buchungen über 80 % des Gesamtbuchwerts ausmachen, wurden ebenfalls nicht berücksichtigt, da diese in der Prüfungspraxis für eine mathematisch-statistische Stichprobenziehung nicht in Frage kämen. Nullsalden wurden entfernt, da deren Wahrscheinlichkeit, im Rahmen des Dollar-Unit-Sampling-Verfahrens in die Stichprobe zu gelangen, sowieso Null beträgt.

Auch wenn letztlich die Zahl der vorliegenden Datensätze absolut gesehen gering ist, relativiert sich die Anzahl der Konten und Saldenlisten bei einem Blick darauf, was bisher in der Literatur an Datensätzen im Zusammenhang mit dem *Dollar Unit Sampling* verwendet wurde. Eine Studie von Horgan (1996) benutzt zum Beispiel Originaldaten für die Simulation von Ziehungsverfahren. In ihrer Arbeit werden jedoch nur zwei

Buchungskonten für die Simulation herangezogen. Fast eine Ausnahmestudie ist die von Neter/John/Leitch (1985), welche im Kern jeweils zehn Konten für die Untersuchung von Fehlerverteilungen in Forderungskonten und für Inventar verwenden. Die meisten Simulationsstudien beschränken sich jedoch auf die Untersuchung von simulierten Grundgesamtheiten, wobei auch hier letztendlich die Anzahl der verwendeten Grundgesamtheiten nach der Ausgestaltung der Gesamtheiten mit unterschiedlichen Fehlermustern nicht (wesentlich) größer als in der vorliegenden Arbeit ist.[128]

Die Ist-Grundgesamtheiten stellen Buchungsgesamtheiten dar, wie sie der Prüfer in einer Jahresabschlussprüfung vorfinden könnte. Sie sind mit bestimmten Fehlern versehen, die im Laufe der Prüfung aufgedeckt werden können. Die Daten wurden vor ihrer Verwendung für die Stichprobensimultion aufgearbeitet, wie es üblicherweise in einer reellen Prüfungssituation der Fall wäre. So wurden auf jedem Konto die Elemente entfernt, deren Buchwert mehr als 1 % des gesamten Buchwertes ausmacht (das sogenannte Topstratum, welches in jedem Fall vollständig geprüft wird).[129] Weiterhin wurden offensichtliche, sich ausgleichende Fehlbuchungen entfernt, um den Gesamtbuchwert nicht künstlich aufzublähen. Wurde im Laufe eines Tages eine Buchung getätigt und diese am selben Tag offensichtlich zurückgebucht (etwa wegen versehentlicher Doppelbuchung), wurde sie aus der Grundgesamtheit entfernt. Für die Saldenlisten wurden nur die Debitoren mit positiven Salden betrachtet, für die Umsatzerlöskonten nur die Habenbuchungen für die Untersuchung herangezogen. Diesem Vorgehen liegt die Überlegung zu Grunde, dass die restlichen Positionen (etwa kreditorische Debitoren und Abgrenzungs- oder Skontibuchungen) entweder aufgrund ihrer Beschaffenheit separat oder aufgrund der Doppik im Rechnungswesen auf einem anderen Konto geprüft würden. Nach den Veränderungen der Konten spiegelt die Datenausgangslage meines Erachtens die Situation in einer reellen Prüfung realistisch wider. Die Anzahl der Positionen auf den Saldenlisten und Buchungskonten rangiert zwischen 471 und 10.754. Die Gesamtbuchwerte reichen von knapp über € 200.000 bis fast € 190.000.000. Auch wenn es sich durchweg um rechtsschiefe Verteilungen handelt, variiert der Grad

[128]Vgl. zum Beispiel Neter (1978), Leitch et al. (1982), McCray (1984), Dworin (1986). Eine Ausnahme bildet Wolz (2003 und 2004).

[129]Vgl. CIPFA (1991), S. 49, sowie Prüfungsstandards mehrerer Wirtschaftsprüfungsgesellschaften.

der Schiefe beträchtlich. Auch die Größe der Buchungen variiert stark: ist die durchschnittliche Buchung auf zwei Konten mit fast gleichem Gesamtbuchwert von ca. neun Millionen Euro auf dem einen nur rund € 2.600 hoch, wird auf dem zweiten Konto ein arithmetisches Mittel von über € 19.000 erreicht. Die Details zu den Eigenschaften der Buchungsgesamtheiten (Gesamtbuchwert, Umfang der Grundgesamtheit, arithmetisches Mittel, Median, Standardabweichung, Schiefe, Minimum sowie Maximum der Buchungen bzw. Salden) können zusammen mit den dazugehörigen Histogrammen im Anhang eingesehen werden.

6.3 Festlegung der Fehler – Soll-Grundgesamtheiten

6.3.1 Vorbemerkungen

Sowohl aus rechtlichen als auch aus wirtschaftlichen Gründen ist für die vorliegende Arbeit keine Vollprüfung der oben beschriebenen Grundgesamtheiten zur Ermittlung der Sollbuchungswerte möglich. Aus diesem Grunde muss auf Annahmen und Erfahrungswerte für die in Buchungsgesamtheiten vorhandenen Fehler zurückgegriffen werden.

Für jede der Ist-Grundgesamtheiten, die im vorherigen Abschnitt beschrieben wurden, werden mehrere Soll-Grundgesamtheiten erzeugt. Dabei sind vier Merkmale für die Fehler festzulegen: Die Fehlerrate, das Verhältnis von Über- und Unterbewertung, der Fehleranteil in einem fehlerhaften Element und die Verteilung von Fehlern. Laut empirischen Studien können die Charakteristika in Abhängigkeit davon, welche Kontenart untersucht wird, variieren. Jedoch sind diese Charakteristika gerade für den Debitoren- und Umsatzerlösbereich, welche hier hauptsächliche untersucht werden, sehr ähnlich.[130]

6.3.2 Fehlerrate

Generell wird in Simulationsstudien eine Fehlerrate von unter zehn Prozent unabhängig von der Kontenart für realistisch erachtet.[131] Empirische Studien zu Fehlerraten bestäti-

[130] Vgl. Ham/Losell/Smielauskas (1985), S. 396 ff., Neter/Johnson/Leitch (1985), S. 491.

[131] Vgl. Wolz (2004), S. 68, Steiger (1998), S. 140, Reneau (1978), S. 671.

gen diese Annahme.[132] Ham et al. fanden speziell für Umsatzerlöse in 90 % und für Debitorensalden in 66 % der Fälle Fehlerraten, die 10 % nicht übersteigen.[133] Für den Großteil der durch sie untersuchten Umsatzkonten lagen die Fehlerraten unter fünf Prozent. Für die folgende Simulationsstudie werden deshalb Fehlerraten von 1, 2, 5 und 10 % angenommen.

6.3.3 Über- und Unterbewertungen

Das Dollar-Unit-Sampling-Verfahren soll auf Prüffelder Anwendung finden, in denen eher Überbewertungen vermutet werden. Für jede oben genannte Fehlerrate wird dementsprechend zuerst angenommen, dass nur Überbewertungen in dem Prüffeld vorliegen und damit implizit die Annahme des Prüfers bestätigt wird. Dieses Vorgehen findet bspw. Bestätigung in der Studie von Horgan (1996), die für die von ihr untersuchten Grundgesamtheiten lediglich Überbewertungen feststellt.[134]

In einem nächsten Schritt werden Unterbewertungen, die 20 % der fehlerhaften Elemente betreffen, eingestreut. In einem letzten Schritt werden auch 40 % Unterbewertungen zugelassen. Die unterschiedlichen Ausprägungen der Zusammensetzung von Über- und Unterbewertungen werden wiederum empirischen Studien entnommen, in denen sich für den Debitorenfall über 20 % und für Umsätze über 40 % der Fehleranteile negativ erwiesen.[135]

6.3.4 Fehleranteil

Für jede oben spezifizierte Fehlerrate und Zusammensetzung von Über- und Unterbewertungen werden zufällig Fehleranteile von bis zu 100 % erzeugt und auf die als fehlerhaft klassifizierten Elemente – sowohl auf Über- als auch auf Unterbewertungen – zufällig verteilt. Hierbei stütze ich mich auf die Fehleranalyse von Johnson et al. (1981), S. 289 ff., in welcher keine Beziehung zwischen Fehleranteil und Höhe des Buchwertes

[132] Hogan (1997), S. 218, Johnson/Leitch/Neter (1981), S. 276.

[133] Ham/Losell/Smieliauskas (1985), S. 397.

[134] Vgl. Horgan (1996), S. 218.

[135] Vgl. Ham/Losell/Smieliauskas (1985), S. 397.

festgestellt werden konnte. Für die Überbewertungen wird der maximale Fehleranteil bei 100 % liegend angenommen. Von dem Fall einer über 100% hinausgehenden Überbewertung (wenn ein ausgewiesener Debitor in Wirklichkeit ein kreditorischer Debitor oder ein Kreditor ist) soll im Rahmen dieser Studie abstrahiert werden. Für die verschiedenen Zusammensetzungen von Über- und Unterbewertungen wird zuerst angenommen, dass der Fehlbtrag bei Unterbewertungen höchstens bei 100 % liegt. Da für Unterbewertungen jedoch nahe liegend ist, dass diese auch über 100%-ige Fehleranteile aufweisen können, werden für den Fall, in dem 20 % der fehlerhaften Elemente Unterbewertungen sind, in einem nächsten Schritt Fehleranteile von bis zu 300 % simuliert.[136]

6.3.5 Fehlerverteilung

Es wird angenommen, dass die Fehler in der Grundgesamtheit gleichverteilt sind. Der Fakt, ob ein Element richtig oder falsch ausgewiesen ist, soll somit als unabhängig von der Größe der Buchung oder des Saldos angenommen werden.[137] Eine Gleichverteilung der Fehler gilt als die konservativste Annahme für das *Dollar Unit Sampling*[138] und bietet bei Unsicherheit über die wahre Fehlerverteilung meines Erachtens deswegen eine gute Ausgangsbasis für den Vergleich.

6.4 Ziehungsmethode und Stichprobenumfang

Als Ziehungsmethode für die Simulation habe ich das *Sieve-Sampling* gewählt. Durch die Anwendung des *Sieve-Samplings* kann ich die Problematik der Doppelziehung von Elementen umgehen. Der dadurch von Stichprobe zu Stichprobe schwankende Stichprobenumfang gefährdet die Ergebnisse nicht, da die Hochrechnungen jeweils anhand derselben gezogenen Stichproben durchgeführt werden. Somit sind die Ausgangsvor-

[136]Nach Johnson/Leitch/Neter (1981), S. 291, werden vor allem für den Bereich des Inventars große Fehleranteile festgestellt.

[137]Alternativ ließe sich sowohl in die Richtung einer negativen (etwa: „Bei kleinen Beträgen werden eher Fehler gemacht, da der Buchhalter die Sachverhalte mit weniger Sorgfalt behandelt“) als auch einer positiven Korrelation zwischen Buchwert und Fehlerhaftigkeit (etwa: „Hinter einem großen Debitorensaldo stehen mehr Geschäftsvorfälle, und je mehr Geschäftsvorfälle zusammengefasst werden, desto wahrscheinlicher wird ein Fehler“) argumentieren.

[138]Vgl. Grimlund (1988), S. 83 oder auch die Ausführungen zum *Cell-Sampling*.

aussetzungen für jedes Hochrechnungsverfahren die gleichen. Es werden Stichprobenumfänge von 100 Elementen angestrebt. Diese Anzahl wird im Durchschnitt auch erreicht.

Der Stichprobenumfang wurde für die Zwecke der Simulation einfach festgelegt und nicht an Erwartungen des Prüfers an das Stichprobenergebnis, die Fehlerhaftigkeit der Grundgesamtheit oder eine Wesentlichkeitsgrenze geknüpft, wie es bei einer Stichprobenumfangsplanung der Fall wäre.[139]

Das *Sieve-Sampling* hat gegenüber anderen Auswahlverfahren den praktisch relevanten Vorteil, dass die einzelne Geldeinheit nicht zum Prüfungselement zurückverfolgt werden muss. Es ist somit immer dann von Vorteil, wenn keine Prüfungssoftware verwendet wird, die die Stichprobenziehung automatisch durchführt, oder wenn die zu prüfenden Beträge nicht in elektronischer Form vorliegen (bzw. sich nicht elektronisch in die Prüfungssoftware einlesen lassen).

6.5 Auswertung der Ergebnisse

6.5.1 Auswertungskriterien

Für den Prüfungsbereich sind in Hinblick auf die Fehlerhochrechnung zwei Faktoren von besonderem Interesse: Auf der einen Seite steht die Verlässlichkeit der berechneten oberen Fehlergrenze in Hinblick auf das angegebene Konfidenzniveau. Auf der anderen Seite ist die Nähe dieser berechneten oberen Fehlergrenze zu dem tatsächlichen Fehler in der Grundgesamtheit von Bedeutung. Diese Faktoren werden im Folgenden anhand zweier Statistiken für die unterschiedlichen Verfahren einer Kontrolle unterzogen.

Für die Verlässlichkeit des Verfahrens bietet sich die Verwendung der relativen Trefferhäufigkeit T an. Diese gibt den Anteil der gezogen Stichproben an, bei denen die berechnete obere Fehlergrenze über dem wahren Fehler liegt.[140] Dieses Vorgehen unterliegt der Annahme, dass der Prüfer insgesamt von einer Überbewertung des Prüffeldes ausgeht. Die untere Fehlergrenze, wie sie der unteren Grenze eines Konfidenzinter-

[139] Vgl. Grimlund (1988), S. 79.

[140] Vgl. Grimlund/Felix (1987), S. 456.

valls entspricht, wird nicht betrachtet. Die relative Trefferhäufigkeit ist ein Schätzer für das Konfidenzlevel. Eine Stichprobenmethode wird als verlässlich eingestuft, wenn die beobachtete relative Trefferhäufigkeit nicht mehr als zwei Standardfehler unter dem genannten Konfidenzlevel liegt.[141] Für 3.750 Wiederholungen sollte die relative Trefferhäufikeit der Verfahren demnach für 95 % bei 0,943, für 80 % bei 0,787 und für das 50 %-Level bei 0,484 liegen, wenn der Standardfehler S nach der Formel

$$S = \sqrt{\frac{\alpha(1-\alpha)}{Q}}$$

berechnet wird, wobei Q für die Anzahl gezogener Stichproben steht (hier $Q = 3.750$).

Das zweite Kriterium wird anhand der Genauigkeit G gemessen. Diese gibt die Differenz zwischen berechneter oberer Fehlergrenze und wahrem Fehler relativ zum wahren Fehler an.

$$G = \left| \frac{obere\ Fehlergrenze\ -\ wahrer\ Fehler}{wahrer\ Fehler} \right|.$$

Je näher G demnach an Null liegt, als umso genauer ist das Verfahren einzuschätzen.

Im Folgenden werden die Ergebnisse für die einzelnen Methoden und im Vergleich dargestellt.

6.5.2 Auswertung für die Durchschnittsfehlermethode

Die Simulationsergebnisse für die Durchschnittsfehlermethode werden in Tabelle 2 dargestellt. Hinter jedem Paar aus relativer Trefferhäufigkeit und Genauigkeit stehen 3.750 ($= 250 \cdot 15$) gezogene Stichproben. Unter Beachtung dessen, dass die relative Trefferhäufigkeit zwei Standardfehler unter der genannten Konfidenzgrenze liegen darf, zeigt sich, dass das gewünschte Konfidenzniveau in sieben von 16 Fällen (auf dem 95%-Niveau) und in zwei von 16 Fällen (auf dem 80%-Niveau) nicht erreicht wird. Auf dem 50%-Niveau liegt die relative Trefferhäufigkeit immer über dem genannten Kon-

[141] Vgl. Horgan (1996), S. 218.

fidenzniveau. Dabei fällt auf, dass es die größten Abweichungen von dem nominalen Konfidenzniveau nach unten für die Fälle gibt, wo Unterbewertungen bis 300 % zugelassen werden. Dominieren Überbewertungen stark, sind die Abweichungen zu dem um zwei Standardfehler verringerten Konfidenzniveau eher gering (zum Beispiel 0,933, wo es 0,943 hätte sein müssen).

Das Verfahren ist genauer, je größer der Anteil an Überbewertungen im Verhältnis zu Unterbewertungen und je größer die Fehlerrate ist. Diese Beobachtung bestätigt die Erwartung, dass jede zusätzlich gefundene Überbewertung die obere Fehlergrenze durch die degressiv wachsende obere Fehlerintensität $\lambda_{x;\alpha}$ erhöht, während von ihr jeweils nur der Punktschätzer für die Unterbewertungen (MLE_U) abgezogen wird.

Als problematisch sehe ich die relativ hohen Werte für das Genauigkeitsmaß. Für Grundgesamtheiten mit einer Fehlerrate von 1 % liegt G bei ca. 4,4, d. h. dass die berechnete Fehlergrenze hier im Schnitt 5,4 mal so hoch wie der wahre Fehler in der Grundgesamtheit ist. In der Prüfung wird meist eine Wesentlichkeitsgrenze von bis zu 5 % festgelegt,[142] was bedeutet, dass eine Grundgesamtheit noch als frei von wesentlichen Fehlern eingestuft werden kann, wenn die (auf das Prüffeld projizierte) obere Fehlergrenze im Verhältnis zum Gesamtbuchwert weniger als 5 % ausmacht. Werden in einer Stichprobe einige wenige Fehler gefunden, wird die Wesentlichkeitsgrenze schnell überschritten, obwohl der wahre Fehler um ein Vielfaches unter der Wesentlichkeitsgrenze liegt. Weiß der Prüfer nun um die Konservativität des Verfahrens, wird er kein klares Prüfungsurteil aufgrund der Stichprobe treffen können. Verlässt er sich dagegen auf die obere Fehlergrenze, lehnt er ein eigentlich ordnungsgemäßes Prüffeld schnell als nicht ordnungsgemäß ab.

Unakzeptabel sind die Werte für die Genauigkeit für kleine Fehlerraten und einen größeren Anteil an Unterbewertungen: Hier reicht G bis zu einem Wert von über 36 und lässt damit kein sicheres Prüfungsurteil mehr zu.

Eine Zusammenfassung der Ergebnisse für die Fehlerreihungsmethode kann der Tabelle 2 entnommen werden.

[142] Vgl. Grant/Depree/Grant (2000), S. 42.

Durchschnittsfehlermethode						
	95 %		**80 %**		**50 %**	
Fehlerrate	T	G	T	G	T	G
	Nur Überbewertungen					
1 %	0,933	4,360	0,902	2,321	0,813	0,841
2 %	0,953	2,301	0,898	1,215	0,626	0,374
5 %	0,944	1,097	0,839	0,575	0,588	0,131
10 %	0,949	0,696	0,837	0,364	0,565	0,066
	Über- und Unterbewertungen 80/20					
1 %	0,943	8,471	0,916	4,533	0,850	1,753
2 %	0,949	3,999	0,899	2,112	0,692	0,674
5 %	0,906	2,423	0,807	1,271	0,590	0,315
10 %	0,942	1,010	0,827	0,535	0,578	0,114
	Über- und Unterbewertungen 60/40					
1 %	0,977	36,208	0,967	19,399	0,851	7,902
2 %	0,945	12,373	0,902	6,570	0,736	2,368
5 %	0,901	4,274	0,794	2,244	0,593	0,614
10 %	0,919	2,941	0,800	1,569	0,578	0,383
	Über- und Unterbewertungen (300) 80/20					
1 %	0,971	30,359	0,914	16,126	0,843	5,999
2 %	0,908	14,944	0,831	7,908	0,700	2,586
5 %	0,884	4,033	0,778	2,123	0,608	0,531
10 %	0,863	2,886	0,748	1,534	0,598	0,342

Tabelle 2: Trefferrate (T) und Genauigkeit (G) für die Durchschnittsfehlermethode auf dem 95-, 80- und 50%-Niveau.

6.5.3 Auswertung für die Fehlerreihungsmethode

Wie erwartet, hält die konservative Fehlerreihungsmethode ihr Konfidenz-Versprechen für alle betrachteten Konfidenzniveaus und Fehlerzusammensetzungen. Diese hohe Verlässlichkeit geht jedoch zu Lasten der Genauigkeit, welche in allen Fällen unter der für die Durchschnittsfehlermethode ausgegebenen liegt. Wie bei der Durchschnittsfehler-

methode kann beobachtet werden, dass sich die Genauigkeit verbessert, je weniger Unterbewertungen im Prüffeld enthalten sind und je höher die Fehlerrate ist.

Es fällt besonders auf, dass die berechnete relative Trefferhäufigkeit meist weit über dem angestrebten Konfidenzniveau liegt. So ist diese für eine angestrebte Sicherheit von 50 % fast immer über 60 % – für das Konfidenzlevel von 80 % liegt sie meist über 90 %. Dies ist vielleicht auf den ersten Blick positiv, da der Prüfer dadurch selten zu der Auffassung gelangen dürfte, dass ein Prüffeld mit wesentlichen Fehlern fehlerfrei ist. Das Risiko der Klassifizierung eines ordnungsgemäßen Prüffeldes als nicht frei von wesentlichen Fehlern ist jedoch groß: Für kleine Fehlerraten ist eine Aussage mit einer nominellen Sicherheit von 95 % kaum möglich. Hier sind die oberen Fehlergrenzen 4,5 bis 48,9 mal so groß wie der tatsächliche Fehler.

Am besten scheint die Methode noch für die Fälle zu funktionieren, in denen in der Grundgesamtheit nur Überbewertungen vorliegen (hierfür wurde sie schließlich auch ursprünglich entwickelt). Auf dem 80%-Niveau werden hier moderate Genauigkeiten von 0,6 bis 3,3 erreicht, wobei die relative Trefferhäufigkeit noch der für das 95%-Niveau entspricht.[143]

Eine Zusammenfassung der Ergebnisse für die Fehlerreihungsmethode befindet sich in Tabelle 3.

[143]Das heißt natürlich nicht, dass der Prüfer generell nur in fünf von hundert Fällen mit seinen Berechnungen den wahren Fehler unterschreitet, wenn er nur eine Irrtumswahrscheinlichkeit von 20 % angegeben hat. Zu einer derartigen Verallgemeinerung sind weitaus mehr Simulationsstudien und variierende Annahmen über Grundgesamtheiten und Fehlerverteilungen notwendig.

Fehlerreihungsmethode						
	95 %		**80 %**		**50 %**	
Fehlerrate	T	G	T	G	T	G
	Nur Überbewertungen					
1 %	1,000	6,353	1,000	3,390	0,959	1,295
2 %	1,000	3,581	0,998	1,901	0,814	0,661
5 %	0,999	1,689	0,954	0,892	0,692	0,254
10 %	0,998	1,049	0,943	0,553	0,653	0,132
	Über- und Unterbewertungen 80/20					
1 %	1,000	11,013	0,999	5,897	0,949	2,338
2 %	1,000	6,172	0,989	3,278	0,827	1,166
5 %	0,999	4,217	0,957	2,231	0,728	0,702
10 %	0,995	1,540	0,937	0,818	0,662	0,216
	Über- und Unterbewertungen 60/40					
1 %	1,000	47,903	0,998	25,676	0,920	10,615
2 %	1,000	18,474	0,985	9,843	0,842	3,764
5 %	0,998	7,136	0,947	3,778	0,714	1,243
10 %	0,988	4,626	0,917	2,470	0,665	0,727
	Über- und Unterbewertungen (300) 80/20					
1 %	0,997	42,397	0,945	22,586	0,882	8,759
2 %	0,983	25,713	0,926	13,686	0,786	5,035
5 %	0,969	6,454	0,881	3,420	0,675	1,048
10 %	0,946	4,528	0,841	2,411	0,643	0,656

Tabelle 3: Trefferrate (T) und Genauigkeit (G) für die Fehlerreihungsmethode auf dem 95-, 80- und 50%-Niveau.

6.5.4 Auswertung für die Momenten-Methode

Abgesehen von zwei geringfügigen Unterschreitungen des Konfidenzniveaus für die 50 % Irrtumswahrscheinlichkeit ist der *Moment-Bound* für den reinen Überbewertungsfall und den Fall, in dem Unterbewertungen ein Fünftel der Fehler ausmachen, für die gegebenen Daten verlässlich. Die Werte für die Genauigkeit sind mit denen der Durchschnittsfehlermethode für diese Fälle vergleichbar und genauer als die durch die Fehler-

reihungsmethode ausgegebenen. Starke Probleme ergeben sich jedoch hinsichtlich der Verlässlichkeit in dem Fall, wo Fehleranteile von bis zu 300 % für die Unterbewertungen zugelassen sind. Auf dem 95%-Niveau ist hier die relative Trefferhäufigkeit inkonsistenter Weise teilweise sogar kleiner als auf niedrigeren Sicherheitsstufen und reicht bis zu nicht akzeptablen Werten wie 0,438 (für eine Fehlerrate von 10 %). Dieses Verhalten wurde bei den anderen Verfahren nicht beobachtet, da diese die „Netto"-Fehlergrenze durch Saldierung der hochgerechneten Überbewertungen und dem Punktschätzer für die Unterbewertungen gewinnen, Über- und Unterbewertungen also separat und unterschiedlich behandelt werden. In der Momenten-Methode erfolgt jedoch eine Hochrechnung sowohl für Über- als auch für Unterbewertungen in einer Prozedur, da der durchschnittliche Fehleranteil aus positiven und negativen Fehleranteilen gebildet wird. Eine Ungleichbehandlung erfolgt nur insofern, dass der durchschnittliche Fehleranteil, sollte er negativ sein, zur Berechnung des hypothetischen Fehleranteils gleich Null gesesetz wird.[144] Durch dieses Vorgehen nehmen die gefundenen Unterbewertungen in der Momenten-Methode mehr Einfluss auf die obere Netto-Fehlergrenze als bei den anderen Verfahren. Je höher das Konfidenzniveau ist, desto weiter erfolgt auch eine Projektion der Unterbewertungen, wodurch die obere Fehlergrenze stärker sinkt als bei den zuvor beschriebenen Verfahren. In der Verrechnung kann dies dazu führen, dass die Trefferhäufigkeit mit steigendem Konfidenzlevel abnimmt und im Durchschnitt die berechnete oberere Fehlergrenze oft sogar unter dem tatsächlichen Fehler liegt.

Dieses Verhalten des *Moment-Bounds* führt zu der Schlussfolgerung, bei beobachteten Unterbewertungen in einem Prüffeld auf die Anwendung dieser Methode zu verzichten. Für Fälle, in denen Überbewertungen klar dominieren, liefert die Methode jedoch deutlich engere obere Fehlergrenzen als die Fehlerreihungsmethode und führt damit seltener zu einer fälschlichen Klassifizierung eines ordnungsgemäßen Prüffeldes als mit wesentlichen Fehlern behaftet.

Tabelle 4 fasst die Ergebnisse für den *Moment-Bound* zusammen.

[144] In der modifizierten Version des *Moment-Bounds* wird einfach der Betrag des durchschnittlichen Fehleranteils für diese Berechnung herangezogen, vgl. Dworin/Grimlund (1986), S. 43.

Moment-Bound						
	95 %		**80 %**		**50 %**	
Fehlerrate	T	G	T	G	T	G
	Nur Überbewertungen					
1 %	1,000	4,755	0,992	2,223	0,658	0,411
2 %	0,999	2,489	0,942	1,138	0,467	0,113
5 %	0,964	1,307	0,847	0,607	0,481	0,024
10 %	0,982	0,875	0,873	0,421	0,493	0,017
	Über- und Unterbewertungen 80/20					
1 %	0,990	8,676	0,990	4,238	0,860	1,078
2 %	0,978	4,380	0,972	2,230	0,624	0,514
5 %	0,981	0,981	0,929	1,728	0,577	0,401
10	0,989	0,989	0,891	0,639	0,539	0,072
	Über- und Unterbewertungen 60/40					
1 %	0,972	37,926	0,972	19,543	0,933	6,548
2 %	0,950	12,792	0,940	6,850	0,788	2,314
5 %	0,908	5,377	0,906	3,064	0,747	1,009
10 %	0,951	3,802	0,938	2,202	0,670	0,674
	Über- und Unterbewertungen (300) 80/20					
1 %	0,877	21,555	0,877	11,861	0,873	5,540
2 %	0,808	8,748	0,794	5,471	0,721	3,274
5 %	0,639	1,773	0,628	0,800	0,666	1,017
10 %	0,438	1,664	0,467	0,522	0,663	0,754

Tabelle 4: Trefferrate (T) und Genauigkeit (G) für den Moment-Bound auf dem 95-, 80- und 50%-Niveau.

6.6 Schlussfolgerungen aus dem Vergleich der Methoden

Auch wenn einer Simulationsstudie mit 240 erzeugten Grundgesamtheiten (15 Buchungskonten mal 16 Fehlermuster) sicherlich keine allgemeingültige Aussage über das Verhalten der Verfahren entnommen werden kann, geben die oben beschriebenen Ergebnisse doch eine grundsätzliche Wirkungsrichtung für die einzelnen Hochrechnungsmethoden an. So kann den Ergebnissen entnommen werden, dass alle beschriebenen

Verfahren nur sehr schlecht für den Fall starker Unterbewertungen (Fehleranteile bis 300 %) einsetzbar sind: Die Momenten-Methode und die Durchschnittsfehlermethode sind in diesem Fall nicht verlässlich. Die Fehlerreihungsmethode – obgleich stets verlässlich – ist in diesen Fällen so ungenau, dass ein verlässliches Prüfungsurteil kaum möglich ist.

Es zeigt sich weiterhin die Tendenz, dass für den Fall dominierender Überbewertungen die Durchschnittsfehlermethode und der *Moment-Bound* genauere Ergebnisse erzielen als die Fehlerreihungsmethode (auch wenn hier minimale Einbußen in der Verlässlichkeit hingenommen werden müssen). Tabelle 5 stellt jeweils den Vorteil der Momenten- und der Fehlerreihungsmethode hinsichtlich der Genauigkeit dar. Als Maß für die Vorteilhaftigkeit V einer Methode wurde jeweils der Quotient des Genauigkeitsmaßes der Fehlerreihungsmethode und der anderen Methode gebildet:

$$V_{D.} = \frac{G_{Str.}}{G_{D.}} \quad \text{bzw.} \quad V_{M.} = \frac{G_{Str.}}{G_{M.}}.$$

	Durchschnittsfehlermethode			**Moment-Bound**		
Sicherheit	95 %	80 %	50 %	95 %	80%	50 %
Fehlerrate	**Nur Überbewertungen**					
1 %	1,457	1,461	1,541	1,336	1,52	3,150
2 %	1,556	1,565	1,768	1,439	1,67	5,835
5 %	1,540	1,551	1,939	1,292	1,46	10,374
10 %	1,507	1,517	1,991	1,200	1,31	7,972
	Über- und Unterbewertungen 80/20					
1 %	1,300	1,301	1,334	1,269	1,391	2,169
2 %	1,543	1,552	1,729	1,409	1,47	2,267
5 %	1,740	1,756	2,229	4,300	1,291	1,749
10 %	1,525	1,529	1,900	1,558	1,279	2,984

Tabelle 5: Vergleich der Genauigkeit der Fehlerhochrechnungsmethoden.

Es wird deutlich, dass die beschriebenen Methoden in jedem der aufgezeigten Fälle genauer sind als die Fehlereihungsmethode. Dabei hebt sich die Momenten-Methode insbesondere für das 50%-Niveau und die Durschschnittsfehleremethode insbesondere für das 95%-Niveau ab. Jedoch sind es gerade diese Konfidenzniveaus, auf denen die jeweils überlegenen Methoden Verlässlichkeitseinbußen zu verzeichnen haben, sodass die Überlegenheit in der Genauigkeit zu relativieren ist.

Einer der Hauptvorteile, die im Dollar-Unit-Samplings-Verfahren gesehen werden, ist seine Überlegenheit in der Anwendung an Grundgesamtheiten mit kleinen Fehlerraten gegenüber traditionellen Stichprobenverfahren, welche für kleine Fehlerraten extrem große Konfidenzintervalle ausgeben.[145] Im Vergleich zu traditionellen Stichprobenverfahren gemessen, trifft diese Aussage sicherlich zu. Werden die obigen Ergebnissen für die Genauigkeit des *Dollar Unit Samplings* für kleine Fehlerraten jedoch isoliert von noch schlechteren Verfahren betrachtet, erscheint das *Dollar Unit Sampling* sogleich viel weniger überzeugend.

Als wichtige Aussage aus der Studie ist zu ziehen, dass Schlussfolgerungen des Prüfers über die Richtigkeit eines Prüffeldes aufgrund der Fehlerreihungsmethode nur mit Vorsicht gezogen werden sollten, da die Methode zwar stets verlässlich ist, die berechneten Fehlergrenzen jedoch häufig Vielfache des wahren Fehlers darstellen. Hieraus könnten folgende Handlungsempfehlung abgeleitet werden: Wird ein Prüffeld aufgrund der Fehlerreihungsmethode als ordnungsgemäß eingestuft, ist das Risiko, dass es in Wirklichkeit mit wesentlichen Fehlern versehen ist, auf allen Konfidenzniveaus sehr gering, sodass keine weiteren Handlungen nötig sind. Liegt die berechnete obere Fehlergrenze jedoch über dem gerade noch tolerierbaren Fehler, sollten die anderen Hochrechnungsmethoden zum Vergleich herangezogen werden. In Hinblick darauf, dass die Anwendung der Fehlerreihungsmethode sehr weit verbreitet ist, ist diese Handlungsempfehlung von breiter praktischer Relevanz.

[145] Vgl. Lillestol (1981), S. 263.

6.7 Mögliche Erweiterungen

Auch wenn versucht wurde, die – relativ unbekannte – Realität in Bezug auf Buchungsgesamtheiten annähernd realistisch wiederzugeben, sind die getroffenen Annahmen zweifelsohne nicht fern von jeder Kritik. Auch wenn es einige empirische Studien zum Fehlerverhalten gibt, die für die vorliegende Arbeit benutzt wurden, ist natürlich nicht gesichert, dass sich die Fehler generell wie in diesen Studien aufgeführt verhalten werden.[146] Vielmehr ist davon auszugehen, dass das Verhalten von Fehlern unter anderem mit Kontenart und -größe, mit Branche und Größe des Unternehmens, dem Charakter und dem Wissen des Jahresabschlusserstellers und sogar der Unternehmenskultur[147] variieren.

Die oberen Annahmen bzgl. Fehlerrate und -verteilung, Fehleranteilsverteilung und Verhältnis von Über- und Unterbewertung könnten demnach noch vielfältig abgeändert werden. Für die Verteilung von Fehleranteilen könnten beispielsweise χ^2-Verteilungen[148] oder Log-Normalverteilungen[149] angenommen werden. 100%-ige Überbewertungen könnten in die Grundgesamtheit gestreut,[150] variierende Fehleranteile je Buchungsschicht[151] simuliert oder die Anzahl der Ausprägungen der Fehleranteile auf einige wenige[152] beschränkt werden.

Die Simulationen sollten ferner auf eine größere Anzahl von unterschiedlichen Buchungsgesamtheiten ausgeweitet werden.

---

[146] Ferner könnten schon die in diesen empirischen Aufsätzen benutzten Methoden in Frage gestellt werden. Ham/Losell/Smieliauskas (1985) verwenden zum Beispiel willkürliche Stichproben und schließen aus diesen auf die Eigenschaften der Grundgesamtheit.

[147] Vgl. Chan/Lin/Mo (2003).

[148] Vgl. Dworin/Grimlund (1984), S. 229, Dworin/Grimlund (1986), S. 45ff.

[149] Vgl. Wolz (2003), S. 149.

[150] Vgl. Horgan (1996), S. 223, Leitch/Neter/Plante/Sinha (1982), S. 391.

[151] Vgl. Reneau (1978), S. 671.

[152] Vgl. Neter (1978), S. 86.

7 Zusammenfassung und Ausblick

In der vorliegenden Arbeit erfolgte eine Beschreibung des Dollar-Unit-Sampling-Verfahrens inklusive seiner unterschiedlichen Ausgestaltungen hinsichtlich der Methoden der Stichprobenziehung und vor allem der diversen Möglichkeiten der Projektion von in der Stichprobe gefundenen Fehlern auf die Grundgesamtheit. Dabei wurde schon durch den Vergleich an einfachen Rechenbeispielen deutlich, dass sich die Varianten der Fehlerhochrechnung in ihrer Konsistenz und Genauigkeit stark unterscheiden.

Ausgehend von diesem intuitiven Rechenvergleich wurden drei für die aktuelle Wirtschaftsprüfungspraxis relevante Verfahren anhand einer Simulationsstudie verglichen. Auf der Grundlage von 60.000 simulierten Stichproben wurden diese drei Verfahren auf drei Konfidenzniveaus beurteilt. Es bestätigte sich hierbei der Eindruck, dass die Fehlerreihungsmethode auf allen Konfidenzniveaus verlässlich ist, wobei für sie jedoch insbesondere für kleine Fehlerraten und beim Auftreten von Unterbewertungen starke Genauigkeitseinbußen zu verzeichnen sind.

Dagegen zeigten die Durchschnittsfehlermethode und die Momenten-Methode auf einigen Konfidenz-Leveln Verlässlichkeitseinbußen. Die Momenten-Methode erwies sich für den Bereich dominanter Unterbewertungen sogar stark unverlässlich und inkonsistent im Bereich der relativen Trefferhäufigkeit. Für den Bereich dominierender Überbewertungen erzielen die Momenten-Methode und die Durchschnittsfehlermethode jedoch durchweg höhere Genauigkeiten als die in der Praxis überwiegend angewandte Fehlerreihungsmethode.

Auch wenn die vorliegenden Ergebnisse aufgrund der Annahmen für die Simulation nicht frei von Kritik sind, geben sie doch eine grundsätzliche Idee über das Verhalten der einzelnen Projektionsmethoden. Dies hat insofern gravierende Implikationen für die Urteilsbildung in Prüfungsprozessen, da die Methoden aufgrund von hohen Genauigkeitseinbußen kein sinnvolles Prüfungsurteil zulassen. Zwar werden nicht ordnungsgemäße Prüffelder mit großer Sicherheit als solche klassifiziert, das Risiko der Ablehnung ordnungsmäßiger Prüffelder scheint aufgrund der gegebenen Ergebnisse jedoch zu hoch zu sein.

In einem nächsten Schritt sollten die in dieser Arbeit erzielten Ergebnisse mit Hilfe variierender Annahmen bzgl. zum Beispiel der Fehlerverteilungen in den Grundgesamtheiten auf Robustheit getestet werden.

Aus den Genauigkeiten für die einzelnen Verfahren ergeben sich ferner Implikationen für die Stichprobenplanung, auf die in dieser Arbeit nicht eingegangen wurde. Da diese jedoch eine wichtige Säule in Hinblick auf die Effizienz in Prüfungsprozessen darstellt, bedarf sie einer tiefer gehenden separaten Betrachtung.

Letztendlich sind jedoch alle oben genannten weiteren Analysen vergebens, wenn – wie es im Moment in der Prüfungspraxis zumindest partiell der Fall zu sein scheint – die beschriebenen Verfahren von Prüfern ohne Kenntnisse über die zugrunde liegenden Methoden und deren Auswirkungen auf das prüferische Urteil (als *„Black Box“*) angewandt werden.

Literatur

[1] American Institute of certified Public Accountants (AICPA) (1999): Audit Sampling. Auditing Practice Release, AICPA, New York.

[2] Anderson, T.W./Samuels, S. M. (1967): Some inequalities among binomial and Poisson Probabilities. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

[3] Apostolou, Barbara/Alleman, Francine (1991): Internal Audit Briefings, Internal Audit Sampling. Institute of Internal Auditors, USA.

[4] Bamberg, Günter/Baur, Franz: Statistik. Oldenbourg Verlag, 12. Auflage, München, Wien.

[5] Barron, Orie/Groomer, Micheal S./Swink, Morgan (1998): Setting tolerable misstatements when auditing aggregate accounts. In: Decision Science. S. 1005-1033.

[6] Berg, Nathan (2006): A Simple Bayesian Procedure for Sample Size Determination in an Audit of Property Value Appraisals. In: Real Estate Economics, Vol. 34, No. 1, S. 133-155.

[7] Biaggio, Loris (1987): Der Einsatz von mathematisch-statistischen Modellen durch den Revisor unter Berücksichtigung der EDV-Unterstützung. Schweizerische Treuhand- und Revisionskammer, Zürich.

[8] Blocher, Edward: Sampling for Integrated Audit Objectives - A Comment. In: the Accounting Review, July, S. 766-772.

[9] Boockholdt, James L./Chang, Stanley Y./Finley, David R. (1992): Using DUDS to detect Fraud. In: The Internal Auditor, 8/1992, S. 59-64.

[10] Braun, Frank (1996): Gebührendruck und Prüfungsqualität bei Pflichtprüfungen mittelständischer Unternehmen. In: Betriebsberater, Heft 19, S. 999-1001.

[11] Bredeck, Hans Martin (1993): Rechnergestütze Stichprobenverfahren im Prüfungswesen. IDW-Verlag GmbH, Düsseldorf.

[12] CaseWare International Inc. (2003): Monetary Unit Sampling Technical Specification, IDEA Toronto, Ohio.

[13] Chan, K. Hung/Lin, Kenny Z./Mo, Phyllis lai Lan (2003): An Empirical Study on the Impact of Culture on Audit-Detected Accounting Errors. In: Auditing – A Journal of Practice & Theory, S. 281-295.

[14] The Chartered Institute of Public Finance and Accountancy (1995): Statistics for Audit – A Guide to Statistical Sampling for Auditors. The Chartered Institute of Public Finance and Accountancy, London.

[15] Coenenberg, Adolf (2003): Jahresabschluss und Jahresabschlussanalyse. 19. Auflage, Stuttgart, Schäfer Pöschel Verlag.

[16] Domschke, Wolfgang/Drexl, Andreas (2002): Einführung in Operations Research. 5. Auflage, Springer Verlag, Kiel.

[17] Dusenbury, Richard B./ Reimers, Jane L./Wheeler, Stephen W. (2000): The Audit Risk Model: An Empirical Test for Conditional Dependencies among Assessed Component Risks. In: Auditing: A Journal of Practice & Theory, Vol. 19, No. 2, S. 105-120.

[18] Dworin, Lowell/Grimlund, Richard A. (1984): Dollar Unit Sampling for Accounts Receivable and Inventory. In: The Accounting Review, April, S. 218-241.

[19] Dworin, Lowell/Grimlund, Richard A. (1986): Dollar-Unit Sampling: A Comparison of the Quasi-Bayesian and Moment Bounds. In: The Accounting Review, January, S. 36-57.

[20] Eggenberger, Hans (1986): Mathematisch-statistische Stichprobenverfahren in der Revision. Schweizerische Treuhand- und Revisionskammer, Zürich.

[21] Felix, William L./Grimlund, Richard A./Koster, Frank J./Roussey, Robert S. (1990): Arthur Andersen's New Monetary Unit Sampling Approach. In: Auditing: A Journal of Practice & Theory, Vol. 9, No. 3.

[22] Fellingham, John C./Newman, Paul D./Patterson, Evelyn R. (1989): Sampling Information in Strategic Audit Settings. In: Auditing: A Journal of Practice & Theory, Spring, Vol. 8, No. 2.

[23] Gill, R. D. (1983): The Sieve Method as an Alternative to Dollar-Unit Sampling: The Mathematical Background. Mathematisch Centrum, Amsterdam.

[24] Grant, C. Terry/Depree, Chauncey M. Jr./Grant, Gerry (2000): Earnings Management and the Abuse of Materiality. In: Journal of Accountancy, Vol. 190/3.

[25] Grimlund, Richard A./ Felix, William L. (1987): Simulation Evidence and Analysis of Alternative Methods of Evaluation Dollar-Unit Samples. In: The Accounting Review, Vol. 62, No. 3.

[26] Guy, Dan M./Carmichel, D. R./ Whittington, Ray (2002): Audit Sampling. 5. Auflage, John Wiley & Sons, Inc., USA.

[27] Hall, Thomas W./Herron, Terri L./Pierce, Bethane Jo (2006): How Reliable Is Haphazard Sampling? In: The CPA Journal, January, S. 26-27.

[28] Hall, Thomas W./Hunton, James E./Jo Pierce, Bethane: Sampling Practices of Auditors in Public Accounting, Industry and Government. In: Accounting Horizons, Vol. 16, No. 2, S. 125-136.

[29] Hall, Thomas W./Pierce, Bethane Jo/Ross, W. R. (1989): Planning Sample Sizes for Stringer-Method Monetary Unit and Single-Stage Attribute Sampling Plans. In: Auditing: A journal of Practice & Theory, Vol. 8, No. 2, S. 64-89.

[30] Hall, Thomas W./Herron, Terri L./Pierce, Bethane Jo/Witt, Terry J. (2001): The Effectiveness of Increasing Sample Size to Mitigate the Influence of Population Characteristics in Haphazard Sampling. In Auditing: A Journal of Practice & Theory, Vol. 20, No. 1, S. 169-185.

[31] Ham, Jane/Losell, Donna/Smieliauskas, Wally (1985): An Empirical Study of Error Characteristics in Accounting Populations. In: The Accounting Review, Vol. 60, No. 3, S. 387-406.

[32] Hauptfachausschuss des Instituts der Wirtschaftsprüfer (1988): Zur Anwendung stichprobengestützter Prüfungsmethoden bei der Jahresabschlußprüfung. HFA 1/1988, S. 1-8.

[33] van Heerden, A. (1961): Steekproeven als meiddelvan accountants controle. In: Maandblad Voor Accountancy en Bedrijfshuishoudkunde, S. 453-475.

[34] Hitzig, Neal B. (1998): Detecting and Estimating Misstatement in Two-Step Sequential Sampling with Probability Proportional to Size. In: Auditing: A Journal of Practice & Theory, Vol. 17, No. 1, S. 54-69.

[35] Hitzig, Neal B. (2004a): Elements of Sampling: The Population, the Frame, and Sampling Unit. In: The CPA Journal, Vol. 11, S. 30-34.

[36] Hitzig, Neal B. (2004b): Statistical Sampling Revisited. In: The CPA Journal, Vol. 5, S. 30-35.

[37] Horgan, Jane M. (1996): the Moment Bound with Unrestricted Random, Cell and Sieve Sampling of Monetary Units. In: Accounting and Business Research, Vol. 26, No. 3, S. 215-223.

[38] Horgan, Jane M. (1997): Stabilizing the Sieve Sample Size Using PPS. In: Auditing: A Journal of Practice and Theory, Vol. 16, No. 2, S. 40-52.

[39] Horgan, Jane M. (1998): Stabilitzed Sieve Sampling: A Point-Estimator Analysis. In: Journal of Business & Economic Statistics, January, Vol. 16, No.1, S. 42-51.

[40] Horgan, Jane M. (1999): Hands on! Here comes the Bug! In: Accountancy Ireland, Vol. 31, No. 2, S. 20-21.

[41] Hyde, Gerald (2007): Enhanced Audit Testing: Modern Sampling and Analysis Techniques Can Add Considerable Efficiency to Traditional Audit Work. In: Internal Auditor, Vol. 64, No. 4, S. 65-69.

[42] International Auditing and Assurance Standards Board (2007): ISA 530, Audit Samplig Exposure Draft. July,International Federation of Accountants.

[43] Johnson, Johnny R./Leitch, Robert A./Neter, John: Characteristics of Errors in Accounts Receivable and Inventory Audits. In: The Accounting Review, Vol. 56, No. 2, S. 270-293.

[44] Johnson, N. L./Kotz, S. (1970): Distributions in Statistics: Continuous Univariate Distributions. Boston, Houghton Mifflin Co.

[45] Kaplan, Robert S. (1975): Sample Size Computations for Dollar-Unit Sampling. In: Journal of Accounting Research, Vol. 13, 126-133.

[46] Kirk, Brad (2000): Delivery Speed, Accuracy, Compliance. In: Internal Auditor, S. 25-27.

[47] Lahiri, D. B. (1951): A Method of Sample Selection Providing Unbiased Ratio Estimators. In: Bulletin of the International Statistics Institute, 33, S. 133-140.

[48] Leslie, Donald A./Teiltlebaum, Albert D./Anderson, Rodney J. (1979): Dollar Unit Sampling – A Practical Guide for Auditors, Toronto.

[49] Lillestol, Jostein (1981): A Note on Computing Upper Error Limits in Dollar-Unit Sampling. In: Journal of Accounting Research, Vol. 19, No. 1, S. 263-267.

[50] Lindgens, Ursula (1999): Der Markt für Prüfungsleistungen – Anmerkungen aus Sicht der Praxis. In: Richter, Martin (Hrsg.): Theorie und Praxis der Wirtschaftsprüfung II.

[51] Lück, Wolfgang/Lexer, Matthias (2004): Lexikon der Betriebswirtschaft. 6. Auflage, München, Oldenbourg Verlag.

[52] Marten, Kai-Uwe (2003): Stichproben im Rahmen der Jahresabschlussprüfung. In: kapitalmarktorientierte Rechnungslegung, 10, S. 444-448.

[53] Matthews, Derek (2006): From Ticking to Clicking: Changes in Auditing Techniques in Britain from the 19th Century to the Present. In: Accounting Historians Journal, Vol. 33, No. 2, S. 63-102.

[54] McCray, John H. (1984): A Quasi-Bayesian Audit Risk Model for Dollar Unit Sampling. In: The Accounting Review, January, S. 35-51.

[55] Needleman, Ted (2002): Midrange Accounting Software Continues to Improve. In: Accounting Today, 18. März - 7. April, S. 24-25.

[56] Neter, John/Johnson, Johnny R. Leitch, Robert A. (1985): Characteristics of Dollar-Unit taints and Error Rates in Accounts Receivable and Inventory. In: The Accounting Review, July, S. 488-499.

[57] Neter, John/Leitsch, Robert A./Fienberg, Stephen E. (1978): Dollar Unit Sampling: Multinomial Bounds for Total Overstatement and Understatement Errors. In: The Accounting Review, January, S. 77-93.

[58] Neter, John/Johnson, Johnny R. Leitch, Robert A. (1985): Characteristics of Dollar-Unit taints and Error Rates in Accounts Receivable and Inventory. In: The Accounting Review, July, S. 488-499.

[59] Plante, Robert/Neter, John/Leitch, Robert (1984): A Lower Multinominal Bound for the Total Overstatement Error in Accounting Populations. In: Management Science, Vol. 30, No. 1, S. 37-50.

[60] Power, M. K. (1992): From Common Sense to Excersise: Reflections on the Prehistory of Audit Sampling. In: Accounting, Organizations and Society, Vol. 17, No. 1, S. 37-62.

[61] Reneau, J. Hal (1978): CAV Bounds in Dollar Unit Sampling: Some Simulation Results. In: The Accounting Review, July, S. 669-680.

[62] Rohrbach, Kermit John (1986): Monetary unit Acceptance Sampling. In: Journal of Accounting Research, Vol. 24, No. 1, S. 127-150.

[63] Rohrbach, Kermit John (1997); Sample size determination using the augmented variance estimator, Vol. 16, No. 1, 124-137.

[64] Steele, Anthony (1992): Audit Risk and Audit Evidence: The Bayesian Appoach to Statistical Auditing. Academic Press, Hrsg. Harcourt/Brace/Jovanovich, London, San Diego, New York, Boston, Sydney, Tokyo, Toronto.

[65] Steiger, Andreas (1998): Die Anwendung heterograder Schätzstichproben bei der handelsrechtlichen Jahresabschlussprüfung. In: Europäische Hochschulschriften, Reihe V Volks- und Betriebswirtschaft, Peter Lang Verlag, Frankfurt am Main, Berlin Bern, New York, Paris, Wien (Dissertation)

[66] Stephan, F.F.: Some Statistical Problems Involved in Auditing and Inspection. In: Proceedings of the American Statistic Association, S. 404.

[67] Stringer, K. W. (1963): Practical Aspects of Statistical Sampling in Auditing. In: Proceedings of the American Statistic Association, S. 405-411.

[68] Wampler, Bruce/McEacharn, Michelle (2005): Monetary Unit Sampling Using Microsoft Excel. In: The CPA Journal, S. 36-40.

[69] Wheeler, Stephen/Dusenbury, Richard/Reimers, Jane (1997): Projecting Sample Misstatements to Audit Populations: Theoretical, Professional, and Empirical Considerations. In: Decision Sciences, Spring, Vol. 28, No. 2, S. 261-278.

[70] Wolz, Mathias (2003): Wesentlichkeit im Rahmen der Jahresabschlussprüfung, Monographie, Düsseldorf. Habilitationsschrift eingereicht an der Universität Essen (2002) unter dem Titel „Materiality, Prüfungsrisiko und Prüfungsumfang – Ein Beitrag zur Steigerung der Effizienz des Prüfungsprozesses“.

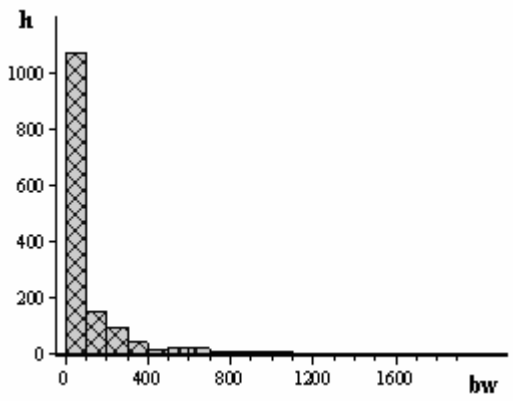
[71] Wolz, Mathias (2004): Dollar Unit Sampling - Ein modifiziertes Verfahren zur Beurteilung über- und unterbewerteter Prüffelder. In: Betriebswirtschaftliche Forschung und Praxis, 1/2004, 60-80.

[72] Wurst, J./Neter, John/Godfrey, J. (1989): Comparision of Sieve Sampling with Random and Cell Sampling of Monetary Units. In: Statistician, Vol. 38, No. 2, S. 235-249.

[73] v. Wysocki, Klaus (2002): Prüfungstheorie, meßtheoretischer Ansatz. In: Ballwieser, W./Coenenberg, A./v. Wysocki, K. (Hrsg.): Handwörterbuch der Rechnungslegung und Prüfung, Stuttgart, Sp. 1886-1899.

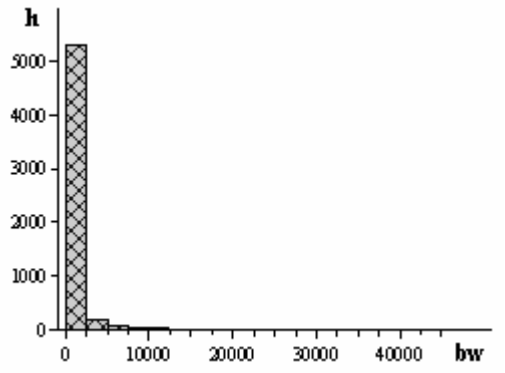
Appendix: Charakteristika der verwendeten Daten

bw – Buchwert einer Position in Euro
h – absolute Häufigkeit der Buchungen einer Buchwertklasse



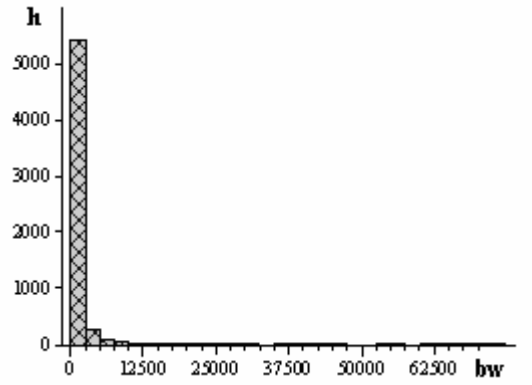
Debitoren 1

Anzahl der Salden	1.496
Gesamtbuchwert	€ 201.444,14
Arithmetisches Mittel	€ 134,66
Median	€ 29,95
Standardabweichung	€ 268,04
Schiefe	3,87
Maximum	€ 2.061,29
Minimum	€ 1,00



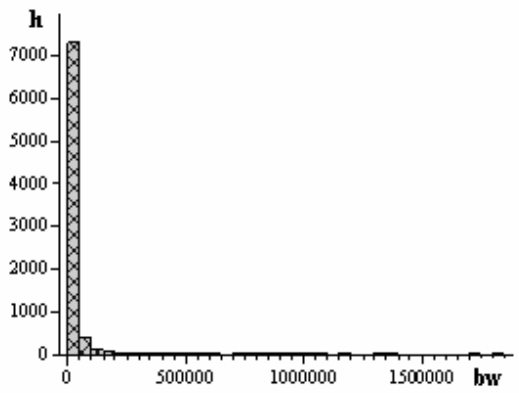
Debitoren 2

Anzahl der Salden	5.683
Gesamtbuchwert	€ 5.036.442,84
Arithmetisches Mittel	€ 886,23
Median	€270,05
Standardabweichung	€2.780,03
Schiefe	9,12
Maximum	€ 49.416,19
Minimum	€ 1,19



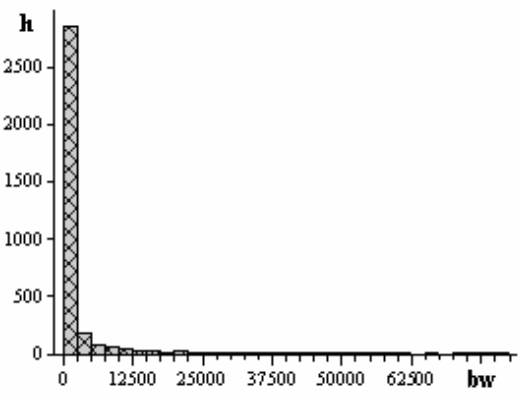
Debitoren 3

Anzahl der Salden	5.957
Gesamtbuchwert	€ 7.580.758,13
Arithmetisches Mittel	€ 1.272,58
Median	€ 374,87
Standardabweichung	€ 4.216,09
Schiefe	10,19
Maximum	€ 72.652,71
Minimum	€ 1,02



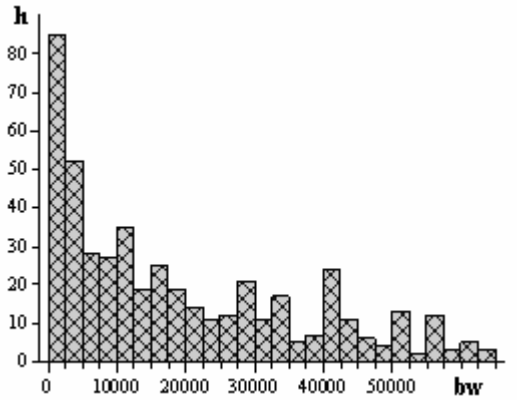
Debitoren (Soll)

Anzahl der Buchungen	7.952
Gesamtbuchwert	€ 189.715.755,07
Arithmetisches Mittel	€ 23.857,62
Median	€ 10.095,47
Standardabweichung	€ 73.399,40
Schiefe	11,88
Maximum	€ 1.843.936,00
Minimum	€ 1,65



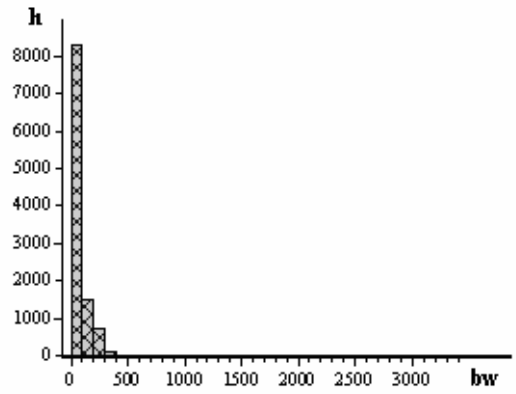
Umsatz 1

Anzahl der Buchungen	3.418
Gesamtbuchwert	€ 8.986.304,53
Arithmetisches Mittel	€ 2.629,11
Median	€ 105,32
Standardabweichung	€ 8.131,86
Schiefe	5,38
Maximum	€ 79.224,55
Minimum	€ 0,12



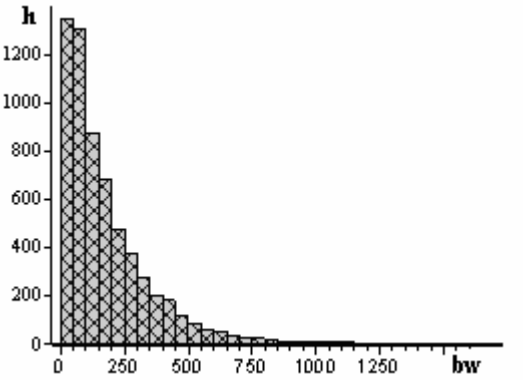
Umsatz 2

Anzahl der Buchungen	471
Gesamtbuchwert	€ 8.975.809,89
Arithmetisches Mittel	€ 19.056,92
Median	€ 13.706,00
Standardabweichung	€ 17.394,45
Schiefe	0,81
Maximum	€ 64.000,00
Minimum	€ 0,96



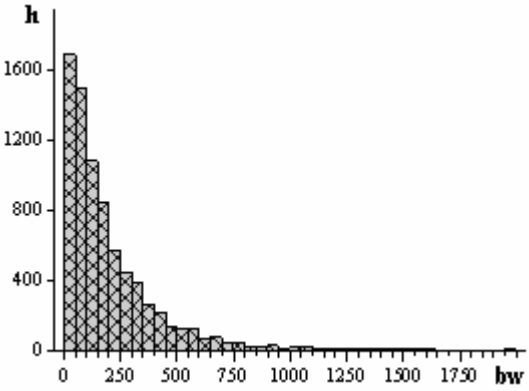
Umsatz 3

Anzahl der Buchungen	10.754
Gesamtbuchwert	€ 777.304,25
Arithmetisches Mittel	€ 72,28
Median	€ 36,54
Standardabweichung	€ 116,61
Schiefe	11,95
Maximum	€ 3779,99
Minimum	€ 0,43



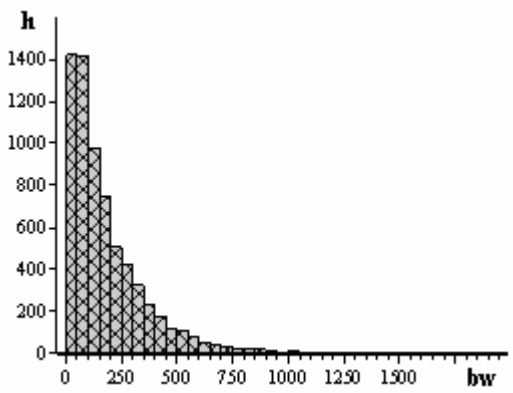
Umsatz 4

Anzahl der Buchungen	6.198
Gesamtbuchwert	€ 1.115.665,96
Arithmetisches Mittel	€ 180,00
Median	€ 125,00
Standardabweichung	€ 179,75
Schiefe	2,21
Maximum	€ 1.655,82
Minimum	€ 0,01



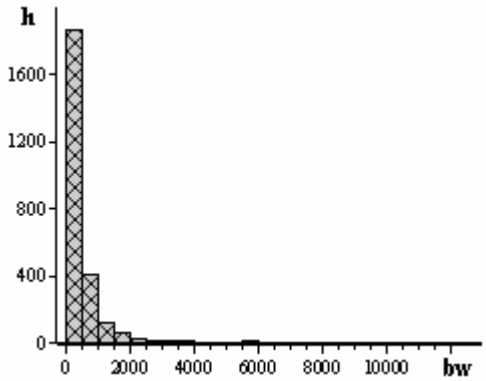
Umsatz 5

Anzahl der Buchungen	7.702
Gesamtbuchwert	€ 1.486.749,63
Arithmetisches Mittel	€ 193,03
Median	€ 129,31
Standardabweichung	€ 199,38
Schiefe	2,27
Maximum	€ 1.965,52
Minimum	€ 0,01



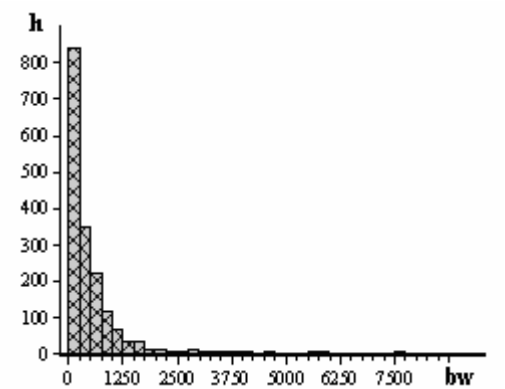
Umsatz 6

	6.764
Anzahl der Buchungen	€ 1.238.167,05
Gesamtbuchwert	€ 183,05
Arithmetisches Mittel	€ 129,22
Median	€ 180,07
Standardabweichung	2,21
Schiefe	€ 1.922,41
Maximum	€ 0,04
Minimum	



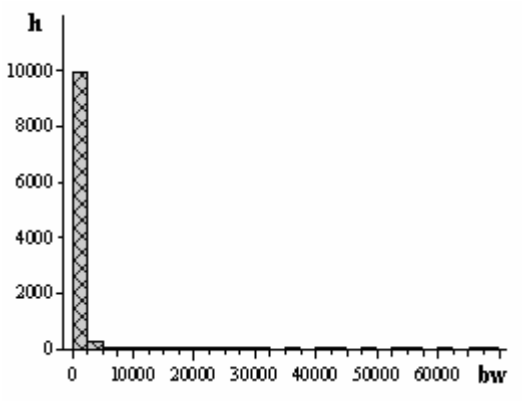
Umsatz 7

Anzahl der Buchungen	2.571
Gesamtbuchwert	€ 1.337.223,77
Arithmetisches Mittel	€ 520,12
Median	€ 196,00
Standardabweichung	€ 1.075,03
Schiefe	5,84
Maximum	€ 13.372,00
Minimum	€ 1,90



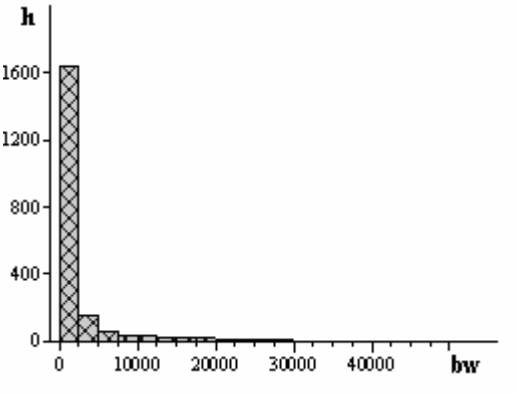
Umsatz 8

Anzahl der Buchungen	1.795
Gesamtbuchwert	€ 1.123.548,47
Arithmetisches Mittel	€ 625,93
Median	€ 281,30
Standardabweichung	€ 1.116,87
Schiefe	4,21
Maximum	€ 9.585,00
Minimum	€ 0,95



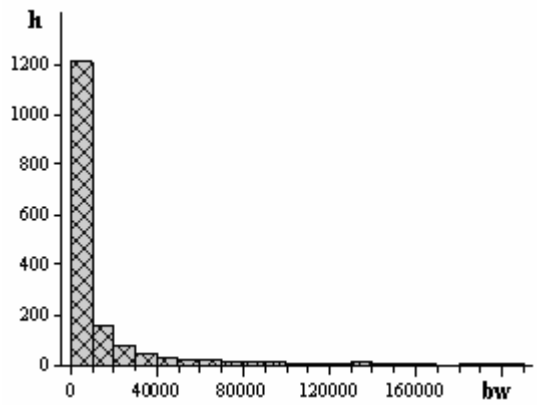
Umsatz 9

Anzahl der Buchungen	10.314
Gesamtbuchwert	€ 7.236.099,31
Arithmetisches Mittel	€ 701,58
Median	€ 280,50
Standardabweichung	€ 2.994,13
Schiefe	15,86
Maximum	€ 68.000,00
Minimum	€ 0,95



Roh-, Hilf und Betriebsstoffe

Anzahl der Buchungen	2.030
Gesamtbuchwert	€ 5.424.780,64
Arithmetisches Mittel	€ 2.672,31
Median	€ 489,52
Standardabweichung	€ 6.566,53
Schiefe	4,41
Maximum	€ 54.051,00
Minimum	€ 0,43



Kreditoren (Soll)

Anzahl der Buchungen	1.648
Gesamtbuchwert	€ 23.188.476,25
Arithmetisches Mittel	€ 14.070,68
Median	€ 2.497,06
Standardabweichung	€ 30.134,25
Schiefe	3,58
Maximum	€ 206.884,20
Minimum	€ 7,42

UNIVERSITÄT POTSDAM

Wirtschafts- und Sozialwissenschaftliche Fakultät

STATISTISCHE DISKUSSIONSBEITRÄGE

Herausgeber: Hans Gerhard Strohe

ISSN 0949-068X

Nr. 1 1995 Strohe, Hans Gerhard: Dynamic Latent Variables Path Models
- An Alternative PLS Estimation -

Nr. 2 1996 Kempe, Wolfram: Das Arbeitsangebot verheirateter Frauen in den neuen und alten Bundeslõndern
- Eine semiparametrische Regressionsanalyse -

Nr. 3 1996 Strohe, Hans Gerhard: Statistik im DDR-Wirtschaftsstudium zwischen Ideologie und Wissenschaft

Nr. 4 1996 Berger, Ursula: Die Landwirtschaft in den drei neuen EU-Mitgliedsstaaten Finnland, Schweden und Österreich
- Ein statistischer Überblick -

Nr. 5 1996 Betzin, Jörg: Ein korrespondenzanalytischer Ansatz für Pfadmodelle mit kategorialen Daten

Nr. 6 1996 Berger, Ursula: Die Methoden der EU zur Messung der Einkommenssituation in der Landwirtschaft
- Am Beispiel der Bundesrepublik Deutschland -

Nr. 7 1997 Strohe, Hans Gerhard/Geppert, Frank: Algorithmus und Computerprogramm für dynamische Partial Least Squares Modelle

Nr. 8 1997 Rambert, Laurence / Strohe, Hans Gerhard: Statistische Darstellung transformationsbedingter Veränderungen der
Wirtschafts- und Beschäftigungsstruktur in Ostdeutschland

Nr. 9 1997 Faber, Cathleen: Die Statistik der Verbraucherpreise in Rußland
- Am Beispiel der Erhebung für die Stadt St. Petersburg -

Nr. 10 1998 Nosova, Olga: The Atractiveness of Foreign Direct Investment in Russia and Ukraine - A Statistical Analysis

Nr. 11 1999 Gelaschwili, Simon: Anwendung der Spieltheorie bei der Prognose von Marktprozessen

Nr. 12 1999 Strohe, Hans Gerhard / Faber, Cathleen: Statistik der Transformation - Transformation der Statistik.
Preisstatistik in Ostdeutschland und Rußland

Nr. 13 1999 Müller, Claus: Kleine und mittelgroße Unternehmen in einer hoch konzentrierten Branche am Beispiel der Elektrotechnik.
Eine statistische Langzeitanalyse der Gewerbezählungen seit 1882

Nr. 14 1999 Faber, Cathleen: The Measurement and Development of Geogian Consumer Prices

Nr. 15 1999 Geppert, Frank / Hübner, Roland: Korrelation oder Kointegration
- Eignung für Portfoliostrategien am Beispiel verbriefter Immobilenanlagen -

UNIVERSITÄT POTSDAM

Wirtschafts- und Sozialwissenschaftliche Fakultät

STATISTISCHE DISKUSSIONSBEITRÄGE

Herausgeber: Hans Gerhard Strohe
ISSN 0949-068X