Gesellschaft
für Qualitätsmanagement
und Statistik mbH

Statistische Modellierung von Gleichwertigkeitsuntersuchungen im Rahmen des Bundesbodenschutzgesetzes anhand von unterschiedlichen Messverfahren anorganischer und organischer Schadstoffparameter

von

PD Dr. habil. Steffen Uhlig

Auftraggeber

Fachbeirat Verfahrens- und Methoden für Bodenuntersuchungen

FKZ 360 13 002 (Umweltbundesamt)

Fachliche Begleitung: Dr. D. Lück, BAM

quo data
Siedlerweg 20
01465 Dresden-Langebrück

Tel.: +49-35201-70387
Fax: +49-35201-80687
www.quodata.de

Inhaltsverzeichnis

Teil I: Äquivalenznachweis durch Ringversuche für Bodenanalysen

1. Gleichwertigkeit und Äquivalenz von Bestimmungsverfahren bei Ringversuchen

1.1 Begriffsabgrenzung: Gleichwertigkeit und Vergleichbarkeit

Gleichwertigkeit und Äquivalenz von Bestimmungsverfahren sind Begriffe, die sich auf Eigenschaften der Messmethoden[1] in Verbindung mit Probenvorbereitungsschritten beziehen. Vergleichbarkeit hingegen ist ein verfahrensspezifischer Kennwert, dessen Zweck nicht im Vergleich mehrerer Verfahren besteht, sondern im Vergleich zweier Messungen, die unter unterschiedlichen Bedingungen, jedoch mit der gleichen Verfahren realisiert wurden. Dem Begriff der Vergleichbarkeit liegt die Vergleichstandardabweichung zugrunde, und diese wiederum charakterisiert die typischerweise auftretenden Schwankungen zwischen verschiedenen Einzelwerten, die von unterschiedlichen Labors ermittelt werden. Dabei stellt sich die Frage, unter welchen Umständen die Differenz zwischen zwei Einzelwerten noch als zufallsbedingt aufgefasst werden kann. Diese Differenz ist – unter der Normalverteilungsannahme – wiederum normalverteilt, mit dem Erwartungswert 0 und der Standardabweichung $\sigma_R\sqrt{2}$. Dass die Differenz betragsmäßig größer wird als $\sigma_R \times 1{,}96\sqrt{2} = \sigma_R \times 2{,}77$, passiert unter der Normalverteilungsannahme nur in ca. 5% aller Fälle. Deshalb kann in einem solchen Fall gefolgert werden, dass möglicherweise noch andere Einflüsse aufgetreten sind, welche die große Differenz der beiden Werte erklären. Möglicherweise wurden bei den Messungen unterschiedliche Messverfahren verwendet, und wenn dies der Fall ist, sollte für die Berechnung der Vergleichbarkeit nicht einfach die Vergleichstandardabweichung der Referenzverfahren herangezogen werden. Sofern deutlich unterschiedliche Wiederfindungsraten auftreten, müssten diese Unterschiede in der Vergleichstandardabweichung berücksichtigt werden.

1.2 Der klassische Ansatz

Will man anhand von Messergebnissen nachweisen, dass zwei Messverfahren gleichwertig sind, stellt sich das Problem, dass es aus prinzipiellen Gründen unmöglich ist, den

[1] Der Begriff des Messverfahrens oder Analysenverfahrens umfasst in diesem Bericht die Kombination aus Mess- bzw. Analysenmethode und Probenvorbereitungsschritten.

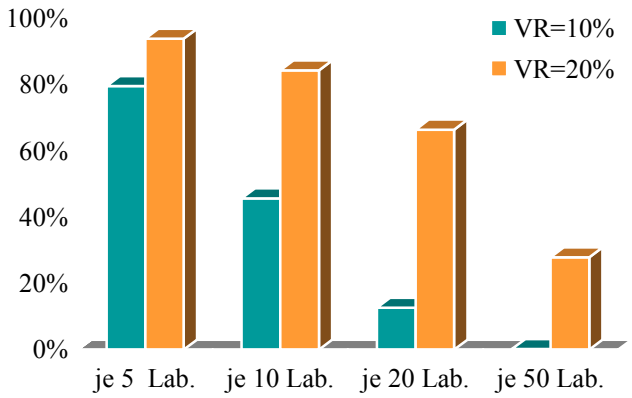
statistischen Nachweis völliger Gleichwertigkeit zu erbringen, da jedes Verfahren mit einem zufälligen Messfehler behaftet ist. Die Erfahrung zeigt, dass selbst unter (scheinbar) identischen Bedingungen keine völlige Reproduzierbarkeit der Messergebnisse gewährleistet ist, und dies gilt bei Bodenanalysen insbesondere deshalb, weil hier aufgrund einer Vielzahl möglicher Einflussfaktoren (Heterogenität des Bodens, Bodenmatrix, Extraktionsverfahren, Analysenverfahren und Detektion) typischerweise mit vergleichsweise hohen Unsicherheiten zu rechnen ist.

Um diese Unsicherheiten berücksichtigen zu können, verwendet man im klassischen Ansatz einen statistischen Test, bei dem in der Nullhypothese unterstellt wird, dass die Messverfahren gleichwertig sind. Es wird dann geprüft, ob die beiden Verfahren signifikant unterschiedlich sind. Um zu überprüfen, ob insbesondere hinsichtlich der Wiederfindungsrate signifikante Unterschiede bestehen, kann ein t-Test eingesetzt werden, bei dem die Differenz der beiden Wiederfindungsraten durch den zugehörigen Standardfehler dividiert wird[2]. Sofern die resultierende Prüfgröße den kritischen Wert der t-Verteilung übersteigt, kann zum vorgegebenen Signifikanzniveau α die Nullhypothese identischer Wiederfindungsraten abgelehnt werden, so dass davon ausgegangen werden muss, dass die beiden Verfahren signifikant unterschiedlich sind. Das Prüfungsergebnis ist bei dieser Vorgehensweise in sehr großem Maße von der Anzahl der Messwiederholungen und von den verschiedenen Unsicherheitskomponenten abhängig, d.h. bei einer geringen Anzahl von Messungen wird die Nullhypothese der Gleichwertigkeit in der Regel beibehalten, während bei einer großen Anzahl von Messungen signifikante Abweichungen festgestellt werden. Um diesen Zusammenhang zu demonstrieren, wurde im Rahmen einer Simulationsstudie die Häufigkeit ermittelt, mit der unter unterschiedlichen Rahmenbedingungen der Nachweis der Unterschiedlichkeit der Messverfahren erfolgt. Abhängig ist diese Häufigkeit zunächst von der Differenz der Wiederfindungsraten: Sofern diese sehr ähnlich sind, wird die Häufigkeit eines signifikanten Nachweis der Unterschiedlichkeit nahe beim vorgegebenen Signifikanzniveau α liegen, also in der Regel bei 5%. Bei größeren Unterschieden der Wiederfindungsraten wird ein solcher Nachweis naturgemäß häufiger auftreten. In dem in Abbildung 1 beschriebenen Simulationsbeispiel wurde unterstellt, dass die wahre absolute Differenz der in Prozent ausgedrückten Wiederfindungsraten bei 10% liegt. Weiterhin wurde unterstellt, dass je 5, 10, 20 oder 50 Laboratorien jeweils eine Messung durchführen, wobei die relative Vergleichstandardabweichung bei 10% bzw. 20% liegt. Dargestellt sind die unter

[2] Sofern die Ergebnisse von verschiedenen Konzentrationsniveaus vorliegen, kann alternativ ein F-Test verwendet werden

dieser Voraussetzungen ermittelten Häufigkeiten, in denen die Wiederfindungsraten als gleichwertig (d.h. nicht signifikant unterschiedlich) ermittelt werden. Es zeigt sich, dass diese Häufigkeit bei geringen Vergleichstandardabweichungen durchweg höher liegt. Weiterhin ist die Abhängigkeit von der Anzahl der Laboratorien augenfällig: Mit nur 5 Laboratorien wird offenbar sehr viel seltener ein statistisch signifikanter Unterschied festgestellt als mit 50 Laboratorien. Somit ist das daraus resultierende Gleichwertigkeitskriterium in hohem Maße in nicht wünschenswerter Weise von der Anzahl der Laboratorien abhängig. Je mehr Messungen und je kleiner die Streuung, desto unwahrscheinlicher ist somit ein Nachweis der Gleichwertigkeit. Dies steht in Widerspruch zu den praktischen Erfordernissen.

Abbildung 1: Häufigkeit der Erfüllung der Gleichwertigkeit hinsichtlich der Wiederfindungsrate (= keine statistisch signifikanten Unterschiede) bei einer wahren absoluten Differenz von 10%.



Da dass Prüfergebnis außerdem in starkem Maße von den jeweiligen Unsicherheitskomponenten abhängig ist, ist der Einsatz des konventionellen Ansatzes nicht zu empfehlen.

1.3 Das Prinzip der Äquivalenz

Um die in dem vorigen Abschnitt genannten Mängel zu vermeiden, liegt es nahe, den statistischen Test so umzustellen, dass in der Nullhypothese nicht unterstellt wird, dass die beiden Messverfahren gleichwertig sind, sondern dass sie unterschiedlich sind. Die dem Äquivalenzprinzip zugrundeliegende und dem medizinischen Bereich entnommene Idee besteht darin, dass zwei (Behandlungs-)Verfahren dann als äquivalent angesehen werden, wenn die Abweichungen zwischen den Verfahren eine vorgegebene Schranke signifikant unterschreiten. Dieses Prinzip kann auch auf Messverfahren angewandt werden.

Gleichwertigkeit bezieht sich dabei nicht nur auf die systematische, mittlere Abweichung, sondern auch auf den Umfang der zufälligen Abweichungen.

Die Wirkungsweise dieses Äquivalenzprinzips kann anhand der Wahrscheinlichkeiten demonstriert werden, mit denen unter unterschiedlichen Rahmenbedingungen der Nachweis der Gleichwertigkeit der Messverfahren erfolgt. Diese Wahrscheinlichkeit ist für das in diesem Bericht beschriebene Verfahren zunächst von der vorgegebenen tolerierten Abweichung der beiden Wiederfindungsraten abhängig. Beläuft sich diese auf 20%, ergeben sich für den Nachweis der Äquivalenz die in Abbildung 2 dargestellten Häufigkeiten.

Abbildung 2: Häufigkeit der Nachweis der Äquivalenz bei identischer wahrer Wiederfindungsrate (= keine statistisch signifikanten Unterschiede) und einer vorgegebenen tolerierten Abweichung von 20%.

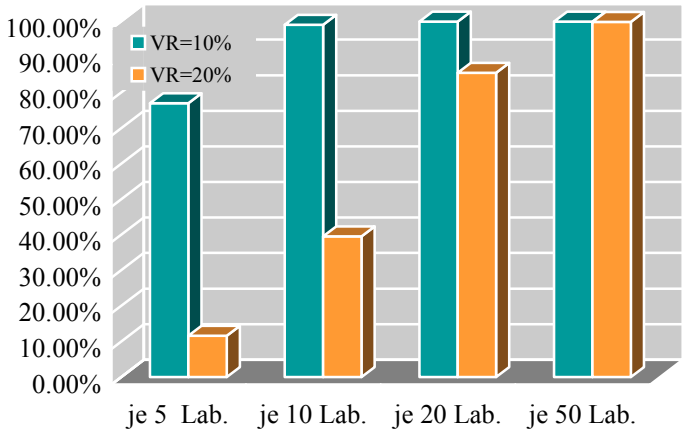


Abbildung 2 zeigt, dass der Nachweis der Äquivalenz hinsichtlich der Wiederfindungsrate umso wahrscheinlicher ist, je mehr Messungen vorliegen und je kleiner die Vergleichstandardabweichung ist. Dies entspricht den praktischen Anforderungen.

Im Rahmen von Ringversuchen ist zu gewährleisten, dass neben der Abweichung des Gesamtmittelwertes auch die Standardabweichung unter Wiederhol- und Vergleichbedingungen ein vorgegebenes Maß nicht überschreitet. Diese drei Aspekte werden in den Abschnitten 1.3, 1.4 und 1.5 eingehend betrachtet. Es werden darin die allgemeinen Äquivalenzanforderungen formuliert, die dann im Rahmen von Ringversuchen überprüft werden können. Die genannten Kriterien werden im Hinblick auf Bodenanalysen erarbeitet, so dass die spezifischen Bedingungen und Fehlerquellen bei Bodenanalysen (Methoden- bzw. Matrixeinfluss etc.) berücksichtigt werden können.

1.4 Notation

In den Abschnitten 1 und 2 wird unterstellt, dass entweder eine oder mehrere Proben p=1,...,P im Rahmen eines Ringversuchs durch unterschiedliche Labors l=1,...,L analysiert werden, welche jeweils eine von mehreren Verfahren m=1,...,M eingesetzt haben. m=1 bezeichnet dabei das Referenzverfahren. Die theoretischen, unbekannten Ringversuchskennwerte für Gesamtmittelwert, Wiederhol- und Vergleichstandardabweichung werden bezeichnet mit:

μ_{mp} = theoretischer Gesamtmittelwert für Verfahren m und Probe p

$\sigma_{r,mp}$ = theoretische Wiederholstandardabweichung für Verfahren m und Probe p

$\sigma_{R,mp}$ = theoretische Vergleichstandardabweichung für Verfahren m und Probe p.

Es handelt sich bei diesen Parametern um theoretische Größen, die durch die im Ringversuch ermittelten Daten mehr oder weniger genau angenähert bzw. geschätzt werden können. Es wird ferner unterstellt, dass für die Teilpopulationen der Ringversuchsergebnisse, die sich bei Aufteilung auf die verschiedenen Messverfahren ergeben, jeweils ein robuster Schätzwert für den Mittelwert μ_{mp} sowie Wiederhol- und Vergleichstandardabweichung $\sigma_{r,mp}$ und $\sigma_{R,,mp}$ vorliegt. Ferner wird angenommen, dass Schätzwerte für den jeweiligen Standardfehler angegeben sind.

1.5 Äquivalenz in Bezug auf die Wiederfindung

Das Äquivalenzprinzip besagt, dass zwei Messverfahren dann als gleichwertig anzusehen sind, wenn ihre Abweichungen ein gewisses Maß statistisch signifikant unterschreiten. Bezogen auf die Wiederfindungsrate bedeutet dies, dass die relative Differenz der Gesamtmittelwerte von Referenzverfahren und Verfahren m einen Toleranzwert Δ_{WFR} unterschreitet:

$$(1.1) \quad \frac{\left|\mu_{mp} - \mu_{1p}\right|}{\mu_{1p}} < \Delta_{WFR} \ .$$

Das Kriterium ist probenbezogen, so dass auch seine Überprüfung für jede einzelne Probe zu erfolgen hat. Sofern jedoch davon auszugehen ist, dass in der zugrundeliegenden Probenpopulation nur geringfügige, zufällige Unterschiede in der Wiederfindungsrate auftreten, liegt es nahe, ein verallgemeinertes Kriterium zu betrachten, bei dem die Abweichungen im quadratischen Mittel über alle untersuchten Proben betrachtet werden:

$$(1.2) \quad \left|\frac{1}{P}\sum_{p=1}^{P}\frac{\mu_{mp} - \mu_{1p}}{\mu_{1p}}\right| < \Delta_{WFR} \ .$$

Für P=1 entspricht das letztgenannte Kriterium dem zunächst vorgestellten Kriterium (1.1).

Es ist festzuhalten, dass die Verwendung eines probenübergreifenden Kriteriums (1.2) die Gefahr in sich birgt, dass Verfahrensunterschiede, die nur bei spezifischen Proben wirksam sind, sich gegenseitig aufheben. Es ist deshalb zu gewährleisten, die betrachtete Probenpopulation als hinreichend homogen angesehen werden kann, so dass derartige probenspezifische Verfahrensfehler nicht zu erwarten sind. Ist dies nicht gewährleistet, ist die Überprüfung der Äquivalenz für jede der untersuchten Proben separat durch (1.1) zu führen. Der Äquivalenznachweis bezieht sich dann nur auf den jeweiligen Probentyp.

1.6 Äquivalenz in Bezug auf die Wiederholbarkeit

Die Forderung der Äquivalenz beinhaltet nicht nur die Äquivalenz in Bezug auf systematische Abweichungen, sondern auch in Bezug auf zufällige Abweichungen unter Wiederholbedingungen. Daraus ergibt sich die Anforderung, dass die Wiederholstandardabweichungen des Referenzverfahrens, $\sigma_{r,1p}$, sowie des Vergleichverfahrens, $\sigma_{r,mp}$, sich nur um einen gewissen Grad $\Delta_r > 1$ unterscheiden dürfen. Versteht man den Begriff der Äquivalenz im reflexiven, gegenseitigen Sinne, bedeutet dies, dass

$$(1.3) \quad \frac{1}{\Delta_r} < \frac{\sigma_{r,mp}}{\sigma_{r,1p}} < \Delta_r \text{ oder gleichwertig } \left| \ln \frac{\sigma_{r,mp}}{\sigma_{r,1p}} \right| < \ln \Delta_r$$

gelten muss. Dies bedeutet, dass der relative Unterschied zwischen den Methoden kleiner sein soll als die vorgegebene Toleranz. Dabei ist sowohl der Fall zu betrachten, dass die Vergleichsmethode m eine höhere Wiederholstandardabweichung aufweist, als auch jener Fall, dass die Referenzverfahren stärkere Streuungen verursacht. Wenn der letztere Fall für den Äquivalenznachweis nicht relevant ist (da die Validität des Referenzverfahrens bereits als nachgewiesen gelten kann), genügt es, Äquivalenz in einer Richtung auf der Basis des Kriteriums

$$(1.4) \quad \frac{\sigma_{r,mp}}{\sigma_{r,1p}} < \Delta_r$$

nachzuweisen. Das Kriterium ist probenbezogen, so dass auch seine Überprüfung für jede einzelne Probe zu erfolgen hat. Wenn –gemäß des in Abschnitt 1.3 beschriebenen Kriteriums– eine probenübergreifende Betrachtungsweise zulässig ist, kann in Analogie zu Kriterium (1.2) eine probenübergreifende Betrachtung der Äquivalenz der Wiederholstandardabweichungen auf der Basis des folgenden Kriteriums (1.5) erfolgen:

$$(1.5) \quad \frac{1}{P} \sum_{p=1}^{P} \ln \frac{\sigma_{r,mp}}{\sigma_{r,1p}} < \ln \Delta_r \ .$$

Es ist darauf hinzuweisen, dass der maximale Quotient Δ_r der beiden theoretischen Wiederholstandardabweichungen auf der Grundlage der jeweiligen praktischen Anforderungen und in Abhängigkeit von der zu erwartenden Präzision der Ergebnisse festzulegen ist. Sinnvolle Werte für Δ_r liegen im Bereich von 1,2 - 1,5.

1.7 Äquivalenz in Bezug auf die Vergleichbarkeit

Die Forderung der Äquivalenz beinhaltet neben der Berücksichtigung zufälliger Abweichungen unter Wiederholbedingungen auch die zufälligen Abweichungen unter Vergleichbedingungen, d.h. mit unterschiedlichem Gerät, unterschiedlichen Operatoren und unterschiedlichen Labors. Ähnlich wie bei der Wiederholstandardabweichungen ergibt sich die Anforderung, dass sich die Vergleichstandardabweichungen des Referenzverfahrens, $\sigma_{R,1p}$, sowie des Vergleichverfahrens, $\sigma_{R,,mp}$, sich nur um einen gewissen Grad $\Delta_R > 1$ unterscheiden dürfen. Äquivalenz im reflexiven, gegenseitigen Sinne bedeutet somit, dass

$$(1.6) \quad \frac{1}{\Delta_R} < \frac{\sigma_{R,mp}}{\sigma_{R,1p}} < \Delta_R \text{ oder gleichwertig } \left| \ln \frac{\sigma_{R,mp}}{\sigma_{R,1p}} \right| < \ln \Delta_R$$

gelten muss. Wenn die Validität des Referenzverfahrens bereits als nachgewiesen gelten kann, genügt der Nachweis auf der Basis des Kriteriums

$$(1.7) \quad \frac{\sigma_{R,mp}}{\sigma_{R,1p}} < \Delta_R \ .$$

Das Kriterium ist probenbezogen, so dass auch seine Überprüfung für jede einzelne Probe zu erfolgen hat. Wenn eine probenübergreifende Betrachtungsweise zulässig ist, kann in Analogie zu Kriterium (1.5) eine probenübergreifende Betrachtung der Äquivalenz der Vergleichstandardabweichungen auf der Basis des folgenden Kriteriums erfolgen:

$$(1.8) \quad \frac{1}{P} \sum_{p=1}^{P} \ln \frac{\sigma_{R,mp}}{\sigma_{R,1p}} < \ln \Delta_R \ .$$

Gilt P=1, ist Kriterium (1.8) äquivalent zu Kriterium (1.7).

2. Äquivalenznachweis von Bestimmungsverfahren bei Ringversuchen

Soll ein Äquivalenznachweis auf der Basis von Ringversuchen erbracht werden, stellt sich zunächst das Problem, dass laborspezifische Unterschiede – unabhängig von Verfahrensunterschieden – einen erheblichen Einfluss auf die Ergebnisse haben können. Es ist auch nicht auszuschließen, dass aufgrund laborspezifischer Fehler Ausreißer auftreten, die bei Anwendung klassischer Methoden, die auf der Normalverteilung basieren, zu Fehleinschätzungen führen können. Es erscheint daher unumgänglich, dass zum Äquivalenznachweis robuste Methoden eingesetzt werden, die so beschaffen sind, dass auch noch mit einigen Ausreißerlabors ein sinnvoller Äquivalenznachweis möglich ist. Die zur robusten Schätzung von Mittelwert und Standardabweichungen erforderliche Methodik ist im Grundsatz verfügbar, doch fehlen geeignete Kennwerte, mit der die statistische Signifikanz einer Abweichung zwischen zwei robusten Mittelwerten oder zwei robusten Standardabweichungen erfasst werden kann. Gegenstand des Abschnittes 2.1 ist daher zum einen die Übersichtsdarstellung der benötigten Schätzverfahren, und zum anderen die Erarbeitung der statistischen Kennwerte, mit der der darauf aufbauende Äquivalenztest realisiert werden kann. Abschnitt 2.2 befasst sich mit der Durchführung der Äquivalenztests selbst. Grundlage des robusten Schätzverfahrens ist die Q-Methode in Verbindung mit dem Hampel-Schätzer. Es wird folgende Notation verwendet: y_{ji} bezeichnet das Messergebnis der i-ten Messung von Labor j, mit $j = 1,...,J$ und $i = 1,...,n_j$. Dabei ist es unerheblich, ob Mehrfachmessungen vorliegen, d.h. $n_j \geq 2$, oder ob alle Labors nur jeweils eine Messung vorgenommen haben, d.h. $n_j = 1$ für alle $j = 1,...,J$.

2.1 Bestimmung der Ringversuchskenndaten

2.1.1 Vergleichstandardabweichung s_R

Zunächst berechnet man die Funktion:

$$(2.1) \quad H_1(x) = \frac{1}{\binom{J}{2}} \sum_{1 \leq j_1 < j_2 \leq J} \frac{1}{n_{j_1} \cdot n_{j_2}} \sum_{k_1=1}^{n_{j_1}} \sum_{k_2=1}^{n_{j_2}} 1\left\{ \left| y_{j_1 k_1} - y_{j_2 k_2} \right| \leq x \right\}$$

Die Sprungstellen dieser Funktion werden mit $x_1,...,x_k$ bezeichnet (mit $x_1 < x_2 < ... < x_k$). Weiterhin definiert man die Funktion:

$$(2.2)\quad G_1(x_i)=\begin{cases}0{,}5\cdot\left(H_1(x_i)+H_1(x_{i-1})\right) & \text{falls} \quad i\geq 2\\ 0{,}5\cdot H_1(x_1) & \text{falls} \quad i=1 \text{ und } x_1>0\\ 0 & \text{falls} \quad i=1 \text{ und } x_1=0\end{cases}$$

für alle Sprungstellen x_i. Zwischen den Sprungstellen definiert man diese Funktion durch lineare Interpolation. Um durch Rundungsfehler verursachte Verzerrungen zu verhindern, setzt man:

$$(2.3)\quad p=0{,}25+0{,}75H_1(0)$$

und berechnet die Vergleichstandardabweichung s_R aus:

$$(2.4)\quad s_R=\frac{G_1^{-1}(p)}{\sqrt{2}\Phi^{-1}(0{,}5+0{,}5p)}$$

Dabei bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung.

Asymptotisch ist das Schätzverfahren normalverteilt und erwartungstreu. Für J Labore kann die Varianz des Schätzfehlers approximiert werden durch

$$(2.5)\quad Var[s_R]=\frac{\sigma_R^2}{2J}\left[\frac{1}{0{,}823}+\frac{7{,}516}{J}-\frac{18{,}75}{J^2}\right],\ J\geq 4\,.$$

Diese Funktion wurde mittels einer Simulationsstudie unter Annahme der Normalverteilung und der Annahme, dass keine Messwiederholungen vorliegen, berechnet. Der Ausdruck in der Klammer entspricht dem Reziprokwert der Effizienz des Schätzverfahrens. Für den Fall mit Messwiederholungen reduziert sich die Varianz geringfügig, so dass der angegebene Ausdruck auch für diesen Fall verwendet werden kann.

2.1.2 Bestimmung der Wiederholstandardabweichung s_r

Sofern Mehrfachmessungen vorliegen, kann die Q-Methode auch zur Schätzung der Wiederholstandardabweichung s_r verwendet werden. Die Wiederholstandardabweichung dient dabei primär informativen Zwecken und wird für die hier beschriebene Eignungsprüfung nicht benötigt. Grundlage der Wiederholstandardabweichung sind nicht die Differenzen zwischen den Labors, sondern die Differenzen innerhalb der Labors. Die zugehörige empirische Verteilungsfunktion der Intra-Labor-Differenzen hat die folgende Gestalt:

$$(2.6)\quad H_2(x)=\frac{1}{J}\sum_{j=1}^{J}\frac{2}{n_j(n_j-1)}\sum_{1\leq i_1<i_2\leq n_j}1\left\{\left|y_{ji_1}-y_{ji_2}\right|\leq x\right\}$$

Die Sprungstellen dieser Funktion werden mit $x_1,\ldots,x_K$ bezeichnet. Weiterhin definiert man die Funktion:

$$(2.7)\quad G_2(x_i)=\begin{cases}0{,}5\cdot\left(H_2(x_i)+H_2(x_{i-1})\right) & \text{falls} \quad i\geq 2\\ 0{,}5\cdot H_2(x_1) & \text{falls} \quad i=1 \text{ und } x_1>0\\ 0 & \text{falls} \quad i=1 \text{ und } x_1=0\end{cases}$$

für alle Sprungstellen x_i. Zwischen den Sprungstellen definiert man diese Funktion durch lineare Interpolation. Um durch Rundungsfehler verursachte Verzerrungen zu verhindern, setzt man:

$$(2.8)\quad p=0{,}5+0{,}5H_2(0)$$

und berechnet die Wiederholstandardabweichung s_r aus:

$$(2.9)\quad s_r=\frac{G_2^{-1}(p)}{\sqrt{2}\Phi^{-1}(0{,}5+0{,}5p)}\ .$$

Dabei bezeichnet Φ wiederum die Verteilungsfunktion der Standardnormalverteilung.

Asymptotisch ist das Schätzverfahren normalverteilt und erwartungstreu. Für J Labore kann die Varianz des Schätzfehlers auf der Basis der asymptotischen Verteilung wie folgt approximiert werden.

$$(2.10)\quad Var[s_r]=\frac{\sigma_r^2}{2\times e_w\times(N-J)},\ \text{mit}\ N=\sum_{j=1}^{J}n_j\ \text{für}\ J\geq 4\ .$$

Dabei bezeichnet e_w die asymptotische Effizienz für den Fall, dass alle Labors jeweils w Messwiederholungen realisieren. Für 2 bis 5 Messwiederholungen sind diese Werte in der folgenden Tabelle 1 wiedergegeben.

Tabelle 1: Asymptotische Effizienz der robusten Wiederholstandardabweichung

W	2	3	4	5
e_w	0,3675	0,463	0,521	0,557

2.1.3 Bestimmung des Mittelwertes $\hat{\mu}$

Im folgenden bezeichnet y_j das arithmetische Mittel der Messungen von Labor j,

$$(2.11)\quad y_j=\frac{1}{n_i}\sum_{i=1}^{n_j}y_{ji}$$

und im Falle ohne Mehrfachmessungen bezeichnet y_j das Messergebnis selbst, mit $j = 1,...,J$. Dann wird der robuste Mittelwert $\hat{\mu}$ gemäß der Rechenvorschrift von Hampel aus der Bestimmungsgleichung

(2.12) $$\sum_{j=1}^{J}\psi\left(\frac{y_j-\hat{\mu}}{s_R}\right)=0$$

mit

(2.13) $$\psi(x)=\begin{cases}0 & x\le -4{,}5\\ -4{,}5-x & -4{,}5<x\le -3\\ -1{,}5 & -3<x\le -1{,}5\\ x & -1{,}5<x\le 1{,}5\\ 1{,}5 & 1{,}5<x\le 3\\ 4{,}5-x & 3<x\le 4{,}5\\ 0 & x>4{,}5\end{cases}$$

berechnet. Die Lösung wird in endlich vielen Berechnungsschritten, also nicht iterativ, unter Ausnutzung der Eigenschaft, dass ψ im Argument $\hat{\mu}$ stückweise linear ist, exakt berechnet. Dabei ist zu beachten, dass die Stützstellen der linken Seite von Gleichung (2.12) – hier als Funktion von $\hat{\mu}$ aufgefasst – wie folgt lauten:

$y_j + ks_R$ mit k=-4,5, -3, -1,5, 0, 1,5, 3 und 4,5.

Es ist jene Lösung zu verwenden, die dem Median am nächsten kommt. Sofern dies nicht zu einem eindeutigen Ergebnis führt, wird als Lageparameter der Median selbst verwendet.
Asymptotisch ist das Schätzverfahren normalverteilt und erwartungstreu. Für J Labore kann die Varianz des Schätzfehlers auf der Basis der asymptotischen Verteilung konservativ wie folgt approximiert werden.

(2.14) $$Var[\hat{\mu}]=\frac{\sigma_R^2}{0{,}95\times J}\text{, für } J\ge 4\ .$$

2.2 Methodik des Äquivalenznachweis

Im folgenden bezeichnet J_{mp} die Anzahl der Labore, welche im Rahmen des Ringversuchs mit Methode m die Probe p gemessen haben.

2.2.1 Nachweis der Äquivalenz bezüglich der Wiederfindungsrate

Statistische Methodik im Falle einer Probe

Der statistische Nachweis des Kriteriums (1.1) $\dfrac{\left|\mu_{mp}-\mu_{1p}\right|}{\mu_{1p}}<\Delta_{WFR}$ erfolgt mit einem vom t-Test abgeleiteten Verfahren. Hierzu bildet man die Prüfgröße

(2.15) $$T_1 = \frac{\frac{\hat{\mu}_{mp} - \hat{\mu}_{1p}}{\hat{\mu}_{1p}}}{\frac{\sqrt{Var\left[\hat{\mu}_{mp}\right] + Var\left[\hat{\mu}_{1p}\right]}}{\hat{\mu}_{1p}}} \quad .$$

Der Nachweis der Äquivalenz kann dann als erfüllt gelten, wenn diese Prüfgröße hinreichend klein ist, wenn also $|T_1| < k$ für einen geeigneten kritischen Wert k gilt. Letzterer sollte so beschaffen sein, dass das vorgegebene Signifikanzniveau α nicht überschritten wird, d.h. k ist so festzulegen, dass unter der Nullhypothese: $\frac{|\mu_{mp} - \mu_{1p}|}{\mu_{1p}} \geq \Delta_{WFR}$ die Wahrscheinlichkeit dafür, dass $-k < T_1 < k$ gilt, nicht größer als α wird, d.h.

(2.16) $$P(-k < T_1 < k) = P(T_1 < k) - P(T_1 < -k) \leq \alpha \quad .$$

Diese Wahrscheinlichkeit nimmt unter der eingeschränkten Nullhypothese $\frac{\mu_{mp} - \mu_{1p}}{\mu_{1p}} = \Delta_{WFR}$ ihren maximalen Wert an. Unter dieser Voraussetzung $\frac{\mu_{mp} - \mu_{1p}}{\mu_{1p}} = \Delta_{WFR}$ kann die Verteilung der Zufallsvariable T_1 durch eine nicht-zentrale t-Verteilung mit

(2.17) $$df_1 = \min\{J_{mp}, J_{1p}\} - 1$$

Freiheitsgraden und dem Nichtzentralitätsparameter

(2.18) $$\delta_1 = \frac{\Delta_{WFR}}{\frac{\sqrt{Var\left[\hat{\mu}_{mp}\right] + Var\left[\hat{\mu}_{1p}\right]}}{\mu_{1p}}}$$

approximiert werden. Durch den gewählten Freiheitsgrad ist ein konservatives Testverhalten gewährleistet, d.h. das tatsächliche Signifikanzniveau überschreitet nicht die Signifikanzgrenze.

Aus (2.16) ergibt sich für den kritischen Wert $k = k(\alpha, df_1, \Delta_{WFR}, \delta_1)$ die Bedingung

(2.19) $$F_{t(df_1,\delta_1)}(k) - F_{t(df_1,\delta_1)}(-k) = \alpha \quad .$$

Dabei bezeichnet $F_{t(df_1,\delta_1)}$ die Verteilungsfunktion der nicht-zentralen t-Verteilung mit df_1 Freiheitsgraden und dem Nichtzentralitätsparameter δ_1.

Es ergibt sich die folgende Testentscheidung: Der statistische Nachweis des Kriteriums (1.1) kann zum Signifikanzniveau α als erbracht gelten, wenn

(2.20) $$\left|\frac{\hat{\mu}_{mp} - \hat{\mu}_{1p}}{\hat{\mu}_{1p}}\right| < \frac{\sqrt{Var\left[\hat{\mu}_{mp}\right] + Var\left[\hat{\mu}_{1p}\right]}}{\hat{\mu}_{1p}} k(\alpha, df_1, \Delta_{WFR}, \delta_1) \quad .$$

Vorgehensweise bei probenspezifischer Prüfung

1. Festlegung des Signifikanzniveaus α. Typischerweise wird ein Wert von α=1%, 5% oder 10% festgelegt.
2. Festlegung der maximal tolerierten theoretischen relativen Abweichung Δ_{WFR}. Geeignete Werte für Δ_{WFR} liegen im Bereich von 10-20%.
3. Berechnung der empirischen relativen Abweichung $\frac{\hat{\mu}_{mp} - \hat{\mu}_{1p}}{\hat{\mu}_{1p}}$. Wenn der Absolutbetrag dieser Abweichung bereits den vorgegebenen Wert Δ_{WFR} überschreitet, können die weiteren Berechnungen abgebrochen werden, denn dann ist das Äquivalenzkriterium (1.1) nicht erfüllt.
4. Berechnung der Varianzen $Var\left[\hat{\mu}_{mp}\right]$ und $Var\left[\hat{\mu}_{1p}\right]$ gemäß (2.14).
5. Berechnung der Freiheitsgrade df_1 und des Nichtzentralitätsparameters δ_1 gemäß (2.17) und (2.18).
6. Berechnung des kritischen Wertes k aus der Bedingung (2.19). Hierfür ist ein iteratives Vorgehen erforderlich.
7. Berechnung der maximal tolerierten empirischen Abweichung $\frac{\sqrt{Var\left[\hat{\mu}_{mp}\right]+Var\left[\hat{\mu}_{1p}\right]}}{\hat{\mu}_{1p}} k(\alpha, df_1, \Delta_{WFR}, \delta_1)$.
8. Überprüfung des Kriteriums (2.20). Äquivalenz ist dann nachgewiesen, wenn der Absolutbetrag der empirischen relativen Abweichung $\frac{\hat{\mu}_{mp} - \hat{\mu}_{1p}}{\hat{\mu}_{1p}}$ nicht größer ist als die maximal tolerierte empirische Abweichung $\frac{\sqrt{Var\left[\hat{\mu}_{mp}\right]+Var\left[\hat{\mu}_{1p}\right]}}{\hat{\mu}_{1p}} k(\alpha, df_1, \Delta_{WFR}, \delta_1)$.

Beispiel: Arsenbestimmung in Boden mittels AAS und ICP. Probenspezifische Äquivalenz bezüglich der Wiederfindung

In vier Ringversuchen BAM1-BAM4 erfolgte bei jeweils einer Probe durch unterschiedliche Labore die Bestimmung von Arsen mittels AAS (Referenzverfahren) bzw. ICP. Die entsprechend der obigen Vorgehensweise ermittelten Resultate sind in Tabelle 2

wiedergegeben. Das vorgegebene Signifikanzniveau liegt bei 5%, die maximal tolerierte Abweichung der Mittelwerte bei 15%.

Tabelle 2: Ergebnisse der Gleichwertigkeitsprüfung bezüglich Wiederfindung bei probenspezifischer Betrachtung

Probe	p	BAM1	BAM2	BAM3	BAM4
Signifikanzniveau	α x 100%	5%	5%	5%	5%
Maximal tolerierte theoretische Abweichung	Δ_{WFR} x 100%	±15%	±15%	±15%	±15%
Anzahl Labore / AAS	J_{1p}	52	67	56	42
Mittelwert / AAS	$\hat{\mu}_{1p}$	159,3	30,4	10,45	1365
Vergleich-Stdabw. / AAS	$s_{R,1p}$	17,76	3,43	1,612	98,6
Rel.Vergleich-Stdabw. / AAS	$s_{R,1p} / \hat{\mu}_{1p}$ x 100%	11,15%	11,28%	15,43%	7,22%
Standardvarianz / AAS	$Var[\hat{\mu}_{1p}]$	6,3850	0,1848	0,0488	243,6581
Anzahl Labore / ICP	J_{mp}	35	31	15	58
Mittelwert / ICP	$\hat{\mu}_{mp}$	162,5	31,6	11,15	1409
Vergleich-Stdabw. / ICP	$s_{R,mp}$	12,67	2,99	2,238	93,4
Standardvarianz / ICP	$Var[\hat{\mu}_{mp}]$	4,8279	0,3036	0,3515	158,3223
Rel.Vergleich-Stdabw. / ICP	$s_{R,mp} / \hat{\mu}_{1p}$ x 100%	7,80%	9,46%	20,07%	6,63%
Freiheitsgrade	df_1	34	30	14	41
Nichtzentralitätsparameter	δ_1	7,1359	6,5249	2,4774	10,2122
Kritischer Wert	K	5,22	4,64	0,83	8,06
Empirische Abweichung	$\frac{\hat{\mu}_{mp} - \hat{\mu}_{1p}}{\hat{\mu}_{1p}}$ x 100%	2,01%	3,95%	6,70%	3,22%
Maximal tolerierte empirische Abweichung	$\frac{\sqrt{Var[\hat{\mu}_{mp}] + Var[\hat{\mu}_{1p}]}}{\hat{\mu}_{1p}} k$ x 100%	±10,98%	±10,67%	±5,05%	±11,84%
Äquivalenz		ja	ja	nein	ja

Ergebnis dieser Äquivalenzprüfung ist somit, dass bei den in den Ringversuchen BAM1, BAM2 und BAM4 verwendeten Proben bezüglich der Wiederfindungsrate Äquivalenz des ICP-Verfahrens zum Referenzverfahren nachgewiesen werden kann. Hingegen kann bei der

in Ringversuch BAM3 verwendeten Probe kein Nachweis erfolgen. Folgende statistische Ursachen sind hier zu benennen: Zum einen ist die tatsächliche relative Abweichung mit nahezu 7% relativ groß, zum anderen ist die relative Vergleichstandardabweichung bei beiden Verfahren vergleichsweise groß. Nicht zuletzt spielt auch die geringere Anzahl von Laboratorien eine gewisse Rolle. Aus analytischer Sicht wäre hinzuzufügen, dass die in Ringversuch BAM3 verwendete Probe einen deutlich geringeren Arsen-Gehalt als die übrigen Proben aufweist und demzufolge auch deutlich näher an der Bestimmungsgrenze liegt. Abweichungen zwischen verschiedenen Verfahren sind hier vielfach besonders deutlich, während bei höheren Konzentrationen nur geringe Abweichungen der Wiederfindung festzustellen sind.

Statistische Methodik im Falle mehrerer Proben

In analoger Weise wird bei mehreren Proben das Kriterium (1.2) überprüft: Hierzu definiert man die Prüfgröße

$$(2.21)\quad T_2 = \frac{\sum_{p=1}^{P} \frac{\hat{\mu}_{mp} - \hat{\mu}_{1p}}{\hat{\mu}_{1p}}}{\sqrt{\sum_{p=1}^{P} \frac{Var\left[\hat{\mu}_{mp}\right] + Var\left[\hat{\mu}_{1p}\right]}{\hat{\mu}_{1p}^2}}},$$

den Nichtzentralitätsparameter

$$(2.22)\quad \delta_2 = \frac{P\Delta_{WFR}}{\sqrt{\sum_{p=1}^{P} \frac{Var\left[\hat{\mu}_{mp}\right] + Var\left[\hat{\mu}_{1p}\right]}{\hat{\mu}_{1p}^2}}}$$

und die Freiheitsgrade

$$(2.23)\quad df_2 = \left(\sum_{p=1}^{P} \min\left\{J_{mp}, J_{1p}\right\} \right) - P \ .$$

Dabei ist festzuhalten, dass diese Festlegung der Freiheitsgrade auf der Annahme beruht, dass die für die verschiedenen Proben ermittelten relativen Standardabweichungen nicht allzu starken Schwankungen unterliegen. Ansonsten ist eine konservativere Festlegung der Freiheitsgrade erforderlich.

Es ergibt sich folgende Testentscheidung: Der statistische Nachweis des Kriteriums (1.2) kann zum Signifikanzniveau α als erbracht gelten, wenn

$$(2.24)\quad \left| \sum_{p=1}^{P} \frac{\hat{\mu}_{mp} - \hat{\mu}_{1p}}{\hat{\mu}_{1p}} \right| < \sqrt{\sum_{p=1}^{P} \frac{Var\left[\hat{\mu}_{mp}\right] + Var\left[\hat{\mu}_{1p}\right]}{\hat{\mu}_{1p}^2}} k(\alpha, df_2, \Delta_{WFR}, \delta_2) \text{ oder gleichwertig hiermit}$$

$$\left|\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp}-\hat{\mu}_{1p}}{\hat{\mu}_{1p}}\right| < \frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var\left[\hat{\mu}_{mp}\right]+Var\left[\hat{\mu}_{1p}\right]}{\hat{\mu}_{1p}^{2}}}k(\alpha, df_2, \Delta_{WFR}, \delta_2)$$

Dabei ergibt sich der kritische Wert $k = k(\alpha, df_2, \Delta_{WFR}, \delta_2)$ aus der Bedingung

(2.25) $F_{t(df_2,\delta_2)}(k) - F_{t(df_2,\delta_2)}(-k) = \alpha$.

Vorgehensweise bei probenübergreifender Prüfung

1. Festlegung des Signifikanzniveaus α (α=1%, 5% oder 10%). Typischerweise wird ein Wert von α=1%, 5% oder 10% festgelegt.
2. Festlegung der maximal tolerierbaren theoretischen mittleren relativen Abweichung Δ_{WFR}. Geeignete Werte für Δ_{WFR} liegen im Bereich von 10-20%.
3. Berechnung der empirischen mittleren relativen Abweichung $\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp}-\hat{\mu}_{1p}}{\hat{\mu}_{1p}}$. Wenn der Absolutbetrag dieser Abweichung bereits den maximal tolerierbaren Wert Δ_{WFR} überschreitet, können die weiteren Berechnungen abgebrochen werden, denn dann ist das Äquivalenzkriterium (1.1) nicht erfüllt.
4. Berechnung der Varianzen $Var\left[\hat{\mu}_{mp}\right]$ und $Var\left[\hat{\mu}_{1p}\right]$ gemäß (2.14).
5. Berechnung der Freiheitsgrade df_2 und des Nichtzentralitätsparameters δ_2 gemäß (2.22) und (2.23).
6. Iterative Berechnung des kritischen Wertes *k* aus der impliziten Bedingung (2.25).
7. Berechnung der maximal tolerierten empirischen Abweichung $\frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var\left[\hat{\mu}_{mp}\right]+Var\left[\hat{\mu}_{1p}\right]}{\hat{\mu}_{1p}^{2}}}k(\alpha, df_2, \Delta_{WFR}, \delta_2)$.
8. Überprüfung des Kriteriums (2.24). Äquivalenz ist dann nachgewiesen, wenn der Absolutbetrag der empirischen relativen Abweichung $\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp}-\hat{\mu}_{1p}}{\hat{\mu}_{1p}}$ nicht größer ist als die maximal tolerierte empirische Abweichung $\frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var\left[\hat{\mu}_{mp}\right]+Var\left[\hat{\mu}_{1p}\right]}{\hat{\mu}_{1p}^{2}}}k(\alpha, df_2, \Delta_{WFR}, \delta_2)$.

Beispiel: Arsenbestimmung in Boden mittels AAS und ICP. Probenübergreifende Äquivalenz bezüglich der Wiederfindung

Die oben bereits vorgestellten Ringversuchsergebnisse der Bestimmung von Arsen mittels AAS (Referenzverfahren) bzw. ICP sollen nun probenübergreifend ausgewertet werden. Das vorgegebene Signifikanzniveau liegt wiederum bei 5%, die maximal tolerierte Abweichung der Mittelwerte bei 15%. Tabelle 3 zeigt das Ergebnis. Es ist darauf hinzuweisen, dass diese probenübergreifende Auswertung auf vier Proben aus vier unterschiedlichen Ringversuchen basiert. Die verwendete statistische Methodik lässt dies zu, solange gewährleistet ist, dass sich die betrachteten Messverfahren in den vier Ringversuchen nicht verändert haben und solange sich die beiden Messverfahren bei allen vier Proben bzw. Ringversuchen gleich verhalten. Ein Verfahren, mit dem diese Annahme überprüft werden kann, wird in Abschnitt 2.3.3 vorgestellt.

Tabelle 3: Ergebnisse der Gleichwertigkeitsprüfung bezüglich der Wiederfindung bei probenübergreifender Betrachtung

Signifikanzniveau	α x 100%	5%
Maximal tolerierte theoretische Abweichung	Δ_{WFR} x 100%	±15%
Freiheitsgrade	df_2	119
Nichtzentralitätsparameter	δ_2	8,6137
Kritischer Wert	k	6,83
Mittlere empirische Abweichung	$\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp}-\hat{\mu}_{1p}}{\hat{\mu}_{1p}}$ x 100%	3,97%
Maximal tolerierte empirische Abweichung	$\frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var\left[\hat{\mu}_{mp}\right]+Var\left[\hat{\mu}_{1p}\right]}{\hat{\mu}_{1p}^2}k}$ x 100%	±11,90%
Äquivalenz		ja

Bei probenübergreifender Betrachtung zeigt sich, dass die Äquivalenz als nachgewiesen gelten kann. Tatsächlich wäre auch noch bei einer mittleren Abweichung von fast 12% das Äquivalenzkriterium erfüllt. Dies erklärt sich daraus, dass durch Zusammenfassung der vier Proben eine sehr hohe statistische Sicherheit resultiert, so dass die maximal tolerierte

empirische Abweichung nahe an die maximal tolerierte theoretische Abweichung von 15% heranrückt.

2.2.2 Nachweis der Äquivalenz bezüglich Vergleich- und Wiederholstandardabweichung

Im folgenden wird die Methodologie anhand der Vergleichstandardabweichung vorgestellt. Die Vorgehensweise bei der Wiederholstandardabweichung ist analog.

Der statistische Nachweis des Kriteriums (1.7) $\ln\frac{\sigma_{R,mp}}{\sigma_{R,1p}} < \ln\Delta_R$ erfolgt auf Basis der Normalapproximation der logarithmierten Standardabweichungen. Hierzu betrachtet man die empirische Differenz der logarithmierten Vergleichstandardabweichungen

(2.26) $$\ln\frac{s_{R,mp}}{s_{R,1p}} = \ln s_{R,mp} - \ln s_{R,1p}\,.$$

Die Varianz dieses Ausdrucks kann approximativ berechnet werden gemäß

(2.27) $$Var\left[\ln\frac{s_{R,mp}}{s_{R,1p}}\right] = \frac{Var\left[s_{R,mp}\right]}{\sigma^2_{R,mp}} + \frac{Var\left[s_{R,1p}\right]}{\sigma^2_{R,1p}},$$

wobei zur Bestimmung der Standardvarianzen die Formel (2.5) verwendet wird. Aus den beiden Ausdrücken (2.26) und (2.27) bildet man die Prüfgröße

(2.28) $$Z_1 = \frac{\ln\frac{s_{R,mp}}{s_{R,1p}}}{\sqrt{\frac{Var\left[s_{R,mp}\right]}{\sigma^2_{R,mp}} + \frac{Var\left[s_{R,1p}\right]}{\sigma^2_{R,1p}}}}\,.$$

Z_1 ist im Falle identischer theoretischer Vergleichstandardabweichungen asymptotisch standardnormalverteilt. Der Nachweis der Äquivalenz kann dann als erfüllt gelten, wenn diese Prüfgröße hinreichend klein ist, wenn also $Z_1 < z$ für einen geeigneten kritischen Wert *z* gilt. Letzterer sollte so beschaffen sein, dass das vorgegebene Signifikanzniveau α nicht überschritten wird, d.h. *z* ist so festzulegen, dass unter der Nullhypothese: $\ln\frac{\sigma_{R,mp}}{\sigma_{R,1p}} \geq \ln\Delta_R$ die Wahrscheinlichkeit dafür, dass $Z_1 < z$ gilt, nicht größer als α wird, d.h.

(2.29) $$P\left(Z_1 < z \,\middle|\, \ln\frac{\sigma_{R,mp}}{\sigma_{R,1p}} \geq \ln\Delta_R\right) \leq \alpha\,.$$

Diese Wahrscheinlichkeit nimmt unter der eingeschränkten Nullhypothese $\ln\frac{\sigma_{R,mp}}{\sigma_{R,1p}} = \ln\Delta_R$ ihren maximalen Wert an. Unter dieser Voraussetzung kann die Verteilung der Zufallsvariable Z_1 durch eine Normalverteilung mit Varianz 1 und Mittelwert

(2.30) $$d_1 = \frac{\ln\Delta_R}{\sqrt{\frac{Var\left[s_{R,mp}\right]}{\sigma_{R,mp}^2} + \frac{Var\left[s_{R,1p}\right]}{\sigma_{R,1p}^2}}}$$

approximiert werden. Aus (2.29) ergibt sich damit für den kritischen Wert z die Bedingung

(2.31) $\Phi(z - d_1) = \alpha$ oder gleichwertig $z - d_1 = z_\alpha$ und $z = d_1 + z_\alpha = d_1 - z_{1-\alpha}$.

Dabei bezeichnet Φ die Verteilungsfunktion der Standardnormalverteilung und z_α das zugehörige α-Quantil. Es ergibt sich die folgende Testentscheidung: Der statistische Nachweis des Kriteriums (1.7) kann zum Signifikanzniveau α als erbracht gelten, wenn

(2.32) $$Z_1 \le \frac{\ln\Delta_R}{\sqrt{\frac{Var\left[s_{R,mp}\right]}{s_{R,mp}^2} + \frac{Var\left[s_{R,1p}\right]}{s_{R,1p}^2}}} - z_{1-\alpha}$$ oder gleichwertig

$$\ln\frac{s_{R,mp}}{s_{R,1p}} \le \ln\Delta_R - z_{1-\alpha}\sqrt{\frac{Var\left[s_{R,mp}\right]}{s_{R,mp}^2} + \frac{Var\left[s_{R,1p}\right]}{s_{R,1p}^2}}$$

gilt. Dabei wurden die theoretischen Vergleichstandardabweichungen durch die empirischen Vergleichstandardabweichungen ersetzt.

Vorgehensweise bei probenspezifischer Prüfung

1. Festlegung des Signifikanzniveaus α. Typischerweise wird ein Wert von α=1%, 5% oder 10% festgelegt
2. Festlegung des maximal tolerierten theoretischen Quotienten Δ_R. Geeignete Werte für Δ_R liegen im Bereich um 1,3 (1,2-1,5). Mit einem Wert von Δ_R=1,3 darf die theoretische Vergleichstandardabweichung des Vergleichverfahrens höchstens den 1,3-fachen Wert der Vergleichstandardabweichung des Referenzverfahrens betragen. Mit einem Wert von 1,3 wird also sichergestellt, dass z.B. bei einem Vergleichvariationskoeffizienten des Referenzverfahrens von 10% bei nachgewiesener Äquivalenz die zu vergleichende Methode einen Vergleichvariationskoeffizienten aufweist, der signifikant kleiner als 13% ist.

3. Berechnung der empirischen Differenz $\ln\frac{s_{R,mp}}{s_{R,1p}} = \ln s_{R,mp} - \ln s_{R,1p}$. Wenn diese Differenz bereits den vorgegebenen Wert $\ln(\Delta_R)$ überschreitet, können die weiteren Berechnungen abgebrochen werden, denn dann ist das Äquivalenzkriterium (1.7) nicht erfüllt.
4. Berechnung der Varianzen $Var\left[s_{R,1p}\right]$ und $Var\left[s_{R,mp}\right]$ gemäß (2.5).
5. Berechnung der Standardabweichung $\sqrt{\frac{Var\left[s_{R,mp}\right]}{s_{R,mp}^2} + \frac{Var\left[s_{R,1p}\right]}{s_{R,1p}^2}}$.
6. Ermittlung von $z_{1-\alpha}$, d.h. des $(1-\alpha)$-Quantils der Standardnormalverteilung.
7. Überprüfung des Kriteriums (2.32). Äquivalenz ist dann nachgewiesen, wenn die empirische Differenz $\ln\frac{s_{R,mp}}{s_{R,1p}} = \ln s_{R,mp} - \ln s_{R,1p}$ nicht größer ist als die maximal tolerierte empirische Differenz $\ln\Delta_R - z_{1-\alpha}\sqrt{\frac{Var\left[s_{R,mp}\right]}{s_{R,mp}^2} + \frac{Var\left[s_{R,1p}\right]}{s_{R,1p}^2}}$.

Beispiel: Arsenbestimmung in Boden mittels AAS und ICP. Probenspezifische Äquivalenz bezüglich der Vergleichbarkeit

Auf Grundlage der bereits vorgestellten Ringversuchsergebnisse der Ringversuche BAM1-BAM4 zur Bestimmung von Arsen mittels AAS (Referenzverfahren) bzw. ICP soll hier die Äquivalenzprüfung bezüglich der Vergleichstandardabweichung vorgestellt werden. Die entsprechend der obigen Vorgehensweise ermittelten Resultate sind in folgender Tabelle 4 wiedergegeben. Das vorgegebene Signifikanzniveau liegt bei 5%, der maximal tolerierte theoretische Quotient der Vergleichstandardabweichungen bei Δ_R = 1,5.

Tabelle 4: Ergebnisse der probenspezifischen Gleichwertigkeitsprüfung bezüglich der Vergleichbarkeit

Probe	P	BAM1	BAM2	BAM3	BAM4
Signifikanzniveau	α	5%	5%	5%	5%
Maximal tolerierter theoretischer Quotient	Δ_R	1,5	1,5	1,5	1,5
Maximal tolerierte theoretische Differenz	$\ln(\Delta_R)$ x 100%	40,5%	40,5%	40,5%	40,5%
Anzahl Labore / AAS	J_{1p}	52	67	56	42
Mittelwert / AAS	$\hat{\mu}_{1p}$	159,3	30,4	10,45	1365
Vergleich-Stdabw. / AAS	$s_{R,1p}$	17,76	3,43	1,612	98,6
Rel.Vergleich-Stdabw. / AAS	$s_{R,1p} / \hat{\mu}_{1p}$ x 100%	11,15%	11,28%	15,43%	7,22%
Varianz von s_R / AAS	$Var[s_{R,1p}]$	4,1025	0,1162	0,0312	160,1103
Anzahl Labore / ICP	J_{mp}	35	31	15	58
Mittelwert / ICP	$\hat{\mu}_{mp}$	162,5	31,6	11,15	1409
Vergleich-Stdabw. / ICP	$s_{R,mp}$	12,67	2,99	2,238	93,4
Rel.Vergleich-Stdabw. / ICP	$s_{R,mp} / \hat{\mu}_{1p}$ x 100%	7,80%	9,46%	20,07%	6,63%
Varianz von s_R / ICP	$Var[s_{R,mp}]$	3,2438	0,2074	0,2726	100,7029
Standardabweichung der empirischen Differenz	$\sqrt{\frac{Var[s_{R,mp}]}{s_{R,mp}^2} + \frac{Var[s_{R,1p}]}{s_{R,1p}^2}}$	0,1822	0,1818	0,2577	0,1674
(1-α)-Quantil der Standardnormalverteilung	$z_{1-\alpha}$	1,645	1,645	1,645	1,645
Empirische Differenz	$\ln\frac{s_{R,mp}}{s_{R,1p}} = \ln s_{R,mp} - \ln s_{R,1p}$ x 100%	-33,77%	-13,73%	32,81%	-5,42%
Maximal tolerierte empirische Differenz	$\ln\Delta_R - z_{1-\alpha}\sqrt{\frac{Var[s_{R,mp}]}{s_{R,mp}^2} + \frac{Var[s_{R,1p}]}{s_{R,1p}^2}}$ x 100%	10,57%	10,63%	-1,85%	13,01%
Äquivalenz		ja	ja	nein	ja

Für die Ringversuche BAM1, BAM2 und BAM4 liegt die empirische Differenz unterhalb der maximal tolerierten empirischen Differenz, so dass hier Äquivalenz nachgewiesen werden kann. Bei der in Ringversuch BAM3 verwendeten Probe ist die Vergleichstandardabweichung – möglicherweise aufgrund der erheblich niedrigeren Gehalte – für die ICP-Methodik erheblich höher als bei der Referenzmethodik.

Statistische Methodik im Falle mehrerer Proben

Im Falle mehrerer Proben wird das Kriterium (1.8) $\frac{1}{P}\sum_{p=1}^{P}\ln\frac{\sigma_{R,mp}}{\sigma_{R,1p}} < \ln\Delta_R$ wie folgt überprüft.

Da die Prüfgröße

$$(2.33)\quad Z_2 = \frac{\sum_{p=1}^{P}\ln\frac{s_{R,mp}}{s_{R,1p}}}{\sqrt{\sum_{p=1}^{P}\left(\frac{Var\left[s_{R,mp}\right]}{\sigma_{R,mp}^2}+\frac{Var\left[s_{R,1p}\right]}{\sigma_{R,1p}^2}\right)}},$$

unter der Voraussetzung identischer theoretischer Vergleichstandardabweichungen standardnormalverteilt ist, ergibt sich folgende Testentscheidung. Der statistische Nachweis des Kriteriums (1.8) kann zum Signifikanzniveau α als erbracht gelten, wenn

$$(2.34)\quad Z_2 \le \frac{P\ln\Delta_R}{\sqrt{\sum_{p=1}^{P}\left(\frac{Var\left[s_{R,mp}\right]}{s_{R,mp}^2}+\frac{Var\left[s_{R,1p}\right]}{s_{R,1p}^2}\right)}} - z_{1-\alpha} \text{ oder gleichwertig hiermit}$$

$$\frac{1}{P}\sum_{p=1}^{P}\ln\frac{s_{R,mp}}{s_{R,1p}} \le \ln\Delta_R - z_{1-\alpha}\sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(\frac{Var\left[s_{R,mp}\right]}{s_{R,mp}^2}+\frac{Var\left[s_{R,1p}\right]}{s_{R,1p}^2}\right)}$$

Dabei ist $z_{1-\alpha}$ das (1-α)-Quantil der Standardnormalverteilung.

Vorgehensweise bei probenübergreifender Prüfung

1. Festlegung des Signifikanzniveaus α. Typischerweise wird ein Wert von α=1%, 5% oder 10% festgelegt
2. Festlegung des maximal tolerierten theoretischen Quotienten Δ_R. Geeignete Werte für Δ_R liegen im Bereich um 1,3 (1,2-1,5). Mit einem Wert von Δ_R=1,3 darf die theoretische Vergleichstandardabweichung des Vergleichverfahrens höchstens den 1,3-fachen Wert der Vergleichstandardabweichung des Referenzverfahrens betragen. Mit einem Wert von 1,3 wird also sichergestellt, dass z.B. bei einem Vergleichvariationskoeffizienten des Referenzverfahrens von 10% bei nachgewiesener Äquivalenz die zu vergleichende Methode einen Vergleichvariationskoeffizienten aufweist, der signifikant kleiner als 13% ist.

3. Berechnung der mittleren empirischen Differenz $\frac{1}{P}\sum_{p=1}^{P}\ln\frac{s_{R,mp}}{s_{R,1p}}$. Wenn diese Abweichung bereits den vorgegebenen Wert $\ln(\Delta_R)$ überschreitet, können die weiteren Berechnungen abgebrochen werden, denn dann ist das Äquivalenzkriterium (1.8) nicht erfüllt.
4. Berechnung der Varianzen $Var\left[s_{R,1p}\right]$ und $Var\left[s_{R,mp}\right]$ gemäß (2.5).
5. Berechnung der Standardabweichung $\sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(\frac{Var\left[s_{R,mp}\right]}{s_{R,mp}^2}+\frac{Var\left[s_{R,1p}\right]}{s_{R,1p}^2}\right)}$.
6. Ermittlung von $z_{1-\alpha}$, d.h. des $(1-\alpha)$-Quantils der Standardnormalverteilung.
7. Überprüfung des Kriteriums (2.34). Äquivalenz ist dann nachgewiesen, wenn die mittlere empirische Differenz $\frac{1}{P}\sum_{p=1}^{P}\ln\frac{s_{R,mp}}{s_{R,1p}}$ nicht größer ist als die maximal tolerierte empirische Differenz $\ln\Delta_R - z_{1-\alpha}\sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(\frac{Var\left[s_{R,mp}\right]}{s_{R,mp}^2}+\frac{Var\left[s_{R,1p}\right]}{s_{R,1p}^2}\right)}$.

Beispiel: Arsenbestimmung in Boden mittels AAS und ICP. Probenübergreifende Äquivalenz bezüglich der Vergleichbarkeit

Die oben bereits vorgestellten Ringversuchsergebnisse der Bestimmung von Arsen mittels AAS (Referenzverfahren) bzw. ICP sollen nun probenübergreifend ausgewertet werden. Das vorgegebene Signifikanzniveau liegt wiederum bei 5%, der maximal tolerierte Quotient der Vergleichstandardabweichungen bei 1,5.

Tabelle 5: Ergebnisse der probenübergreifenden Gleichwertigkeitsprüfung bezüglich der Vergleichbarkeit

Signifikanzniveau	α x 100%	5%
Maximal tolerierter Quotient Vergleichstandardabweichungen	Δ_R	1,5
Maximal tolerierte theoretische Differenz	$\ln(\Delta_R)$ x 100%	40,5%
Standardabweichung	$\sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(\frac{Var[s_{R,mp}]}{s_{R,mp}^2}+\frac{Var[s_{R,1p}]}{s_{R,1p}^2}\right)}$	0,200
(1-α)-Quantil der Standardnormalverteilung	$z_{1-\alpha}$	1,645
Mittlere empirische Differenz	$\frac{1}{P}\sum_{p=1}^{P}\ln\frac{s_{R,mp}}{s_{R,1p}}$ x 100%	-5,03%
Maximal tolerierte empirische Differenz	$\ln\Delta_R - z_{1-\alpha}\sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(\frac{Var[s_{R,mp}]}{s_{R,mp}^2}+\frac{Var[s_{R,1p}]}{s_{R,1p}^2}\right)}$ x 100%	7,6%
Äquivalenz		ja

Bei probenübergreifender Betrachtung zeigt sich, dass die Äquivalenz bezüglich der Vergleichbarkeit als nachgewiesen gelten kann. Die mittlere empirische Differenz liegt bei -5%, während die maximal tolerierte empirische Differenz bei 7,6% liegt. Vergleicht man die Resultate mit der probenspezifischen Analyse, zeigt sich wiederum sehr deutlich der große Einfluss einer großen Anzahl von Proben bzw. Laboratorien, welche die statistische Sicherheit wesentlich verbessern können.

2.3 Anmerkungen und Empfehlungen

2.3.1 Festlegung der maximal tolerierten relativen Abweichung der Wiederfindung

Damit zwei in Ringversuchen überprüfte Messverfahren als gleichwertig angesehen werden können, ist ein Äquivalenznachweis bezüglich Mittelwert, Wiederhol- und Vergleichstandardabweichung erforderlich. Grundlage dieses Nachweis sind die maximal zulässigen Abweichungen Δ_R, Δ_r und Δ_{WFR}, die geeignet festgelegt werden müssen, wobei ein Ausgleich zwischen dem analytisch Möglichen und den praktischen Anforderungen gefunden werden sollte.

Grundsätzlich sollte gewährleistet sein, dass der Einsatz verschiedener äquivalenter Verfahren nicht zu einer deutlichen Vergrößerung der resultierenden Vergleichstandardabweichung führt. So sollte Δ_{WFR} keinesfalls größer als die relative Referenzvergleichstandardabweichung sein. Wenn die Vergleichstandardabweichung $s_{1R} = s_{mR}$ beider Methoden gleich ist und wenn b den durch Anwendung dieser Methode auftretenden Bias kennzeichnet, verschlechtert sich die Vergleichbarkeit zwangsläufig, wenn beide Methoden eingesetzt werden. Wenn q den Anteil jener Labore bezeichnet, welche die Vergleichmethode anwenden, ergibt sich für die resultierende relative Vergleichstandardabweichung

$$(2.35) \qquad \frac{\sqrt{\left(\frac{\sigma_{1R}}{\mu_1}\right)^2 + \left(\frac{b}{\mu_1}\right)^2 q(1-q)}}{1+\frac{bq}{\mu_1}} .$$

Dieser Ausdruck wird maximal für q=0,5, d.h. wenn 50% aller Laboratorien die Vergleichmethode anwenden, erhöht sich die relative Vergleichstandardabweichung um den Faktor

$$(2.36) \qquad \frac{\sqrt{1+\frac{1}{4}\left(\frac{b}{\sigma_{1R}}\right)^2}}{1+\frac{b}{2\mu_1}} .$$

Wenn nun vorausgesetzt wird, dass der maximale Bias b nicht größer als die Vergleichstandardabweichung des Referenzverfahrens sein darf, erhöht sich die resultierende relative Vergleichstandardabweichung um maximal $\sqrt{1{,}25}$=11,8%. Dieser Wert erscheint vertretbar: Solange Δ_{WFR} nicht größer als die relative Vergleichstandardabweichung des Referenzverfahrens ist, ist keine unnötige Aufblähung der resultierenden Vergleichstandardabweichung zu befürchten.[3]

[3] Es ist allerdings zu prüfen, ob das Kriterium der Vergleichbarkeit angepasst werden sollte: Wenn für die Referenzverfahren die Vergleichstandardabweichung bei 30% liegt und die Abweichung der Wiederfindungsrate bei $\Delta_{WFR} = 0{,}2 = 20\%$ liegt, müsste – wenn ein äquivalentes Messverfahren zugelassen wird - die relative Vergleichbarkeit von 0,831 auf den Wert 0,3 x 2,77 +0,2=1,031 angehoben werden.

2.3.2 Festlegung der maximal tolerierten relativen Abweichung bei Wiederhol- und Vergleichbarkeit

Die Grenzen Δ_R und Δ_r sind unter Abwägung der möglichen Teilnehmerzahl, des Signifikanzniveaus sowie der jeweiligen analytischen Anforderungen an die Wiederhol- und Vergleichstandardabweichung sehr sorgfältig festzulegen. So erscheint ein Wert von $\Delta_R = \Delta_r = 1{,}3$ aus statistischer Sicht mit den Anforderungen der Praxis im Einklang. Allerdings erschwert eine solche Festlegung den Äquivalenznachweis zumindest dann, wenn Wiederhol- und Vergleichstandardabweichung des zu prüfenden Verfahrens nicht kleiner sind als bei der Referenzverfahren. Sofern vergleichbare Proben ggf. im mehreren Ringversuchen untersucht worden sind, liegt es nahe, die Ergebnisse zu kombinieren, um praktikable Toleranzgrenzen festlegen zu können.

2.3.3 Wann ist eine probenübergreifende Äquivalenzprüfung zulässig?

Wie oben bereits erwähnt, setzt eine probenübergreifende Analyse voraus, dass probenspezifische Methodeneffekte vernachlässigbar klein sind. Ansonsten ist zu befürchten, dass das Prüfergebnis weniger die Leistungsfähigkeit des Verfahrens selbst reflektiert, sondern nur auf die jeweilige Probenauswahl zurückzuführen ist. Um dies hinsichtlich der Wiederfindung zu überprüfen, kann folgende Vorgehensweise verwendet werden:

Die Differenz zwischen der probenspezifischen relativen Abweichung $\frac{\hat{\mu}_{ms} - \hat{\mu}_{1s}}{\hat{\mu}_{1s}}$ bei Probe *s* und der mittleren relativen Abweichung $\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp} - \hat{\mu}_{1p}}{\hat{\mu}_{1p}}$ sollte nur zufällig um den Nullpunkt schwanken. Die Varianz dieser Differenz lässt sich wie folgt berechnen:

$$(2.37)\quad Var\left[\frac{\hat{\mu}_{ms} - \hat{\mu}_{1s}}{\hat{\mu}_{1s}} - \frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp} - \hat{\mu}_{1p}}{\hat{\mu}_{1p}}\right]$$

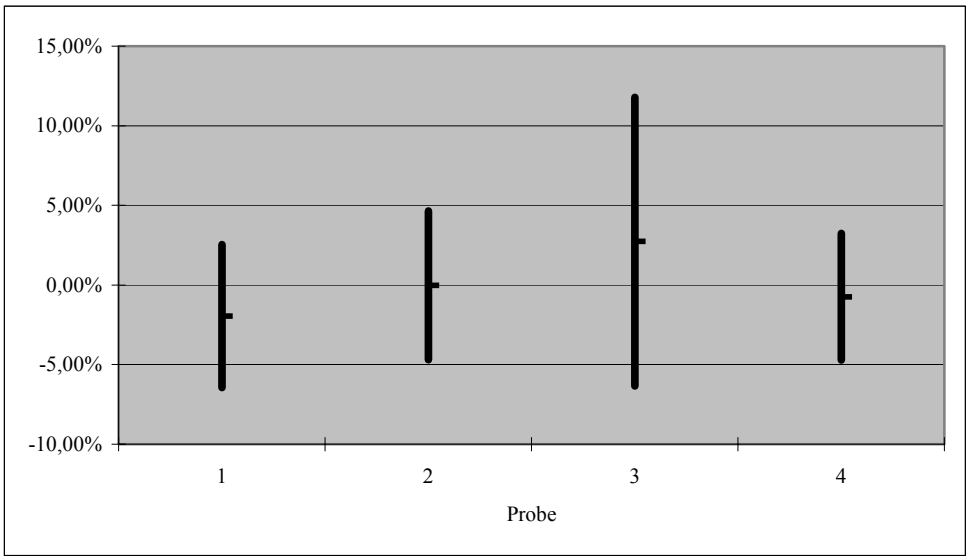
$$\approx \left(1 - \frac{1}{P}\right)^2 \frac{1}{\mu_{1s}^2}\left(Var\left(\hat{\mu}_{ms}\right) + Var\left(\hat{\mu}_{1s}\right)\right) + \frac{1}{P^2}\sum_{\substack{p=1 \\ p\neq s}}^{P}\frac{1}{\mu_{1p}^2}\left(Var\left(\hat{\mu}_{mp}\right) + Var\left(\hat{\mu}_{1p}\right)\right).$$

Man kann nun ein Konfidenzintervall für die betrachtete Differenz bilden, worin die Differenz selbst den Mittelpunkt bildet und – im Falle eines Konfidenzniveaus von 95% – die Endpunkte durch die mit 1,96 multiplizierte Standardabweichung der Differenz bestimmt sind. Solange jedes dieser Konfidenzintervalle den Nullpunkt überdeckt, kann davon ausgegangen werden, dass probenspezifische Methodeneffekte nicht nachweisbar sind.

Beispiel

Für die Ringversuchsergebnisse des obigen Beispiels ergeben sich für die vier Proben die in Abbildung 3 dargestellten Konfidenzintervalle. Alle vier Konfidenzintervalle überdecken den Nullpunkt, so dass ein probenspezifischer Verfahrenseffekt nicht nachgewiesen werden kann. Somit erscheint eine probenübergreifende Auswertung gerechtfertigt. Hinzuweisen ist darauf, dass trotz der vergleichsweise hohen Teilnehmerzahlen die Länge der Konfidenzintervalle beträchtlich ist. Eine Ursache hierfür liegt in dem Umstand, dass sich die vergleichsweise hohe Unsicherheit im Ringversuch BAM3 aufgrund der Differenzenbildung auch in den übrigen Ringversuchen niederschlägt.

Abbildung 3: 95%-Konfidenzintervall für den probenspezifischen Methodenbias



Teil II: Äquivalenznachweis durch In-House-Experimente für Bodenanalysen

3. Was bedeutet Gleichwertigkeit und Äquivalenz von Bestimmungsverfahren bei In-House-Experimenten?

Der Nachweis der Äquivalenz umfasst – wie in Abschnitt 1 bereits dargelegt – immer mehrere Aspekte, die jedoch – je nach Datenbasis – möglicherweise nur eingeschränkt erfasst werden können und somit spezifisch interpretiert werden müssen. Bei In-House-Experimenten lässt sich, da die Experimente definitionsgemäß immer im gleichen Labor durchgeführt werden, kein Nachweis bezüglich der zufälligen Fehler unter Vergleichbedingungen erbringen, mit der Folge, dass alle Aussagen hinsichtlich der Äquivalenz sich nur auf das jeweilige Labor beziehen können. Um dennoch die geforderte analytische Sicherheit zu gewährleisten, ist es deshalb ratsam, die In-house Analyik mittels zertifizierter Matrix-Referenzmaterialien durchzuführen. Die sich daraus ergebenden Konsequenzen und auch die Vorteile von In-House-Experimenten gegenüber Ringversuchen werden in diesem Abschnitt vorgestellt.

3.1 Statistisches Modell

In den Abschnitten 3 und 4 wird unterstellt, dass mehrere Proben p=1,...,P in mehreren Runs i=1,...,I_{mp} (zeitlich zusammenhängenden Messläufen) und möglicherweise mehreren Messwiederholungen (oder mit unterschiedlicher Aufstockung) j=1,..., J_{mpi} analysiert werden, wobei jeweils eine von mehreren Messverfahren m=1,...,M eingesetzt wird. m=1 bezeichnet wiederum das Referenzverfahren, und μ_p den theoretischen Mittelwert (d.h. den wahren Gehalt) für Probe p. Weiterhin bezeichnet μ_{mp} den theoretischen Mittelwert für Methode m. Handelt es sich bei Probe p um ein zertifiziertes Matrix-Referenzmaterial, kann der wahre Gehalt μ_p durch den zugehörigen Referenzwert $\hat{\mu}_p$ approximiert werden. Dieser Referenzwert wird nicht notwendiger mittels des Referenzverfahrens „1“ ermittelt, so dass deren theoretischer Mittelwert μ_{1p} in der Regel nicht mit dem Referenzwert $\hat{\mu}_p$ übereinstimmt. Für die zugehörigen Messwerte Y_{mpij} wird das folgende statistische Modell unterstellt:

(3.1) $$Y_{mpij} = \mu_p + \delta_{mp} + \alpha_{mi} + \varepsilon_{mpij}$$

mit dem systematischen Methodenbias δ_{mp}, dem Runeffekt α_{mi} bei Methode m in Run i, sowie der zufälligen Messabweichung ε_{mpij}. Die beiden letztgenannten Variablen werden als stochastisch unabhängige und normalverteilte Zufallsvariablen aufgefasst, mit dem Erwartungswert 0 und den Standardabweichungen $\sigma_{run,r}$ und σ_{rm}. Bildet man auf beiden Seiten der Gleichung (3.1) den Erwartungswert, ergibt sich die Gleichung

(3.2) $\mu_{mp} - \mu_p = \delta_{mp}$.

σ_{mr} ist die Standardabweichung der laborinternen relativen Messabweichungen unter Wiederholbedingungen und entspricht daher der relativen Wiederholstandardabweichung. Aus der Summe der beiden Varianzkomponenten lässt sich die relative Standardabweichung für die Messabweichungen unter laborinternen Vergleichbedingungen

(3.3) $\sigma_{Rm} = \sqrt{\sigma^2_{run,m} + \sigma^2_{rm}}$

ableiten.

Es ist zu beachten, dass die Standardabweichungen für die Messabweichungen in Modell (3.1) unabhängig vom Probenindex *p* sind, d.h. probenübergreifend einheitlich festgelegt sind. Auf eine solche Festlegung kann nicht verzichtet werden, da im Gegensatz zu Ringversuchen die bei In-House-Experimenten ermittelten Daten in der Regel keine probenspezifische Festlegung ermöglichen. Wenn jedoch die Varianzen nicht als konstant angenommen werden können, empfiehlt sich für die Modellberechnungen eine geeignete lineare Transformation der Messwerte: Sofern ein näherungsweise proportionaler Zusammenhang zwischen den Messabweichungen und den gemessenen Konzentrationen vorliegt, sollten die Berechnungen nicht auf der Basis der Messwerte selbst durchgeführt werden. In diesem Falle sollte zuvor eine Division durch den jeweiligen Referenzwert vorgenommen werden, d.h. anstelle von Y_{mpij} wird $\widetilde{Y}_{mpij} = \frac{Y_{mpij}}{\mu_p}$ betrachtet. In diesem Fall ergibt sich aus Modell (3.1) folgendes vereinfachtes Modell:

(3.4) $\widetilde{Y}_{mpij} = 1 + \widetilde{\delta}_{mp} + \widetilde{\alpha}_{mi} + \widetilde{\varepsilon}_{mpij}$.

Die auf der rechten Seite stehenden Ausdrücke für den Methodenbias und die beiden Komponenten der Messabweichungen beziehen sich nunmehr nicht mehr auf die relativen, sondern auf die absoluten Abweichungen. Dies ist bei der Interpretation der Resultate entsprechend zu berücksichtigen.

3.2 Notation

3.2.1 Varianzen und Mittelwerte bei Vorliegen von Messwiederholungen

Ein Run (Analysenserie) umfasst jeweils eine oder mehrere unabhängige Einzelmessungen, die unter einheitlichen Bedingungen kurz hintereinander durchgeführt werden. Unterschiedliche Runs werden in der Regel unter variierenden Bedingungen und nicht am gleichen Tag durchgeführt. Die Anzahl der Methoden wird mit M bezeichnet, die Anzahl der Proben mit P. Die Anzahl der Runs bei Methode m und Probe p wird mit I_{mp} bezeichnet, die Anzahl der zugrundeliegenden Messwiederholungen für Methode m bei Probe p und Run i mit $J_{mpi.}$. Weiterhin steht

$$(3.5) \quad \overline{Y}_{mpi} = \frac{1}{J_{mpi}} \sum_{j=1}^{J_{mp}} Y_{mpij}$$

für das arithmetische Mittel bei Probe p in Run i, sowie

$$(3.6) \quad s_{mpi} = \sqrt{\frac{1}{J_{mpi} - 1} \sum_{j=1}^{J_{mp}} \left(Y_{mpij} - \overline{Y}_{mpi} \right)^2}$$

für die zugehörige Standardabweichung. Daraus ermittelt man die Wiederholstandardabweichung

$$(3.7) \quad s_{rm} = \sqrt{\frac{1}{df_{rm}} \sum_{i,p} s_{mpi}^2} \text{ , mit den Freiheitsgraden } df_{rm} = \sum_{i,p} \left(J_{mpi} - 1 \right) .$$

wobei über alle Runs i=1,...,I_{mp} und alle Proben p=1,...,P summiert wird.

Aus dem arithmetischen Mittel für Probe p über alle Runs i=1,...,I_{mp},

$$(3.8) \quad \hat{\mu}_{mp} = \frac{1}{\sum_{i=1}^{I_{mp}} J_{mpi}} \sum_{i=1}^{I_{mp}} J_{mpi} \overline{Y}_{mpi} ,$$

ermittelt man weiter

$$(3.9) \quad s_{mp} = \sqrt{\frac{1}{I_{mp} - 1} \sum_{i=1}^{I_{mp}} J_{mpi} \left(\overline{Y}_{mpi} - \hat{\mu}_{mp} \right)^2}$$

und daraus die probenspezifische Run-Standardabweichung

$$(3.10) \quad s_{run,mp} = \sqrt{\left(s_{mp}^2 - s_{rm}^2 \right) / \widetilde{J}_{mp}}$$

mit

$$(3.11) \quad \widetilde{J}_{mp} = \frac{1}{I_{mp} - 1} \left[\sum_{i=1}^{I_{mp}} J_{mpi} - \frac{\sum_{i=1}^{I_{mp}} J_{mpi}^2}{\sum_{i=1}^{I_{mp}} J_{mpi}} \right] .$$

Durch Mittelung über alle Proben erhält man die Run-Standardabweichung

(3.12) $s_{run,m} = \sqrt{\frac{1}{P}\sum s^2_{run,mp}}$

sowie die In-House-Vergleichstandardabweichung

(3.13) $s_{Rm} = \sqrt{s^2_{run,m} + s^2_{rm}}$

mit den Freiheitsgraden

(3.14) $df_{Rm} = \sum_{p=1}^{P}\left(I_{mp} - 1\right)$.

Mit diesen Bezeichnungen ergeben sich die Varianzen der Schätzfehler unter Normalverteilung approximativ wie folgt:

(3.15) $Var\left[\hat{\mu}_{mp}\right] = \frac{\sum_{i=1}^{I_{mp}} J^2_{mpi}}{\left(I_{mp}\bar{J}_{mp}\right)^2}\sigma^2_{run,m} + \frac{1}{I_{mp}\bar{J}_{mp}}\sigma^2_{r,m}$,

mit

(3.16) $\bar{J}_{mp} = \frac{1}{I_{mp}}\sum_{i=1}^{I_{mp}} J_{mpi}$,

(3.17) $Var\left[s_{rm}\right] = \frac{\sigma^2_{rm}}{2df_{rm}}$,

(3.18) $Var\left[s_{Rm}\right] = \frac{\sigma^2_{Rm}}{2df_{Rm}}$.

3.2.2 Varianzen und Mittelwerte ohne Messwiederholungen

Sofern in jedem Messlauf nur ein Einzelwert gemessen wird und unterschiedliche Messläufe nur an unterschiedlichen Tagen unter unterschiedlichen Bedingungen realisiert werden dürfen, ergibt sich ein vereinfachtes Rechenschema. Dabei ist zu beachten, dass eine Ermittlung der Wiederholstandardabweichung naturgemäß nicht möglich ist. Das hier beschriebene Rechenschema leitet sich aus den unter Abschnitt 3.2.1 beschriebenen Berechnungsvorschriften ab, wenn für alle Methoden *m*, Proben *p* und Runs *i* unterstellt wird, dass J_{mpi} =1 gilt. Weiterhin wird auf die Mitführung des Index j verzichtet, d.h. Y_{mpij} entspricht Y_{mpi}.

Aus dem arithmetischen Mittel für Methode *m* und Probe *p* über alle Runs i=1,...,I_{mp},

(3.19) $\hat{\mu}_{mp} = \frac{1}{I_{mp}}\sum_{i=1}^{I_{mp}} Y_{mpi}$,

ermittelt man die probenspezifische Run-Standardabweichung

$$(3.20)\quad s_{run,mp} = \sqrt{\frac{1}{I_{mp}-1}\sum_{i=1}^{I_{mp}}\left(Y_{mpi}-\hat{\mu}_{mp}\right)^2}\ .$$

Durch Mittelung über alle Proben erhält man die Run-Standardabweichung

$$(3.21)\quad s_{run,m} = s_{Rm} = \sqrt{\frac{1}{P}\sum s_{run,mp}^2}\ ,$$

die hier zugleich der In-House-Vergleichstandardabweichung entspricht. Die Anzahl der zugrundeliegenden Freiheitsgrade wird mit

$$(3.22)\quad df_{Rm} = \sum_{p=1}^{P}\left(I_{mp}-1\right)$$

bezeichnet. Mit diesen Bezeichnungen ergeben sich die Varianzen der Schätzfehler unter Normalverteilung approximativ wie folgt:

$$(3.23)\quad Var\left[\hat{\mu}_{mp}\right] = \frac{\sigma_{Rm}^2}{I_{mp}}\ ,$$

$$(3.24)\quad Var\left[s_{Rm}\right] = \frac{\sigma_{Rm}^2}{2df_{Rm}}\ .$$

3.3 Äquivalenz in Bezug auf die Wiederfindung

In Analogie zum Äquivalenzbegriff bei Ringversuchen wird auch bei In-House-Experimenten gefordert, dass die relative Differenz der Gesamtmittelwerte von Referenzverfahren und Methode m einen Toleranzwert Δ_{WFR} unterschreitet:

$$(3.25)\quad \left|\delta_{mp}\right| = \frac{\left|\mu_{mp}-\mu_p\right|}{\mu_p} < \Delta_{WFR}\ .$$

Das Kriterium ist probenbezogen, so dass auch seine Überprüfung für jede einzelne Probe zu erfolgen hat. Ebenso wie bei Ringversuchen kann auch ein verallgemeinertes Kriterium verwendet werden, bei dem die mittlere Abweichung über alle untersuchten Proben betrachtet wird:

$$(3.26)\quad \left|\frac{1}{P}\sum_{p=1}^{P}\delta_{mp}\right| = \left|\frac{1}{P}\sum_{p=1}^{P}\frac{\mu_{mp}-\mu_p}{\mu_p}\right| < \Delta_{WFR}\ .$$

Es ist wiederum festzuhalten, dass die Verwendung eines probenübergreifenden Kriteriums (3.26) die Gefahr in sich birgt, dass Methodenunterschiede, die nur bei spezifischen Proben wirksam sind, nicht mehr auffällig werden. Sofern letztere nicht vernachlässigbar sind, ist die Überprüfung der Äquivalenz für jede der untersuchten Proben separat durchzuführen. Der

Äquivalenznachweis bezieht sich dann nur auf den jeweiligen Probentyp. Für P=1 entspricht das letztgenannte Kriterium (3.26) dem zunächst vorgestellten Kriterium (3.25). Die folgenden Betrachtungen werden daher für das allgemeinere Kriterium (3.26) angestellt.
Zu bemerken ist ferner, dass die Abschätzung des wahren Gehaltes μ_p mittels des Referenzwertes $\hat{\mu}_p$ erfolgen sollte, nicht jedoch mit dem in der In-House-Studie berechneten Mittelwert, der einen systematischen Fehler aufweisen könnte.

3.4 Äquivalenz in Bezug auf die laborinterne Wiederholbarkeit

In Analogie zum Äquivalenzbegriff für Ringversuche ergibt sich die Anforderung, dass sich die laborinternen Wiederholstandardabweichungen des Referenzverfahrens, σ_{r1}, sowie des Vergleichverfahrens, σ_{rm} , nur um einen gewissen Grad $\Delta_r > 1$ unterscheiden dürfen. Im Sinne der reflexiven Äquivalenz bedeutet dies, dass

(3.27) $$\frac{1}{\Delta_r} < \frac{\sigma_{rm}}{\sigma_{r1}} < \Delta_r \text{ oder gleichwertig } \left|\ln \frac{\sigma_{rm}}{\sigma_{r1}}\right| < \ln \Delta_r$$

gelten muss. Dies bedeutet, dass der relative Unterschied zwischen den Methoden kleiner sein soll als die vorgegebene Toleranz. Da in der Regel die Validität des Referenzverfahren bereits als nachgewiesen gelten kann, genügt es vielfach, Äquivalenz in einer Richtung auf der Basis des Kriteriums

(3.28) $$\frac{\sigma_{rm}}{\sigma_{r1}} < \Delta_r$$

nachzuweisen. Das Kriterium ist probenübergreifend, so dass eine Verallgemeinerung im Sinne des Kriteriums (1.5) nicht erforderlich ist.

3.5 Äquivalenz in Bezug auf die laborinterne Vergleichbarkeit

In Analogie zur Äquivalenz in Bezug auf die laborinterne Wiederholstandardabweichung ergibt sich die Anforderung, dass sich die laborinternen Vergleichstandardabweichungen des Referenzverfahrens, σ_{R1}, sowie des Vergleichverfahrens, σ_{Rm} , nur um einen gewissen Grad $\Delta_R > 1$ unterscheiden dürfen. Dies bedeutet, dass

(3.29) $$\frac{1}{\Delta_R} < \frac{\sigma_{Rm}}{\sigma_{R1}} < \Delta_R \text{ oder gleichwertig } \left|\ln \frac{\sigma_{Rm}}{\sigma_{R1}}\right| < \ln \Delta_R$$

gelten muss. Damit wird gefordert, dass der relative Unterschied der Vergleichstandardabweichungen kleiner sein soll als die vorgegebene Toleranz. Zum Nachweis der Äquivalenz in einer Richtung ergibt sich das Kriterium

(3.30) $\frac{\sigma_{Rm}}{\sigma_{R1}} < \Delta_R$.

4. Äquivalenznachweis bei In-House-Analysen

Soll ein Äquivalenznachweis auf der Basis von In-House-Analysen erbracht werden, werden alle Untersuchungen in demselben Labor durchgeführt. Damit ist in der Regel ein besserer Zugriff auf die Daten möglich und zudem sind wesentlich homogenere Messbedingungen gegeben. Somit kann auf die Anwendung robuster statistischer Verfahren verzichtet werden. Die grundsätzliche Methodik entspricht dem Vorgehen beim Äquivalenznachweis durch Ringversuche, wobei allerdings zu beachten ist, dass an die Stelle der Laboratorien die Messläufe (Runs) treten.

Ausreißerverdächtige Messwerte sollten zuvor mittels geeigneter Testverfahren, z.B. mit dem Grubbstest, identifiziert werden. Sofern die Ursache für den Ausreißerwert ermittelt und ausgeschaltet werden kann, kann der betreffende Wert aus dem Datensatz entfernt werden.

Eine wesentliche Voraussetzung für einen tragfähigen Äquivalenznachweis auf der Basis von In-House-Experimenten besteht darin, dass zertifizierte Matrix-Referenzmaterialien in die Messungen einbezogen werden, die erst eine Bewertung hinsichtlich eines systematischen Fehlers erlauben. Dies bedeutet, dass sich die entsprechenden Varianzen für das Referenzverfahren aus den Angaben des Herstellers des Referenzmaterials ergeben. Wird für den Referenzwert ein Vertrauensintervall der Breite b_p angegeben, kann aus diesem die Varianz des Mittelwertes wie folgt abgeleitet werden: Sofern ein Vertrauensintervall zum Konfidenzniveau $1-\alpha$ auf Basis von N Einzelwerten ermittelt wurde, gilt für die Breite dieses Vertrauensintervalls $b_p = 2t_{N-1,1-\alpha/2}\sqrt{Var\left[\hat{\mu}_p\right]}$, d.h.

$$(4.1) \quad Var\left[\hat{\mu}_p\right] = \frac{b_p^2}{4t_{N-1,1-\alpha/2}^2} \ .$$

Dabei entspricht $Var\left[\hat{\mu}_p\right]$ der durch N dividierten empirischen Varianz der N Einzelwerte. Beachtet man, dass sich der t-Wert für ein 95%-Konfidenzniveau bei mittleren Stichprobengrößen im Bereich von 2 bewegt, kann nach einer Faustformel $Var\left[\hat{\mu}_p\right] = \frac{b_p^2}{16}$ berechnet werden.

4.1 Nachweis der Äquivalenz bezüglich der Wiederfindung unter Verwendung zertifizierter Referenzmaterialien

4.1.1 Statistische Methodik

Ist ein Referenzwert für die untersuchte Probe bekannt, erfolgt der Äquivalenznachweis nicht unter Bezugnahme auf das Referenzverfahren, sondern auf den betreffenden Referenzwert. Damit und unter Verwendung der Freiheitsgrade unter Vergleichbedingungen erhält man aus dem Prüfkriterium (2.24) für Ringversuche das abgewandelte Prüfkriterium für In-House-Analysen auf Basis einer oder mehrerer Proben:

$$(4.2) \quad \left| \frac{1}{P} \sum_{p=1}^{P} \frac{\hat{\mu}_{mp} - \hat{\mu}_p}{\hat{\mu}_p} \right| < \frac{1}{P} \sqrt{\sum_{p=1}^{P} \frac{Var\left[\hat{\mu}_{mp}\right] + Var\left[\hat{\mu}_p\right]}{\hat{\mu}_p^2}} k(\alpha, df_2, \Delta_{WFR}, \delta_2),$$

mit dem Nichtzentralitätsparameter

$$(4.3) \quad \delta_3 = \frac{P\Delta_{WFR}}{\sqrt{\sum_{p=1}^{P} \frac{Var\left[\hat{\mu}_{mp}\right] + Var\left[\hat{\mu}_p\right]}{\hat{\mu}_p^2}}}$$

den durch die Anzahl der Runs bestimmten Freiheitsgraden

$$(4.4) \quad df_3 = \left(\sum_{p=1}^{P} I_{mp} \right) - P,$$

und dem kritischen Wert $k = k(\alpha, df_3, \Delta_{WFR}, \delta_3)$, der implizit durch die Bedingung

$$(4.5) \quad F_{t(df_3,\delta_3)}(k) - F_{t(df_3,\delta_3)}(-k) = \alpha$$

definiert ist.

4.1.2 Vorgehensweise

1. Festlegung des Signifikanzniveaus α. (α=1%, 5% oder 10%).
2. Festlegung der als maximal tolerierbaren theoretischen mittleren relativen Abweichung Δ_{WFR} (10-20%).
3. Berechnung der empirischen mittleren relativen Abweichung $\frac{1}{P} \sum_{p=1}^{P} \frac{\hat{\mu}_{mp} - \hat{\mu}_p}{\hat{\mu}_p}$. Wenn der Absolutbetrag dieser Abweichung bereits den maximal tolerierbaren Wert Δ_{WFR} überschreitet, können die weiteren Berechnungen abgebrochen werden, denn dann ist das Äquivalenzkriterium (3.6) nicht erfüllt.
4. Berechnung der Varianzen $Var\left[\hat{\mu}_{mp}\right]$ gemäß Abschnitt 3.2.1 bzw. 3.2.2. Ermittlung der Varianzen $Var\left[\hat{\mu}_p\right]$ anhand der Angaben des Herstellers des Referenzmaterials.

5. Berechnung der Freiheitsgrade df_3 und des Nichtzentralitätsparameters δ_3 gemäß (4.3) und (4.4).
6. Iterative Berechnung des kritischen Wertes k aus der impliziten Definition (4.5).
7. Berechnung der maximal tolerierten empirischen Abweichung $\frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var[\hat{\mu}_{mp}]+Var[\hat{\mu}_p]}{\hat{\mu}_p^2}}k(\alpha, df_2, \Delta_{WFR}, \delta_2)$.
8. Überprüfung des Kriteriums (4.2). Äquivalenz ist dann nachgewiesen, wenn der Absolutbetrag der empirischen relativen Abweichung $\frac{\hat{\mu}_{mp}-\hat{\mu}_p}{\hat{\mu}_p}$ nicht größer ist als die maximal tolerierte empirische Abweichung $\frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var[\hat{\mu}_{mp}]+Var[\hat{\mu}_p]}{\hat{\mu}_p^2}}k(\alpha, df_2, \Delta_{WFR}, \delta_2)$.

4.1.3 Beispiel

Zur Bestimmung der Summe der 16 EPA-PAK-Gehalte wurden verschiedene Extraktionsmethoden und Extraktionsmittel kombiniert und jeweils 4 unabhängige Bestimmungen durchgeführt, die in diesem Beispiel als Einzelbestimmungen von vier Messläufen interpretiert werden sollen. Die Messergebnisse sind in der folgenden Tabelle 6 wiedergegeben.

Tabelle 6: Bestimmung der Summe der 16 EPA-PAK-Gehalte bei 4 Proben mit 13 verschiedenen Methoden.

Abkürzungen: ACN = Acetonitril, Ac/PE = Aceton/Petrolether, W/NaCl/Ac/PE = Wasser/Natriumchlorid/Aceton/Petrolether

Boden 1 / trocken						**Boden 2**					
	Durchgang				**MW**		**Durchgang**				**MW**
Ultraschall	**1**	**2**	**3**	**4**	**mg/kg**	**Ultraschall**	**1**	**2**	**3**	**4**	**mg/kg**
Toluol	57,50	54,69			56,10	Toluol	6,51	6,27	6,34	6,91	6,51
ACN	56,20	53,75	54,80	52,25	54,25	ACN	8,23	8,72	8,43	8,26	8,41
Ac/PE	53,86	58,24	54,44	61,89	57,11	Ac/PE	7,78	7,18	7,39	7,18	7,38
ASE						**ASE**					
Toluol	63,18	65,63	58,47	62,33	62,40	Toluol	8,80	8,92	8,48	8,5	8,68
ACN	58,23	58,21	61,34	59,63	59,35	ACN	8,95	8,79	8,48	8,68	8,73
Ac/PE	57,44	62,84	59,92	65,88	61,52	Ac/PE	9,24	9,29	8,84	8,99	9,09
Soxhlet						**Soxhlet**					
Toluol	64,88	68,02	59,33	53,86	61,52	Toluol	7,30	7,39	6,96	7,18	7,21
ACN	53,44	57,49	53,82	57,91	55,67	ACN	8,19	7,62	7,84	7,92	7,89
Ac/PE	57,52	59,39	61,80	62,65	60,34	Ac/PE	6,66	6,99	6,70	6,8	6,79
Schütteln						**Schütteln**					
Toluol	50,52	51,18	49,08	54,89	51,42	Toluol	4,88	4,56	4,80	4,98	4,81
ACN	48,11	49,14	48,03	46,55	47,96	ACN	7,50	8,19	7,93	7,87	7,87
Ac/PE	59,68	50,53	56,69	52,58	54,87	Ac/PE	6,40	6,61	6,51	6,64	6,54
VDLUFA						**VDLUFA**					
W/NaCl/Ac/PE	61,17	72,15	65,46	61,47	65,06	W/NaCl/Ac/PE	6,21	6,03	6,10	6,29	6,16

Boden 1 / feucht						**Boden 3**					
	Durchgang				**MW**		**Durchgang**				**MW**
Ultraschall	**1**	**2**	**3**	**4**	**mg/kg**	**Ultraschall**	**1**	**2**	**3**	**4**	**mg/kg**
Toluol	46,86	48,22	53,42	50,64	49,79	Toluol	162,11	162,94	163,71	166,69	163,86
ACN	54,62	57,36	55,51	58,00	56,37	ACN	163,38	172,68	169,50	168,12	168,42
Ac/PE	59,46	61,87	60,32	59,24	60,22	Ac/PE	179,61	177,93	179,61	172,78	177,48
ASE						**ASE**					
Toluol	63,96	58,61	62,13	59,58	61,07	Toluol	170,14	171,76	170,79	174,13	171,71
ACN	58,96	59,25	57,71	62,83	59,69	ACN	167,25	170,02	167,36	171,79	169,11
Ac/PE	62,01	59,44	59,65	57,75	59,71	Ac/PE	174,37	176,03	163,62	168,22	170,56
Soxhlet						**Soxhlet**					

Toluol	62,49	65,83	65,16	63,55	64,26	Toluol	175,26	178,46	168,26	171,15	173,28
ACN	60,48	60,44	61,24	63,78	61,49	ACN	174,40	173,43	164,01	171,47	170,83
Ac/PE	57,17	63,92	57,65	58,50	59,31	Ac/PE	163,42	167,96	162,78	164,45	164,65
Schütteln						**Schütteln**					
Toluol	48,67	50,45	47,48	47,75	48,59	Toluol	164,52	161,62	158,28	168,83	163,31
ACN	50,51	51,66	51,30	51,80	51,32	ACN	161,90	166,27	173,20	168,67	167,51
Ac/PE	54,09	53,15	52,65	52,44	53,08	Ac/PE	180,28	176,13	178,99	176,28	177,92
VDLUFA						**VDLUFA**					
W/NaCl/Ac/PE	49,32	49,88	50,78	47,91	49,47	W/NaCl/Ac/PE	174,14	170,42	161,43	162,38	167,09

Als Referenzverfahren dient im folgenden jeweils das VDLUFA-Verfahren. Um die Daten vergleichbar zu machen, werden gemäß der in Abschnitt 3.1 beschriebenen Vorgehensweise alle Messwerte durch den Mittelwert des VDLUFA-Verfahrens für die jeweilige Probe dividiert.

Die auf der Basis dieser normierten Messwerte ermittelten Kennwerte sind in den folgenden Tabellen eingetragen.

Tabelle 7: Ergebnisse der probenübergreifenden Gleichwertigkeitsprüfung bezüglich Wiederfindung für die Extraktionsmethode Ultraschall.

		Toluol	ACN	Ac/PE
Signifikanzniveau	α	5%	5%	5%
Maximal tolerierte theoretische Abweichung	Δ_{WFR}	±15%	±15%	±15%
Freiheitsgrade	df_3	10	12	12
Nichtzentralitätsparameter	δ_3	7,97	9,48	8,53
Kritischer Wert	k	5,48	6,78	6,02
Mittlere empirische Abweichung	$\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp}-\hat{\mu}_p}{\hat{\mu}_p}$	-2,35%	8,68%	8,90%
Maximal tolerierte empirische Abweichung	$\frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var[\hat{\mu}_{mp}]+Var[\hat{\mu}_p]}{\hat{\mu}_p^2}k}$	10,31%	10,73%	10,59%
Äquivalenz		ja	ja	ja

Tabelle 8: Ergebnisse der probenübergreifenden Gleichwertigkeitsprüfung bezüglich Wiederfindung für die Extraktionsmethode ASE.

		Toluol	ACN	Ac/PE
Signifikanzniveau	α	5%	5%	5%
Maximal tolerierte theoretische Abweichung	Δ_{WFR}	±15%	±15%	±15%
Freiheitsgrade	df_3	12	12	12
Nichtzentralitätsparameter	δ_3	8,67	9,41	8,42
Kritischer Wert	k	6,14	6,73	5,93
Mittlere empirische Abweichung	$\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp}-\hat{\mu}_p}{\hat{\mu}_p}$	15,75%	13,69%	16,24%
Maximal tolerierte empirische Abweichung	$\frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var[\hat{\mu}_{mp}]+Var[\hat{\mu}_p]}{\hat{\mu}_p^2}k}$	10,62%	10,72%	10,57%
Äquivalenz		nein	nein	nein

Eine Vergleichbarkeit war auch nicht zu erwarten, da bei ASE meist höhere Ausbeuten zu erwarten sind.

Tabelle 9: Ergebnisse der probenübergreifenden Gleichwertigkeitsprüfung bezüglich Wiederfindung für die Extraktionsmethode Soxhlet.

		Toluol	ACN	Ac/PE
Signifikanzniveau	α	5%	5%	5%
Maximal tolerierte theoretische Abweichung	Δ_{WFR}	±15%	±15%	±15%
Freiheitsgrade	df_3	12	12	12
Nichtzentralitätsparameter	δ_3	7,30	9,09	8,62
Kritischer Wert	k	5,03	6,46	6,1
Mittlere empirische Abweichung	$\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp}-\hat{\mu}_p}{\hat{\mu}_p}$	11,30%	10,06%	5,35%
Maximal tolerierte empirische Abweichung	$\frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var[\hat{\mu}_{mp}]+Var[\hat{\mu}_p]}{\hat{\mu}_p^2}k}$	10,34%	10,66%	10,61%
Äquivalenz		nein	ja	ja

Tabelle 10: Ergebnisse der probenübergreifenden Gleichwertigkeitsprüfung bezüglich Wiederfindung für die Extraktionsmethode Schütteln.

		Toluol	ACN	Ac/PE
Signifikanzniveau	α	5%	5%	5%
Maximal tolerierte theoretische Abweichung	Δ_{WFR}	±15%	±15%	±15%
Freiheitsgrade	df_3	12	12	12
Nichtzentralitätsparameter	δ_3	9,39	9,56	9,06
Kritischer Wert	k	6,71	6,84	6,45
Mittlere empirische Abweichung	$\frac{1}{P}\sum_{p=1}^{P}\frac{\hat{\mu}_{mp}-\hat{\mu}_p}{\hat{\mu}_p}$	-11,75%	1,39%	1,08%
Maximal tolerierte empirische Abweichung	$\frac{1}{P}\sqrt{\sum_{p=1}^{P}\frac{Var[\hat{\mu}_{mp}]+Var[\hat{\mu}_p]}{\hat{\mu}_p^2}k}$	10,71%	10,74%	10,68%
Äquivalenz		nein	ja	ja

Die hier vorgestellten Ergebnisse basieren auf einer probenübergreifenden Auswertung. Diese setzt voraus, das probenspezifische Effekte vernachlässigbar sind. Das dies nicht der Fall zu sein scheint, zeigen die im folgenden dargestellten Konfidenzintervalle für den probenspezifischen Methodeneffekt. Besonders hervorzuheben ist der Umstand, dass dieser probenspezifische Effekt gleichermaßen alle Methoden zu betreffen scheint. Es ist also davon auszugehen, dass die hier ermittelten Ergebnisse nur bedingt aussagefähig sind, da starke probenspezifische Methodeneffekte vorliegen. Im vorliegenden Falle ist allerdings davon auszugehen, dass die hier vorgestellten probenspezifischen Methodeneffekte möglicherweise Tageseffekte sind, da die zu einer Probe und einem Verfahren gehörigen Messungen nicht an unterschiedlichen Tagen wiederholt wurden. Somit erscheint eine probenspezifische Äquivalenzprüfung auch nicht sinnvoll, zumal hierfür die Anzahl der Freiheitsgrade zu gering ist.

Abbildung 4: 95%-Konfidenzintervall des probenspezifischen Methodenbias für die Probe *Boden 1/trocken*

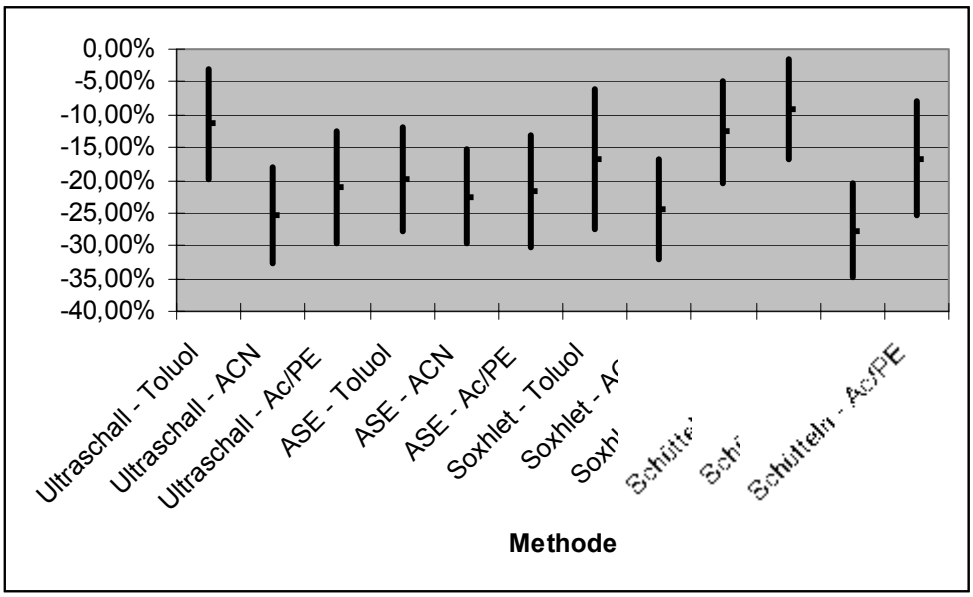


Abbildung 5: 95%-Konfidenzintervall des probenspezifischen Methodenbias für die Probe *Boden 1/feucht*

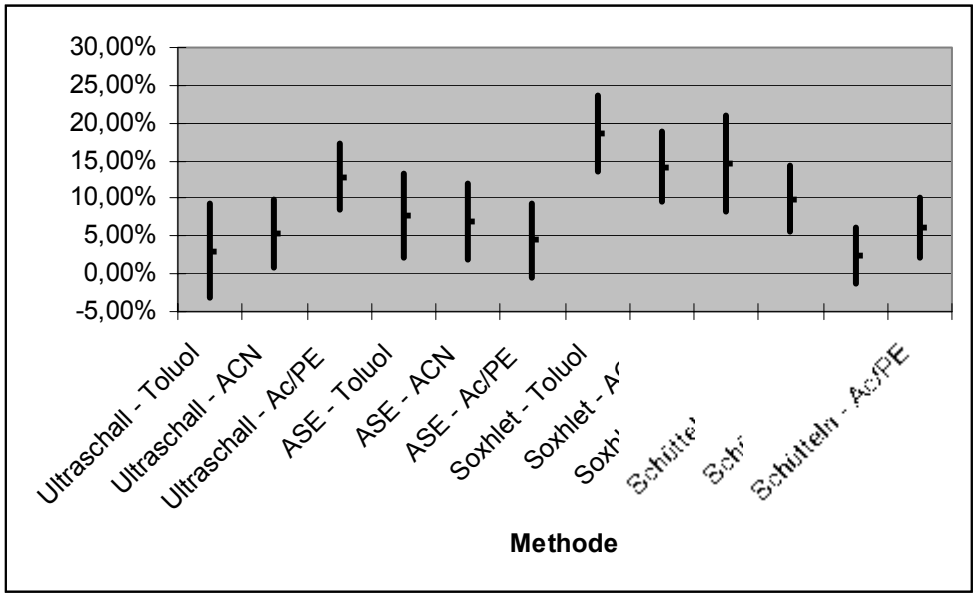


Abbildung 6: 95%-Konfidenzintervall des probenspezifischen Methodenbias für die Probe Boden 2

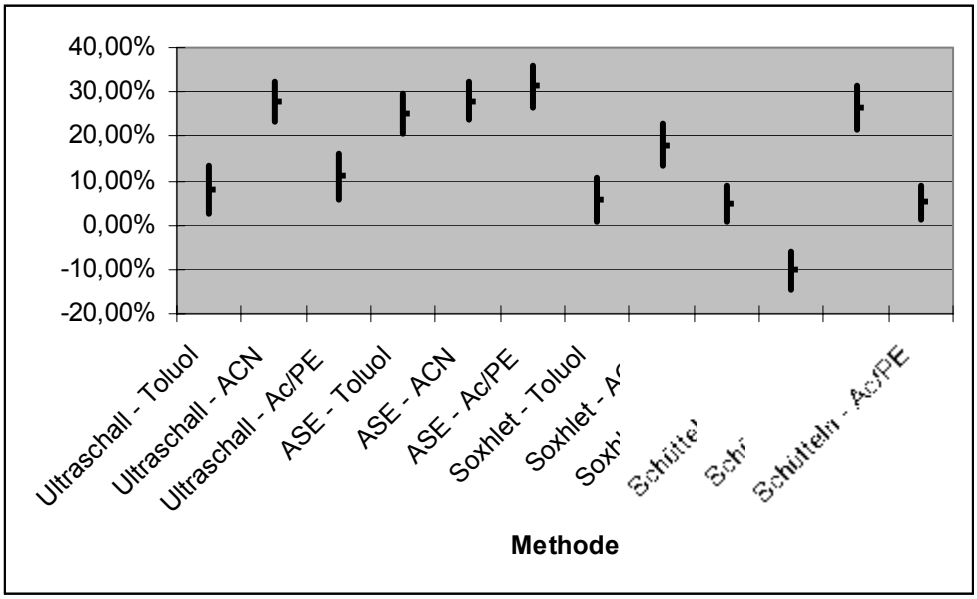
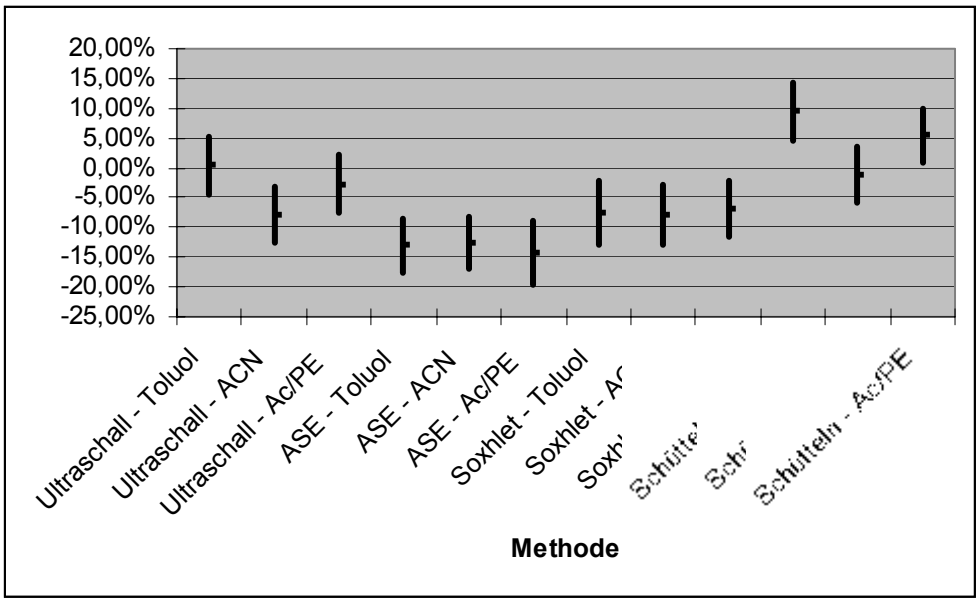


Abbildung 7: 95%-Konfidenzintervall des probenspezifischen Methodenbias für die Probe *Boden 3*



4.2 Nachweis der Äquivalenz bezüglich der Wiederhol-und In-House-Vergleichbarkeit

Die Methodologie entspricht weitgehend der Vorgehensweise für eine Äquivalenzprüfung auf der Grundlage von Ringversuchen, so dass auf eine detaillierte Beschreibung verzichtet werden kann.

4.3 Empfehlungen

Die Überlegungen zur der Festlegung maximal zulässigen Abweichungen Δ_R, Δ_r und Δ_{WFR} aus Abschnitt 2.3 gelten grundsätzlich auch im Rahmen von In-House-Validierungen. Grundsätzlich besteht das Ziel der Festlegung maximal zulässiger Abweichungen darin, dass sich die Messunsicherheit nicht oder nur unwesentlich erhöht.

Es ist dabei zu beachten, dass aufgrund der Möglichkeit, Ergebnisse zu „poolen“, und aufgrund sehr unterschiedlicher Versuchsanordnungen die Kriterien stark variieren können. Es wird daher empfohlen, Standard-„Designs“ für In-House-Studien zu erarbeiten und hierfür die entsprechenden Äquivalenzkriterien zu explizieren.

5. Weitere Anmerkungen

Wichtig zu bemerken ist, dass alle Aussagen nur für einen bestimmten Matrixtyp gemacht werden und nur für eine gewisse Bandbreite unterschiedlicher Konzentrationen Gültigkeit besitzen. Wie groß diese Bandbreite ist und wie eng der jeweilige Matrixtyp gefasst werden muss, ist aus chemisch-analytischer und bodenkundlicher Sicht zunächst und explizit festzulegen.

Zu beachten ist auch, dass niemals alle Verfahren zugleich auf Äquivalenz untersucht werden können, sondern dass jeweils ein paarweiser Vergleich erfolgt. Dies ist auch schon deshalb erforderlich, weil eine globale Aussage, die für alle Verfahren gleichermaßen gilt, nicht hilfreich ist. Wenn z.B. 9 von 10 Methoden sehr gut mit dem Referenzverfahren übereinstimmen und deshalb nur bei einem Verfahren größere Abweichungen festzustellen sind, werden letztere insgesamt kaum auffallen. Dies macht eine systematische Vorgehensweise erforderlich: Zunächst ist davon auszugehen, dass ein Verfahren Referenzcharakter hat, d.h. es gibt eine Methode, welche i.d.R. DIN-Methode ist bzw. welche umfassend validiert wurde. Dann besteht das Ziel der Untersuchung in dem Nachweis, dass alle weiteren Verfahren äquivalent zum Referenzverfahren sind, d.h. es werden der Reihe nach alle weiteren Verfahren auf Äquivalenz überprüft.

6. Zusammenfassung

Das Ziel einer Äquivalenzprüfung besteht in dem Nachweis, dass sich die zu prüfende Methode von der Referenzverfahren nur in einem geringem, zu tolerierenden Ausmaß unterscheidet. Diese Bemessung erfolgt anhand von Validierungskriterien wie Wiederfindungsrate, Vergleich- und Wiederholstandardabweichung. Die statistische Grundlage dieser Bemessung ist ein Modell der Messunsicherheit. Dadurch kann gewährleistet werden, dass sich die Messunsicherheit auch nach Einführung äquivalenter Methoden nicht oder nicht wesentlich erhöht.

Eine Äquivalenzprüfung erfolgt entweder mit Hilfe von Ringversuchen oder auf Basis einer In-House-Studie. Im erstgenannten Fall muss unterstellt werden, dass mittels der Referenzverfahren eine exakte Bestimmung der wahren Konzentration möglich ist, d.h. weitere Methodenfehler, die z.B. durch die Matrix verursacht sein können, bleiben unberücksichtigt. Im zweiten Falle, d.h. bei einer In-House-Studie, erfolgt der Nachweis der Äquivalenz anhand von zertifiziertem Referenzmaterial. Dadurch ist gewährleistet, dass Methodenfehler vermieden werden. Allerdings ist in diesem Falle das Ergebnis der Äquivalenzprüfung auch nur für das jeweilige Labor gültig, da nicht gewährleistet ist, dass andere Labore dieselbe „Labor“-Methode verwenden.

7. Verzeichnis der verwendeten Größen

b_p	Breite des Vertrauensintervalls für den (zertifizierten) Referenzgehalt bei Probe p
df	Anzahl der Freiheitsgrade
df_{rm}	Anzahl der Freiheitsgrade bei Verfahren m zur Ermittlung der Wiederholstandardabweichung
df_{Rm}	Anzahl der Freiheitsgrade bei Verfahren m zur Ermittlung der Vergleichstandardabweichung
e_w	Effizienz der robusten Wiederholstandardabweichung
G_1	Empirische Verteilungsfunktion der Interlabordifferenzen mit Stetigkeitskorrektur
G_2	Empirische Verteilungsfunktion der Intralabordifferenzen mit Stetigkeitskorrektur
H_1	Empirische Verteilungsfunktion der Interlabordifferenzen
H_2	Empirische Verteilungsfunktion der Intralabordifferenzen
I_{mp}	Anzahl Runs bei Probe p und Verfahren m
i	Run-Index
J_{mpi}	Anzahl Messwiederholungen bei Probe p und Verfahren m in Run i
J_{mp}	Anzahl der Labore bei Probe p und Verfahren m
$\bar{J}_{mp}$	Mittlere Anzahl der Messwiederholungen bei Probe p und Verfahren m über alle Runs
m	Messverfahrensindex: 1=Referenzverfahren; m=VergleichsVerfahren
M	Anzahl der Messverfahren
P	Probenindex
P	Anzahl der Proben
s_{rm}	Wiederholstandardabweichung bei Verfahren m
s_{Rm}	Vergleichstandardabweichung bei Verfahren m
T_1	Prüfgröße, die unter der Nullhypothese t-verteilt ist
$t_{df-1,1-\alpha/2}$	Quantil der t-Verteilung mit df-1 Freiheitsgraden
Var	Varianzoperator
Y_{mpij}	Messwert bei Verfahren m, Probe p, Run i und Wiederholung j
$z_{1-\alpha}$	Quantil der Standardnormalverteilung
Z_1	Prüfgröße, die unter der Nullhypothese standardnormalverteilt ist
α	Signifikanzniveau
α_{mi}	Runeffekt bei Verfahren m in Run i
δ_{mp}	Verfahrensbias für Verfahren m bei Probe p
Δ_R	Tolerierte Abweichung der Vergleichstandardabweichung

Δ_r	Tolerierte Abweichung der Wiederholstandardabweichung
Δ_{WFR}	Tolerierte Abweichung der Wiederfindungsrate
ε_{mpij}	Messabweichung bei Verfahren m für Probe p in Run i bei der Wiederholung j
$\hat{\mu}_{mp}$	Gesamtmittelwert der Messungen von Verfahren m bei Probe p (Ringversuch: robuster Mittelwert; In-House Untersuchung: arithmetischer Mittelwert)
$\hat{\mu}$	Robuster Mittelwert
$\hat{\mu}_p$	Referenzgehalt bei Probe p
μ_{mp}	Theoretischer Mittelwert für Verfahren m bei Probe p
μ_p	Wahrer Gehalt bei Probe p
σ_R	Vergleichstandardabweichung
σ_{rm}	Standardabweichung der relativen Messabweichungen bei Verfahren m unter Wiederholbedingungen
σ_{Rm}	Standardabweichung der relativen Messabweichungen bei Verfahren m unter In-House Vergleichbedingungen
$\sigma_{r,mp}$	Wiederholstandardabweichung bei Verfahren m für Probe p
$\sigma_{R,m}$	Vergleichstandardabweichung bei Verfahren m für Probe p
$\sigma_{run,m}$	Standardabweichung des relativen Runeffektes α_{mi}
Φ	Verteilungsfunktion der Standardnormalverteilung
ψ	Einflussfunktion des robusten Schätzers für den Mittelwert

8. Literatur

Graf, Henning, Stange, Wilrich (1987) Formeln und Tabellen der angewandten mathematischen Statistik. Springer-Verlag.

Jülicher, B., Gowik, P. und Uhlig, S. (1998) Assessment of detection methods in trace analysis by means of a statistically based in-house validation concept. The Analyst.

Jülicher, B., Gowik, P. und Uhlig, S. (1999) A top-down in-house validation based approach for the investigation of the measurement uncertainty using fraction factorial experiments. The Analyst.

Müller, Chr. und Uhlig, S. (2001) Estimation of variance components with high breakdown point and high efficiency. Biometrika.

Uhlig, S. und Lischer, P. (1999) Statistically based performance characteristics in laboratory performance studies. The Analyst.

9. Abbildungsverzeichnis

10. Tabellenverzeichnis