Fachausschuss Mineralöl- und Brennstoffnormung – FAM im
Normenausschuss Materialprüfung (NMP) des DIN

Leitfaden

Beschreibung Ringversuchsauswertung

FAM Fachausschuss Mineralöl- und Brennstoffnormung – Große Elbstraße 131 – D – 22767 Hamburg

Tel.: +49 40 639 004 61/62 ● Fax.: +49 40 639 004 50 ● Dr. Jürgen Fischer Email: juergen.fischer@fam-hamburg.de

In diesem Dokument wird im Interesse der Lesbarkeit grundsätzlich die männliche Form von Funktionsbezeichnungen verwendet; dies schließt die weibliche Form ein.

FAM LF 003

Inhalt

1. Ringversuchsauswertung

Die Auswertung von Ringversuchen mit dem Ziel, Präzisionsdaten zu erzeugen und die Leistungsfähigkeit von Prüfmethoden und Laboratorien zu überprüfen, erfolgt beim FAM nach den statistisch abgesicherten und international anerkannten Rechenverfahren der EN ISO 4259-Normenserie, die weltweit für Mineralölerzeugnisse angewandt wird. Die bisher einteilige Norm wurde in den letzten Jahren überarbeitet; derzeit besteht sie aus zwei Teilen, ein dritter Teil ist in Vorbereitung. Die Neufassung beschreibt in Teil 1 die Durchführung und Auswertung von Ringversuchen und in Teil 2 die Grenzwertsetzung auf Basis statistischer Daten und die Behandlung von Streitfällen wegen nicht übereinstimmender Laborergebnisse.

Mit Hilfe der statistischen Auswertung wird versucht, durch die Analyse einer möglichst großen Menge empirisch erfasster Daten einen Zusammenhang zwischen der Theorie und der Realität herzustellen. Das Ergebnis der statistischen Auswertung von Ringversuchen ist dabei gemeinhin eine Aussage zur Präzision eines Prüfverfahrens. Aus entsprechend angelegten Programmen wie dem FAM-Ringversuch lassen sich ebenso Informationen zur Eignung einer Prüfmethode in der Praxis und zur Qualität von Prüflaboratorien gewinnen. Hierzu ist allerdings eine ausreichende Anzahl an Teilnehmern je Prüfmethode notwendig; wenn weniger als fünf Labore teilnehmen, ist eine sinnvolle Aussage des Ringversuchs nicht mehr möglich.

Im Folgenden ist im Wesentlichen die Auswertung der FAM-Ringversuche mit zwei Proben und einer Vielzahl von Laboren beschrieben, die zu den sogenannten „Proficiency Test“-Programmen (Eignungstests) gehören. „Echte“ Ringversuche werden zur Ermittlung der Präzision eines Prüfverfahrens benötigt. Sie müssen den gesamten Anwendungsbereich abdecken und genügend Proben und Teilnehmer haben, um eine statistisch abgesicherte Aussage treffen zu können. Leider ist es nicht möglich, auf statistische Fachbegriffe zu verzichten; für Erklärungen und bei weitergehendem Interesse sei hier auf einschlägige Fachliteratur verwiesen.

2. Randbedingungen

Die meisten statistischen Methoden setzen voraus, dass die auszuwertenden Daten normalverteilt sind, d.h. einer Gauß'schen Normalverteilung entsprechen: rund um den am häufigsten vorkommenden Wert liegen weitere Werte, die aber mit größerem Abstand vom Mittelwert immer weniger häufig werden. Da diese Verteilung im Idealfall einer Glocke ähnelt, spricht man auch von der Gauß’schen Glockenkurve.

Die Prüfung auf Vorliegen einer Normalverteilung erfolgt im Rahmen der FAM-Auswertung mit dem Anderson Darling Test. In vielen Fällen stellt sich dabei heraus, dass Abweichungen von der Normalverteilung besonders am Rand der Verteilung auftreten und daher stark von den erkannten oder (insbesondere bei nur wenigen Teilnehmern) auch nicht erkannten Ausreißern beeinflusst werden. Einige weitere Einzelheiten sind in den nachfolgenden Abschnitten beschrieben.

3. Prüfung auf Ausreißer

Ein Ausreißer ist ein Ergebnis, das signifikant von den übrigen Werten abweicht. Ausreißer, die nicht eliminiert werden, können das Gesamtergebnis, insbesondere bei wenigen Teilnehmern oder Proben, stark beeinflussen, z.B. durch Verzerrung des Mittelwertes. Aus diesem Grund werden vor der eigentlichen Auswertung Prüfungen durchgeführt, die das Ziel haben, Ausreißer zu erkennen und zu eliminieren. Leider versagen diese Tests bei entsprechend heterogenen Ergebnissen mitunter; um möglichst viele Ausreißer zu erkennen, werden in der EN ISO 4259-1 mittlerweile vier verschiedene Prüfungen durchgeführt, die nachfolgend kurz erläutert werden.

a) GESD-Test
Der GESD (Generalized Extreme Studentized Deviate)-Test wurde neu in die EN ISO 4259-1 aufgenommen und ist in der Lage, Ausreißer zuverlässiger zu identifizieren. Der GESD-Test erfasst Ausreißer auf rein statistischer Basis und wird üblicherweise den anderen Prüfungen vorgeschaltet. Der GESD prüft dabei sowohl hinsichtlich der Differenz des Wertepaares eines Labors als auch hinsichtlich des Mittelwertes im Vergleich zum Probenmittelwert. Wird ein Ausreißer bezüglich eines Wertepaares identifiziert, wird nur derjenige Wert eliminiert, der am weitesten vom Probenmittelwert entfernt ist.

GESD-Ausreißer werden in der Auswertung wie folgt gekennzeichnet:
- g1, g2: Ausreißer bzgl. der Differenz eines Wertepaares; Wert 1 bzw. Wert 2 eliminiert;
- G: Ausreißer bzgl. der Differenz der Labormittelwertes vom Probenmittelwert.

b) Cochran-Test
Für die Datenmatrix der Ergebnisse (im Standardfall Doppelbestimmungen für je zwei Proben und N Laboratorien) werden entsprechend EN ISO 4259-1 zunächst die Ausreißer bezüglich der Wiederholbarkeit mit Hilfe des Cochran-Tests auf einem Signifikanzniveau von 1% ermittelt, d.h. die Resultate von Doppelbestimmungen, die sich bezüglich der Wiederholbarkeit deutlich von den anderen Ergebnispaaren unterscheiden, werden eliminiert.

Beispiel:

In einem Ringversuch werden von den Teilnehmern folgende Ergebnispaare (1. und 2. Messung) abgeliefert:

A: 3/5, B: 4/5, C: 2/4, D: 3/4, E: 1/7. Obwohl der Mittelwert zu den anderen Ergebnissen passt, wird das Ergebnis von Labor E als Cochran-Ausreißer eliminiert.

EN ISO 4259 empfiehlt im Grundsatz das Ausscheiden von nicht mehr als 10% Cochran-Ausreißern. In nicht besonders häufigen Fällen mit einem höheren Ausreißeranteil erfolgt eine gesonderte statistische Begutachtung der Ergebnisse und ggf. auch eine spezielle Untersuchung oder Umfrage bezüglich der Einhaltung der analytischen Parameter für die jeweilige Prüfmethode.

Cochran-Ausreißer werden in der Auswertung mit „C“ gekennzeichnet.

c) Hawkins-Test
Zur Ermittlung von Ausreißern bezüglich der Vergleichbarkeit wird der Hawkins-Test, ebenfalls auf dem 1% Signifikanzniveau, eingesetzt. Dieser Test identifiziert Ergebnisse, die so weit von den anderen Ergebnissen entfernt liegen, dass sie als nicht zur Gesamtheit der Ergebnisse zugehörig betrachtet und eliminiert werden.

Beispiel:
In einem Ringversuch werden von den Teilnehmern folgende Ergebnispaare (1. und 2. Messung) abgeliefert:

A: 10/12, B: 11/12, C: 21/23, D: 10/11, E: 9/11. Der Mittelwert von Labor C liegt so weit von den anderen Ergebnissen entfernt, dass das Ergebnis von Labor C als Hawkins-Ausreißer eliminiert wird.

Auch hier erfolgt bei mehr als 10 % Ausreißern eine spezielle Begutachtung.

Hawkins-Ausreißer werden häufig nicht erkannt, wenn der Datensatz stark streut oder wenn „entgegengesetzte“ Ausreißer vorhanden sind. Im obigen Beispiel würde ein Ergebnis von Labor F mit 2/3 den ersten Ausreißer möglicherweise kompensieren.

Hawkins-Ausreißer werden in der Auswertung mit „H“ gekennzeichnet.

d) Cook-Distanz
Auch die Cook-Distanz wurde neu in die EN ISO 4259-1 aufgenommen. Sie ist ein Maß für den Einfluss, den eine einzelne Probe auf das gesamte Modell nimmt. Für die Auswertung der Ringversuche heißt das, dass bestimmt wird, wie stark sich das Ergebnis der Auswertung durch Berücksichtigung einzelner Werte verändern würde; ein Ergebnis, dass eine signifikante Veränderung z.B. einer Regressionsgerade bewirken würde, ist mit hoher Wahrscheinlichkeit ein Ausreißer.

Zusätzlich zu den statistischen Ausreißern können andere zu eliminierende Werte vorkommen, wie z.B. nicht quantifizierbare Angaben der Form "≤ 0,3" (mit "x" markiert) oder in seltenen Fällen noch manuell eliminierte Ausreißer, gekennzeichnet mit "m". Für unverdächtige, "intakte" Ergebnisse wird je Probe ein "+" verzeichnet.

Die mathematische Beschreibung der o.g. Prüfungen würde den Rahmen dieses Dokumentes sprengen. Bei Interesse sei auf einschlägige Fachliteratur verwiesen.

4. Transformation

Die Transformation kann notwendig sein, wenn die Präzision des Prüfverfahrens vom Niveau des Prüfergebnisses abhängt, was durchaus häufiger vorkommt. Die Transformation ist dabei eine legitime mathematische bzw. statistische Operation und verändert den Datensatz und damit die Ergebnisse nicht; EN ISO 4259-1 gibt eine ganze Anzahl möglicher Transformationen vor, die je nach Struktur

der Daten angewendet werden können. Wenn die Transformation gelingt, kann mit den transformierten Daten wie üblich weitergerechnet werden.

Bei Ringversuchen, die mit lediglich zwei Proben je Prüfung durchgeführt werden, ist eine Transformation (mit nur zwei Stützpunkten) häufig weder zuverlässig noch sinnvoll. Die Auswertung erfolgt daher im Regelfall ohne eine Prüfung auf notwendige Transformation der Daten. Erst für den Fall eines umfangreichen Ringversuches zur vollgültigen Präzisionsermittlung für ein Prüfverfahren mit einer größeren Probenanzahl wird die Prüfung auf die Notwendigkeit einer Transformation vorgenommen.

5. Weitere Überprüfungen

Da die Ausreißertests mit den aktuell im Ringversuch ermittelten Daten durchgeführt werden, liegt vermutlich bei den als Ausreißer identifizierten Ergebnissen ein Fehler bei der Durchführung der Analyse oder ein technischer Fehler/Gerätefehler vor. Als weitere Möglichkeit sind Übertragungsfehler bei der Weitergabe der Ergebnisse denkbar. Für alle diese Fälle gilt, dass die Prüfung nicht notwendigerweise normkonform vorgenommen wurde.

Neben der rein statistischen Betrachtung sollen von den Beteiligten (Anwender, zuständige Ausschüsse, Normobleute) insbesondere bei großer und ggf. wiederholter Abweichung von den in der Norm publizierten Präzisionsangaben auch weitere Fehlerquellen überprüft und ggf. beseitigt werden. Dies sind z.B.:

- systematische Abweichungen von dem in der Norm beschriebenen Procedere in einzelnen Laboren;
- von den Proben stammende Einflüsse (Inhomogenität, Verwechslung etc.);
- mangelhaft beschriebene Details oder Unklarheiten im Prüfverfahren;
- Einflüsse durch fehlerhaftes Prüfgerät, Drift, Kalibrierung, unsaubere Lösemittel, usw.
- sonstige bei Durchführung der Prüfung gemachte Beobachtungen.

6. Grafische Darstellung

Von den vielen möglichen grafischen Darstellungsmöglichkeiten wird hier lediglich der für zwei Proben am häufigsten verwendete Diagrammtyp "XY-Diagramm" beschrieben. Darin werden je Labor die Labormittelwerte für die eine Probe auf der X-Achse, und für die andere Probe auf der y-Achse dargestellt.

Diese Darstellung entspricht im Wesentlichen dem so genannten Youden-Diagramm und erlaubt eine recht gute Beurteilung der systematischen und zufälligen Abweichungen von den jeweiligen Probenmittelwerten; zufällige Abweichungen ergeben das Bild eines „Schrotschusses“ mit um den Mittelwert streuenden Werten (Abb. 1).

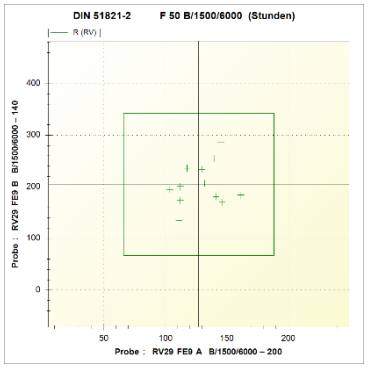

Abb. 1: X-/Y-Diagramm, zufällige Abweichung

Liegen die Werte für ein Prüflabor dagegen im oberen rechten oder unteren linken Quadranten, ist eine systematische Abweichung anzunehmen, die betroffenen Labore messen dann relativ gesehen systematisch hohe oder niedrige Werte.

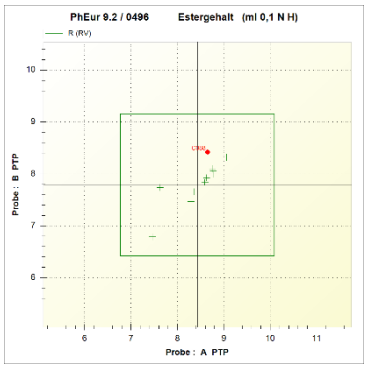

Abb. 2: X-/Y-Diagramm, systematische Abweichung

Die Grafik enthält zusätzlich ein besonders markiertes Quadrat mit den Abmessungen "$Mittelwert \pm \frac{1}{\sqrt{2}} \times Vergleichbarkeit$". Akkreditierte Prüflaboratorien können bei dieser Prüfung die erfolgreiche Ringversuchsteilnahme nachweisen, wenn Ihre Ergebnisse innerhalb dieses Quadrates liegen.

Die Interpretation der Daten anhand des XY-Diagramms ist allerdings nur für jeweils zwei Proben sinnvoll. Beim Vorliegen einer größeren Anzahl von Proben eignet sich diese Grafik nicht besonders, weil dann alle denkbaren paarweisen Probenkombinationen betrachtet werden müssten. Für die Begutachtung der Ergebnisse von mehreren Proben eignet sich daher die komprimierte Darstellung in einem multivariaten "Bi-Plot" besser.

7. Beurteilung der Ringversuchsvergleichbarkeit über das Varianzverhältnis (F-Wert)

Der Begutachtung der im Ringversuch ermittelten Vergleichbarkeit, R_{RV}, und deren Vergleich mit der Präzisionsangabe aus der Prüfvorschrift (R_{Norm}), soweit diese vorliegt) kommt eine zentrale Bedeutung zu.

Von besonderem Interesse sind diejenigen Prüfparameter, für die $R_{RV} > R_{Norm}$ ist. Es stellt sich die Frage, ob R_{RV} in einem statistisch signifikanten Ausmaß von R_{Norm} abweicht. Der Statistiker beantwortet eine solche Fragestellung üblicherweise mit Hilfe eines Hypothesentests. Im vorliegenden Fall werden in einem sogenannten F-Test die Varianzen der beiden zugrunde liegenden Standardnormalverteilungen miteinander verglichen.

Grundsätzlich liefert ein statistischer Hypothesentest keine mit 100% Sicherheit zutreffenden Aussagen, sondern lediglich Hinweise. Es besteht eine gewisse Wahrscheinlichkeit – die sogenannte Irrtumswahrscheinlichkeit -, dass die Schlussfolgerung falsch ist, die man aus dem Hypothesentest zieht. Diese Irrtumswahrscheinlichkeit wird für den F-Test gemäß DIN EN ISO 4259-3 mit 5% angenommen (Details siehe DIN EN 4259-3).

Im Folgenden werden die Rechenschritte beschrieben, die für das Verständnis der RV-Auswertung notwendig sind.

Der Prüfwert von F ergibt sich rechnerisch als Quotient der beiden zu vergleichenden Varianzen,

$$\boldsymbol{F = \frac{s^2_{R(RV)}}{s^2_{R(Norm)}} = \left(\frac{R_{RV} \times t(95\%,\ f_{Norm} = 30)}{R_{Norm} \times t(95\%,\ f_{RV})}\right)^2 = \left(\frac{R_{RV} \times 2{,}042}{R_{Norm} \times t(95\%,\ f_{RV})}\right)^2}$$

mit

F	F-Wert, Prüfwert für den Hypothesentest
$s_{R(RV)}$	Vergleichsstandardabweichung wie aus dem RV ermittelt
$s_{R(Norm)}$	Vergleichsstandardabweichung, berechnet aus der im Prüfverfahren angegebenen Vergleichbarkeit auf Basis von 30 Freiheitsgraden;
RRV	Vergleichbarkeit wie aus dem Ringversuch ermittelt;
RNorm	Vergleichbarkeit wie im Prüfverfahren angegeben;
t (95%, fRV)	Quantil der Student t-Verteilung auf einem Vertrauensniveau von 95% für f Freiheitsgrade. Dieser Wert kann mit einer gängigen Tabellenkalkulation berechnet oder einschlägiger Literatur entnommen werden.

Wenn F, berechnet nach obiger Formel < 1 ist, dann wird der Kehrwert von F als Prüfwert von F angegeben.

Die Interpretation dieses Prüfwerts und somit das Ergebnis des Hypothesentests ergeben sich aus einem Vergleich mit einem kritischen F-Wert F_{krit}, den man aus der DIN EN ISO 4259-3 entnehmen kann. Nur wenn $F > F_{krit}$ ist, besteht ein begründeter Verdacht, dass die Vergleichbarkeit des Ringversuchs signifikant von derjenigen der Norm abweicht. Die Abweichung betrifft sowohl $R_{RV} > R_{Norm}$ als auch $R_{RV} < R_{Norm}$.

Ist $F > F_{krit}$ und $R_{RV} > R_{Norm}$, dann kann dies am RV, an der Probe, an den Labors oder auch an einer zu optimistischen, nicht realistischen Präzision des Prüfverfahrens liegen.

Ist $F > F_{krit}$ und $R_{RV} < R_{Norm}$, dann kann dies an der Probe oder auch an einer nicht realistischen Präzision des Prüfverfahrens, bezogen auf die gängige Verfahrensweise, liegen.

Bei wiederholten signifikanten Abweichungen von der publizierten Präzision R_{Norm} sind die Ursachen zu ermitteln, und es sind Maßnahmen zur Überprüfung und Überarbeitung des Prüfverfahrens zu treffen, mit dem Ziel, die Situation zu verbessern.

8. Anderson-Darling-Test

Für die Anwendung von Ringversuchsergebnissen in der täglichen Praxis (z.B. Mittelwerte, "wahre Werte", Streuungen, Konfidenzintervalle usw.) müssen bei der Auswertung drei wesentliche Schlüsselbedingungen erfüllt sein:

- Die Ringversuchsproben sind homogen und im statistischen Sinne voneinander unabhängig;
- Die teilnehmenden Labore sind repräsentativ für die Gesamtheit aller Labore mit vergleichbaren Eigenschaften;
- Die Messergebnisse entsprechen in adäquater Weise einer Normalverteilung.

Für die letztgenannte Anforderung kann der Anderson-Darling-Test zur Prüfung ohne subjektive Beeinflussung eingesetzt werden. Dazu wird ein Prüfwert **$[AD^{2*}]$** berechnet und gegen eine Tabelle mit kritischen Werten verglichen. Die Berechnung des Prüfwertes wird hier nicht im Detail beschrieben; dazu wird auf die einschlägige Literatur verwiesen.

Die Ergebnisse werden als nicht normalverteilt betrachtet, wenn die kritischen Prüfwerte überschritten werden. Bei AD-werten < 0,75 ist kein Hinweis auf Verletzung der Normalität erkennbar. Bei Werten > 1,03 ist ein starker Hinweis auf Nicht-Normalität gegeben.

Der Anderson-Darling Test ermöglicht mit nur einem Kennwert eine schnelle und übersichtliche Prüfung auf Einhaltung der Normalverteilung und ist damit der Begutachtung von Histogrammen insbesondere bei kleineren Datenmengen (etwa <50 Ergebnisse) deutlich überlegen, kann allerdings erst oberhalb von Teilnehmerzahlen von etwa 10 - 12 Laboren als ausreichend aussagekräftig betrachtet werden. Bei zwangsweise stark gerundeten Ergebnissen mit nur wenigen "Treppenstufen" (Farbzahl, Cloud Point) kann der Anderson-Darling Test nicht eingesetzt werden.

9. Z–Scores – Beurteilung der Labore

Die Beurteilung der Qualität der Messergebnisse eines Labors in einem Ringversuch erfolgt besonders im gesetzlich geregelten Bereich sehr häufig mit Hilfe der sogenannten "Z – Scores".

Z-Scores sind, mathematisch ausgedrückt, die Absolutwerte der auf den Probenmittelwert bezogenen Abweichungen der Labormittelwerte, normiert auf die Standardabweichung. Mit dieser Normierung können die Laborergebnisse weitestgehend unabhängig von der Höhe des Ergebnisses begutachtet werden.

Die "Z – Scores" werden unter Auslassung der als Ausreißer (GESD, Cochran und Hawkins gemäß EN ISO 4259-1) erkannten Ergebnisse gemäß Gleichung (1) berechnet.

$$|Z| = abs[\frac{Labormittelwert - Probenmittelwert]}{Standardabweichung} \quad (1)$$

Die in Gleichung (1) angegebene Standardabweichung wird gemäß Gleichung (2) aus der im Ringversuch ermittelten Vergleichbarkeit (R_{RV}) berechnet.

$$Standardabweichung = \frac{Vergleichbarkeit}{1{,}96 \times \sqrt{2}} \quad (2)$$

Für Bewertungen anhand dieser "Z – Scores" ergeben sich aus der analytischen Praxis heraus die nachfolgenden Anhaltspunkte:

- $0 \leq |Z| < 1$ • gutes Ergebnis;
- $1 \leq |Z| < 2$ • zufriedenstellendes Ergebnis;
- $2 \leq |Z| < 3$ • fragwürdiges Ergebnis (nicht erfolgreich);
- $3 \leq |Z|$ • sehr fragwürdiges Ergebnis (meist ein identifizierter Ausreißer).

Bei |Z| -Scores > 2 wird dem Labor eine Überprüfung des Prüfverfahrens (z.B. der Kalibrierung) empfohlen, und bei |Z| - Scores > 3 sollte das Labor unbedingt eine umfangreiche Fehlersuche vornehmen. Allerdings gilt auch hier wie bei der Betrachtung der Präzision, dass nur bei Teilnahme einer ausreichenden Anzahl von Teilnehmern ein aussagekräftiges Ergebnis erzielt wird; bei weniger als fünf Teilnehmern ist die Auswertung der Z-Scores nicht mehr sinnvoll und kann ganz entfallen.