Limitation of standard deviation to express variability

TEILLET B.[1]; VERRON T.[1]; CAHOURS X.[1]; COLARD S.[2]; PURKIS S.[2]

[1] *SEITA Imperial Tobacco Group, 48 rue Danton, 45404 Fleury-les-Aubrais, France*

[2] *Imperial Tobacco Limited, PO Box 525, Winterstoke Road, Bristol BS99 1LQ, U.K.*

Introduction

In the framework of growing regulations regarding tobacco products, increased requirements are needed for reporting of analytical figures (e.g. FDA Draft guidance for Reporting Harmful and Potentially Harmful Constituents (HPHC) in Tobacco Products and Tobacco Smoke [1]). Manufacturers are required to report data on HPHC using unspecified testing protocols that provide reproducible results based on multiple measurements. FDA strongly recommends that the standard deviation of the mean is provided to report variability as well as the number of replicate measurements made (7 replicates for all HPHC and 20 for Nicotine and carbon monoxide).

The objective of this study was to evaluate the consistency of the short term standard deviation to describe the variability of measurements across laboratories (including different methodologies) as well as the "improvement" effect of the number of replicates.

Experimental

We have re-visited the Coresta 2006 joint experiment dataset [2] that included the 18 HPHC smoke constituents, as well as different sources of variability across measurements: short term variability (replicates), medium term variability (periods) and lab-to-lab variability (laboratories).

CORESTA Joint Experiment 2006 dataset:
The 20 participating laboratories each applied its usual collection and measurement methods to determine the smoke constituents yields, under the ISO smoking regime. The testing protocol required the analysis of the 2R4F and 1R5F Kentucky reference cigarettes performing 5 replicates (run over one or two consecutive days) in 3 independent experiments (run with a minimum of one week or longer in between each experiment).
Removal of extreme value:
For this study, we excluded extreme values on the basis of Cochran's test (for extreme variance) and Grubb's test (for extreme individual value or mean). The extreme value's removal was stopped at 22% of the total number of participating laboratories (this is IUPAC rules usually applied for collaborative studies [3]).

1 Evaluation of the equivalence of data generated across laboratories:
Analysis of variance (ANOVA) was performed for each reference cigarette with laboratory code as factor. When the P-value was significant, a multiple range test (Newman-Keuls at 95% confidence interval) was performed to classify laboratory means taking into account the standard deviation associated with the number of replicates of each laboratory. If the means are significantly different, groups will be generated. The same product analyzed across laboratories should be seen as just one group.
2 Evaluation of the effect of the number of replicates on variability:
The dataset was considered by changing the number of replicates. ANOVA (laboratory as factor) was performed with different numbers of replicates: 15 (5 replicates x 3 time periods) or 5 or 3 or 2 replicates (from the 1st time period).
3 Evaluation of the relative composition of total variance (replicate, time period and lab-to-lab):
A hierarchical ANOVA was performed with laboratory and time period as factors, the residual represented the replicate factor.

1 Equivalence of data generated across laboratories

Constituent	Number of significant different group	
	Ky-1R5F	Ky-2R4F
Ammonia	7	4
Acetaldehyde	6	8
Acrolein	7	7
Crotonaldehyde	5	5
Formaldehyde	8	8
Acrylonitrile	6	11
Benzene	5	4
1,3-Butadiene	5	5
Isoprene	6	5
Toluene	7	7
4-Aminobiphenyl	7	5
1-Aminonaphthalene	8	8
2-Aminonaphthalene	5	5
Benzo[a]pyrene	6	5
NNK	6	5
NNN	7	5

Table I: Multiple range test

ANOVA with laboratory code as factor provided **significant differences between the data generated across laboratories for the same product** (Ky-1R5F or Ky-2R4F) cigarettes for all smoke HPHC constituents.

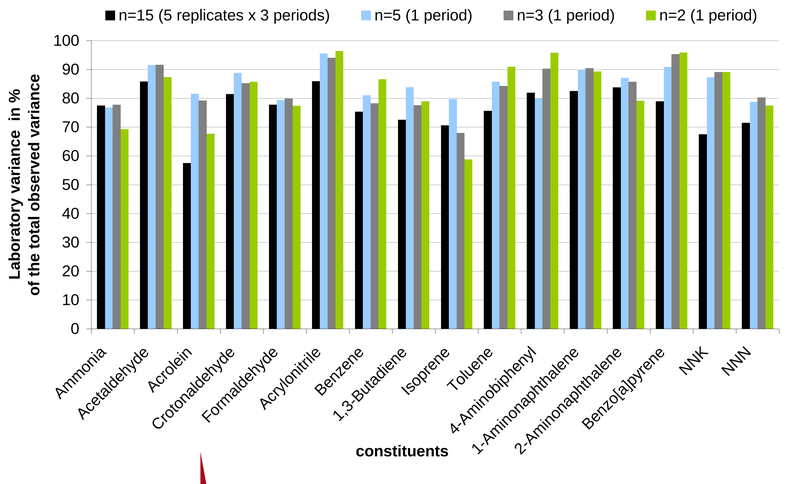
Table I presents the number of resulting significant groups from the multiple range test per HPHC smoke constituents observed for both reference cigarettes. **Ky-2R4F is seen as different groups (from 4 to 11) as well as Ky-1R5F (from 5 to 8).**

Each product can be falsely differentiated due to laboratory to laboratory variability for every analyte studied.

2 Effect of the number of replicates on variability across laboratory

Figure 1 presents the lab-to-lab variability in relative to the total variability per HPHC smoke constituents observed for the Ky-2R4F cigarette. Similar results were observed for the low tar Ky-1R5F product (not shown).

Figure 1: Ky-2R4F cigarette, lab-to-lab variance in relative to the total variance

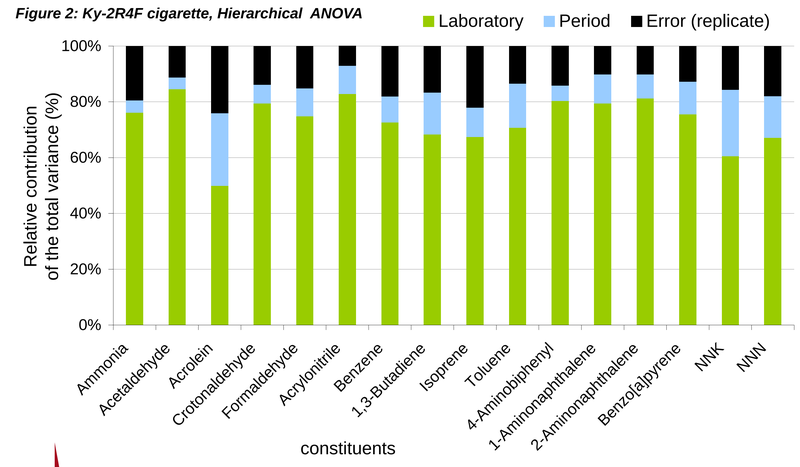


A higher number of replicates does not improve the lab-to-lab variability.

3 Composition of total variance (replicate, time period and lab-to-lab)

Figure 2 presents the relative composition of the total variance per HPHC smoke constituents observed for the Ky-2R4F cigarette. Similar results were observed for the low tar Ky-1R5F product (not shown).

Figure 2: Ky-2R4F cigarette, Hierarchical ANOVA



The short term standard deviation (replicate) represents a very low part of the total variance: from 5 to 25%.

Conclusion

On the basis of the data from the 2006 Coresta Joint Experiment where each laboratory applied their usual method, results show the **significant limitation of standard deviation to express variability of measurement across laboratories.**

Using the standard deviation to express variability **generated false differentiation for the same product** (~ 10 mg or 2 mg tar products) across laboratories. **The lab-to-lab variability was the major component of total variability whatever the number of replicates**.
Standard deviation is useful only to describe the precision of the measurement; it contributed only a low percentage from 5% to 25% of the total variability.

Expressing the full variability as the standard deviation ignores the lab-to-lab variability (bias in trueness across laboratories) which is critical when no reference materials with assigned values are available.

In the context of reporting data for a regulatory purpose, especially for product equivalence or comparison, **it is crucial to consider the Reproducibility** to avoid misleading information (for variability in its full definition or accuracy see ref [4]).

References
[1] Draft guidance for industry Reporting Harmful and Potentially Harmful Constituents in Tobacco Products and Tobacco Smoke Under Section 904(a)(3) of the Federal Food, Drug, and Cosmetic Act published in March 2012.
[2] Determination of "Hoffmann Analytes" in Cigarette Mainstream Smoke. The Coresta 2006 Joint Experiment, Michael Intorp & All, Beitr. Tabakforsch. Int. 24 (2009) 161–202
[3] International Union of pure and applied chemistry, Analytical, applied and clinical chemistry divisions, Protocol for the design, conduct and interpretation of collaborative studies, Pure Appl. Chem. 60 (1988) 855–964.
[4] ISO 5725-1 1994 Accuracy (trueness and precision) of measurement methods and results - Part 1 General principles and definitions