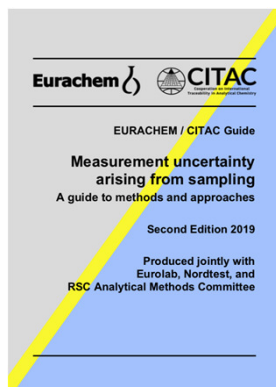
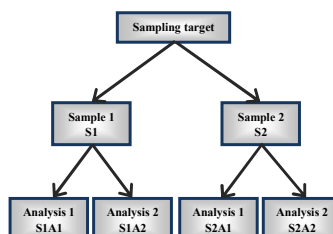


Comparing Uncertainty Values – are they really different?



*Eurachem/Eurolab Workshop,
Uncertainty from sampling and
analysis for accredited laboratories
November 2019, Berlin*



Peter Rostron

US University
of Sussex



Analytical Methods Trust



Overview

- Why compare uncertainties?
- 2 methods of comparison – F-test, Confidence limit comparison
- Example 1: Comparing uncertainties due to heterogeneity between 2 different PXRF beam sizes
- Data with outliers – robust ANOVA
 - Bootstrapping approach to estimating confidence limits of uncertainties for non-normal data using the duplicate method
 - Validation of the bootstrapping method (computerised simulations)
 - Example 2: Application of the bootstrapping method (UfS from SPT versus Duplicate methods)
- Conclusions / Further work

Why/How Compare Uncertainties

Sometimes useful to compare U estimates from different methods, e.g.

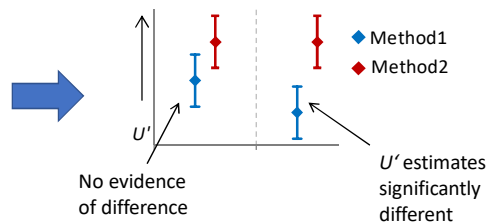
- Duplicate method / Sampling proficiency test
- ICP/AAS
- Beam diameters in Portable X-ray Fluorescence



Methodology:

1. F-Test on variance ratio – assumes normal distribution

2. Calculate confidence intervals (CIs) on U (or U') estimates

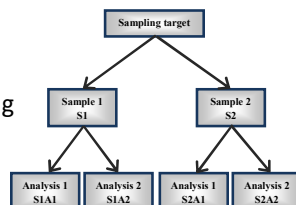


Example 1 - Heterogeneity estimation

Heterogeneity is interesting in its own right, e.g. because it:-

- Varies between analytes
- Contributes to U of certified values for CRMs
 - First published from Laser Ablation ICP-MS (Jochum *et al*, 2011)
- Is limiting factor for UFS

Heterogeneity (U_{het}) can be **quantified** using duplicate method



Often using *in situ* measurement devices



Example 1 uses PXRF to quantify heterogeneity at mm scale

Example 1 – Experimental Objective

SdAR sediment reference materials - Set of 3 RMs, Low[L2], Medium[M2], High[H1]

Intended for use in the calibration of field portable XRF instruments, as well as reference materials in laboratory analysis*

SdAR-H1 Metalliferous sediment
 SdAR-M2 Metal-rich sediment
 SdAR-L2 Blended sediment



Objective: Estimate the uncertainty of the reference value of these RMs for a range of elements when analyzed using small test portion masses

Specifically when these test portion mass may be below the minimum of 0.2 g
 - as specified by SdAR data sheets

SdAR-H1 – Metalliferous sediment

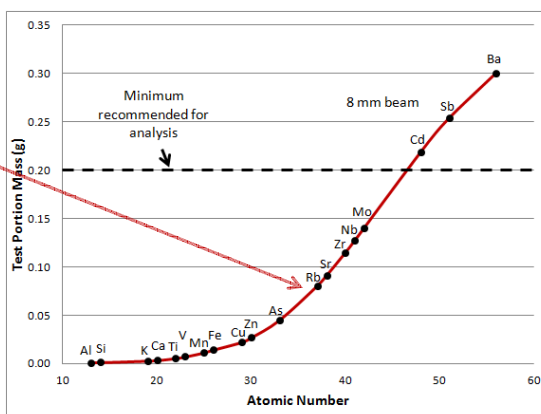
Reference values							
<i>Assigned value elemental/oxide mass concentration fractions and uncertainties from the GeoPT35a report on a dried (105 °C) basis</i>							
Oxide / element	Reference value g 100g ⁻¹	Uncertainty g 100g ⁻¹	n	Element	Reference value mg kg ⁻¹	Uncertainty mg kg ⁻¹	n
SiO ₂	65.45	0.18	71	La	44.9	1.0	60
TiO ₂	0.560	0.004	79	Li	50.5	2.5	37
Al ₂ O ₃	11.83	0.07	76	Lu	0.398	0.012	40
Fe ₂ O ₃ T	6.45	0.04	79	Mo	64	3	60
MnO	0.515	0.005	79	Nb	21.9	0.9	60
				Sr	22.7	1.0	60

Rostron P. and Ramsey, M.H. (2017) Geostandards and Geoanalytical Research, 41, 3, 359-473
 *IAGEO Limited <http://iageo.com/sdar-reference-materials/>

Example 1 - Rationale

PXRF - Modelling of test-portion mass suggests:

- For 8mm beam: Test-portion mass < minimum recommended (0.2g) for all but 3 elements



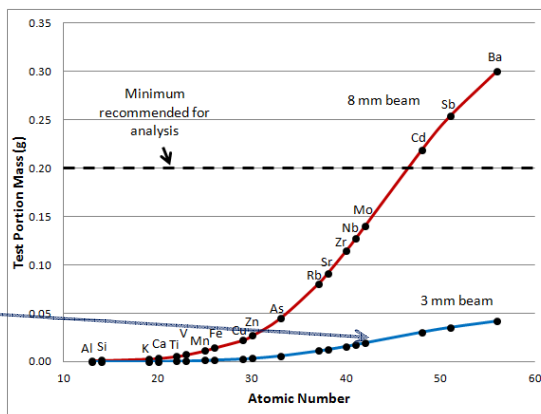
Rostron P. and Ramsey, M.H. (2017) Geostandards and Geoanalytical Research, 41, 3, 359-473

Example 1 - Rationale

PXRF - Modelling of test-portion mass suggests:

- For 8mm beam: Test-portion mass < minimum recommended (0.2g) for all but 3 elements



For 3 mm beam: Test-portion mass of all elements << 0.2 g



Rostron P. and Ramsey, M.H. (2017) Geostandards and Geoanalytical Research, 41, 3, 359-473

Heterogeneity estimation - Duplicate Method (mm scale)

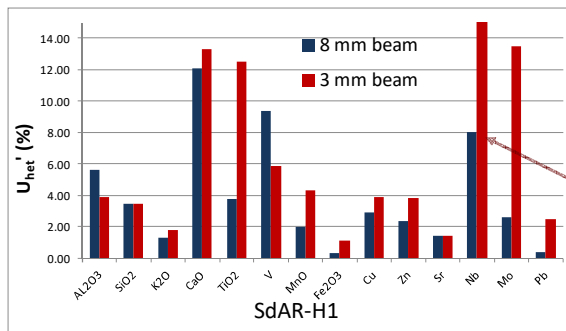
METHOD

- Estimate heterogeneity for two small beam sizes of PXRF (3mm and 8mm)
- Sample duplicates by placing PXRF on two sides of 6 pressed powder pellets of each RM → 
- Analytical duplicate as two readings without repositioning the PXRF → 
- Classical ANOVA applied to balanced design in usual way
- Between-sample duplicate variance used as estimate of (UfS), specifically U_{het}'
- U_{het}' can be added into U of certified value when mass of test portion is small (e.g. Beam measurements may be <10mg) and specified minimum is much larger (e.g. 200mg in this case)

Rostron P. and Ramsey, M.H. (2017) Quantifying heterogeneity of small test portion masses of geological reference materials by PXRF: implications for uncertainty of reference values. Geostandards and Geoanalytical Research, 41, 3, 359-473., DOI: 10.1111/ggr.12162.

Heterogeneity estimation – RESULTS (1)

Concentrations of 19 analytes measured with PXRF at 2 typical beam diameters



- U_{het} ranges from <1% to 39% for different analytes
- Generally higher for smaller 3mm beam size (10/14 elements)

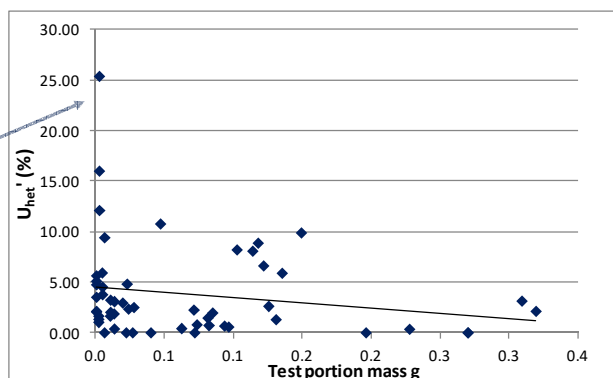
Rostron P. and Ramsey, M.H. (2017) Quantifying heterogeneity of small test portion masses of geological reference materials by PXRF: implications for uncertainty of reference values. *Geostandards and Geoanalytical Research*, 41, 3, 359-473., DOI: 10.1111/ggr.12162. onlinelibrary.wiley.com/doi/10.1111/ggr.12162/full



Heterogeneity estimation – RESULTS (2)

Concentrations of 19 analytes measured with PXRF at 2 typical beam diameters

U_{het} trend larger for smaller test portion mass (all 3 SdAR RMs shown)



Rostron P. and Ramsey, M.H. (2017) Quantifying heterogeneity of small test portion masses of geological reference materials by PXRF: implications for uncertainty of reference values. *Geostandards and Geoanalytical Research*, 41, 3, 359-473., DOI: 10.1111/ggr.12162. onlinelibrary.wiley.com/doi/10.1111/ggr.12162/full



Comparing U_{het}' values (SdAR-H1) using F-ratio

Define U_{het}' ratio: $U_{het}' \text{ Ratio} = \frac{U_{het}' 3mm}{U_{het}' 8mm}$

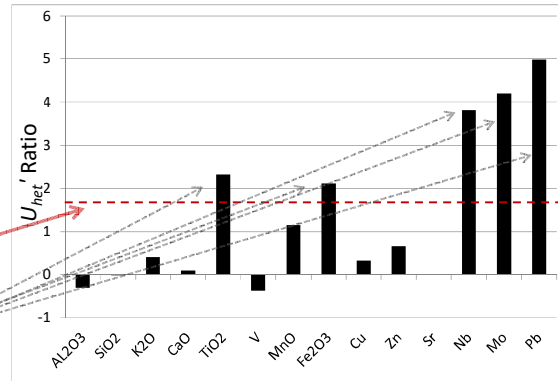
F-Ratio: $F = \frac{S_{3mm}^2}{S_{8mm}^2}$

Measurements made on 2 sides of 6 pellets,
D.F. = (2*6)-1

$F_{Critical 0.05(1),11,11} = 2.818$

$U_{het}' \text{ Critical Ratio} = \frac{S_{3mm}}{S_{8mm}} = \sqrt{F_{Critical}} = \sqrt{2.818} = 1.7$

5 elements with U_{het}' ratio exceeding this critical value considered to show significantly greater heterogeneity when measured using the 3 mm beam size

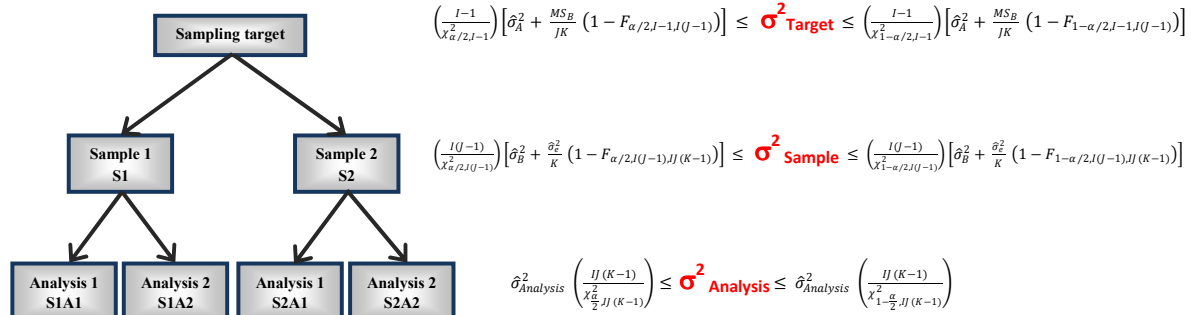


Rostron P. and Ramsey, M.H. (2017) Geostandards and Geoanalytical Research, 41, 3, 359-473

Comparing U_{het}' values (SdAR-H1) using Confidence Limits

Normally distributed data

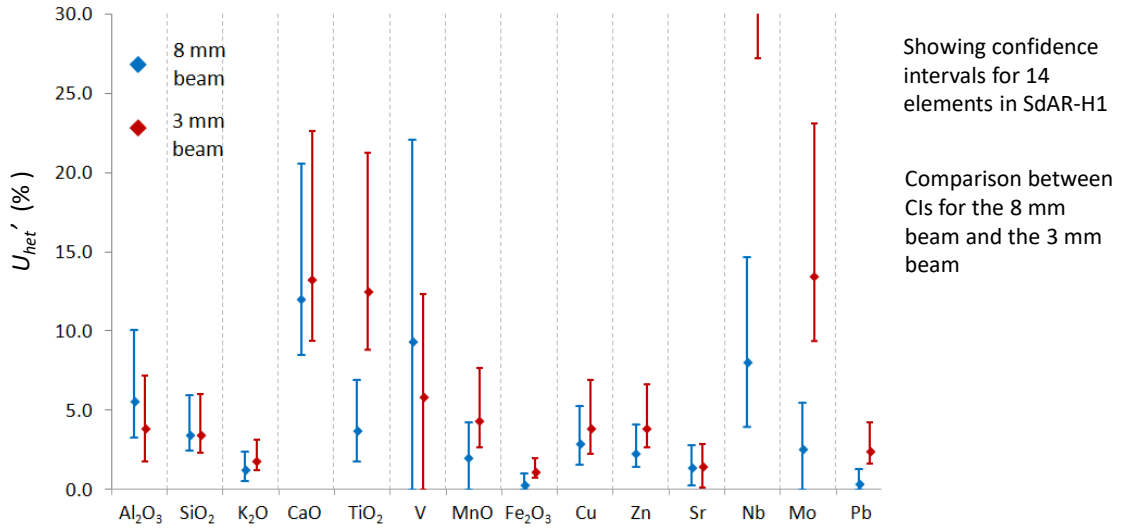
CIs Can be estimated using published probability model



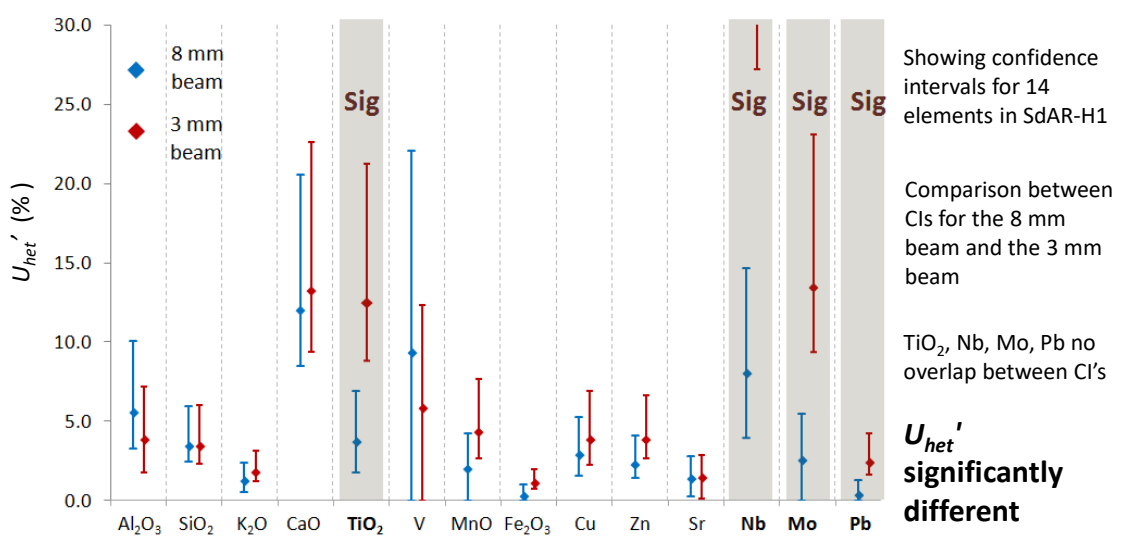
I = number of targets with variance σ^2_A , J = number of samples with variance σ^2_B , and K is the number of analyses with variance. MS_B = mean square of the middle (sampling) level from the ANOVA.

Williams JS (1962) A confidence interval for variance components. Biometrika 49:278-281; Graybill FA (1976) Theory and Application of the Linear Model. Duxbury Press

Comparing U_{het}' values (SdAR-H1) CI's (RANOVA2)

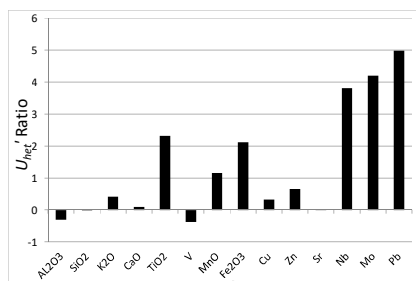


Comparing U_{het}' values (SdAR-H1) CI's (RANOVA2)

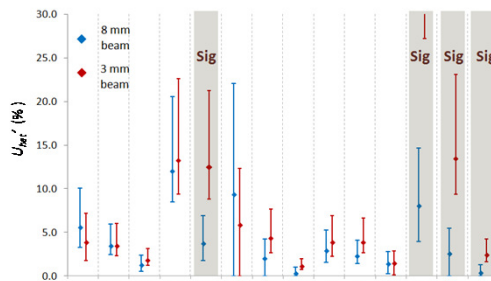


Method comparison

F-Ratio



Confidence Intervals



Both methods: Same conclusions for 4 elements

U_{het}' Fe₂O₃ not found to be significantly different using the confidence interval method
- comparing 95% CIs is often conservative method)

When outlying values present – Robust ANOVA

- F-test on variance ratio more powerful test for normally distributed data / classical ANOVA
- In practice: Often a small proportion (i.e. <10%) of outlying values exist in the frequency distributions of the analytical, within-sample and between-sample variability¹
- Robust ANOVA gives more reliable estimate of the variances of the **underlying** populations (See example in Appendix A1 of the Eurachem UFS guide¹)
 - F-test **not** reliable with outlying variances
 - Formulaic approach will **not** provide reliable CIs
- CI estimates can be made using **bootstrapping** method
 - This has been designed and validated for implementation in **RANOVA2**²

¹Ramsey, M.H., Ellison, S.L.R. (eds.) (2007). Eurachem/EUROLAB/CITAC/Nordtest/AMC Guide: Measurement uncertainty arising from sampling: a guide to methods and approaches Eurachem (2007).

²RANOVA2 - Excel program, free download from the AMC Software page on the Analytical Methods Committee section of the website of the Royal Society of Chemistry

Robust ANOVA - Bootstrapping

Computer-intensive method – re-sampling of observed data (with replacement)

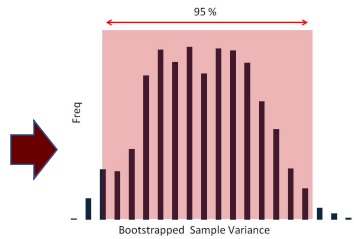
Large number of independent bootstrap samples generated (e.g. 2000)

	S1A1	S1A2	S2A1	S2A2
→	0.31	0.29	0.27	0.33
	0.42	0.44	0.42	0.42

Bootstrap sample – dataset of the same size and structure as observed data set
 - Random sampling (with replacement) from observed data

Statistic of interest (e.g. variance) calculated for each bootstrap sample

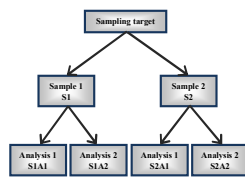
Confidence intervals derived from distribution of results



Allows estimates of CIs for variances produced by robust ANOVA

Validation of bootstrapping in RANOVA

Simulated 50,000 normally distributed balanced designs



Mean target value (μ_{Target})	75.8
Between-target standard deviation (σ_{Target})	73.2
Between-sample standard deviation (σ_{Sample})	27
Between-analyses standard deviation (σ_{Analysis})	20.4

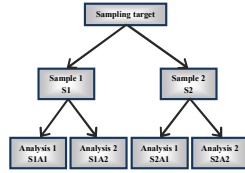
Confidence Intervals (CIs) calculated and averaged for 50,000 simulations

	True σ	Classical ANOVA			Robust ANOVA		
		σ	CI (Math)	%cov	σ	CI (Bootstrap)	%cov
Target	73.2	73.2	(62.9, 86.2)	95.9	73.3	(61.8, 88.7)	95.4
Sample	27.0	27.0	(22.3, 32.6)	95.9	27.1	(22.1, 33.6)	94.5
Analysis	20.4	20.4	(18.6, 22.6)	95.1	20.4	(18.3, 23.1)	94.8

Rostron PD, Fearn T, Ramsey MH (2019) Confidence intervals for robust estimates of measurement uncertainty (submitted for publication)

Validation of bootstrapping in RANOVA

Simulated 50,000 normally distributed balanced designs



Mean target value (μ_{Target})	75.8
Between-target standard deviation (σ_{Target})	73.2
Between-sample standard deviation (σ_{Sample})	27
Between-analyses standard deviation (σ_{Analysis})	20.4

Coverage percentages estimated by counting the number of times the CI contained the true value of the input parameter.

Classical ANOVA (using published formulae)

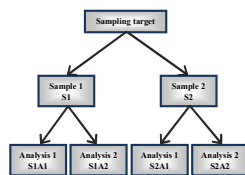
• CIs slightly conservative (> 95% coverage)

	True σ	Classical ANOVA		Robust ANOVA		
		$\hat{\sigma}$	CI (Math)	$\hat{\sigma}$	CI (Bootstrap)	%cov
Target	73.2	73.2	(62.9, 86.2)	73.3	(61.8, 88.7)	95.4
Sample	27.0	27.0	(22.3, 32.6)	27.1	(22.1, 33.6)	94.5
Analysis	20.4	20.4	(18.6, 22.6)	20.4	(18.3, 23.1)	94.8

Rostron PD, Fearn T, Ramsey MH (2019) Confidence intervals for robust estimates of measurement uncertainty (submitted for publication)

Validation of bootstrapping in RANOVA

Simulated 50,000 normally distributed balanced designs



Mean target value (μ_{Target})	75.8
Between-target standard deviation (σ_{Target})	73.2
Between-sample standard deviation (σ_{Sample})	27
Between-analyses standard deviation (σ_{Analysis})	20.4

Classical ANOVA (using published formulae)

• CIs slightly conservative (> 95% coverage)

	True σ	Classical ANOVA		Robust ANOVA		
		$\hat{\sigma}$	CI (Math)	$\hat{\sigma}$	CI (Bootstrap)	%cov
Target	73.2	73.2	(62.9, 86.2)	73.3	(61.8, 88.7)	95.4
Sample	27.0	27.0	(22.3, 32.6)	27.1	(22.1, 33.6)	94.5
Analysis	20.4	20.4	(18.6, 22.6)	20.4	(18.3, 23.1)	94.8

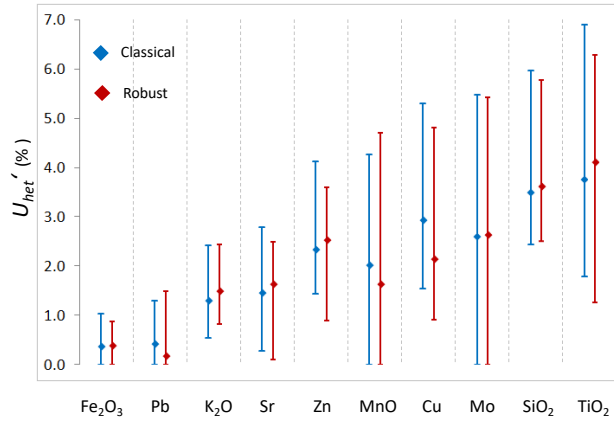
Robust method (using bootstrapping): Good approximations to method using published formulae

Rostron PD, Fearn T, Ramsey MH (2019) Confidence intervals for robust estimates of measurement uncertainty (submitted for publication)

Bootstrapped CIs - application to real (SdAR) data

Comparison of confidence limits U_{het} ' Classical & Robust ANOVA

- Shows 10 elements analysed by PXRF (8 mm beam) in candidate RM SdAR-H1
- **Classical ANOVA CIs** calculated using *formulae* (published method)
- **Robust ANOVA CIs** estimated by *bootstrapping*
- CIs from bootstrapping **very close approximates** to published methods – valid to use bootstrapping



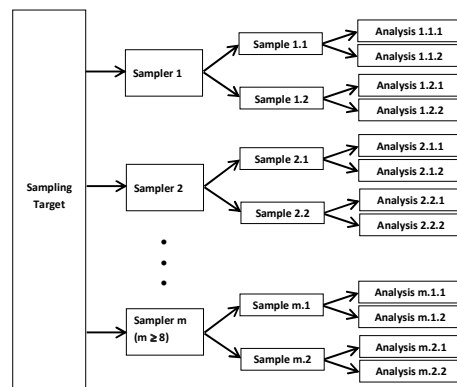
Example 2 – SPT for moisture in butter

Sampling Proficiency Test - Described in earlier lecture)

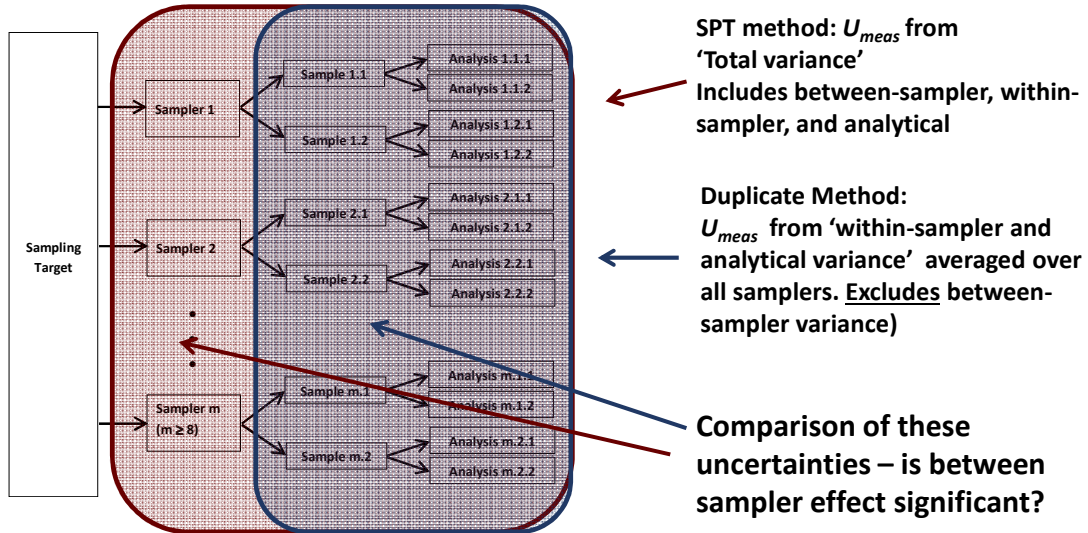


Multiple samplers each apply whatever sampling protocol they consider appropriate (to achieve stated objective)

Balanced design across all of the different samplers includes 'between sampler' bias



Example 2 – SPT for moisture in butter



Example 2 – SPT for moisture in butter

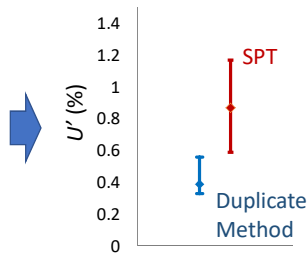
Robust ANOVA: U' on concentration of moisture in butter

Duplicate Method gives $U' = 0.39\%$
 SPT (n=9) gives $U' = 0.87\%$

- SPT U' larger x 2.2
- Includes bias between-samplers
- Is effect significant?

CIs calculated by bootstrapping

	U' (%)	CI
Sampling	0.17	(0.00, 0.44)
Analysis	0.35	(0.29, 0.49)
Meas	0.39	(0.33, 0.56)
Total SD	0.87	(0.57, 1.15)



No overlap = strong evidence of difference between uncertainty estimates

Significant effect of using different samplers – Sampler BIAS

Conclusions (1)

- Uncertainty estimates are not true values – they have confidence intervals (CIs)
- Sometimes useful to compare uncertainties between methods
 - e.g. SPT vs duplicate method
 - AAS vs ICP
- F-test can be used if normally distributed data

Conclusions (2)

- Alternative approach – calculate CIs and compare for overlap
 - *Normally distributed data*: CIs estimated using math model
 - *Data with outlying values*: Requires bootstrapping method
- Bootstrapping method of estimating CIs on uncertainties has been devised and validated for robust ANOVA (*submitted for publication*)
- Intention to make available for $n*2*2$ balanced design in program **RANOVA-CI**

Possible Future Work

- Comparison of CIs often conservative approach
- Hence F-Test recommended for normally distributed data
- Potential for further work on more reliable comparison of CIs produced by bootstrapping

Acknowledgements

With thanks to:

Professor Tom Fearn
Department of Statistical Science, University College London

Financial Support from The Analytical Methods Trust,
RSC Analytical Methods Committee



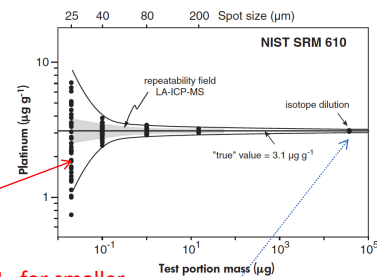
Excluded slides

Effect of Heterogeneity on U_{CV} of Beam RMs

- Uncertainty on certified value of RM (U_{CV}) is specified for a minimum mass
- Doesn't apply to lower mass of test portion used in beam measurements e.g.
 - in the *mg* range for PXRF, μg – *ng* for LA-ICP-MS, *pg* range for SIMS
- At these smaller scales nothing is truly homogeneous for all analytes, even glasses.
 - e.g. **Pt** by **LA-ICP-MS** in **NIST 610 Glass** (Jochum *et al.*, 2011)

Table 8.
Summary of compositional data for NIST SRM 610–611. Dc

Analyte	Ov. mean	Type of data	Uncertainty (U) at 95% CL			
			Test portion mass			
			mg range	1 μg	0.1 μg	0.02 μg
		LA Spot Size	80 μm	40 μm	25 μm	
Pt	3.12	IV	0.08	0.46	0.97	5.5



Larger U_{CV} for smaller test portion mass

Small U_{CV} for bulk (mg) test portion mass

MOU15

Updated uncertainty on reference values (U_{RV})

Greater U on the certified or reference value (as mass < specified minimum)

U'_{HET} can be added to U of certified value for small bean size

i.e. small test portion mass

$$U_{RV\ 8mm} = \sqrt{(U_{RV}^2 + U_{HET\ 8mm}^2)}$$



Assumes that within-bottle heterogeneity does not contribute significantly to the published value of U_{RV} at the minimum recommended mass of 0.2 g.

	RV		Pellets	
	g 100g ⁻¹	U _{RV}	U _{HET 8 mm}	U _{RV 8 mm}
AL2O3-L2	11.58	0.05	0.57	0.57
AL2O3-M2	12.47	0.06	0.64	0.65
AL2O3-H1	11.83	0.07	0.80	0.80
	mg kg ⁻¹		-	
As-M2	76	5	7.8	9.3
As-H1	396	24	0.0	24.0
Ba-L2	809	10	25.2	27.1
Ba-M2	990	12	20.0	23.4

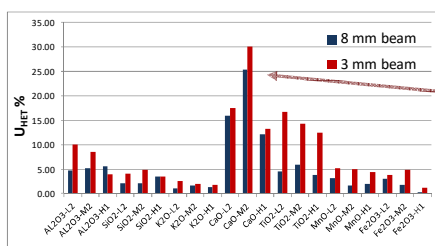
$U_{RV\ 8mm}$ ten times larger for light element like Al

- Dominated by $U_{HET\ 8mm}$

$U_{RV\ 8mm}$ only slightly larger for heavier element like As

- Dominated by U_{RV}

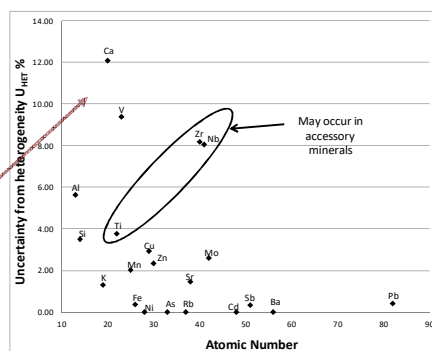
Heterogeneity estimation - RESULTS



U'_{HET} ranges from <1% to 30% for different analytes

- higher for smaller 3mm beam size

- Elements with lower atomic number (in PXRF) have:
 - smaller test portion mass (e.g. <10 mg), hence **Larger U'_{HET}**



Rostron P. and Ramsey, M.H. (2017) Quantifying heterogeneity of small test portion masses of geological reference materials by PXRF: implications for uncertainty of reference values. *Geostandards and Geoanalytical Research*, 41, 3, 359-473., DOI: 10.1111/ggr.12162. onlinelibrary.wiley.com/doi/10.1111/ggr.12162/full

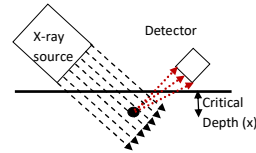
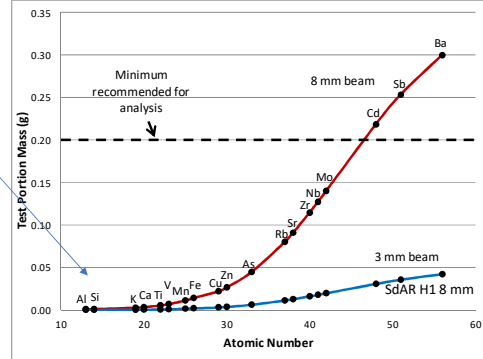
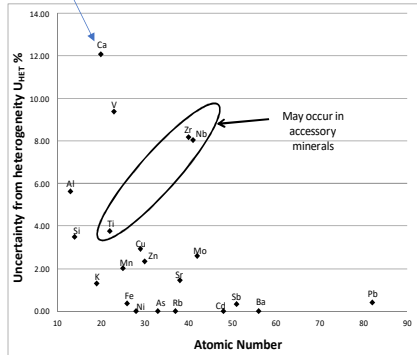
Slide 31

MOU15 Delete columns for powders (as extra complication)

Microsoft Office User, 03/10/2019

Heterogeneity estimation – RESULTS(2)

- Elements with lower atomic number (in PXRF) have:
 - smaller test portion mass (e.g. <10 mg), hence
 - Larger U'_{HET}



Rostron P. and Ramsey, M.H. (2017) Geostandards and Geoanalytical Research, 41, 3, 359-473

