

Statistical Aspects of Incremental Sampling

Kelly Black

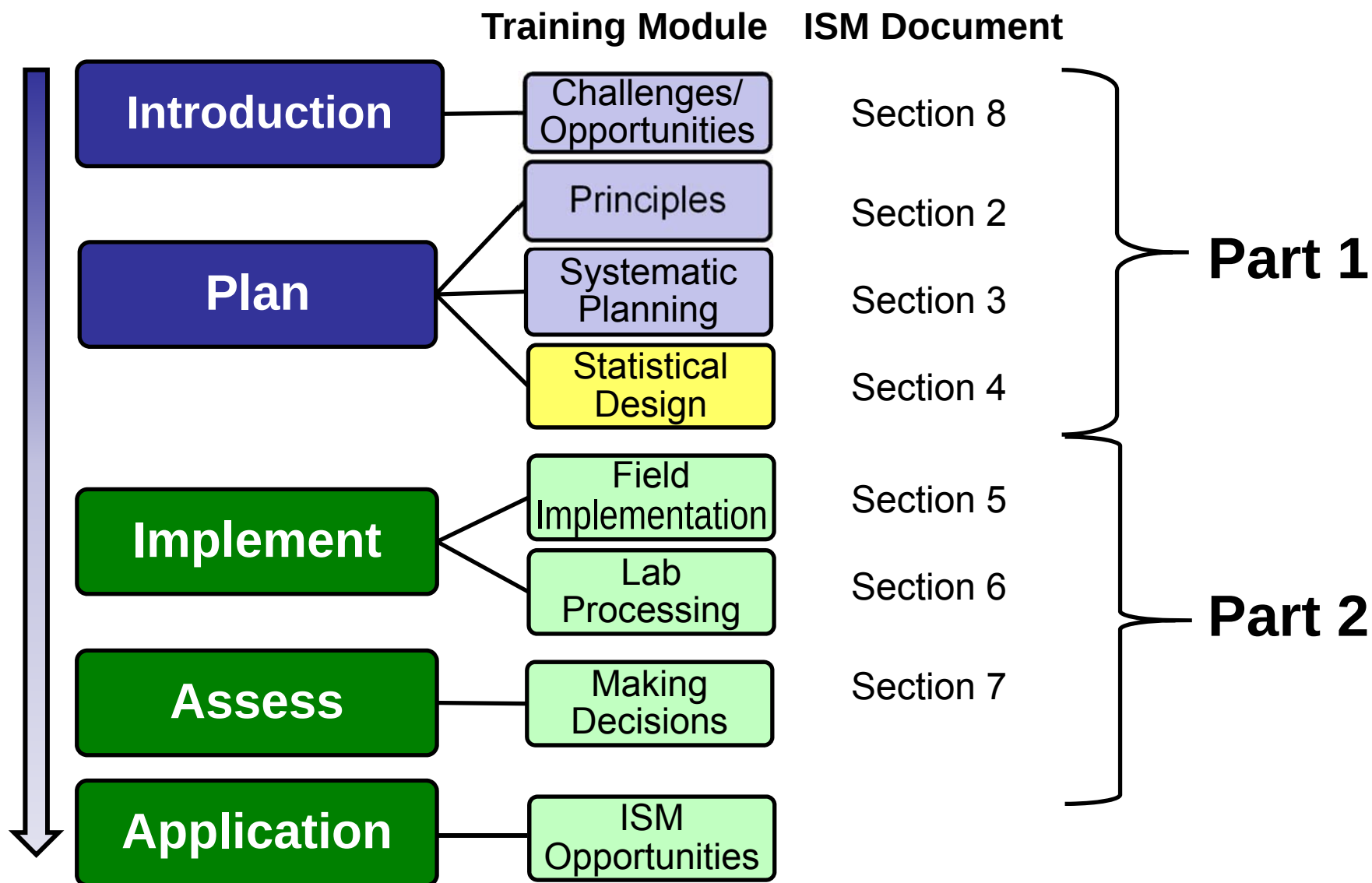
Neptune and Company, Inc.

kblack@neptuneinc.org

EMDQ – La Jolla, CA
March 26, 2012

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 26 MAR 2012		2. REPORT TYPE		3. DATES COVERED 00-00-2012 to 00-00-2012	
4. TITLE AND SUBTITLE Statistical Aspects of Incremental Sampling				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Neptune and Company, Inc,1435 Garrison St Ste 110,Denver,CO,80215				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES Presented at the 9th Annual DoD Environmental Monitoring and Data Quality (EDMQ) Workshop Held 26-29 March 2012 in La Jolla, CA. U.S. Government or Federal Rights License					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 28	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

ISM Document and Training Roadmap



Questions – Data Analysis

1. Does a single ISM sample provide a *reasonable* estimate of the mean?

Section 4.2.1

2. Can a 95UCL be calculated with ISM data?

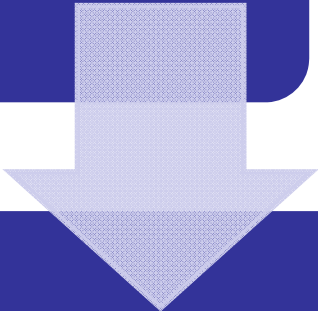
Section 4.2.2

95UCL = 95% Upper Confidence Limit of the mean

Questions – Sampling Design

3. What sampling design should I use?

Section 4.3.4.2



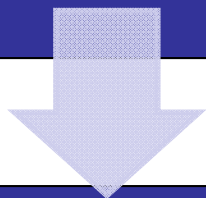
4. Is it reasonable to assume that concentrations are similar across DUs?

Section 4.4.2

Questions – Comparison of Data

5. Can background and site data be compared using ISM?

Sections 4.4.3.3 and 7.2.4



6. Can ISM data be compared to discrete or composite data?

95UCL = 95% Upper Confidence Limit of the mean

1. Does a single ISM provide a reasonable estimate of the mean?

Answer:

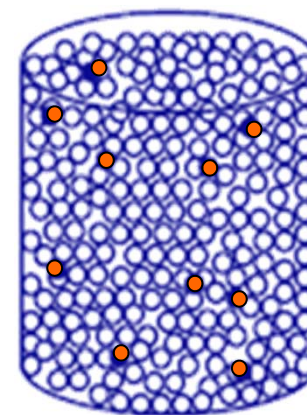
It depends how much error we are willing to accept.

► Why would someone collect just 1 ISM?

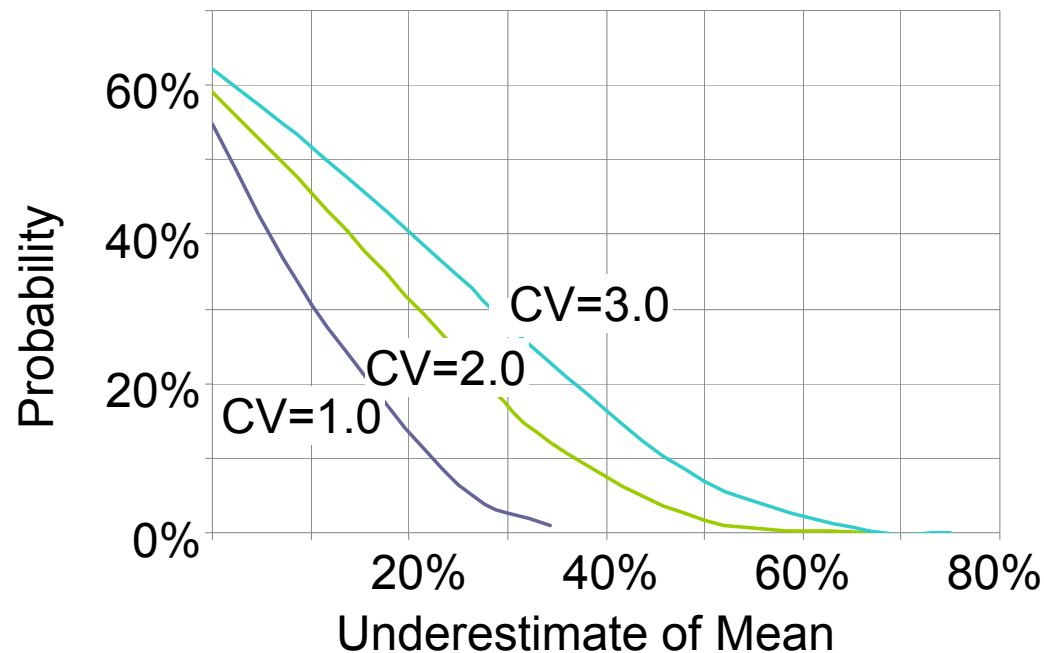
- UCL not required by regulator
- Save time and expense
- Assumption that more samples wouldn't change the decision.

For example

- Variance among individual increments is low
- Mean of DU is far above or below an action level



1(b). How “badly” might I underestimate the mean?



CV	Frequency	Magnitude	True Mean	Estimate
1	1 out of 4	15%	400 ppm	≤ 360 ppm
2	1 out of 4	25%	400 ppm	≤ 320 ppm
3	1 out of 4	30 - 60%	400 ppm	160 - 280 ppm

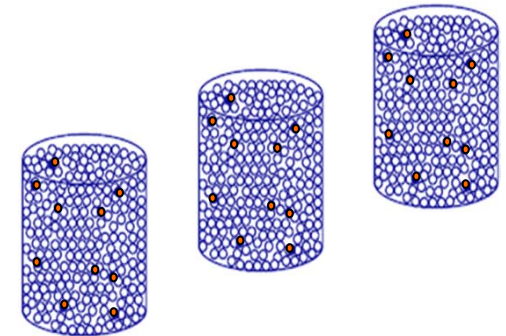
***Coefficient of variation (CV) = St Dev / mean**

2. Can a 95UCL be calculated?

Answer:

Yes, even with as few as 3 ISM samples.

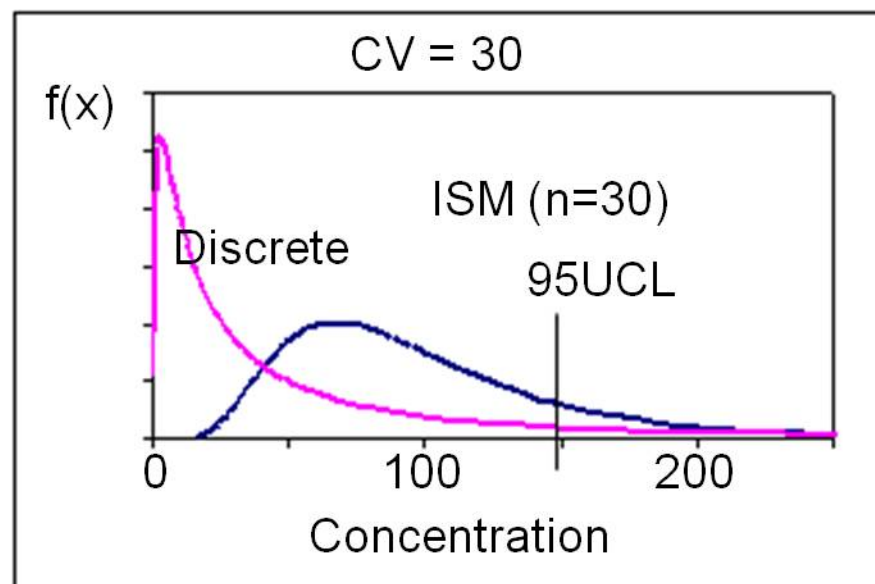
- ▶ Need at least 3 replicates ($r \geq 3$)
- ▶ Fewer methods are available than we are used to with discrete sampling:
 - Student's *s-t*
 - Chebyshev
- ▶ Each ISM result provides an estimate of the mean ("x-bar")
- ▶ Parameter estimates are calculated directly from ISM data



2(b). How do I choose a UCL method?

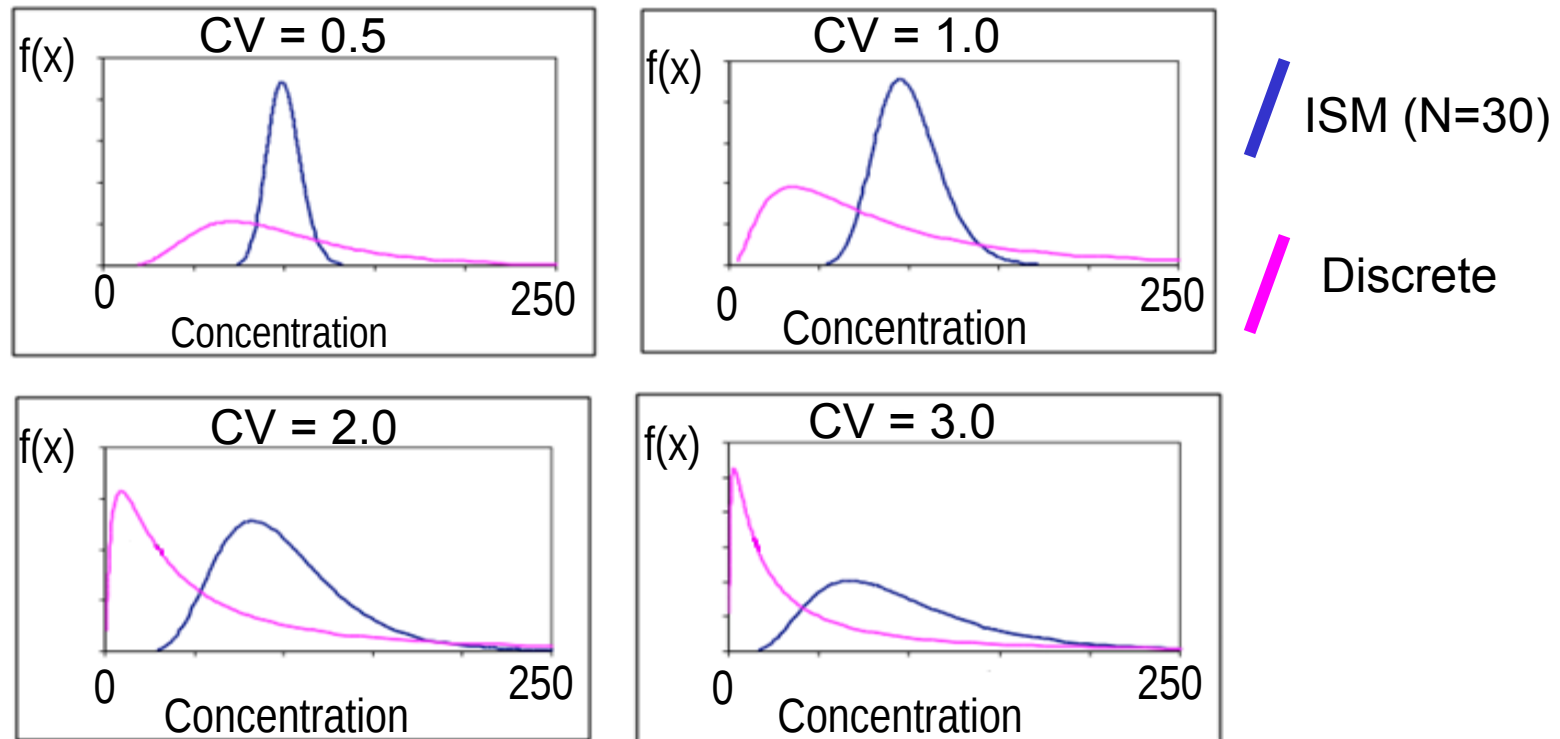
- ▶ Consider performance measures (informed by simulation study)

- Coverage (probability $\text{UCL} > \text{mean}$)
- Magnitude of difference between UCL and mean



- ▶ Recognize the key to performance is variability
 - Distribution of discretes $\neq$ Distribution of ISM results

Distribution of Means (ISM Replicates)



- ▶ ISM distribution variance is smaller
- ▶ ISM distribution shape becomes more non-normal with increasing CV of discrete distribution

Coverage Probabilities

UCL Method	Dispersion Among Individual Increments		
	Low (CV <1.5 or GSD <3)	Medium (1.5 < CV < 3 or 3 < GSD < 4.5)	High (CV >3 or GSD >4.5)
Student's-t	Yes	No	No
Chebyshev	Yes	Yes	Maybe

- ▶ Both methods provide desired 95% coverage when variability is low
- ▶ Chebyshev has more consistent 95% coverage for medium and high variability
- ▶ Increasing r (>3) and n (>30) provides marginal improvement

How much does Chebyshev overestimate?

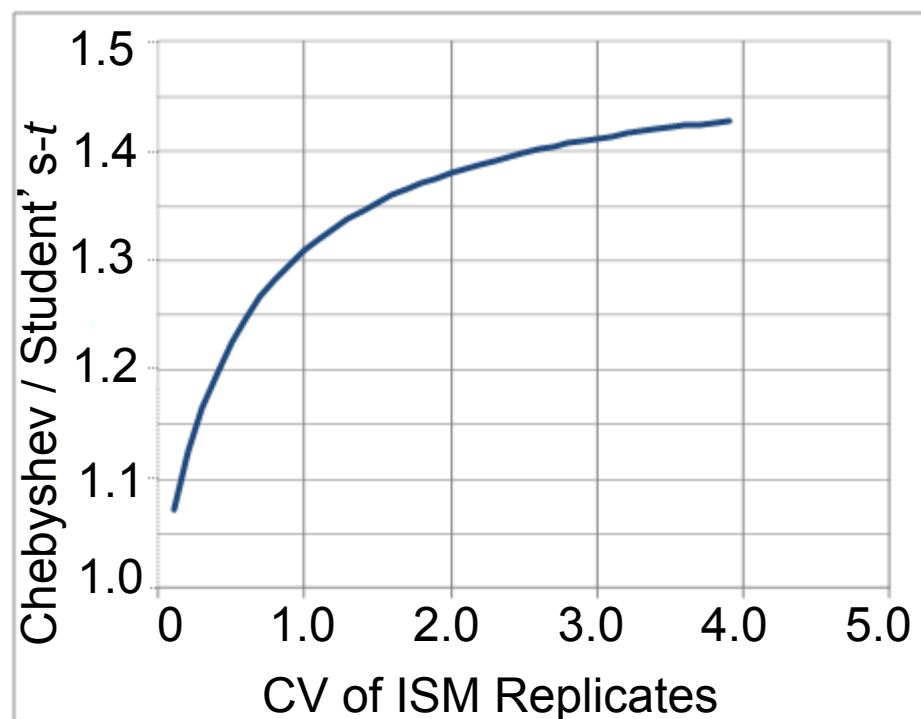
- ▶ Chebyshev will tend to yield 10-45% higher UCLs than Student's-t depending on the CV of the replicates
- ▶ Example: Student's-t = 100 ppm, Chebyshev = 110 -145 ppm

Student's-t

$$UCL = \bar{X} + t_{1-\alpha, r-1} \times \frac{s_{\bar{x}}}{\sqrt{r}}$$

Chebyshev

$$UCL = \bar{X} + \left(\sqrt{\frac{1}{\alpha}} - 1 \right) \frac{s_{\bar{x}}}{\sqrt{r}}$$



2(c). Can I use ProUCL to calculate the 95UCL?

Answer:

No. However, there are other tools available to calculate a 95UCL from ISM data.

- ▶ ProUCL is designed to work with discrete sample data.
- ▶ ISM replicates are fundamentally different from discrete samples, and data are typically available for only a few replicates.
- ▶ The ITRC guidance document has a link to a calculator that works for ISM data.

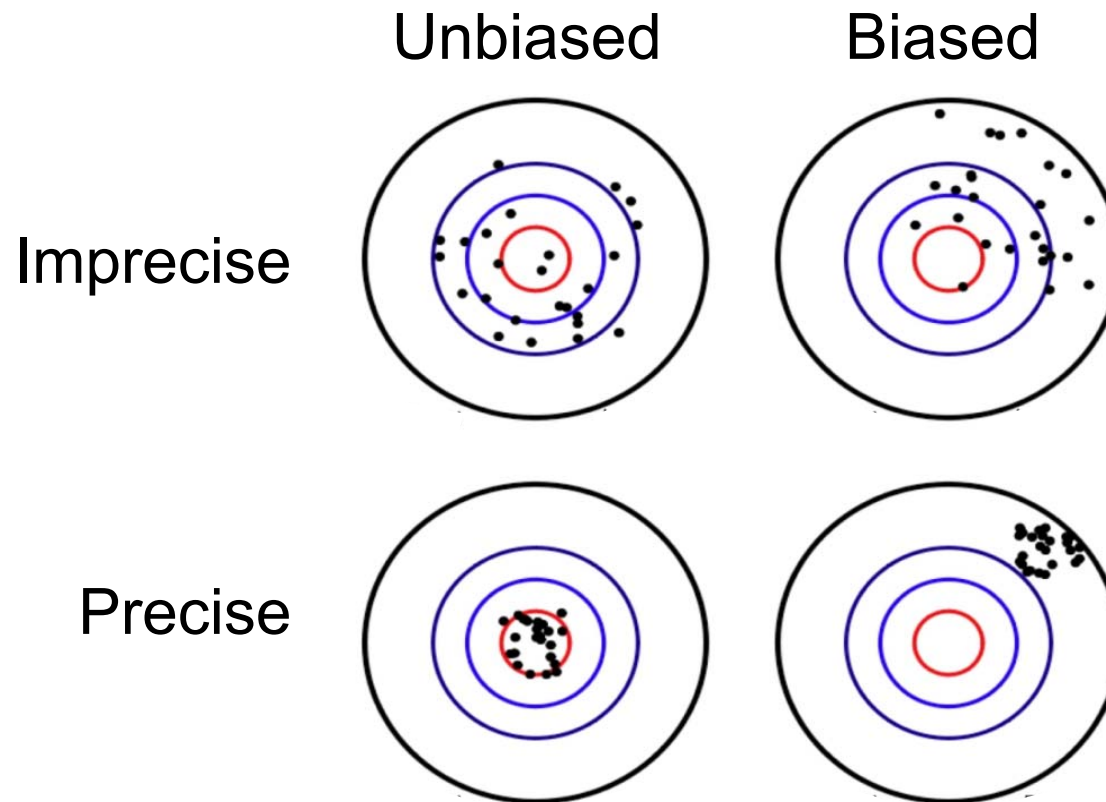
2(d). What does the variability in ISM reveal?

Answer:

High variability can indicate lab error. Low variability does not ensure that the results are sufficiently accurate to avoid decision error.

- ▶ RSD is the ratio of statistics calculated from ISM replicates
 - $RSD = SD / \text{mean}$
- ▶ If the goal is to make sure that the mean is not underestimated, a 95UCL should be calculated regardless of whether the RSD is high or low

2(d). What does the variability in ISM reveal?

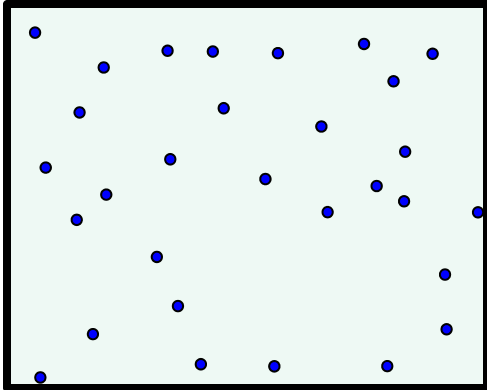


- Accuracy reflects both bias and precision (reproducibility)

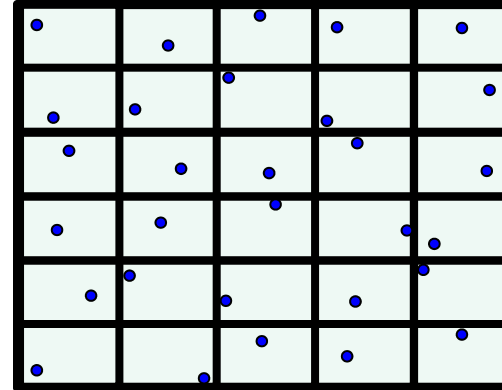
3. Is there a preferred ISM sampling design?



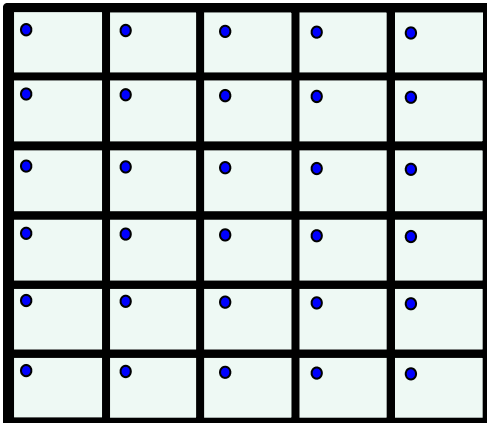
3. Is there a preferred ISM sampling design?



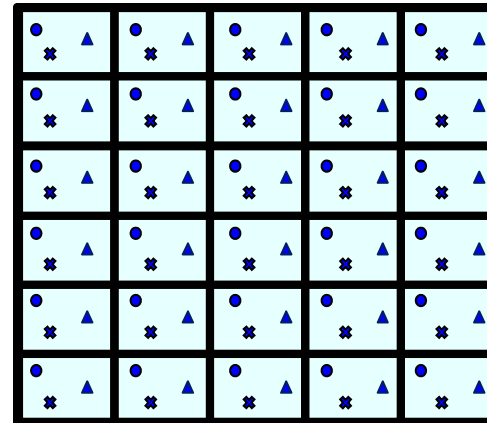
Simple Random



Random within Grid



Systematic



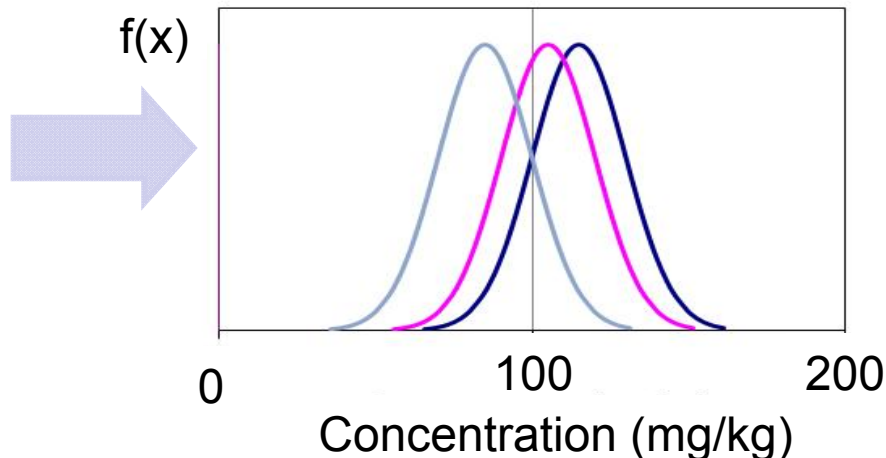
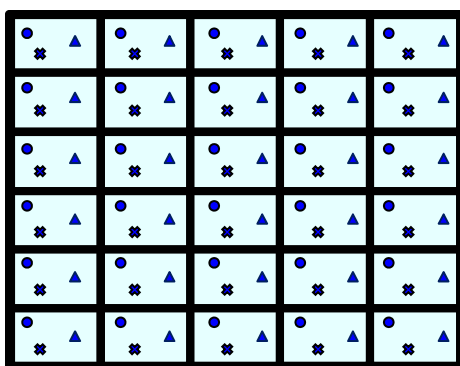
Systematic (3 replicates)

3. Is there a preferred ISM sampling design (continued)?

Answer:

Each random sampling design yields unbiased estimates of the mean and is an acceptable approach in most situations.

- ▶ Systematic random sampling is most often used because it is the easiest to implement



3(b). How many increments?

Answer:

$n = 30$

- ▶ As the number of increments increases:
 - spatial coverage improves
 - variability between ISM results decreases
 - 95UCL will tend to be closer to the mean
- ▶ Size of DU can be a consideration – large DUs may require more increments

10 20 **30** 40 50 60 70 80 90 100

3(c). How many replicates?

Answer:

$r = 3$

- ▶ For most DUs, 3 is sufficient.
- ▶ Minimum number to calculate standard deviation (and 95UCL) of ISM results
- ▶ More replicates will produce a 95UCL closer to the actual mean, but may not be cost-effective unless the result is near the action level



4. Can I extrapolate results across DUs?

- ▶ Unsampled DU – extrapolate estimate of mean
- ▶ DU with 1 ISM – extrapolate estimate of variability
 - Standard deviation (SD)
 - Coefficient of variation (CV)



DU-1 [?] DU-2



4(b). Extrapolation of the Mean

Answer:

Assumes the mean concentration in the unsampled DU(s) is the same as the sampled DU.

DU-1 $\stackrel{?}{=}$ DU-2



► DU-1:

- Mean = 100
- SD = ?

► DU-2:

- Assumed Mean = 100



4(c). Extrapolation of the Variance

Answer:

Assumes that the variability is similar in all DUs.

DU-1 [?] = DU-2



► DU-1:

- Mean = 100
- SD = 50
- $CV = 50/100 = 0.5$



► DU-2:

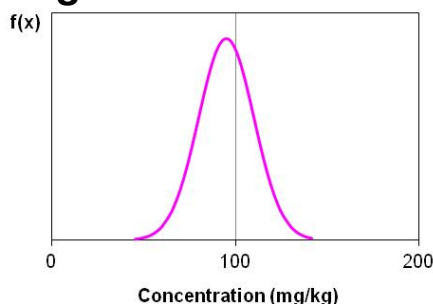
- Mean = 400
- Assumed $CV = 0.5 = SD / 400$, therefore
SD = 200

24 5. Can background and site ISM data be compared?

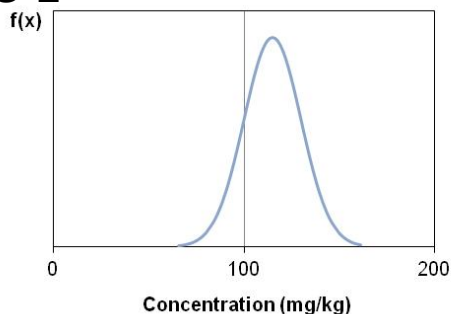
Answer:

Yes, but statistical tools for comparison are limited.

Background

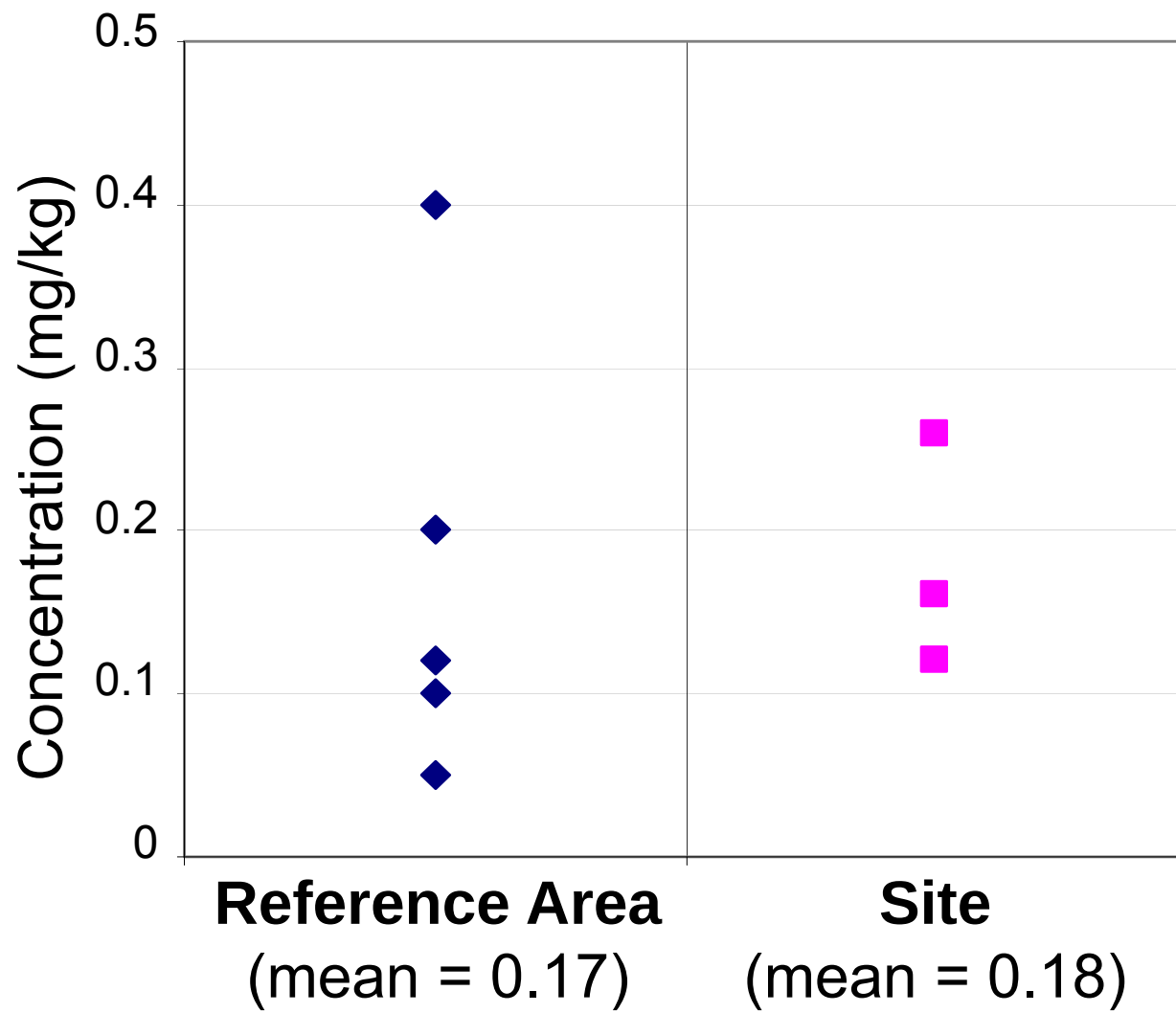


DU-1



- ▶ Is the site mean equal to background?
- ▶ Hypothesis testing is limited
- ▶ Cannot test upper tails with ISM data

5. Example Background Comparison



26 6. Can ISM and discrete or composite data be compared?

Answer:

Yes, if the sample support is known.

- ▶ Measurement variability is a function of sample support
- ▶ Comparisons are possible with caution and consideration of all differences in sampling and analysis methods and meaning of data
 - Sample collection
 - Sample prep
 - Analysis method
 - ISM estimates mean

Summary: Statistical Design

- ▶ Mean or 95UCL from ISM data may be used to make decisions
- ▶ 3 replicate samples provide adequate information for a 95UCL
- ▶ Systematic random sampling is most commonly used
- ▶ About 30 increments per ISM sample is usually sufficient
- ▶ Extrapolation of the mean or variance can be very uncertain
- ▶ Comparisons between ISM site and background data are possible, with caution
- ▶ Comparisons between ISM and other data are possible, with caution

ISM Part 1 Summary and Part 2 Preview

