Why 30 ?

ASTM D02, Dec./ Orlando

Statistics Seminar

Presented by: Alex Lau, Chairman, D02.94

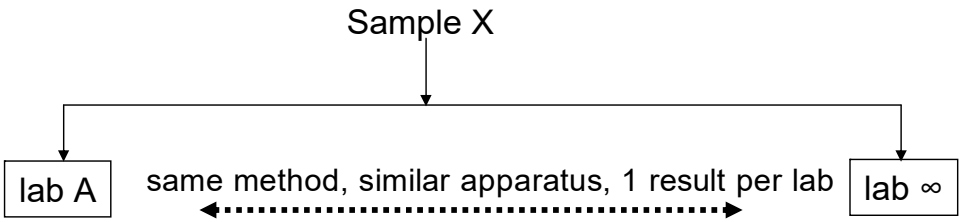
What this seminar is about

This brief seminar will provide :

- a very simple explanation on the use of statistics to estimate *population parameters*
- specific focus on the sample standard deviation statistic (*s*), where "*degrees of freedom*" will be explained
- why the minimum degree of freedom (df) of 30 is specified for r and R statistics in ASTM D6300

What is a 'population' ?

Within the context of this seminar and ASTM Standard Test Method → it is the 'universe' of results obtained for a specific material using the same STM from an infinite number of labs

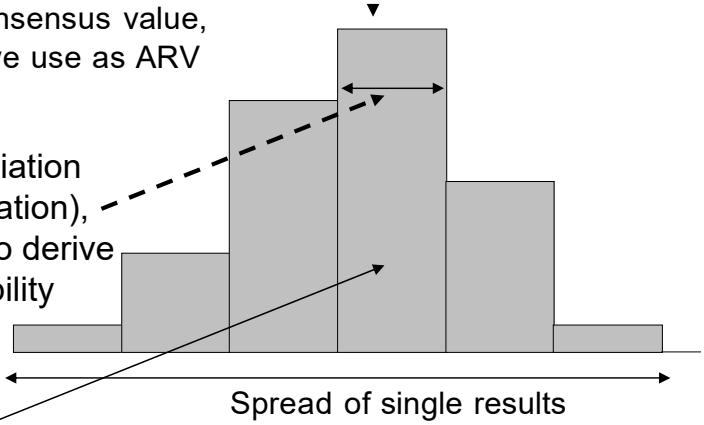


What are 'population parameters ?

Two that we care about the most are: μ, σ

μ → mean (of the population), *aka* Consensus value, which we use as ARV

σ → standard deviation (of the population), which we use to derive Reproducibility



Histogram of results from the 'hypothesized' universe of results

(in Statistics jargon we call this the "target population" of interest)

If we had infinite resources, patience, and time ….

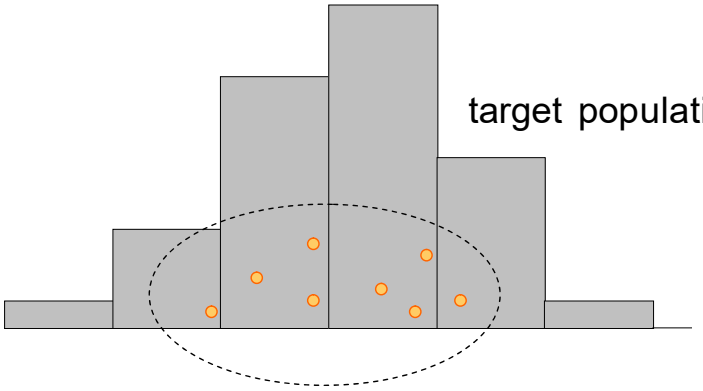
We would collect each and every result from the 'population', then, we can calculate the *exact* values for μ, and σ

…but, we don't (have infinite resources …..)

Statistics come to the rescue

- We can take a *random sample of adequate size* from the 'target population', do some math, and come up with 'statistics' to 'estimate' the desired population parameters:
  - sample average $\bar{x}$ is an estimator of μ
  - sample standard deviation (s) is an estimator of σ

How to use statistics to estimate parameter values



do the math

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \hat{\mu}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \hat{\sigma}$$

So, what constitutes

adequate sample size ?

No free ride …

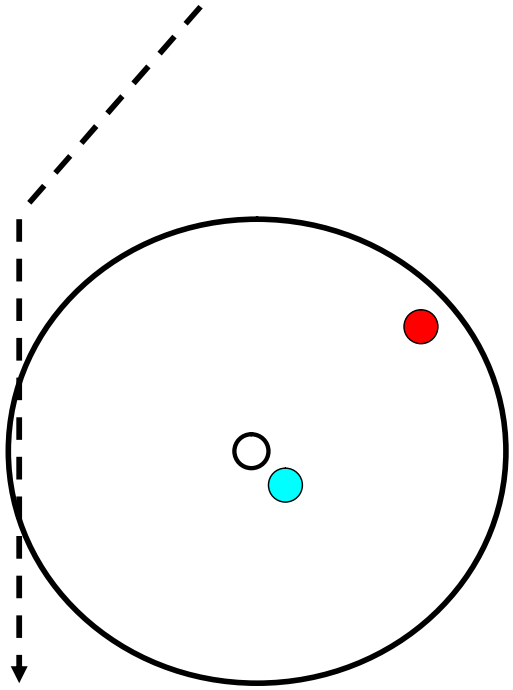
- because only a limited amount of data is used, these estimates have variability themselves due to random sampling →
  - which means when you do it again, you will most likely get a numerically different answer
- this is known as variability of the sample statistics
- furthermore, the variability is a function of the sample size

Variability of the sample std dev (*s*)

The most common question posed to statisticians:

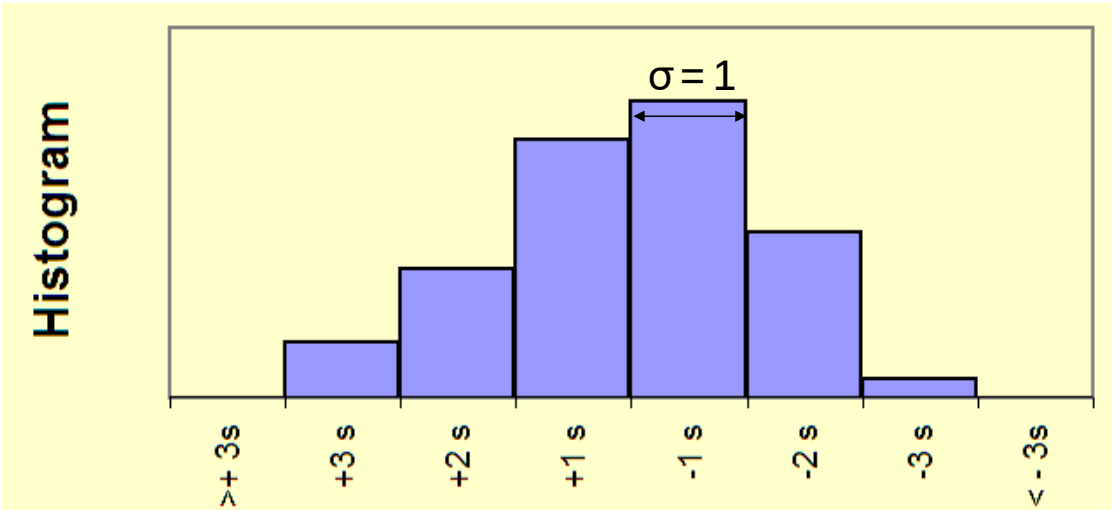
What is the minimum no. of data points required to estimate 'sigma' ($\hat{\sigma}$) ?

Re-phrasing that question, how many 'shots' would you like to see before you would be comfortable to stand here ?

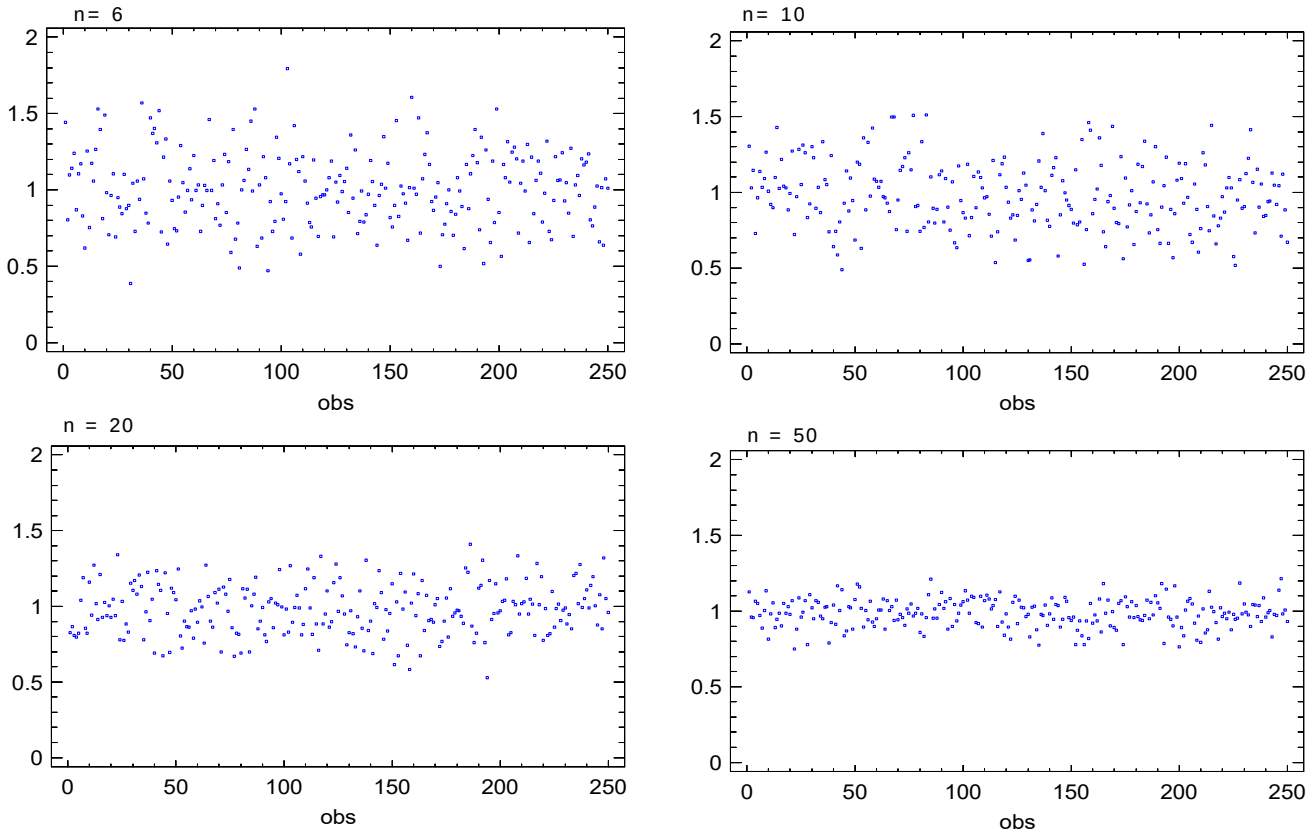


To visualize the variability of *s*, let's do this:

1. repeatedly take samples of various sizes (n) from the following large (10,000) dataset which came from a Normal process with a *true* std dev of 1
2. calculate the sample std dev (*s*) for each 'sample'
3. plot the numerical values of each *s* for each sample size



variability of sample std dev statistic (*s*) versus sample size n



Learning from previous slides

Numerical variation of the std dev statistic (*s*), calculated from repeated sampling, is related to the sample size n:

☞ the larger the n, the less the variation

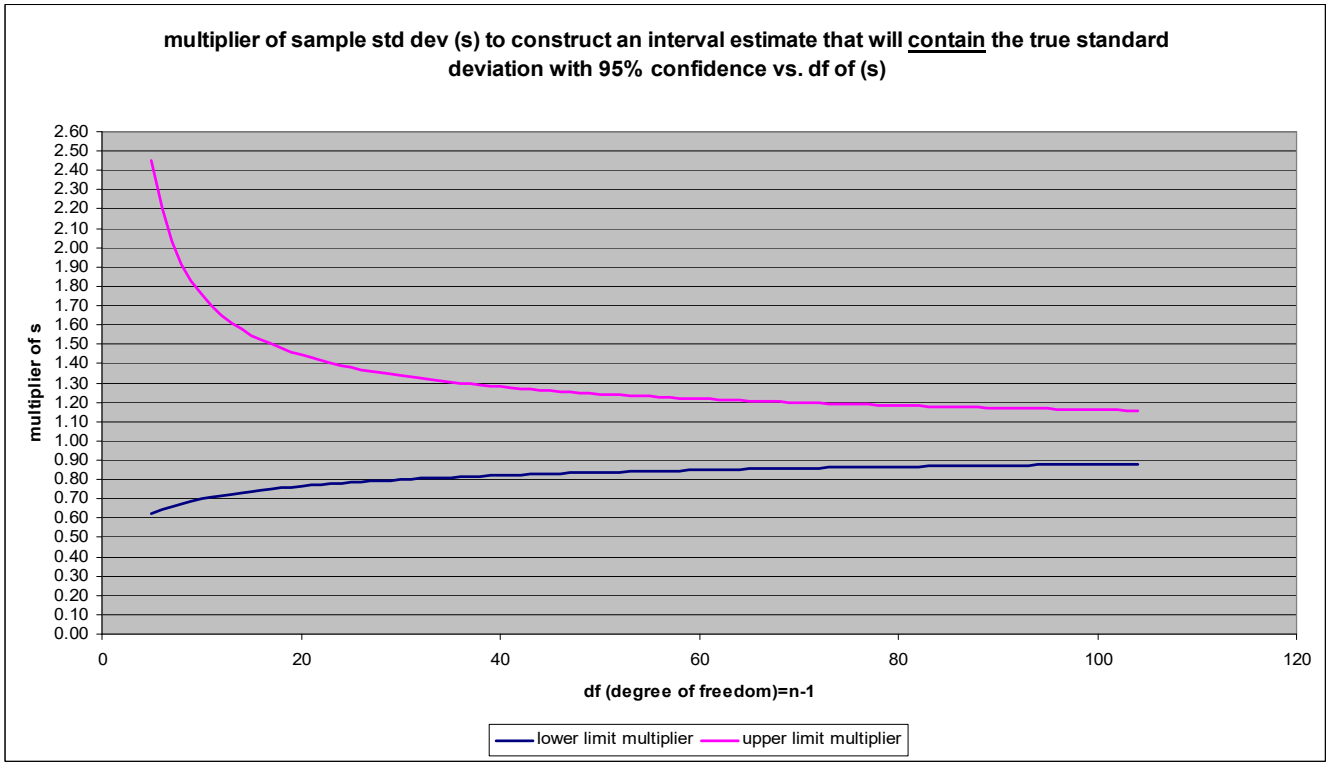
Therefore, std dev estimate that is calculated from a 'small' dataset is highly variable from dataset to dataset !

Hence, decisions using a std dev estimate based on 'small' dataset is highly unreliable.

‘degrees of freedom’ for *s*

- this is a metric that can be viewed as the ‘quality’ or ‘reliability’ associated with the sample standard deviation statistic (*s*)
- the ‘reliability’ or ‘quality’ is gauged by the margin of error of the statistic in estimating the desired parameter (σ)
- mathematically , df = n - 1

visualizing ‘margin of error’ of *s*



'Pooling' of sample standard deviations from multiple 'sampling'

- sample standard deviation statistics obtained from repeated sampling of the same target population can be combined via a process called 'pooling' to improve the *reliability* of the estimate
- pooling is just a fancy term for 'weighted average'
- fundamental assumption → the target population parameter remains constant over multiple samplings

a single ILS to estimate r and R

- this is just a 'one-time' snapshot 'sampling' of the 'target population' to arrive at *estimates* for r and R
- these estimates have the 'margin of errors' as presented in the previous slides

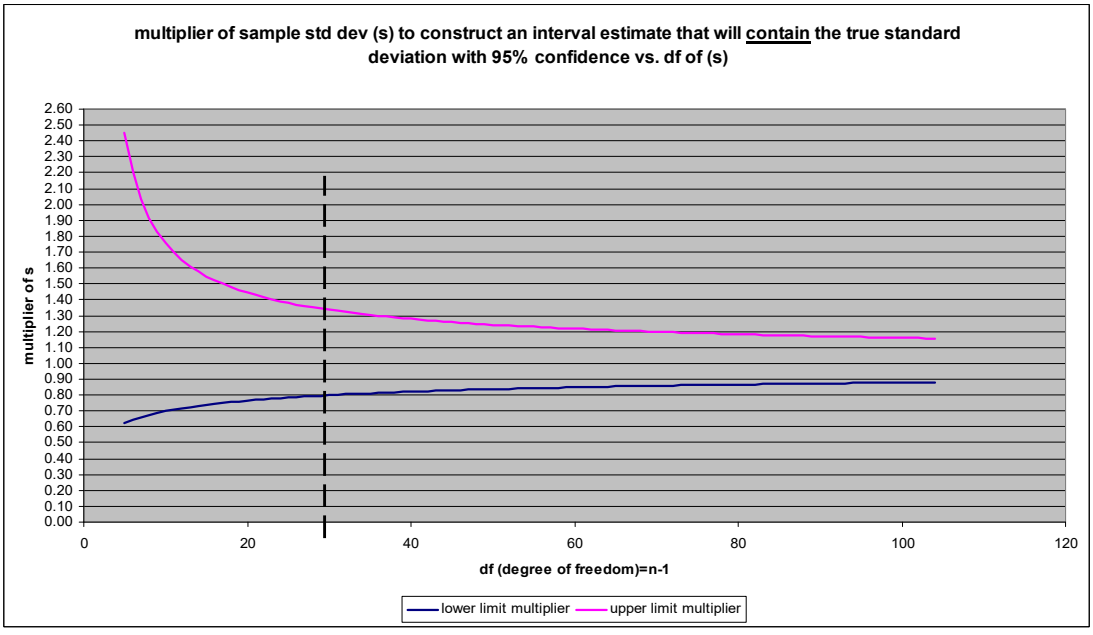
D6300 Requirements for r and R

8.4.4 *Number of Laboratories and Degrees of Freedom for Final Precision Estimates:*

8.4.4.1 The final statement of precision of a test method shall be based on acceptable test results from at least six (6) laboratories and at least thirty (30) degrees of freedom for *R* and *r*.

So, why 30 ?

- 30 df is the *de facto* accepted point of 'diminishing return'
- the nominal upper margin of error is no more than 35%



principal df drivers for r and R in an ILS

- for r, it's no. of samples
- for R, it 'depends'

df for R depends on contribution of "between-lab" bias towards R

- the ILS design and analysis technique in D6300 enables break out of the 'between-lab bias' component
- if between-lab bias is 'dominant', there will be a significant loss of df for R
- in the 'limiting case', the df for R will approach (no. of labs − 1)

Message for ILS Designers

- do *not* be a 'minimalist' and roll the dice
- strive for at least 10 labs (16 is my preference)
- if you can't find enough willing participants, it speaks to the real need (or, lack of) for the STM

Guidance from D300

6.4 *Planning the Interlaboratory Program:*

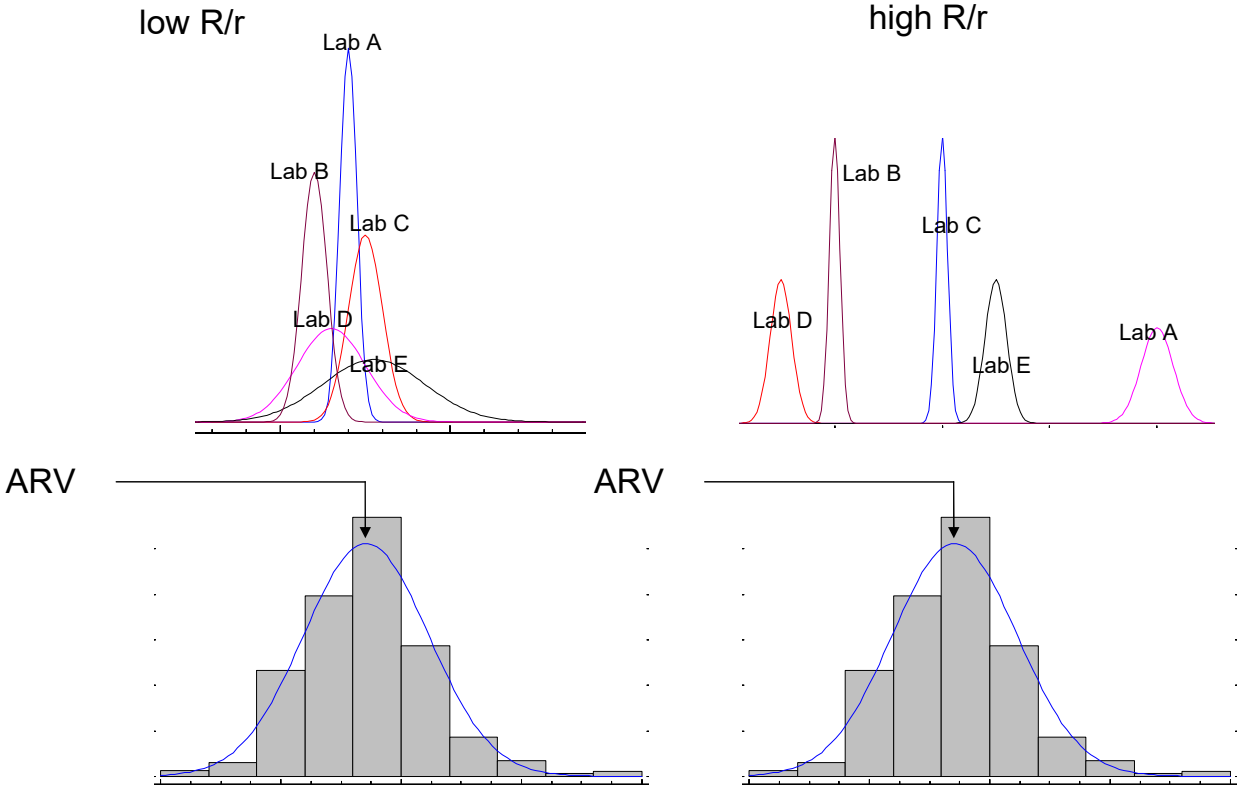
6.4.1 There shall be at least six (6) participating laboratories, but it is recommended this number be increased to eight (8) or more in order to ensure the final precision is based on at least six (6) laboratories and to make the precision statement more representative of the qualified user population.

6.4.2 The number of samples shall be sufficient to cover the range of the property measured, and to give reliability to the precision estimates. If any variation of precision with level was observed in the results of the pilot program, then at least six samples, spanning the range of the test method in a manner than ensures the leverage (h) of each sample (see Eq 2) is less than 0.5 shall be used in the interlaboratory program.

General rule-of-thumb for judging magnitude of 'between-lab' bias

- contribution of between-lab bias towards R can be *roughly* judged by the R/r ratio

Visualizing R/r ratio



What does a very large R/r ratio mean ?

It means:

- the degree of agreement in the execution of a Standard Test Method (STM) *within* a lab is significantly better than the degree of agreement in the execution of the test method *between* laboratories

But, what does it 'really mean' ?

What large R/r ratio really means

→ it means the method developers did not do a very good job in the "S" of "TM", where "S" stands for "Standard"

→ in other words, the test method needs more 'between-lab' standardization to bring the 'between-lab agreement closer

2 KPI's for *any* Standard Test Method

→ the 'precision quality' of any Standard Test Methods can be judged by the following 2 KPI's *used together (i.e.: not individually)* :

1. signal-to-noise ratio (S/N) for r

2. R/r ratio

Signal-to-noise (S/N) ratio for r

- this metric judges the ability of the test method to be repeated in a single lab relative to the level (signal) of measurand:

- it is the simple ratio of : $\frac{\bar{x}}{r}$

- min. acceptable S/N is 3.6

Why 3.6 min. for S/N

Because →

– when you do the math, this is the LOQ

R/r ratio

- this metric judges the adequacy of between-lab standardization, or, agreement between labs
- the ideal case for R/r is 1: which means the dominant component of disagreement (within-lab or between-lab) is similar in magnitude
  – what this really means is the method is so well-standardized that the dominant contributor towards R and r are within-lab noise

Reality

reality: R/r ratio is rarely 1

- strive for 2 or less;
- raise your 'eyebrow' if it's 3.5 - 5;
- frown if it's > 5
- back to the drawing board if it's >10

'Qualification' protocol → an effective approach to improve between-lab agreement

Some examples of 'Qualification' protocol:

- TSF in octane test methods
- reference oil in D2887
- reference standards in D5191

Message to Method Developers

Have I gotten your attention ?

Message to Test Method 'customers'

"Insanity" is: → keep doing the same thing and expect a different outcome…"

Only you, the *paying* customer, can *drive* change

So, if you want a better test method, speak up, and, pitch in → "… if you want a helping hand, look in your wrist …"

- actively participate in method development task groups
- actively support ILS's by *participation*

Work-in-Progress (CS94 is doing our part)

Date: January 5, 2016
To: Coordinating Subcommittee 94
Tech Contact: W James Bover, 908-451-0564
Work Item #: WK52522

Ballot Action: New Standard Guide for Evaluating Test Method Capability and Fitness for Use

Rationale: This proposed standard guide is intended to provide guidance to D02 test method and specification developers regarding the evaluation of the inherent quality of the precision and the adequacy of between-laboratory standardization for a new or modified test method and the evaluation the fitness for use of a test method reproducibility relative to its intended use in specifications.

Standard Guide for
Evaluating Test Method Capability and Fitness for Use[1]
An American National Standard

Questions ?