Draft-geib-ippm-metrictest

Pick up RFC2330 philosophy and merge with draft bradner-metrictest

Expectation of draft-geib and philosophy RFC2330:
Two IPPM metric implementations measuring simultaneously along an identical path, should result in the same measurement.

“The same measurement” expressed in statistical terms is:
Two probing processes (using IPPM implementations) generate samples from the same underlying distribution (due to network conditions along the shared path).

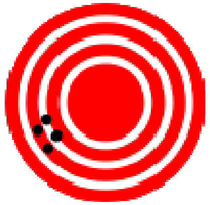
Standard statistical “goodness of fit” test are applied to compare two distributions.

A “goodness of fit test” is proposed by RFC2330 to compare a real probing process against an ideal Poisson distribution. This is a “calibration”.

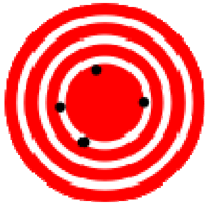
Draft–geib proposes to apply a standard method of statistics suggesting that
two different samples were drawn from the same underlying distribution.
It’s another kind of “goodness of fit” test, but not a calibration.

Draft-geib-ippm-metrictest
Prior work: RFC2330 repeatability (precision)

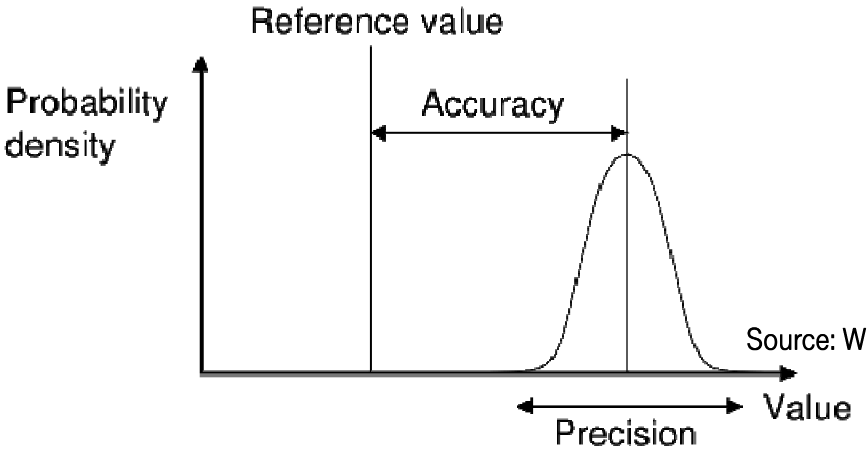
RFC2330: „A methodology for a metric should [be] repeatable: if the methodology is used multiple times under identical conditions, the same measurements should result in the same measurements.”



High precision,
low accuracy



High accuracy,
low precision



Source: Wikipedia

By measuring a metric multiple times, probes are drawn from an underlying (and unknown) distribution due to networking conditions.

Draft-geib-ippm-metrictest
Prior work: RFC2330 goodness of fit

RFC2330: „A methodology for a given metric exhibits continuity if, for small variations in conditions, it results in small variations in the measurements.”

Draft-geib: Using a different metric implementation under otherwise identical (network) conditions should introduce only a “small variation”.

The sample distribution of metric implementation A is taken as the „given“ distribution against which the sample distribution of metric implementation B is compared by a goodness of fit test (proposal: Anderson-Darling 2-sample test).

RFC2330 provides guidelines on testing for goodness of fit for calibration (quotes):

- Summarizing measurements by histograms, the EDF is preferred.
- IPPM goodness-of-fit tests are done using 5% significance.
- Anderson-Darling EDF tests are recommended and implemented in the appendix.

EDF: Empirical distribution function

Draft-geib-ippm-metrictest
Prior work: RFC2330 self-consistency

RFC2330: „A fundamental requirement for a sound measurement methodology is that measurement be made using as few unconfirmed assumptions as possible."

Draft-geib: Then ensure "identical (networking) conditions" during a metric test as far as possible and pay respect to measurement errors and statistical errors.

What does it mean in practice? Some ideas, further discussion needed:

- Two metric implementations should measure within the same IP tunnel, if possible.
- Avoid evaluation close to (or beyond) the resolution and measurement error of the environment.
- Minimum sample size: Within any compared sample, at least five singletons of the compared metric must be present.
- Compare metric samples over identical time intervals where applicable.

Draft-geib-ippm-metrictest
A brief comparison with draft-bradner

Both follow the same basic idea to verify that two metric implementations measure the same under identical conditions.

Draft-bradner proposes a test based on sample means and standard deviations.

In addition, a goodness of fit test that measures if distributions of two samples deviate or may correspond to a common underlying distribution provides a broader decision basis and often higher confidence. That‘s what draft–geib proposes.

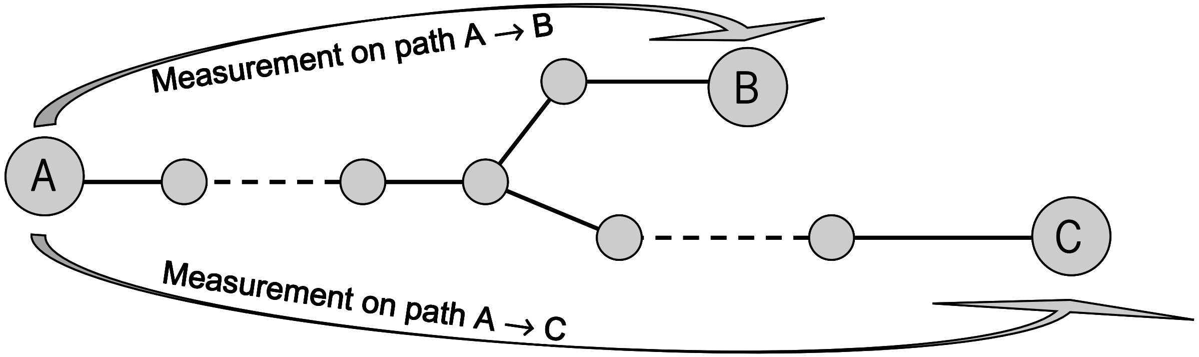
The following slides show Anderson-Darling tests of two instances of the same metric implementation *partially sharing a network path* (note that this gives a hint how comparable network conditions may be established).
No tunneling has been used.

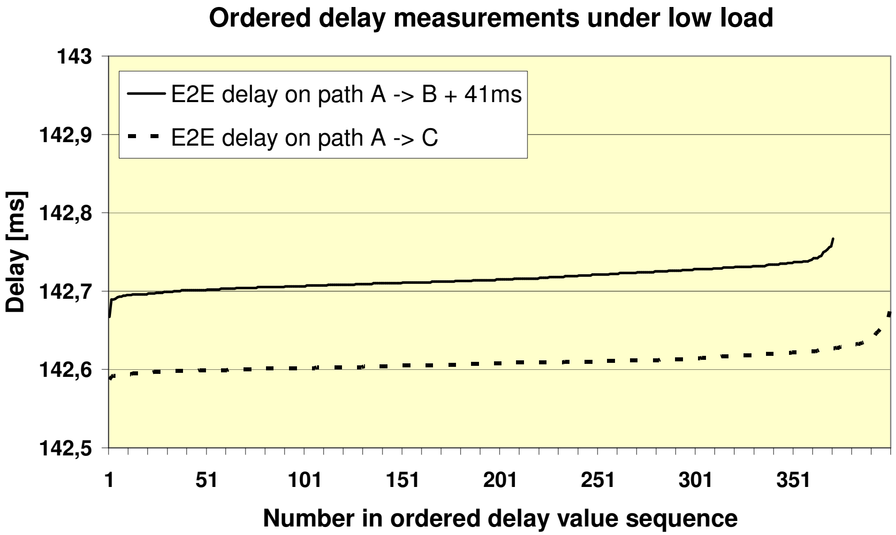
Example: E2E delay measurement comparison

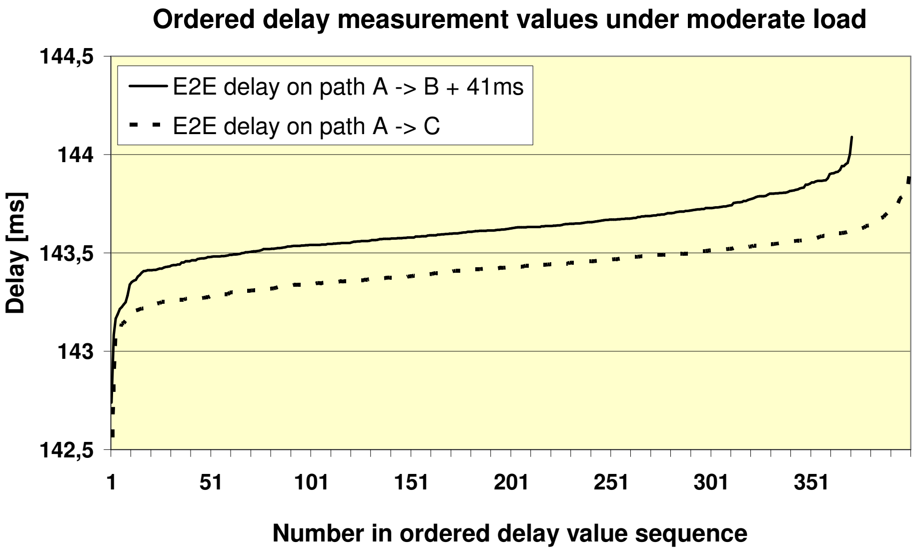
Samples in the same time frame on both paths AB, AC via a broadband access network taken by identical implementations (PERFAS-tool incl. GPS)

Measurement during

- a low load period
  see graphs on the left
- a moderate load period
  see graphs on the right



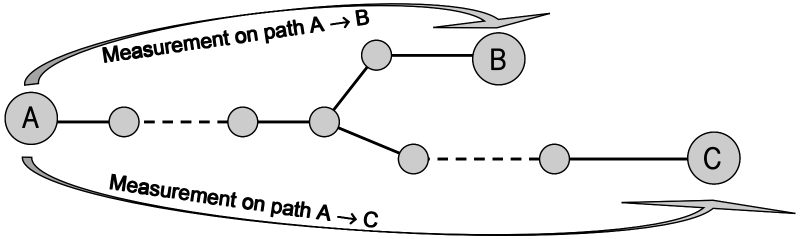




Example: E2E delay measurement comparison

Results of measurement statistics:

[ms]	Mean	Standard Deviation
Moderate load AB	**101.4**	**0.142**
Moderate load AC	**142.6**	**0.148**
Low load AB	**100.6**	**0.012**
Low load AC	**141.7**	**0.013**



Anderson-Darling 2-sample test: Measure of difference in delay distributions

$$\left[\; A^2 = \frac{n+m-1}{(n+m)^2}\left[\frac{1}{n}\sum_{j=1}^{L} h_j \frac{((n+m)F_j - nH_j)^2}{H_j(n+m-H_j) - \frac{(n+m)h_j}{4}} + \frac{1}{m}\sum_{j=1}^{L} h_j \frac{((n+m)G_j - mH_j)^2}{H_j(n+m-H_j) - \frac{(n+m)h_j}{4}}\right] \;\right]$$

Critical value, 5% level

⇓

For moderate load (AB+41.2ms) ↔ AC: $A^2 = 0.465 < 1.993 \Rightarrow$ may be ident.

For low load (AB+41.1ms) ↔ AC: $A^2 = 3.778 > 1.993 \Rightarrow$ not identical

(Accuracy limit of the measurement tool may cause deviating distributions)

Comparing each of the four distributions to a Gaussian distribution, the Anderson-Darling measure is >3-fold beyond the 5% level critical value

E2E delay measurement comparison

Final Remarks

Consistent statistical metrics often can be taken on operational WANs

- in the example even when comparing different paths
- reports e.g. from the Sprint measurement architecture indicate low delay variability in backbones (Fraleight et al., IEEE Networks'03)

For valid comparison, knowledge of the network conditions has to be included.

WAN network conditions are expected to be less tunable than in a lab but more realistic; problems with accuracy, variability, rare events etc. have to be addressed more or less in any network environment.

Test results should clearly indicate when network conditions make it infeasible to measure and compare a considered metric.