Issues in Metric Selection

Alain Crolotte
Problem Statement

• Requirement for a single number
• Arithmetic mean potentially dominated by a large value (a priori issue)
• Solution
  – Throw away one?
  – Another Metric?
Characteristics of Central Tendency

- Arithmetic mean
  \[ m = \frac{1}{n} \sum x_i \]
- Geometric mean
  \[ g = \left( \prod x_i \right)^{1/n} \]
- Harmonic mean
  \[ h = \frac{1}{\frac{1}{n} \sum \frac{1}{x_i}} \]
The $\phi$–average

$$\phi(M_\phi) = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$$

$$m_r = \frac{1}{n} \sum x_i$$

$$s = \left(\sqrt{x_1} + \sqrt{x_2} + \ldots \sqrt{x_n}\right)^2$$
The Geometric Mean

- Used in Statistics and Economics
- Treats relative variations equally

\[
\frac{\Delta g}{g} = \frac{1}{n} \frac{\Delta x_i}{x_i}
\]

- One zero observation point brings the geometric mean to zero!
Avoiding the geometric mean pitfall

• The $a$-displaced average

$$\log(g_a + a) = \frac{1}{n} \sum \log(x_i + a)$$

• The TPC-D power metric – geometric mean but replace the small observations by the max observation divided by 1000
Pitfall cannot be avoided

• TPC-D pre-joined techniques penalized heavily by UF1 and UF2
• Pre-aggregation results in small tables that can be updated at virtually no cost
• Example: all queries 100 sec. – with pre-aggregation Q1 goes to 0.2 sec.
• Arithmetic mean: 100 -> 95 [-5%]
• Geometric mean: 100 -> 72 [-28%]
TPC-D 1999

- Hyper-inflation of power metric
- Benchmark retired – TPC-H starts
- TPC-H does not allow explicit materialization
- Same metric but problem did not appear
Conclusion

• Use arithmetic mean in DSS for a single-stream metric
• It is simple
• It represents meaningful physical quantities (its inverse is a real rate)
• In general use it for application involving quantities that require the additive property