2 On Metrics and Measurements

Rainer Böhme[1] and Felix C. Freiling[2]

[1] Technische Universität Dresden, Germany
[2] University of Mannheim, Germany

The following chapter attempts to define the notions of metric and measurement which underlie this book. It further elaborates on general properties of metrics and introduces useful terms and concepts from measurement theory, without being overly formal.

2.1 On Measurement

In many cases, *to measure* means to attach a number to an object, i.e., to represent some aspect of the object in a quantitative way. For example, scientists can measure the temperature and the humidity of a location at a certain time by coding observations (temperature, humidity) related to the object (location) with numbers. More generally, a measurement function assigns an element of a set to an object, where the specific element is chosen depending on an observation. The set must not necessarily comprise numbers but can also consist of unordered symbols. For example, classifying the weather today as "rainy", "dry", "foggy" etc. is also regarded as measurement. However, not every assignment of numbers to objects is considered as measurement. For example, the matriculation number of a student is not a measurement because the number is chosen regardless of the student's attributes (here we ignore that higher matriculation numbers may be an indicator of later admission).

In the setting of this book we usually want to measure *attributes* of systems or parts thereof, such as methods or processes. As system can be complex, there are many different measurable attributes. Any form of measurement is an *abstraction*: it reduces the complexity of one or several attributes of the original system to a single symbol. The main purposes of this form of abstraction are to *classify* and *compare* systems.

It is important to stress the difference between an attribute and its measurement. For example, the complexity of a software system is an attribute which can be measured in many different ways. However, the difference between an attribute and its measurement sometimes blurs because measurements are also taken to *define* the attributes.

Measurement is closely connected to the notion of a metric. In the course of this book we will use the term *metric* for a precisely defined method which is used to associate an element of an (ordered) set V to a system S. This definition is used in the area of software quality. In other areas, the term metric only refers to the set V, which contains *indicator values* that answer certain questions asked about a system. As we will see later, our understanding neither corresponds to the strict mathematical definition of a metric (where it is a generalisation of the notion of a distance).

In general, a metric can be formalised as a function M that takes a particular system from the set of systems S and maps it to an element of V:

$$M : S \mapsto V$$

I. Eusgeld, F.C. Freiling, and R. Reussner (Eds.): Dependability Metrics, LNCS 4909, pp. 7–13, 2008.

For example, M may be the assignment of a distance between two measurement points of a system. Then V is the set of real numbers, a totally ordered set. The set V can also be a discrete set like the set of natural numbers in the "lines of code" metric for software. The set V must also not necessarily be totally ordered; it can also be a partially ordered set or an unordered set like in the classification example above where V consists of the elements $\{\mathrm{foggy}, \mathrm{rainy}, \mathrm{dry}\}$ etc.

Attributes can have certain properties which should be reflected in their metrics. For example, the complexity of a software package can be categorised as "low" or "high". Some attributes are meaningful in the context of composed systems. For example, the attribute "size of a program" can be measured in lines of code. Given two programs x and y we can define their composition z as the concatenation of x and y. The metric "lines of code" reflects additivity in the following sense: the sizes of program x and program y together sum up to the size of their composition z. Similarly, some attributes allow to state relations between systems. Taking the "size" metric lines of code again, it is possible to say that some program is twice as large as another program.

Determining a suitable metric for an attribute of a system is not always easy. A good metric should reflect the relevant properties of the attribute in a homomorphic way. This means that certain statements which can be made for a certain attribute of systems should be reflected in the measurements of that attribute. In particular, two properties should hold:

- Any sensible *relation* between systems regarding a particular attribute should be reflected by a corresponding relation between the measurements of this attribute. For example, a system x which is more complex than a system y should be ordered appropriately if some complexity metric c is used, i.e., $c(x) > c(y)$ should hold.
- Any meaningful *operation* on attributes of a system should have a corresponding operation on the measurements of that attribute. Assume there exists an addition operation ("plus") for the "size" of programs. If the size of program x "plus" the size of program y equals the size of program z, then this should be reflected in the appropriate metric for size. For example, lines of code is an appropriate metric if the "plus" operator refers to concatenation of source code.

Any relations or operations on measurements which do not have a corresponding relation or operation on attributes must not be used to process the measurements.

2.2 On Scales

The result of measurements is data, which is further processed or analysed to answer questions of interest to the researcher or practitioner. A useful approach to classify types of data is given in the notion of *scales*. The term scale refers to the range V of a metric, and the relation between elements within V. The most commonly used typology of scales goes back to Stevens [459], who defined a hierarchy of four different types of scales based on the invariance of their meaning under different classes of transformation. He further proposed to derive permissible procedures for data analysis and statistical inference depending on the scale level.

Nominal Scale

The simplest type of scale is the *nominal scale* (also known as *categorial scale*). With a nominal scale, V is an unordered discrete set. Classifications usually employ the nominal scale, for example when classifying computers according to their operation system ($V = \{\mathrm{Windows}, \mathrm{Unix}, \mathrm{OS}/2\}$). Measurements on a nominal scale can be compared for identity or distinction and a number of measurements can be aggregated by counting the frequencies in each class (or combination of classes if data from more than one scale are analysed at a time).

Nominal scales can be transformed into other nominal scales by applying a bijective mapping, i.e., a $1:1$ correspondence between the elements of both scales V_1 and V_2. If more than one category in V_1 is mapped to a single element in V_2 then the transformation loses information and thus is irreversible. It might still be useful to apply such a transformation to aggregate data and increase the number of observations in each (combined) category.

The special case where V consists of two elements only is called *dichotomic scale* (examples: "yes"/"no", "0"/'1'", "male"/"female").

Ordinal Scale

The *ordinal scale* differs from the nominal scale in that V is a discrete *ordered* set. Examples for ordinal scales include severity measures for earthquakes or grades given to students in examinations. In contrast to the nominal scale, two measurements on the same ordinal scale can be compared with operators "less than" or "greater than". This allows the data analyst to create ranks and compute rank correlations. Ordinal scales also allow for simple models of measurement error and they can be included as dependent variables in regression models (ordinal logit or probit models).

Two ordinal scales can be transformed into each other by applying a bijective mapping f which preserves the ordering relation (monotonic mapping), i.e., if $a < b$ on one scale then $f(a) < f(b)$ on the other scale. As ordinal scales are one step higher in the hierarchy than nominal scales, a downgrading (with information loss) to the nominal level is always possible.

Interval Scale

The *interval scale* is an extension of the ordinal scale where the distance between adjacent elements in V is both meaningful and constant (equidistance). Interval scales therefore support the difference operator, so that the difference between two points on the same scale can be compared to the difference between two other points.

Interval scale A can be transformed into interval scale B by linear transformations (adding/subtracing a constant, multiplying/dividing by a constant) since the relative distance between any two scale points is not changed.

The standard example for an interval scale is the measurement of temperature. For example, let scale A be the scale of measurement in degrees Fahrenheit and scale B be

the measurement in degrees Celsius. To transform a measurement in scale A into scale B we can then use the formula:

$$f(x) = \frac{5}{9}(x - 32)$$

Other applications for the interval scale include multi-point rating scales in questionnaires when the scale points are labeled with increasing numbers or are not annotated at all. If more detailed annotations are given then the semantic difference between any two scale points may vary and thus the resulting data should be treated as ordinal rather than interval.

A number of measurements can be summarised with statistics of location (mean), scale (variance) and higher moments. Moreover, interval scales allow for continuos distribution error models, such as Gaussian measurement errors. This implies that the entire class of parametric statistics can be applied to data on interval scales. Again, interval scales can be converted to ordinal (and nominal) scales by cutting the scale at some breakpoints and assigning the observations to the categories between.

Ratio Scale

The *ratio scale* is an extension of the interval scale where the origin (value of 0) is defined in a natural way. Examples for ratio scales are length, mass, pressure, time duration or monetary value. Additional possible operations for analysis are multiplication of a measurement with a constant factors, taking logs and finding roots (among others). Therefore statistical measures such as the geometric mean and the coefficient of variation are defined for ratio scales only. Transformations between different ratio scales can be achieved by simply multiplying measurements with a scaling factor. For example, converting a length metric in metres into a length in imperial feet is done by using a scaling factor of 3.2808.

Interval and ratio scales are sometimes subsumed to *cardinal scales*.

Summary

Table 1 shows an overview of the different scale types discussed so far. Nominal and ordinal scale are usually referred to as *qualitative scales*, whereas interval and ratio scale are called *quantitative scales*. From a measurement point of view it is recommended to collect data at the highest possible scale level. In particular metrics with at least a quantitative scale are useful because they enable parametric statistics, which are more powerful than non-parametric methods. The term *parametric* refers to a distribution assumption where inference can be made on the parameters of the distribution rather than on individual data points. For example, a comparison of means from two sequences of measurements is a parametric method because the mean is a parameter of the distribution of the data.

As a final remark: although very popular, Steven's typology of scale levels has been criticised for not being comprehensive (it is easy to find pathological examples that do not fit well in one of the four categories) and for imposing unnecessary restrictions to data analysis by adhering to strict mathematical standards, which are difficult to meet

Table 1. The hierarchy of scale levels

Scale level	Examples	Operators	Possible analyses
Quantitative scales			
Ratio	size, time, cost	$*, /, \log, \sqrt{}$	geometric mean, coefficient of variation
Interval	temperature, marks, judgement expressed on rating scales	$+, -$	mean, variance, correlation, linear regression, analysis of variance (ANOVA), ...
Qualitative scales			
Ordinal	complexity classes	$<, >$	median, rank correlation, ordinal regression
Nominal	feature availability	$=, \neq$	frequencies, mode, contingency tables

with real data [484]. Nevertheless, even the critics acknowledge that the typology provides simple guidance and protects naive data analysts from errors in applying statistics. Therefore we deem it useful to keep in mind when designing and discussing dependability metrics.

2.3 On Mathematical Metrics and Norms

In mathematics, a metric is a precisely defined term. It is the abstraction of a *distance*. Formally, a metric d on a set X is a function which assings a "distance" value (a real number) to pairs of elements from X:

$$d : X \times X \mapsto R$$

The function d must satisfy several conditions to be a mathematical metric, i.e., for all $x, y, z \in X$ must hold:

- every distance is non-negative, i.e., $d(x, y) \geq 0$,
- the distance is zero for identical inputs, i.e., $d(x, x) = 0$,
- the distance is symmetric, i.e., $d(x, y) = d(y, x)$, and
- the triangle inequality holds, i.e., $d(x, z) \leq d(x, y) + d(y, z)$.

For example, consider the (two-dimensional) Euclidian distance where X is the set of coordinates in a two-dimensional space, i.e., $X = R \times R$. The Euclidian distance d_E calculates the distance of two points in X. Given two elements (x_1, y_1) and (x_2, y_2) of X, d_E is defined as

$$d_E = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

It is easy to see that the conditions above hold for the Euclidian distance.

Somewhat closer to the notion of metric defined in section 2.1 is the mathematical notion of a *norm*. In mathematics, a norm is an abstraction of a positive length or size. A norm is a function p that maps an element of a set X to the real numbers:

$$p : X \mapsto R$$

The set X is usually a multi-dimensional vector space. To be called a norm, p must satisfy the following conditions for all $x, y \in X$:

- the norm is always be positive, i.e., $p(x) \geq 0$,
- the norm is scalable, i.e., $p(ax) = |a|p(x)$ for some scalar a,
- the triangle inequality holds, i.e., $p(x + y) \leq p(x) + p(y)$,
- the norm is zero for the zero vector only, i.e., $p(x) = 0$ if and only if x is the zero vector.

Standard examples are the (two-dimensional) Euclidian norm p_E which assigns a length to a (two-dimensional) vector. More precisely, $X = R \times R$ and for any $(x, y) \in X$ p_E is defined as follows:

$$p_E = \sqrt{|x|^2 + |y|^2}$$

Another notation for $p_E(\boldsymbol{x})$ is $||\boldsymbol{x}||_2$.

Another well-known norm is is the Taxicab (or Manhatten) norm, which assigns to a vector the "length" if you would take a taxi in a rectangular street grid from the origin point to the point described by the vector. Here, $X = N \times N$ (where N denotes the set of natural numbers) and for any element $(x, y) \in X$ the Manhatten norm is defined as $x + y$.

There is a close relationship between norms and metrics in the sense that every norm implicitly defines a metric and special types of metrics implicitly define a norm. For example, given a norm p on a set X, the construction $d : X \times X \mapsto R$ with $d(x, y) = p(x - y)$ satisfies all the properties of a metric in the mathematical sense.

It is obvious that the mathematical definition of a metric requires the properties of a ratio scale in Steven's terminology. Since by far not every measurable aspect satisfies these conditions, we will use the term metric in a less rigourous way throughout this book.

2.4 Classification of Metrics

This volume presents a large number of metrics for measuring the dependability of systems. Metrics can be classified by a number of aspects, most importantly by the way they are constructed and the attributes of systems they represent. However, there are also some possibilities to classify metrics according to their abstract properties, such as:

- *Scale level*, and hence the granularity of V, as discussed in section 2.2.
- *Construction:* a metric can be derived in different ways, which results in the difference between *analytical* vs. *empirical* metrics. An analytical metric measures the system by analysing its structure or its properties using *models* of the system. An empirical metric measures by observing the real behavior of the system.
- *Directness:* it is possible to distinguish *direct* vs. *indirect* metrics. A direct metric measures the system itself, whereas an indirect metric measures the effects of the system onto another system. For example, the stock market price of the share of company X is an indirect metric of the expected performance of company X. Analytical metrics can also be regarded as indirect metrics.

– *Obtrusiveness:* metrics that require a modification of the system for the purpose of being measured are called *obtrusive* metrics, as opposed to *unobtrusive* ones, which can be taken without touching (i.e. influencing) the system.

2.5 On the Quality of Metrics

It is easy to define metrics, but much harder to find meaningful ones. An important quality of a metric is whether it reflects the attributes in question in a homomorphic way (see section 2.1). This can be regarded as a notion of *validity* of a metric. Closely related is the question of the *granularity* of a metric, i.e., does it allow to distinguish all systems that differ in their respective attributes?

There are also practical considerations when defining a metric: its *availability* and its *cost*. Is it always possible to compute the metric for a given system? An empirical performance metric of a production system is obviously much harder to collect than an analytical one, because the production system will not always be available for benchmarking or the costs to conduct the measurement are much higher.

Finally, a very desirable quality of a metric is its *stability*: different people measuring the attribute in question should roughly get the same results. This property is sometimes called *scale reliability* which should not be confused with the attribute of dependability called *reliability*.