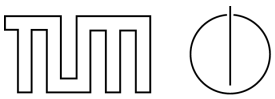
Software Quality Management
Summer Term 2010
Dr. S. Wagner
L. Heinemann, S. Islam
04.06.2010



Technische Universität München
Fakultät für Informatik

Work Sheet 1: Scales and Aggregation

Scales

Assign the measure examples to the correct scale types.

- Number of defects
- Defect types
- Effort in person-hours
- Rating of ease of use between 1 and 5
- Requirements IDs
- Lines of code
- Cyclomatic complexity
- Response time
- Maintenance hours
- Training hours for users
- Recovery time
- Probability that an attacker breaks the system
- Workload/time
- Number of clicks

Aggregation Theory

Informally, aggregation is the problem of combining n-tuples of elements all belonging to a given set into a single element often of the same set. In mathematical aggregation, this set can, for example, be the real numbers. Then an aggregation operator A is a function that assigns an y to any n-tuple $(x_1, x_2, \ldots, x_n)$:

$$A(x_1, x_2, \ldots, x_n) = y \tag{1}$$

The literature defines additional properties that are requirements for a function to be called an *aggregation operator*. However, these properties are not all compatible. Yet, there seem to exist some undisputed properties that must be satisfied. For simplification, the sets that aggregation operators are based on are usually defined as $[0, 1]$, i.e., the real numbers between 0 and 1. However, other sets can be used and by normalisation to this set it can be shown that the function is an aggregation operator. Additionally, the following must hold:

$$A(x) = x \quad \text{identity when unary} \tag{2}$$

$$A(0, \ldots, 0) = 0 \wedge A(1, \ldots, 1) = 1 \quad \text{boundary conditions} \tag{3}$$

$$\forall x_i, y_i : x_i \leq y_i \Rightarrow$$

$$A(x_1, \ldots, x_n) \leq A(y_1, \ldots, y_n) \qquad \text{monotonicity} \tag{4}$$

The first condition obviously only is relevant for unary aggregation operators, i.e., the tuple that needs to be aggregated only has a single element. Then we expect the result of the aggregation to be that element. The boundary condition cover the extreme cases of the aggregation operator. With only minimal input there must be the minimum output and vice-versa. Finally, we expect that an aggregation operator is monotone. If all values stayed the same or increased we want the aggregation result also to increase or at least stay the same.

Apart from these three conditions, there is a variety of further properties that an aggregation operator can have. We only introduce three more that are relevant for aggregation operators of software measures.

The first condition that introduces a very basic classification of aggregation operators is *associativity*. An operator is associative if the results keep the same no matter in what packages the results are computed. This has interesting effects on the implementation of the operator as associative operators are far easier to compute. Formally for an associative aggregation operator A_a the following holds:

$$A_a(x_1, x_2, x_3) = A_a(A_a(x_1, x_2), x_3) = A_a(x_1, A_a(x_2, x_3)) \tag{5}$$

The next interesting property is *symmetry*. This is also known as *commutativity* or *anonymity*. If an aggregation operator is symmetrical, the order of the input arguments has no influence on the results. For every permutation σ of $1, 2, \ldots, n$ the operator A_s must satisfy:

$$A_s(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)} = A_s(x_1, x_2, \ldots, x_n) \tag{6}$$

The last property we look at because it holds for some of the operators relevant for software measures is *idempotence*. It is also known as *unanimity* or *agreement*. Idempotence means that if the input consists of only equal values, it is expected that the result is also this value.

$$A_i(x, x, \ldots, x) = x \tag{7}$$

Aggregation Operators

Grouping

A very high level aggregation is to define a set of groups, probably with a name each, and assign the inputs to the groups. This allows a very quick comprehension and easy communication about the results. However, the information loss is rather large.

Rescaling. An often used technique to be able to overlook the large amount of information provided by various metrics is to change the scale type by grouping the individual values. This is usually done from higher scales such as ratio scales to ordinal or nominal scales. For example, we could define a certain threshold value. Above the group is *red*, below it is *green*. This is useful for all purposes apart from trend analysis where it can be applied only in a few cases. It is not idempotent in general and it depends on the specifics of the rescaling whether it is symmetrical.

Cluster Analysis. Another, more sophisticated way, to find regularities in the input is cluster analysis. It does basically the same thing as the rescaling described above but with finding the groups using clustering algorithms. The *K-means* [**?**] algorithm is a common example of such algorithms. It works with the idea that the input are points scattered over a plain and there is a distance measure that can express the space between the points. The algorithms then works out which points should fall into the same cluster. This aggregator is not associative and not idempotent.

Central Tendency

The central tendency describes what colloquially is called the average. There are several aggregation operators that can be used for determining this average of an input. They depend on the scale type of the measures the are aggregating. All of them are not associative but idempotent.

Mode. The mode is the only way for analysing the central tendency for measures in a nominal scale. Intuitively, it gives the value that occurs most often in the input. Hence, for inputs with more than one maximum, the mode is not uniquely defined. If the result is then defined by the sequence of inputs, the mode is not symmetrical. The mode is useful for assessing the current state of a system and for comparisons w.r.t. measures in a nominal scale. For $n_1, \ldots, n_k$ being the frequencies of the input values, the mode M_m is defined as

$$M_m(x_1, \ldots, x_k) = x_j \Leftrightarrow n_j = \mathsf{max}(n_1, \ldots, n_k). \tag{8}$$

Median. The median is the central tendency for metrics in an ordinal scale. An ordinal scale allows to enforce an order on the values and hence a value that is in the middle can be found. The median ensures that at most 50% of the values are smaller and at most 50% are greater or equal. The median is useful for assessing the current state and comparisons. The median $M_{0.5}$ is defined as

$$M_{0.5}(x_1, \ldots, x_k) = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)} & \text{otherwise} \end{cases} \tag{9}$$

The median of measures in ordinal scale, the division by 2 is not possible. Hence, in this case there are two medians.

Mean. For measures in interval, ratio, or absolute scale, the mean value is defined. There are mainly three instances of means: arithmetic, geometric, and harmonic mean. The arithmetic mean is what usually is considered as average. It can be used for assessing the current state, predictions and comparisons. The arithmetic mean M_a is defined as follows:

$$M_a(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{10}$$

For trend analysis, the geometric mean can be used. It is necessary when measures are relative to another metric. For example, the growth rates of the size of several releases could be aggregated using the geometric mean. The geometric mean M_g is defined as

$$M_g(x_1, \ldots, x_n) = \sqrt[n]{\prod_{i=1}^{n} x_i}. \tag{11}$$

As it uses the product it actually has the absorbent element 0 [**?**]. Finally, the harmonic mean needs to be used when different sources are combined and hence weights need to harmonise the values. An example would be when, in order to analyse the reliability of a system, the fault densities of the components are weighted based on their average usage. Given the weights w_i for all the inputs x_i, the harmonic mean M_h is given by

$$M_h(x_1, \ldots, x_n) = \frac{w_1 + \ldots + w_n}{\frac{w_1}{x_1} + \ldots \frac{w_n}{x_n}}. \tag{12}$$

Dispersion

In contrast to the central tendency, the dispersion gives an impression about how scattered the inputs are over their base set. Hence, we look at extreme values and their deviation from the central tendency.

Variation Ratio. This ratio is given by the proportion of cases which are not the mode. This is the only way for nominal measures to have a measure of dispersion. It indicates whether the data is in balance. This is useful for hot spot identification. The variation ratio V is defined again using $(n_1, \ldots, n_k)$ as the frequencies of $(x_1, \ldots, x_k)$ by

$$V(x_1, \ldots, x_k) = 1 - \frac{\mathsf{max}(n_1, \ldots, n_k)}{k}. \tag{13}$$

Maximum and Minimum. Very useful operators for various analysis situations are the maximum and minimum of a set of measures. They can be used with measures of any scale apart from nominal. They are useful for identifying hot spots and for comparisons. They both are associative and symmetrical. The maximum *max* and the minimum *min* are defined as follows:

$$\forall x_i.y \geq x_i \Rightarrow \mathsf{max}(x_1, \ldots, x_n) = y \tag{14}$$

$$\forall x_i.y \leq x_i \Rightarrow \mathsf{min}(x_1, \ldots, x_n) = y \tag{15}$$

Range. The range is the standard tool for analysing the dispersion. Having defined the maximum and the minimum above, it is easy to compute. It is given by the highest value minus the lowest value in the input, which is only possible in interval scale or higher. It can be useful for assessing the current state and comparisons. This operator is neither idempotent nor associative. The range R is simply defined as

$$R(x_1, \ldots, x_n) = \mathsf{max}(x_1, \ldots, x_n) - \mathsf{min}(x_1, \ldots, x_n) \tag{16}$$

Median Absolute Deviation. This dispersion measure is useful for interval and ratio metrics. It is calculated as the average deviation of all values from the median. This again can be used for current state analyses and comparisons when the median is the most useful measure for the

central tendency. The median absolute deviation is not associative but symmetrical. It is defined as

$$D(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} |x_i - M_{0.5}(x_1, \ldots, x_n)|. \tag{17}$$

Variance and Standard Deviation. For other scales, the most common dispersion measures are the variance and the standard deviation. It is always the best choice for analysing the dispersion when the mean is the best aggregator for the central tendency. The standard deviation has the advantage over the variance that it has the same unit as the measure to be aggregated. Hence, it is easier to interpret We use it again for analysing the current state and for comparisons. It is not associative but symmetrical. The variance S^2 and the standard deviation S are defined as follows:

$$S^2(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i_1}^{n} (x_i - M_a(x_1, \ldots, x_n))^2 \tag{18}$$

$$S(x_1, \ldots, x_n) = \sqrt{S^2(x_1, \ldots, x_n)} \tag{19}$$