On the Definition and Use of Aggregate Indices for Nominal, Ordinal, and Other Scales

Sandro Morasca

Dipartimento di Scienze Politiche, della Cultura e dell'Informazione
Università degli Studi dell'Insubria
via Valleggio, 11 - 22100 Como (Italy)
sandro.morasca@uninsubria.it

Abstract

It is not uncommon in software engineering measurement to deal with attributes measured with nominal or ordinal scales. Also, it has long been debated whether it is possible to find ordinal scales for the structural complexity of software code. In this paper, we address two problems: (1) the definition of concentration and dispersion indices for nominal scales; (2) the conditions under which the comparisons of arithmetic means or geometric means are meaningful for scales that are ordinal or not even ordinal.

1. Introduction

Nominal and ordinal scales are often used in software engineering measurement applications, even though they are considered to be less information bearing than interval or ratio ones. Some pieces of information are intrinsically nominal, like the programming language of a software module. In other cases, ordinal scales are used because it would not make much sense to use ratio or interval scales. For instance, software failures are classified on an ordinal criticality scale during debugging. Using a ratio scale would entail precise knowledge of, say, the economic damage produced by a failure, which is hardly ever possible to assess. Thus, nominal and ordinal scales sometimes provide the only pieces of information available for some software product or process attributes. So, it is important to extract as much information as possible from them.

Aggregate indices play a special role when dealing with data measured with any kind of scale, because they provide a concise idea about the data set at hand. Various kinds of indices have been used for statistical populations, and the best known ones are indices of central tendency (e.g., the mean) and indices of dispersion (e.g., the standard deviation). The actual index used in a measurement application clearly depends on the kind of scale used. For instance, using the mean for nominal data or the standard deviation for ordinal data may lead to meaningless statements and results.

In this paper, we address issues related to aggregate indices for nominal and ordinal scales. Specifically, we first provide properties and example indices that show how it is possible to define concentration indices for nominal scales. Our properties generalize properties that are the counterpart of those used to define dispersion indices (e.g., Shannon's information content). Then, contrary to conventional wisdom, we show that it may be meaningful to compare the mean values of ordinal scales *in some circumstances*, and we prove a theorem that provides the necessary and sufficient condition to this end. Finally, we extend this result to scales that are not even ordinal.

The remainder of this paper is organized as follows. Section 2 introduces Measurement Theory, where we describe two irregular scale types, in addition to regular scales. Section 3 describes a proposal for characterizing and defining concentration indices for nominal scales. Section 4 characterizes the cases in which arithmetic and even geometric means may be used for ordinal scale, while Section 5 generalizes these results to the irregular scales introduced in Section 2. Section 6 summarizes the results presented in the paper and outlines future work in this field.

2. Measurement Theory

We first describe the basic concepts of Measurement Theory used in the remainder of the paper (Section 2.1). Then, we discuss some of the usual regular scale types (Section 2.2), and we introduce two irregular scale types (Section 2.3) that may be found in software measurement.

Proceedings of the 10th International Symposium on Software Metrics (METRICS'04)
1530-1435/04 $ 20.00 IEEE

2.1. Basics of Measurement Theory

Measurement Theory [6, 10] separates the "intuitive," empirical knowledge on a specified attribute of a specified set of entities, captured via the so-called Empirical Relational System (Definition 1), from the "quantitative," numerical knowledge about the attribute, captured via the so-called Numerical Relational System (Definition 2).

Definition 1 (Empirical Relational System) *Given an attribute, let*

- E *denote the set of entities for which we would like to measure the attribute*
- $R_1, \ldots, R_y$ *denote* y *empirical relations capturing our intuitive knowledge on the attribute: each* R_i *has an arity* n_i*, so* $R_i \subseteq E^{n_i}$*; we write* $(e_1, \ldots, e_{n_i}) \in R_i$ *to denote that tuple* $(e_1, \ldots, e_{n_i})$ *is in relation* R_i*; if* R_i *is a binary relation, we use the infix notation* $e_1 R_i e_2$
- $o_1, \ldots, o_z$ *denote* z *empirical binary operations on the entities that describe how the combination of two entities yields another entity, i.e.,* $o_j : E \times E \rightarrow E$*; we use an infix notation, e.g.,* $e_3 = e_1 o_j e_2$.

An Empirical Relational System is is an ordered tuple $ERS = (E, R_1, \ldots, R_y, o_1, \ldots, o_z)$.

For example, to study the control-flow complexity (attribute) of program segments (set of entities E), as a possible empirical binary relation we may use $less\_complex\_than \subseteq E \times E$, i.e., $e_1 less\_complex\_than e_2$ represents the fact that e_1 is less complex than e_2. An operation may be concatenation, i.e., $e_3 = e_1; e_2$. The relations R_i need not be "complete" in any way. For instance, suppose that R_i is a binary relation. Given two entities e_1, e_2, we may have $\neg(e_1 R_i e_2) \wedge \neg(e_2 R_i e_1)$.

The Empirical Relational System does not make use of any kind of measurement values, which are introduced by the Numerical Relational System.

Definition 2 (Numerical Relational System) *Given an attribute, let*

- V *denote the set of values with which we would like to measure the attribute*
- $S_1, \ldots, S_y$ *denote* y *relations on the values: each* S_i *has the same arity* n_i *of* R_i
- $\bullet_1, \ldots, \bullet_z$ *denote* z *numerical binary operations on the values, so each* $\bullet_j$ *has the form* $\bullet_j : V \times V \rightarrow V$*; we use an infix notation, e.g.,* $v_3 = v_1 \bullet_j v_2$.

A Numerical Relational System is an ordered tuple $NRS = (V, S_1, \ldots, S_y, \bullet_1, \ldots, \bullet_z)$.

We have chosen to represent V as a set of "values" and not necessarily numbers for greater generality and because in some cases numbers are not really needed (e.g., for nominal or ordinal measures as described below). In our segment complexity example, $V = Re_{0+}$ may be the set of nonnegative real numbers, a binary relation may be "$>$", and a binary operation–for instance–may be '+'.

The Numerical Relational System in itself does not provide any information about the entities and the attribute. The Empirical Relational System and the Numerical Relational System are linked by a measure (Definition 3), which associates entities and values.

Definition 3 (Measure) *A function* $m : E \rightarrow V$ *is said to be a measure.*

Not all measures are sensible ones, since any $m \in V^E$ is a measure. Given program segments e_1, e_2, e_3 such that $e_1 less\_complex\_than e_2$ and $e_2 less\_complex\_than e_3$, a measure m may be such that $m(e_1) < m(e_2)$ and $m(e_3) < m(e_2)$. A sensible measure must be consistent with the empirical knowledge about the attribute, as follows.

Definition 4 (Representation Condition) *A measure must satisfy the two conditions*

$$\begin{aligned} &\forall i \in 1 \ldots n, \forall (e_1, \ldots, e_{n_i}) \in E^{n_i} \\ &(e_1, \ldots, e_{n_i}) \in R_i \Leftrightarrow (m(e_1), \ldots, m(e_{n_i})) \in S_i \qquad (1) \\ &\forall j \in 1 \ldots m, \forall (e_1, e_2) \in E \times E \\ &(m(e_1 o_j e_2) = m(e_1) \bullet_j m(e_2)) \qquad (2) \end{aligned}$$

In our segment complexity example, the Representation Condition states that $e_1 less\_complex\_than e_2 \Leftrightarrow m(e_1) < m(e_2)$ and $(m(e_1; e_2) = m(e_1) + m(e_2)$. This leads to the concept of scale (Definition 5).

Definition 5 (Scale) *A scale is a triple* (ERS, NRS, m)*, where* ERS *is an Empirical Relational System,* NRS *is a Numerical Relational System, and* m *is a measure that satisfies the Representation Condition.*

The definition of scale restricts sensible measures to be a subset $M(ERS, NRS) \subseteq V^E$ of the set of possible measures. In what follows, we assume that measures satisfy the Representation Condition, so we use the terms "scale" and "measure" interchangeably. It is well-known that $M(ERS, NRS)$ is in general a set, i.e., given ERS and NRS, more than one legitimate measure may be built. As a consequence, the actual values obtained via measures have no actual information content for scales (with the exception of so-called absolute scales [6, 10], which we do not address in this paper). For example, the fact that an object weighs 2 does not convey any information. The value kg 2 is certainly more informative, but (1) it conveys the same information as g 2000, and (2) it is actually a statement that

involves the weight of the object and the weight of a reference object that weighs 1 kg. Thus, the value itself has little interest. Only what is invariant across scales is of interest, e.g., the fact that the ratio between the weights of two object is the same regardless of the scale used. Invariant properties of scales are called meaningful statements and provide the real information content of a scale.

Definition 6 (Meaningful Statement) *A statement* $S(m)$ *that depends on a measure* m *is* meaningful *if its truth value does not change across all scales, i.e.,* $\forall m \in M(ERS, NRS)(S(m)) \vee \forall m \in M(ERS, NRS)(\neg S(m))$.

Thus, it is meaningful to say that an object is twice as heavy as another. Instead, suppose we can just tell if a software failure is more critical than another, so we can classify failures with a 5-value criticality measure m'_{cr} with values $Range(m'_{cr}) = \{1, 2, 3, 4, 5\}$, where 1 is least severe and 5 is most severe. It it meaningless to say that criticality 2 failures are twice as severe as a criticality 1 failure, as the truth value of this statement depends on the specific choice of values. If we choose another scale m''_{cr} with values $Range(m''_{cr}) = \{7, 19, 34, 981, 4365\}$, the truth value of the statement changes.

2.2. Regular scale types

It *may* be possible to map one scale into another. In the weight example, any proportional transformation provides a legitimate scale with a different weight unit. In the failure criticality example, we can map one scale into another by applying a monotonically increasing transformation. This leads to the definition of admissible transformation.

Definition 7 (Admissible Transformation) *Given a scale* (ERS, NRS, m)*, the transformation of scale* f *is admissible if* $m' = f \circ m$ *(i.e.,* m' *is the composition of* f *and* m*) and* (ERS, NRS, m') *is a scale.*

The set of admissible transformations depends on the kind of invariant statements that need to be preserved. The scales for which admissible transformations exist are called *regular* and may be classified according to the set of admissible transformations they can undergo.

Nominal scales. The values of these scales are labels–not necessarily numbers–for equivalence classes (which we call "categories" from this point on) in which the entities are partitioned, with no notion of order among the categories. The invariant property states that the actual labels used do not matter, as long as different labels are used for different categories. Formally, $\forall e_1, e_2 \in E$

$$\forall m \in M(ERS, NRS)(m(e_1) = m(e_2)) \vee \\ \forall m \in M(ERS, NRS)(m(e_1) \neq m(e_2))$$

Thus, the set of categories is the information that needs to be preserved, so nominal scales can be transformed into other nominal scales via one-to-one transformations.

Ordinal scales. The values of these scales are labels–not necessarily numbers–for categories in which the entities are classified, with a *total order* across values. The invariant property states that the actual labels used do not matter, as long as the order of the values that label different categories is preserved. Formally, $\forall e_1, e_2 \in E$

$$\forall m \in M(ERS, NRS)(m(e_1) > m(e_2)) \vee \\ \forall m \in M(ERS, NRS)(m(e_1) = m(e_2)) \vee \\ \forall m \in M(ERS, NRS)(m(e_1) < m(e_2))$$

Thus, the ordering across the categories needs to be preserved, so ordinal scales can be transformed into other scales via strictly monotonic transformations.

Interval scales. Each entity is associated with a numerical value. The invariant property states that the actual values used do not matter, as long as the ratios between all pairs of differences between values are preserved. Formally, by denoting the set of real number by Re, $\forall e_1, e_2, e_3, e_4 \in E$

$$\exists k_1, k_2 \in Re, \forall m \in M(ERS, NRS) \\ k_1(m(e_1) - m(e_2)) = k_2(m(e_3) - m(e_4))$$

An interval scale m' can be transformed into another interval scale m'' only via linear transformations $m'' = am' + b$, with $a > 0$, i.e., we can change the origin of the values (by changing b) and the unit of measurement (by changing a).

Ratio scales. Each entity is associated with a numerical value. The invariant property states that the actual values used do not matter, as long as the ratios between the all pairs of values are preserved. Formally, $\forall e_1, e_2 \in E$

$$\exists k_1, k_2 \in Re, \forall m \in M(ERS, NRS) \\ k_1 m(e_1) = k_2 m(e_2)$$

A ratio scale m' can be mapped into another ratio scale m'' only via proportional transformations $m'' = am'$, with $a > 0$, i.e., we can change the measurement unit by changing a.

The description of scale types has important practical consequences. Some mathematical operations may not be applied to measures of certain types, e.g., summing numerical values of ordinal or even interval measures. Suppose that m' is an interval measure. If statement $m'(e_3) = m'(e_1) + m'(e_2)$ was meaningful, it would have the same truth value as $m''(e_3) = m''(e_1) + m''(e_2)$ under the transformation $m'' = am' + b$. By replacing m'' with $am' + b$, we have $am'(e_3) + b = am'(e_1) + b + am'(e_2) + b$, from which we obtain $b = 0$, which is a *specific* value of b.

Also, based on the type of a scale, it is commonly believed that different indices of central tendency should be used, i.e., the mode for nominal scales, the mode and the median for ordinal scales, the mode, the median, and the

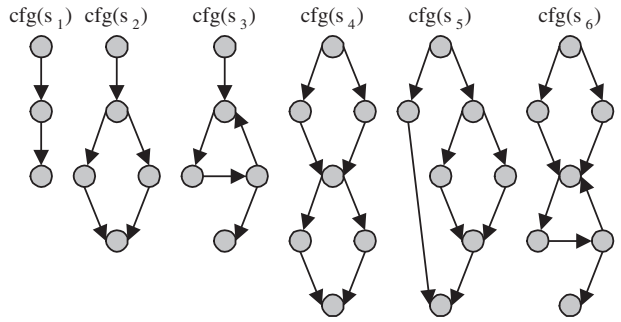

Figure 1. Control Flow Graphs.

arithmetic mean for interval scales, and the geometric mean as well for ratio and absolute scales. We elaborate on this in Sections 4 and 5.

2.3. Two irregular scale types

Meaningful statements are probably the central aspect of Measurement Theory. The meaningful statements associated with the scale types of Section 2.2 have one common characteristic, i.e., a universal quantification on all the entities that appear in them. For instance, the invariant property for interval scales begins with "$\forall e_1, e_2, e_3, e_4 \in E \ldots$" However, these are *specific* kinds of invariant statements. Suppose that we know that we believe that program segment e_{min} is the least complex program segment. The corresponding invariant statement states that $\forall e \neq e_{min} \in E, \forall m \in M(ERS, NRS)(m(e) > m(e_{min}))$. In this statement, entity e is universally quantified, but e_{min} is not.

On a related note, an admissible transformation may not always exists [10]. For instance, in our program complexity example, let $ERS = (E, R)$, with $E = \{e_1, e_2, e_3\}$ and $R = \{(e_1, e_2), (e_1, e_3)\}$, and let $NRS = (Re, \ll)$, where $\ll$ is a binary relation on Re such that $x \ll y$ if and only if $x < y - 1$. Let m' be such that $m'(e_1) = 0$, $m'(e_2) = 2$, $m'(e_3) = 2$, and m'' such that $m''(e_1) = 0$, $m''(e_2) = 2.1$, $m''(e_3) = 2$. Both m' and m'' are legitimate scales, but there is no admissible transformation that transforms m' into m''. Scales of this kind are called *irregular*. The invariant statement preserved in these scales is $m(e_1) \ll m(e_2) \wedge m(e_1) \ll m(e_3)$. Now, suppose we chose a Numerical Relational System like $NRS = (Re, <)$, and suppose that m'' is a scale. The Representation Condition (Definition 4) requires that also $e_3 Re_2$, which is not in our empirical intuition on the ordering of the segments. So, it would be impossible to have a scale only because we are using $NRS = (Re, <)$, but not because of some intrinsic problem in $ERS = (E, R)$.

The situation in which we are not able to provide an empirical total order among entities is not infrequent in software measurement [4]. For instance, suppose we have six program segments $s_1, \ldots, s_6$ that we model with the six control flow graphs $cfg(s_i)$ of Figure 1, and suppose we come up with the order represented in Figure 2, where

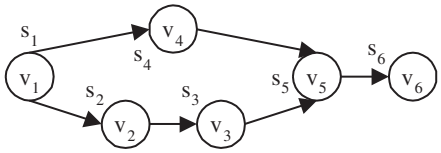

Figure 2. Hierarchy for the graphs of Figure 1.

- each node represents a segment s_i
- the annotation inside a node represents the value associated by some complexity measure m with the control flow graph represented by the node, e.g., $m(s_4) = v_4$
- an arc from s_i to s_j represents the fact that $s_i less\_complex\_than s_j$ and $m(s_i) < m(s_j)$.

For instance, we can say that $s_1 less\_complex\_than s_2$ and $v_1 < v_b$, but we are unable to order s_2 and s_4, and s_3 and s_4, on the entities' side; v_2 and v_4, and v_3 and v_4, on the values' side. Thus, in several cases, the best kind of relation that we can establish among the entities is a *hierarchy*.

Definition 8 (Hierarchy) *Let X be a set and $Q \subseteq X \times X$ a relation. The pair (X, Q) is a hierarchy if and only if Q is*

- *asymmetric, i.e.,* $\forall x, y \in X$, $xQy \Rightarrow \neg yQx$
- *transitive, i.e.,* $\forall x, y, z \in Q$, $xQy \wedge yQz \Rightarrow xQz$

A hierarchy can be modeled by a Directed Acyclic Graph (DAG), like in Figure 2, where it is shown that we have a hierarchy for both the entities (with $X = E$ and $RX = R$) and the values (with $X = V$ and $RX = S$). For completeness only, we now explicitly represent both the Empirical and the Numerical Relational Systems:

- $ERS = (E, R)$, where E is the set of all possible program segments that produce control flow graphs like the ones in Figure 1, and

$$\begin{aligned} R = \{(s_x, s_y) \in E^2 | \forall s_1, s_2, s_3, s_4, s_5, s_6 \in E \\ (s_1 < s_2 \wedge s_2 < s_3 \wedge s_1 < s_4 \wedge \\ s_3 < s_5 \wedge s_4 < s_5 \wedge s_5 < s_6) \wedge \\ (s_x = s_1 \wedge s_y \in \{s_2, s_3, s_4, s_5, s_6\} \vee \\ s_x = s_2 \wedge s_y \in \{s_3, s_4, s_5, s_6\} \vee \\ s_x = s_3 \wedge s_y \in \{s_4, s_5, s_6\} \vee \\ s_x = s_4 \wedge s_y \in \{s_2, s_3, s_5, s_6\} \vee \\ s_x = s_5 \wedge s_y \in \{s_6\})\} \end{aligned} \quad (3)$$

- $NRS = (V, S)$, where $V = Re_{0+}$ and

$$S = \{(x,y) \in Re_{0+}^2 | \forall v_1, v_2, v_3, v_4, v_5, v_6 \in Re_{0+} \\ (v_1 < v_2 \wedge v_2 < v_3 \wedge v_1 < v_4 \wedge \\ v_3 < v_5 \wedge v_4 < v_5 \wedge v_5 < v_6) \wedge \\ (x = v_1 \wedge y \in \{v_2, v_3, v_4, v_5, v_6\} \vee \\ x = v_2 \wedge y \in \{v_3, v_4, v_5, v_6\} \vee \\ x = v_3 \wedge y \in \{v_4, v_5, v_6\} \vee \\ x = v_4 \wedge y \in \{v_2, v_3, v_5, v_6\} \vee \\ x = v_5 \wedge y \in \{v_6\})\} \quad (4)$$

Given the similarities in the structure of R and S, it is not difficult to show that a set of scales can be built. For all scales m, we have $m(e_1) < m(e_2)$, for instance, but there exist three scales m', m'', m''' such that $m'(e_2) < m'(e_4)$, $m''(e_2) = m''(e_4)$, $m'''(e_2) > m'''(e_4)$.

At any rate, given an entity $\hat{e}$, there in general is a subset of entities $\bar{e} \in E$ such that $\forall m \in M(ERS, NRS), m(\hat{e}) = m(\bar{e})$. For instance, if we assess program segment complexity based only on control flow graphs, for any two program segments $\hat{s}$ and $\bar{s}$ modeled by the same control flow graph we have $m(\hat{s}) = m(\bar{s})$ for all measures m. Thus, the set of entities is partitioned in categories: for instance, each node of the hierarchy of Figure 2 actually represents an entire subset of entities, which is the subset of entities that are always associated with a common value by all measures, and the specific entities denoted in Figure 2 are only representative ones for their categories. Also, it is possible to establish (asymmetric and transitive) order relationships between some of the categories, but not all of them. Each category is associated with a value of a measure in a way that is consistent with the order relationships between the entities. K will represent the set of categories into which the set of entities E is partitioned.

We now deal with an important particular kind of hierarchies, i.e., strict weak orders [1].

Definition 9 (Strict Weak Order) *Let X be a set and $Q \subseteq X \times X$ a relation. The pair (X, Q) is a strict weak order if and only if Q is asymmetric, transitive, and the indifference relation IND is an equivalence relation, where IND is the relation among those elements that are not ordered, i.e., $\forall x, y \in X$, we have $xINDy \Leftrightarrow (\neg(xQy) \wedge \neg(yQx))$.*

In a strict weak order, the categories identified as explained above for hierarchies are organized in equivalence classes, and it can be shown that the equivalence classes of categories are indeed totally ordered. We denote by $IND(lev)$ the set of level lev categories, and by $PREC(lev) = \bigcup_{h \in 1...lev-1} IND(lev)$ the set of categories belonging to the levels before lev, with $PREC(1) = \oslash$. The DAG in Figure 3, with the four equivalence classes of categories denoted as levels 1 through 4, represents a strict weak order.

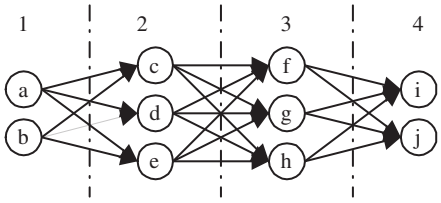

Figure 3. A strict weak order.

Like in the case of measures based on general hierarchies, we use an Empirical Relational System $ERS = (E, R)$ and a Numerical Relational System $NRS = (Re, S)$ that are both strict weak orders. R and S can be built in much the same way as we did in formulae (3) and (4). This is different from using a Numerical Relational System that is a total order, i.e., $NRS = (Re, <)$, as is usually done. Note that the building of an ordinal scale that links $ERS = (E, R)$ to $NRS = (Re, <)$ requires that $ERS = (E, R)$ be a strict weak order [4]. However, because of our choice of a strict weak order for the Numerical Relational System, we are dealing with a different type of scales than ordinal ones.

At any rate, what matters is the use that can be made of scales. In Section 5, we show that the comparison of means may be meaningful even for these irregular scales.

3. Concentration for nominal measures

In software measurement, nominal measures are used for several attributes, including: programming language used to write a software module; type of a system (e.g., real-time, embedded, business application); type of a fault in software code (e.g., dangling pointer, memory overflow, uninitialized variable); type of an error made when coding a module, (e.g., logical, omission, clerical); phase in which the error was made; etc. The measures for all of these attributes convey useful information that can be used before, during, and after development. For instance, based on the type of software system to be developed, different levels of requirements are set before development: a real-time system will probably have higher reliability requirements than a non-critical web application, but also lower usability requirements. The programming language of a module is a fundamental piece of information when assessing its size after development, as the bare knowledge of the number of lines of code is certainly not enough.

The mode is the main index for the "central tendency" of the distribution of a nominal measure. It is one's "best

1 Here we do not provide the "traditional" definition of strict weak order, which is based the concept of "negatively transitive" relation, but we characterize strict weak orders in an equivalent way, based on a necessary and sufficient condition [10]

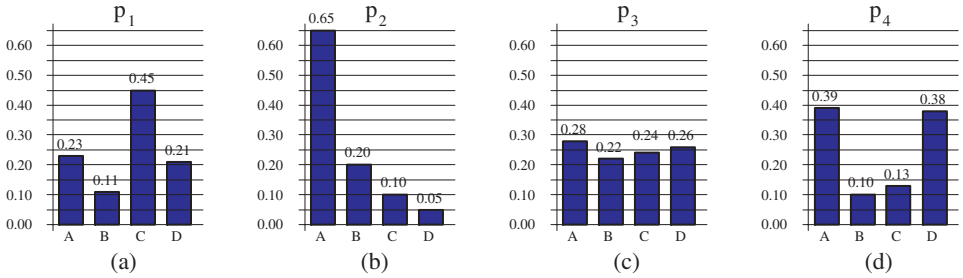

Figure 4. Four fault distributions.

guess" if no other information is available on a set of entities. For instance, take the example in Figure 4(a) where a set of software faults are categorized depending on their type (A - D). The best guess on the type of a newly found fault would be 'C', i.e., the mode of the distribution. Also, if Figure 4(b) contains the distribution of fault types of Java modules, the best guess, *conditional* on the fact that the defect was found in a Java module would be 'A', which still is the mode of the conditional distribution of Figure 4(b).

However, the mode does not provide information about the concentration of the distribution, which may influence our confidence in the knowledge we extract from a distribution. If, instead of the distribution in Figure 4(b), we had the distribution in Figure 4(c), our best guess would still be 'A', but we would probably be less confident in our answer.

On a different, but related note, suppose that we have the distribution in Figure 4(d), where two types of defects are much more frequent than the others. We may reply that we believe that the defect is most likely to be of type 'A' or 'D', but we would be more confident about our reply than in the case of Figure 4(c). This kind of information can also be used in decision making. Suppose that we would like to subcontract the development of a component of an application. We can have a panel of experts vote to decide which contractor should get the job. Suppose we obtain a distribution of votes like the one in Figure 4(c). We would probably be quite uncomfortable when making the final decision. With a distribution like the one in Figure 4(d), we may be more confident about our decision if we choose the contractor that corresponds to the mode of the distribution or we may narrow the alternatives to 'A' and 'D' and then ask our experts to decide between only those two. This makes the decision process faster and smoother.

A few indices have been proposed in the literature mainly to quantify the *dispersion* of the distribution of a nominal measure. Information content H [3] is probably the best known dispersion index, since it has provided the foundations for Information Theory

$$H = -\sum_{v \in V} p(v) \log_2 p(v) \tag{5}$$

We have used $p(v)$ to denote the probability that $m(e) = v$, as the different categories can be identified by means of the different values of the measure. Other entropy measures are the Renyi entropy measure [9]

$$H_\alpha^R = \frac{1}{1-\alpha} \log_2 \sum_{v \in V} p^\alpha(v) \tag{6}$$

with $\alpha > 0$ and the Tsallis entropy measure [12]

$$H_\alpha^T = c \frac{\sum_{v \in V} p^\alpha(v) - 1}{1-\alpha} \tag{7}$$

with $\alpha > 0$ and $c > 0$, which is usually assumed $c = 1$. Gini's dispersion index $G = 1 - \sum_{v \in V} p^2(v)$ is a special case with $\alpha = 2$. Renyi's and Tsallis' entropy measures tend to H as α tends to 0. It is important to note that these indices were defined based on axioms, which were used to substantiate their definition formulae.

These indices have been applied for instance in the building of decision or classification trees. Classification trees use the values of the measures of independent attributes for grouping data points into the categories of a dependent attribute [8, 11]. A classification tree is built by recursively splitting a subset of data points into subsets, each of which is identified by specific values of the nominal measures for the independent attributes. At the beginning of the tree building, the initial "subset" of data points is the entire data set. Continuous independent measures are discretized (with a partial exception in Classification and Regression Trees [2, 1]) and the discretized, nominal version is used. At each splitting step, for a subset of data points, several different measures for the independent attributes may be chosen. The idea is to choose the one that minimizes a combination of the dispersions of the nominal measure used for the dependent attribute measured on the newly created subsets. A subset is not split any further if the dispersion of the distribution of the measure used for the dependent variable is below some specified threshold, or if the subset is too small. H is probably the dispersion measure that has been used to this end, but there is not necessary reason for this.

Here, we focus on *concentration* indices. Concentration may be seen as the "flip side of the coin" with respect to

dispersion. We choose to focus on concentration and not on dispersion because, as we show later in this section, a family of concentration indices may be easily built that satisfy a set of axioms that are the counterpart of the axioms used for dispersion. In what follows, $C(p)$ denotes a concentration index based on the probability distribution p.

Axiom 1 (Symmetry) *$C(p)$ is a symmetric function of the values $p(v)$, i.e., if the values $r(v)$ are a permutation of the values $p(v)$, we have $C(p) = C(r)$.*

Axiom 2 (Impossible events) *An impossible event v, i.e., one with $p(v) = 0$ does not contribute to the concentration index. If we denote by r the probability distribution on the set of events $V - \{v\}$ such that $\forall x \in V - \{v\}, r(x) = p(x)$, we must have $C(p) = C(r)$.*

Axiom 2 makes it impossible to change the value of $C(p)$ by adding fictitious events that cannot possibly occur.

Axiom 3 (Pairwise concentration) *Given a probability distribution p, let r be a probability distribution that coincides with p except for the values of the probabilities associated with two events x and y, i.e., $\forall v \neq x, v \neq y, p(v) = r(v)$. So, $p(x) + p(y) = r(x) + r(y) = P$.*

We have $C(p) \geq C(r)$ if and only if the values $p(x)$ and $p(y)$ are not closer to each other than the values $r(x)$ and $r(y)$, i.e., $|p(x)-p(y)| \geq |r(x)-r(y)|$, which can be rewritten as $|p(x) - P/2| \geq |r(x) - P/2|$.

In other words, the conditional distribution $\{p(x)/P, p(y)/P\}$ is not "closer" to being equiprobable than the conditional distribution $\{r(x)/P, r(y)/P\}$.

From Axioms 1 - 3, we can derive the following two axioms, on the minimum and the maximum values for $C(p)$.

Derived Axiom 1 (Minimum value) *$C(p)$ is minimum when all the events of the probability distribution p are equiprobable, i.e., $\forall v \in V p(v) = 1/|V|$.*

Derived Axiom 2 (Maximum value) *$C(p)$ is maximum when one event is certain, i.e., $\exists v \in V p(v) = 1$.*

Based on Axiom 2, the maximum value of $C(p)$ does not depend on the number of possible events, so, for a given functional form for $C(p)$, it is the same for all distributions.

Function $C(p)$ is not necessarily convex or continuous. However, it can be shown based on Axiom 3 that $C(p)$ is limited for all probability distributions p, except the distributions with one certain event, since, for any other distribution r, it is always possible to find a different distribution p such that $C(p) \geq C(r)$. At any rate, for better tractability, we add the following axiom.

Axiom 4 (Continuity) *$C(p)$ is a continuous function of the values $p(v)$ for $p(v) < 1$.*

Based on these axioms, a number of concentration indices can be defined as a function of the distribution p. Here, we focus on a family of indices. The basic idea is that a discrete distribution p is more concentrated if, when n random drawings are carried out, there is a higher probability that the same value is selected. The probability of selecting the same value n times is $P(p, n) = \sum_{v \in V} p^n(v)$, as it is the sum of the probabilities of obtaining n times each value $v \in V$. Different concentration indices can be defined by weighting these probabilities $P(p, n)$ with a function $w(n) \geq 0$ for each n, and $w(n) > 0$ for at least one n. So, we obtain the function (N denotes the positive integers)

$$C(p) = \sum_{n \in N} w(n) \sum_{v \in V} p^n(v) \tag{8}$$

We assume that the series of formula (8) converges, with the possible exception of the probability distribution with one certain event. Formula (8) shows that $C(p)$ is a symmetrical function of the $p(v)$'s, so Axiom 1 is satisfied. Impossible events cannot provide any contribution to the value of $C(p)$, so Axiom 2 is satisfied. In addition, Axiom 3 is satisfied as well. With the same p and r as in Axiom 3, take a value n and compute the two probabilities $P(p, n) = \sum_{v \in V - \{x,y\}} p^n(v) + p^n(x) + (P - p(x))^n$ and $P(r, n) = \sum_{v \in V - \{x,y\}} r^n(v) + r^n(x) + (P - r(x))^n$. So, $P(p, n) \geq P(r, n)$ means $p^n(x) + (P - p(x))^n \geq r^n(x) + (P - r(x))^n$. We can now study the function $h^n + (P - h)^n$. Through derivations and simple mathematical computations, we can find that this function (I) attains its minimum for $h = P/2$, (II) is symmetrical with respect to $h = P/2$, and (III) has a positive second derivative. So, we can conclude that $h^n + (P - h)^n$ is greater for values that are farther away from $P/2$, i.e., $p^n(x) + (P - p(x))^n \geq r^n(x) + (P - r(x))^n$ if and only if $|p(x) - P/2| \geq |r(x) - P/2|$. Since all the weights $w(n)$ of the probabilities $P(p, n)$ are nonnegative, we obtain that $C(p) \geq C(r)$. We also assume that Axiom 4 is satisfied, as that is not guaranteed by formula (8).

Under these hypotheses, the series of formula (8) converges absolutely, so we can rewrite it as

$$C(p) = \sum_{v \in V} \sum_{n \in N} w(n) p^n(v) \tag{9}$$

Formula (9) allows the computation of $C(p)$ as a sum of Z-transforms. The Z-transform $Z(w; z)$ of a discrete function $w(n)$ is defined as

$$Z(w; z) = \sum_{n \in 0 \ldots \infty} \frac{w(n)}{z^n}$$

Thus, we can use the existing results for Z-transforms by

rewriting formula (9) as

$$C(p) = \sum_{v \in V} (Z(w; \frac{1}{p(v)}) - w(0)) = \sum_{v \in V} Z(w; \frac{1}{p(v)}) - |V| w(0)$$

Now, we show how formula (9) can be used to build a number of concentration indices.

Herfindahl-Hirschman index [7]. The weighting function is $w(2) = 1$ and $\forall n \in N - \{2\}, w(n) = 0$. Thus,

$$C_H(p) = \sum_{v \in V} p^2(v)$$

$C_H(p)$ ranges between $C_{Hmin}(p) = \frac{1}{|V|}$ and $C_{HMax}(p) = 1$. Note that other similar indices may be defined as $C_H(p; j) = \sum_{v \in V} p^j(v)$ with $j \geq 2$, since the choice $j = 2$ is somewhat arbitrary.

Constant weighting function. The weighting function is $w(n) = w, \forall n \in N$. We have

$$C_C(p) = w \sum_{v \in V} \sum_{n \in N} p^n(v) = w \sum_{v \in V} (\frac{1}{1 - p(v)} - 1) = w \sum_{v \in V} \frac{p(v)}{q(v)}$$

where $q(v) = 1 - p(v)$. So, the concentration index is the sum of the odds of each event. $C_C(p)$ ranges between $C_{Cmin}(p) = w\frac{|V|}{|V|-1}$ and ∞, as $C(p)$ is defined except when one event, say v, is certain, since $q(v) = 0$.

Linear weighting function. The weighting function is $w(n) = wn, \forall n \in N$. Through computations, we have

$$C_L(p) = w \sum_{v \in V} \frac{p(v)}{q^2(v)}$$

$C_L(p)$ ranges between $C_{Cmin}(p) = w(\frac{|V|}{|V|-1})^2$ and ∞.

Information-like concentration. Formula

$$C_I(p) = -\sum_{v \in V} p(v) log_2 q(v) = -\sum_{v \in V} p(v) log_2 (1 - p(v)) \quad (10)$$

defines a concentration index:

$$C_I(p) = -\sum_{v \in V} p(v) \sum_{n \in N} (-1)^{n+1} \frac{(-p(v))^n}{n} = \sum_{v \in V} \sum_{n \in N} \frac{p^{n+1}(v)}{n} = \sum_{v \in V} \sum_{n \in 2 \ldots \infty} \frac{p^n(v)}{n-1}$$

So, $w(n) = \frac{1}{n-1} > 0$ if $n > 1$ and $w(1) = 0$. $C_I(p)$ ranges between $C_{Imin} = -log_2 \frac{|V|-1}{|V|}$ and ∞.

	p_1	p_2	p_3	p_4	min	Max
C_H	0.046	0.180	0.0164	0.044	0.20	1
C_C	1.506	2.271	1.338	1.513	1.333	∞
C_L	2.351	5.798	1.792	2.332	1.777	∞
C_I	0.565	1.068	0.4195	0.582	0.415	∞
C_P	0.433	0.476	0.4184	0.435	0.418	0.632

Table 1. Examples of concentration values.

Poisson weighting function. The weighting function is $w(n) = \lambda^n / n! e^{-\lambda}$. We have

$$C_P(p) = \sum_{v \in V} \sum_{n \in N} \frac{\lambda^n}{n!} e^{-\lambda} p^n(v) = e^{-\lambda} \sum_{v \in V} \sum_{n \in N} \frac{(\lambda p(v))^n}{n!} = e^{-\lambda} \sum_{v \in V} (e^{\lambda p(v)} - 1) = \sum_{v \in V} e^{-\lambda q(v)} - e^{-\lambda} |V|$$

$C_P(p)$ ranges between $C_{Pmin} = |V| e^{-\lambda} (e^{\frac{\lambda}{|V|}} - 1)$ and $C_{PMax} = 1 - e^{-\lambda}$. In this case, the number of selections n may also be viewed as a random variable distributed according to a Poisson distribution. Other discrete distributions (e.g., binomial, geometric) may be used as well. Table 1 contains the values of the above indices for the distributions $p_1 - p_4$ in Figure 4 (e.g., $C_C(p_3) = 1.338$). We used $w = 1$ for both C_C and C_L, and $\lambda = 1$ for C_P. The last two columns of Table 1 contain the minimum and the maximum value for each index. As all distributions in Figure 4 have four values, the minimum and maximum value of a concentration index is the same across all distributions.

Table 1 shows that the indices capture concentration differently. For instance, $C_H(p_1) > C_H(p_4)$, but $C_C(p_1) < C_C(p_4)$. This is not unexpected, as formula (8) shows that $C(p)$ can be built as a weighted sum of the probabilities $P(p, n)$, in a similar way to the definition of other indices as weighted sums, where the weights may be different in different applications. The existence of a number of concentration indices allows users to choose the one that best fits their needs. Also, though the lack of a finite range for some indices may seem to be a problem, we note that the existence of a finite range is no condition for a better interpretability of an index, as it is no guarantee that the distribution of the values of the index is uniform in practice.

Axiom 3 characterizes concentration from dispersion indices. It can be shown that a dispersion index $D(p)$ satisfies Axioms 1, 2, 4, and an axiom that closely mirrors Axiom 3, with the *only* difference of replacing $C(p) \geq C(r)$ with $D(p) \leq D(r)$. More specific axioms have also been introduced for the definition of specific dispersion measures. For instance, the following formula describes an addition axiom for the definition of Shannon's H (other dispersion in-

dices have different addition rules)

$$H(p(v), p(x), p(y), \ldots, p(z)) = H(p(v) + p(x), p(y), \ldots, p(z)) + (p(v) + p(x))H(\frac{p(v)}{p(v) + p(x)}, \frac{p(x)}{p(v) + p(x)})$$

At any rate, concentration indices may be used instead of dispersion indices to extract information on the spread of a distribution. Thus, they may also be used in the building of decision trees, by maximizing a combination of the concentrations of the distributions of the nominal measure used for the dependent attribute measured on the newly created subsets at each splitting of the data set.

4. Means for ordinal measures

It is often said that the mean is not a sensible index of central tendency for ordinal scales because its use would lead to meaningless statements. For instance, take a failure criticality measures m, e.g., m''_{cr} or m''_{cr} of Section 2.1. Suppose that we have obtained two sets of failures on two different programs (or on two versions of a program), with frequencies $p(i)$ and $r(i)$ for each failure criticality category, where each category is indexed in increasing order of criticality by an integer number. Comparing the two arithmetic means leads to the following statement

$$E_p(m) = \sum_{i \in 1 \ldots |V|} p(i)m(i) > E_r(m) = \sum_{i \in 1 \ldots |V|} r(i)m(i) \quad (11)$$

which may be true for some ordinal measures and false for others, as simple examples can show.

Surprisingly, even for ordinal measures, statement (11) may be meaningful, i.e., in some cases, statement (11) is always true or always false independent of the ordinal measure chosen. We characterize the general case in which this happens with Theorem 1. Before stating and proving the theorem, we discuss a few cases via the eight frequency distributions in Table 2, starting from an obvious, extreme one. In Table 2, for instance, the frequency of criticality #4 failures according to distribution p_3 is $p_3(4) = 0.25$. We denote the mean of distribution p_k of Table 2 as E_k.

Let us compare the means obtained with distributions p_1 and p_2. Statement (11) becomes $E_1(m) = m(5) > E_2(m) = m(1)$, so it is obviously true for all possible ordinal measures, and thus it is meaningful. Let us now compare the means obtained with distributions p_3 and p_4: statement (11) becomes $E_3(m) = 0.1m(1) + 0.15m(2) + 0.2m(3) + 0.25m(4) + 0.3m(5) > E_4(m) = 0.11m(1) + 0.17m(2) + 0.18m(3) + 0.24m(4) + 0.3m(5)$. So, the question is: is this inequality true (or false) for all possible ordinal measures m, i.e., is it meaningful? Also, is statement $E_4(m) > E_5(m)$ is meaningful? Let us now consider the

	1	2	3	4	5
p_1	0	0	0	0	1
p_2	1	0	0	0	0
p_3	0.1	0.15	0.2	0.25	0.3
p_4	0.11	0.17	0.18	0.24	0.3
p_5	0.08	0.17	0.18	0.27	0.3
p_6	0.08	0.18	0.18	0.26	0.3
p_7	0.17	0.25	0.3	0.18	0.1
p_8	0.15	0.20	0.33	0.22	0.1

Table 2. Failure frequency distributions.

two comparisons $E_3(m) > E_5(m)$ and $E_3(m) > E_6(m)$. Distributions p_5 and p_6 are quite similar if compared to distribution p_3: both p_5 and p_6 have higher frequencies for categories #2 and #4 than distribution p_3; lower frequencies for categories #1 and #3; and the same frequency for category #5. Does this allow us to say that both statements $E_3(m) > E_5(m)$ and $E_3(m) > E_6(m)$ are meaningful? That neither is meaningful? Note that the *median* obtained for all distributions from p_3 to p_6 is the same, i.e., #4. Does meaningfulness of the comparison depend on the value of the median in any way? So, let us take distributions p_7 and p_8, both of which have median equal to #3. Is statement $E_7(m) > E_8(m)$ meaningful? Or, is statement $E_6(m) > E_7(m)$ meaningful? Theorem 1 provides the general answer to this kind of meaningfulness questions.

Theorem 1 *Let m be an ordinal measure with $|V| > 1$ values, and let us denote its values in nondecreasing order by identifying each of the $|V|$ categories with an integer value, so $m(1) < m(2) < \ldots < m(|V|)$. Let $w(1), w(2), \ldots, w(|V|)$ be a set of real numbers (the weights) such that $\sum_{i \in 1 \ldots |V|} w(i) = 0$. We have*

$$\sum_{i \in 1 \ldots |V|} w(i)m(i) > 0 \quad (12)$$

for every possible choice *of an ordinal measure m if and only if for all $i \in 1 \ldots |V|$*

$$\sum_{j \in i \ldots |V|} w(j) \geq 0 \quad (13)$$

and for at least one $i \in 2 \ldots |V|$

$$\sum_{j \in i \ldots |V|} w(j) > 0 \quad (14)$$

Proof. The weighted sum of (12) can be rewritten via a discrete version of the integration rule by parts:
$[m(|V|) - (|V| - 1)][w(|V|)] +$
$[m(|V| - 1) - m(|V| - 2)][w(|V|) + w(|V| - 1)] +$
$[m(|V| - 2) - m(|V| - 3)][w(|V|) + w(|V| - 1) + w(|V| -$

2)] + ...
$[m(1)][w(|V|) + w(|V|-1) + w(|V|-2) + \ldots + w(1)]$
and, since $\sum_{i\in 1\ldots|V|} w(i) = 0$, we have

$$\sum_{i\in 2\ldots|V|} (\sum_{j\in i\ldots|V|} w(j))(m(i) - m(i-1)) \qquad (15)$$

The condition is *sufficient*. In the sum of (15), each term $(\sum_{j\in i\ldots|V|} w(j))(m(i) - m(i-1)) \geq 0$, since we have $(m(i) - m(i-1)) \geq 0$ by hypothesis. Also, we have $(\sum_{j\in i\ldots|V|} w(j))(m(i) - m(i-1)) > 0$ for at least one $i \in 1\ldots|V|$. So, the left-hand side of (12) is positive.

The condition is *necessary*, and we prove it by contradiction. Suppose that condition (14) does not hold, so for all $i \in 2\ldots|V|$, $\sum_{j\in i\ldots|V|} w(j) = 0$. It is immediate that all the terms in (15) become zero. So, condition (14) is necessary. Now, suppose that that condition (13) does not hold, so there exists one value $i = h$ such that $\sum_{j\in h\ldots|V|} w(j) < 0$. We now show that a measure m exists such that (12) does not hold. Suppose that there is the same difference $m(i) - m(i-1) = t$ between the measures for all consecutive pairs of categories, *except for* $m(h) - m(h-1) = u$. Expression (15) can be rewritten as

$$t \sum_{i\neq h\in 2\ldots|V|} (\sum_{j\in i\ldots|V|} w(j)) + u \sum_{j\in h\ldots|V|} w(j)))$$

Now, for any given set of weights $w(1), w(2), \ldots, w(|V|)$, we can choose $u > 0$ large enough and $t > 0$ small enough to make the expression in (15) negative. (*End of proof*) $\diamondsuit$

Theorem 1 can be applied to the comparison of the mean values of ordinal measures by writing $E_p(m) > E_r(m)$ as

$$\sum_{i\in 1\ldots|V|} (p(i) - r(i))m(i) > 0 \qquad (16)$$

The left-hand sides of (16) and (12) coincide by taking $w(i) = p(i) - r(i)$. Now, (13) can be rewritten as

$$\sum_{j\in i\ldots|V|} w(j) \geq \sum_{j\in 1\ldots|V|} w(j) = 0 \qquad (17)$$

i.e., via basic computations

$$\sum_{j\in 1\ldots i-1} w(j) \leq 0 \qquad (18)$$

Thus, based on (18), conditions (13) and (14) actually require that, for all $i \in 2\ldots|V|$

$$\sum_{j\in 1\ldots i-1} p(j) \leq \sum_{j\in 1\ldots i-1} r(j) \qquad (19)$$

with strict inequality for at least one $i \in 2\ldots|V|$.

In Table 3, P_k denotes the cumulative frequency of p_k, where $P_k(i) = \sum_{j\in 1\ldots i-1} p(j)$, i.e., the value in each cell

	1	2	3	4	5
P_1	0	0	0	0	0
P_2	0	1	1	1	1
P_3	0	0.1	0.25	0.45	0.7
P_4	0	0.11	0.28	0.46	0.7
P_5	0	0.08	0.25	0.43	0.7
P_6	0	0.08	0.26	0.44	0.7
P_7	0	0.17	0.42	0.72	0.9
P_8	0	0.15	0.35	0.68	0.9

Table 3. Cumulative failure frequencies.

of Table 3 is the sum of the values of Table 2 that appear in the *previous* columns of the same row. So, Table 3 can help provide the answers to the questions about the comparisons between the means of the distributions in Table 2. For all values i, $P_1(i) \leq P_2(i)$, and for a least one value i, $P_1(i) < P_2(i)$. This confirms the obvious result we already explained above. Likewise, for all values i, $P_3(i) \leq P_4(i)$, and for a least one value i, $P_3(i) < P_4(i)$, so $E_3(m) > E_4(m)$ is meaningful. In much the same way, $E_4(m) > E_5(m)$ and $E_3(m) > E_5(m)$ are meaningful, but $E_3(m) > E_6(m)$ is not, because $P_3(2) > P_6(2)$ and $P_3(3) < P_6(3)$. Finally, $E_7(m) < E_8(m)$ and $E_6(m) > E_7(m)$ are meaningful.

As a consequence to Theorem 1, it is possible to show that, if condition (19) holds, with strict inequality for at least one $i \in 2\ldots|V|$, and $\forall i \in 1..|V|, m(i) > 0$, then even the comparison of geometric means is meaningful, i.e.,

$$\prod_{i\in 1\ldots|V|} (m(i))^{p(i)} > \prod_{i\in 1\ldots|V|} (m(i))^{r(i)} \qquad (20)$$

for all possible choices of m. To show this, let us take the logarithm of both sides of (20). We obtain

$$\sum_{i\in 1\ldots|V|} p(i)\log m(i) > \sum_{i\in 1\ldots|V|} r(i)\log m(i) \qquad (21)$$

As the logarithm is a monotonically increasing function of its argument, $m' = \log m$ is an ordinal measure too. So, Theorem 1 shows that comparison (21) is meaningful and so is comparison (20).

For instance, these results may help us gain more confidence in assessing if there is a difference between the central tendencies of two distributions, e.g., we may be more confident in our belief that there has been, say, a decrease in failure criticality from one version of a program to another, as the "center" of failure distribution has decreased. As the examples of Table 2 show, the median may not be enough to this end: the median of both distributions p_3 and p_4 is on category #4, but we have proved that $E_3(m) > E_4(m)$. However, there is a relationship

between the comparisons of medians and the comparisons of means. In general, if Theorem 1 holds, then we also have $median_p \geq median_r$, so the fact that $median_p \geq median_r$ is a necessary condition for $E_p > E_r$ for all possible ordinal measures. Based on formulae (13) and (14), we have that $\sum_{j \in 1...median_p - 1} p(j) \leq \sum_{j \in 1...median_p - 1} r(j)$, so $median_r$ cannot be greater than $median_p$.

5. Hierarchical ordinal measures

We generalize the results of Section 4 to the two irregular scales of Section 2.3, starting from scales built on strict weak orders (Section 5.1) and then addressing the case of general hierarchies (Section 5.2). In this section, the weight $w(k)$ found in a weighted sum for a category k denotes the difference between the frequencies of category k in two distributions p and r, i.e., $w(k) = p(k) - r(k)$. Also, $m(k)$ denotes the common value that measure m associates with all entities belonging to k.

5.1. Strict weak orders

We suppose that (E, R) and (V, S) are strict weak orders (Definition 9). Many possible total orders are compatible with a strict weak order. For instance, for the example of Figure 3, $(b, a, d, c, e, f, h, g, i, j)$ is a total order compatible with the strict weak order, while $(b, a, d, c, f, e, h, g, i, j)$ is not, because e precedes f in the strict weak order. Unless otherwise noted, by "total order" we mean "total order compatible with the strict weak order" in the rest of this section. As an example, rows p and r of Table 4 contain two frequency distributions for the example of Figure 3 (e.g., the frequency of the category with d according to distribution r is $r(d) = 0.1$), and row w contains the differences between the corresponding frequencies according to the two distributions. Note that p gives higher frequency than r only for 2 out of the 10 values.

If we want $E = \sum_{k \in K} w(k)m(k) > 0$ to be meaningful, we need to make sure that Theorem 1 holds on *all* total orders. For each total order, we apply Statement 1 by requiring that the condition in formula (18) holds for all cumulative weights and with strict inequality for at least one cumulative weight. So, we need to identify all possible cumulative weights of all possible total orders.

Take one total order (e.g., $(b, a, d, c, e, f, h, g, i, j)$) and consider a cumulative weight that contains weights associated with the categories belonging to $IND(lev)$ (e.g., $lev = 3$ in our example, i.e., $IND(3) = \{f, g, h\}$), but not $IND(lev + 1)$, $IND(lev + 2)$, etc. In other words, consider the cumulative weights for a given subsequence of categories in the total order that reaches level lev (e.g., (b, a, d, c, e, f, h) for level $lev = 3$) but goes no further. In this cumulative weight, the weights of *all* the categories in $PREC(lev)$ must appear, since all the categories of $PREC(lev)$ precede all the categories of $IND(lev)$ in *every* total order. So, the sum of the weights of the categories of $PREC(3)$ is $w(a)+w(b)+w(c)+w(d)+w(e)$, to which we add 1, 2, or 3 weights from the set $w(f), w(g), w(h)$, depending on the length of the subsequence considered. In the general case, we add from 1 to $|IND(lev)|$ weights to the sum of the weights of all the categories in $PREC(lev)$.

We now focus only on the "contribution" to the cumulative weight given by categories belonging to $IND(k)$. For instance, in the subsequence (b, a, d, c, e, f, h), portion (f, h) gives the contribution $w(f) + w(h)$ that is added to the sum of preceding weights $w(a)+w(b)+w(c)+w(d)+w(e)$. However, we are not really interested in the *order* with which the categories appear in the last portion of the subsequence. Portion (h, f) would give exactly the same contribution $w(f) + w(h)$. We are interested only in which subset of categories of $IND(lev)$ appears in the portion. Since there is no mandatory order between the categories in $IND(lev)$, all possible nonempty subsets of the sets of $IND(lev)$ may provide a different contribution. So, there are $2^{|IND(lev)|} - 1$ different contributions for each level, which are independently added to the sum of the weights of all the categories in $PREC(lev)$. The algorithm is clearly exponential in the maximum number of categories for each single level. In practical cases, the maximum number of categories at the same level may be not so large as to make the algorithm unusable.

In general, the following statement holds (the proof is omitted as it is similar to the proof of Theorem1).

Theorem 2 *Let the categories be ordered in a strict weak order, and let $\{w(a), w(b), \ldots\}$ be a set of real numbers. We have $\sum_{k \in K} w(k)p(k) > 0$ (with $|K| > 1$) for* every *possible total order of categories compatible with the strict weak order, if and only if, for each level lev, the sum of the weights of the categories in $PREC(lev)$ is less than or equal to the sum of the weights of any nonempty subset of the categories in $IND(lev)$, with strict inequality for at least one nonempty subset of one level. (By convention, the sum of weights of the categories in $PREC(lev)$ is zero.)*

As for our example of Figure 3 and Table 4, it is possible to show that $E_p > E_r$ is meaningful.

5.2. The general case of hierarchies

In Figure 2, we cannot identify levels, i.e., equivalence classes, since an order relationship cannot be established between the category whose representative element is s_4, on the one side, and those whose representative elements are s_2 and s_3, on the other side, so this is not a strict weak order. We can still use Theorem 1 by applying it to all pos-

	a	b	c	d	e	f	g	h	i	j
p	0.05	0.05	0.1	0.15	0.1	0.15	0.1	0.05	0.2	0.05
r	0.06	0.06	0.13	0.18	0.12	0.07	0.1	0.08	0.15	0.05
w	-0.01	-0.01	-0.03	-0.03	-0.02	+0.08	0	-0.03	+0.05	0

Table 4. Strict Weak Order: Frequencies.

sible total orders compatible with the hierarchy, or (equivalently) Theorem 2 on all possible strict weak orders. In the example, we have two possible strict orders that are not total orders: (1) one in which we have an indifference class given by the categories whose representative elements are s_2 and s_4, and (2) one in which we have an indifference class given by the categories whose representative elements are s_3 and s_4. Note that the sums of weights for all the levels before and after these indifference classes do not change, so the sets of weights for the strict weak orders are not disjoint. The worst case might seem to be the one in which there is no order at all among the categories, but we only need that all the single weights are not positive, with at least one of them being negative.

6. Conclusions and future work

In this paper, we have investigated the definition and the use of aggregate indices for two regular scales, i.e., nominal and ordinal scales, and for two irregular scales, based on strict weak orders and hierarchies. In particular, we have characterized the necessary and sufficient condition under which the mean can be used even with ordinal and the two irregular scales. When the necessary and sufficient condition is not satisfied, then comparing two means is not a meaningful statement.

Future work will address

- the definition of irregular scales for software attributes
- the definition of measures of association for the kinds of scales investigated in this paper, in addition to the ones existing for ordinal measures [5].

7. Acknowledgments

The research documented in this paper has been partially supported by MIUR (the Italian Ministry for University, Education, and Research).

The author wishes to thank Stefano Serra-Capizzano for the useful discussions on the means for ordinal measures.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Inc., Monterey, 1984.

[2] L. Briand and J. Wuest, "The Impact of Design Properties on Development Cost in Object-Oriented System, International Software Engineering," IEEE Trans. on Soft. Eng., Vol. SE-27, No. 11, pp. 963 - 986, Nov. 2001.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications, John Wiley and Sons, New York, 1991.

[4] N. Fenton, "Software Measurement: A Necessary Scientific Basis," IEEE Trans. on Soft. Eng., Vol. SE-20, No. 3, pp. 199-206, Mar. 1994.

[5] M. Hollander M., D.A. Wolfe, *Nonparametric Statistical Methods, 2nd Edition*, Wiley-Interscience, New York, 1999.

[6] D. H. Krantz, R. D. Luce, P. Suppes, A. Tversky, *Foundations of Measurement*, Academic Press, New York, 1971.

[7] R.R. Nelson and S.G. Winter, *An Evolutionary Theory of Economic Change*, Belknap Press, Cambridge, Mass. and London, 1982.

[8] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, Vol. 1, pp. 81 - 106, 1986.

[9] A. Renyi, *Probability Theory*, North-Holland, Amsterdam, 1970.

[10] F. S. Roberts, *Measurement Theory*, Addison-Wesley, Reading, 1979.

[11] R.W. Selby and A. A. Porter, "Learning from Examples: Generation and Evaluation of Decision Trees for Software Resource Analysis," IEEE Trans. on Soft. Eng., Vol. SE-14, No. 12, pp. 1743-1757, Dec. 1988.

[12] C. Tsallis, "Possible Generalization of Boltzmann-Gibbs Statistics," Journal of Statistical Physics, Vol. 52, pp. 479-487, 1988.