TECHNICAL REPORT

ISO/TR 14468

First edition
2010-12-15

Selected illustrations of attribute agreement analysis

Illustrations choisies d'une analyse d'accord d'attribut

Reference number
ISO/TR 14468:2010(E)

Contents

Page

Foreword ... iv
Introduction ... v
1 Scope ... 1
2 Normative references ... 1
3 Terms and definitions ... 1
4 Symbols and abbreviated terms ... 3
5 Generic description of attribute agreement analysis ... 3
5.1 Overview of the structure of attribute agreement analysis ... 3
5.2 Overall objectives of attribute agreement analysis ... 3
5.3 Measurement process description ... 4
5.4 Agreement analysis methodology ... 4
5.5 Sampling plan for attribute agreement analysis ... 5
5.6 Data analysis ... 6
5.7 Conclusions and suggestions ... 6
6 Description of Annexes A to E ... 7
Annex A (informative) Liquid crystal display manufacture ... 8
Annex B (informative) Technical support centre triage of issues ... 14
Annex C (informative) Tasting differences in water ... 20
Annex D (informative) Thermistor defects ... 26
Annex E (informative) Assessment of level of disability following a stroke ... 34
Bibliography ... 42

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In exceptional circumstances, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide by a simple majority vote of its participating members to publish a Technical Report. A Technical Report is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TR 14468 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods*, Subcommittee SC 7, *Applications of statistical and related techniques for the implementation of Six Sigma*.

Introduction

The Six Sigma[1)] and statistical International Standards communities share a philosophy of continuous improvement and many analytical tools. The statistical International Standards community arrives at rigorous documents through long-term consensus. The disparities in time pressures, mathematical rigour, and statistical software usage have inhibited exchanges, synergy, and mutual appreciation between the two groups.

This Technical Report takes one specific statistical tool, attribute agreement analysis, develops the topic somewhat generically (in the spirit of International Standards), then illustrates it through the use of five detailed and distinct applications. The generic description focuses on the commonalities across studies designed to assess the agreement of attribute measurements. The annexes, containing five illustrations, follow the basic framework, but also identify the nuances and peculiarities in the specific applications.

1) Six Sigma is a trademark of Motorola, Inc.

Selected illustrations of attribute agreement analysis

1 Scope

This Technical Report assesses a measurement process where the characteristic(s) being measured is (are) in the form of attribute data (including nominal and ordinal data).

This Technical Report provides examples of attribute agreement analysis (AAA) and derives various results to assess closeness of agreement amongst appraisers, such as agreement within appraisers, agreement between appraisers, agreement of each appraiser vs. a standard, and agreement of all appraisers vs. a standard.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 3534-1, *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*

ISO 3534-2, *Statistics — Vocabulary and symbols — Part 2: Applied statistics*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 3534-1, ISO 3534-2, and the following apply.

3.1
measurement system
collection of operations, procedures, devices and other equipment, software, and personnel used to assign a value to the characteristic being measured

[IWA 1:2005[4], 3.1.9]

NOTE In the context of this Technical Report, the personnel refer to the appraiser.

3.2
nominal data
categorical variables that have two or more levels with no natural ordering

3.3
ordinal data
categorical variables that have three or more levels with a natural ordering

3.4
binary data
categorical variables that have two levels with no natural ordering

3.5
agreement within appraiser
extent to which each appraiser agrees with himself or herself on all trials when each appraiser conducts more than one trial

3.6
agreement between appraisers
extent to which all appraisers agree with each other on all trials when more than one appraiser makes one or more appraisals

3.7
agreement of each appraiser vs. standard
extent to which each appraiser agrees with himself or herself as well as with the standard when a known standard is specified

3.8
agreement of all appraisers vs. standard
extent to which all appraisers agree with each other on all trials as well as with the standard when a known standard is specified

3.9
percentage of agreement
P %
agreement, expressed as a percentage, for multiple appraisals by one appraiser or among different appraisers

3.10
kappa
κ
statistic indicating the degree of agreement of the nominal or ordinal assessments made by multiple appraisers when evaluating the same samples

NOTE Kappa statistics are commonly used in cross-tabulation (table) applications and in attribute agreement analysis.

3.11
Fleiss's kappa
statistic used for assessing the reliability of agreement when appraiser(s) are selected at random from a group of available appraisers

3.12
Cohen's kappa
statistic used for assessing the reliability of agreement when the appraiser(s) are specifically chosen and are fixed

3.13
***p*-value**
probability of observing the observed test statistic value or any other value at least as unfavourable to the null hypothesis

[ISO 3534-1:2006, 1.49]

NOTE This concept is used in hypothesis tests to help in deciding whether to reject or fail to reject a null hypothesis.

3.14
***Z*-statistic**
test statistic which follows the standard normal distribution

4 Symbols and abbreviated terms

95 % CI	95 % confidence Interval
AAA	attribute agreement analysis
MSA	measurement system analysis
σ_κ	standard error (SE) of kappa statistic
n	sample size
P %	percentage of agreement
Z	value of the Z-statistic

5 Generic description of attribute agreement analysis

5.1 Overview of the structure of attribute agreement analysis

This Technical Report provides general guidelines on the design, conduct and analysis of studies aiming at evaluating the agreement amongst appraisers when classifying an item into two or more categories (e.g. “good” or “bad”). It describes a procedure with five steps and illustrates the steps with five distinct applications given in Annexes A to E.

The steps given in Table 1 are generic and apply to design and analysis of AAA studies in general. Each of the five steps as well as general agreement analysis methodology are explained in general in 5.2 to 5.7. Specific explanations of the substance of these steps are provided in the examples in Annexes A to E.

Table 1 — Basic steps in attribute agreement analysis

1	State the overall objectives
2	Describe the measurement process
3	Design the sampling plan
4	Analyse the result
5	Provide a conclusion with suggestions

5.2 Overall objectives of attribute agreement analysis

AAA is often used in Six Sigma projects and quality improvement projects. The primary motivation for AAA studies should be clearly stated and agreed upon by all parties. The main purpose of AAA is to evaluate the capability of a measurement system based on attribute data and to judge whether it is acceptable in the context of making correct decisions within a given monitored process. AAA determines how good agreement is among appraisers, and between appraisers and a given recognized “standard”.

AAA is conducted for a variety of reasons, which include, but are not limited to:

a) a lack of consistency in the assessment of a part or unit determined by one appraiser during different trials;

b) a lack of consistency in the assessment of a part or unit determined by different appraisers;

c) the measurement results of a part or unit determined by an appraiser or appraisers exhibiting disagreement with a known standard value for that part or unit;

d) a requirement of quality management standards, e.g. ISO/TS 16949[5].

5.3 Measurement process description

This Technical Report focuses on processes where the characteristic(s) being measured consist(s) of attribute data.

The measurement process should be clearly described before conducting AAA, including appraisers, procedures, the quality characteristic(s) to be measured, measurement conditions, and attribute data type (i.e. nominal, ordinal or binary).

5.4 Agreement analysis methodology

Many measurement processes in industry rely on gauges, weighing instruments, micrometers or other devices that make fairly direct physical measurements of a product characteristic. There are, however, many situations in which quality characteristics are difficult to define and assess, e.g. automobile performance ratings, classification of fabric quality as “good” or “bad”, and ratings of wine colour, aroma and taste on a 1 to 10 scale.

In cases when physical measurements are not possible, subjective classifications or ratings are made by people. In these situations, an AAA is needed where more than one appraiser gives a rating and an evaluation of the agreement between appraisers is made. If the appraisers agree, the possibility exists that the ratings are accurate. If the appraisers disagree, rating usefulness is limited.

The assigned ratings can be nominal, ordinal or binary. Nominal data are categorical variables that have two or more levels with no natural ordering. For example, the levels in a food tasting study may include crunchy, mushy, and crispy. Ordinal data are categorical variables that have three or more levels with a natural ordering, such as strongly disagree, disagree, neutral, agree, and strongly agree. However, distances between categories are unknown. Binary data are categorical variables that only have two levels. For instance, appraisers classify items as “good/bad”, or “go/no go”. It should be noted that binary data actually constitute a special case of nominal data with only two levels. Binary data are widely used in industry and when a standard exists giving the correct value of the unit being measured, misclassification rates can also be employed to assess the performance of a measurement system. A binary measurement system is discussed further in Annex A. Thus, in this Technical Report, nominal data refer to a variable that has three or more possible levels.

No matter what the data type is, percentage of assessment agreement can be utilized to evaluate the agreement of an attribute measurement system. Percentage of agreement quantifies the agreement for multiple ratings within one appraiser or among different appraisers. The percentage of assessment agreement, P %, is actually the point estimate for a population proportion, and is given by

$$P\,\% = \frac{n_{\text{match}}}{n} \times 100\,\%$$

where

n_{match} is the number of agreements among multiple ratings;

n is the number of samples.

For nominal data, the kappa statistic, κ, is most appropriate. It is defined as the proportion of agreement between appraisals after agreement by chance has been removed.

$$\kappa = \frac{P_{\text{obs}} - P_{\text{exp}}}{1 - P_{\text{exp}}}$$

where

P_{obs} is the observed proportion of agreement;

P_{exp} is the expected proportion due to chance agreement.

The value of kappa ranges from −1 to +1. Generally speaking, the higher the value of kappa, the stronger the agreement. If kappa has the value 1, the ratings show perfect agreement (consistency). If kappa is 0, the agreement of the ratings is the same as that expected by chance. In general, kappa values above 0,9 are considered excellent.

Kappa values less than 0,7 indicate that the rating system needs improvement, whereas those less than 0,4 indicate the measurement system capability is inadequate. Typically a kappa value of at least 0,7 is required.

The two most popular kappa statistics are Cohen's kappa, based on the two-way contingency table, and Fleiss's kappa, based on matched pairs. They treat the selection of appraisers differently when calculating the probability of agreement by chance. Cohen's kappa assumes that the appraisers are specifically chosen and are fixed, whereas Fleiss's kappa assumes that the appraisers are selected at random from a group of available appraisers. This leads to two different methods of estimating the probability. Thus kappa, and its standard error (SE), σ_κ, can be calculated with either Fleiss's method or Cohen's method. The test statistic for kappa is

$$Z = \frac{\kappa}{\sigma_\kappa}$$

with the null hypothesis $H_0{:}\kappa = 0$ and the alternative hypothesis $H_1{:}\kappa > 0$.

This is a one-sided test. Under the null hypothesis, Z follows the standard normal distribution. Reject the null hypothesis if the p-value is less than the prespecified value, commonly taken to be 0,05.

Since binary data are a special case of nominal data with only two levels, kappa statistics can also be employed to deal with a binary measurement system.

Kappa statistics do not take into account the magnitude of differences observed in ordinal data. They represent absolute agreement among ratings. Therefore, when examining ordinal data, Kendall's coefficients are the best choice. Two types of Kendall's coefficients are mentioned in this Technical Report, Kendall's coefficient of concordance (also known as Kendall's W) and Kendall's correlation coefficient (also called Kendall’s tau). Both of these coefficients are non-parametric statistics. The former, ranging from 0 to 1, expresses the degree of association among multiple ratings, whereas the latter, ranging from −1 to 1, expresses the degree of association between the known standard and a single rating. Thus, Kendall's coefficient of concordance should be used to evaluate the consistency within appraisers and between appraisers. Furthermore, when the true standard is known, Kendall's correlation coefficient can be employed to assess the following two types of agreements: agreement of each appraiser vs. standard and agreement of all appraisers vs. standard.

5.5 Sampling plan for attribute agreement analysis

In the sampling plan for AAA studies, the subgroup size of parts, the number of appraisers, and the number of trials should be determined. Generally speaking, three to five appraisers are selected to rate more than 20 parts (for multiple attributes, more samples are required to cover all the attributes) with two or three trials. Note that the selected samples should represent the entire production process. For nominal data, the appraiser selection method also determines which kappa statistic should be calculated. If the appraisers are specifically chosen and are fixed, Cohen's kappa is more appropriate. If appraisers are selected at random from a group of available appraisers, Fleiss's kappa is preferred. It is also worth mentioning that Cohen's kappa is based on the two-way contingency table. When the standard is not known, Cohen's kappa can only be calculated if and only if the data satisfy the conditions:

a) within appraiser — there are exactly two trials with an appraiser;

b) between appraisers — there are exactly two appraisers each having one trial.

In the process of measurement for AAA, randomization is a very important consideration. Randomization means the parts should be measured by the appraiser in a random order.

Table 2 shows a basic layout of an AAA with three appraisers, three repetitions, and 20 items measured by each appraiser.

Table 2 — Layout of a generic attribute agreement analysis design

Item number	Appraiser A			Appraiser B			Appraiser C			Standard
	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	Trial 1	Trial 2	Trial 3	
1										
2										
3										
20										

5.6 Data analysis

The following four types of agreement need be taken into consideration:

a) within appraisers, which means that each appraiser agrees with himself or herself on all trials;

b) between appraisers, which means that all appraisers agree with each other on all trials;

c) each appraiser vs. standard, which means that each appraiser agrees with himself or herself as well as with the standard;

d) all appraisers vs. standard, which means that all appraisers agree with each other on all trials as well as with the standard.

It is quite obvious that the type of agreement c) is no less than the first one a) since it adds a constraint, namely, agreeing with the standard. The condition is quite similar for the fourth and the second types of agreements. Obviously, the fourth kind of agreement is the smallest of the four. And for each type of agreement, two types of kappa statistics are generally adopted, those of Cohen and Fleiss. Also, for nominal data with three or more categories, two types of kappa coefficients can be calculated. First, one can compute an overall kappa, which is an assessment of raters' agreement across all categories. Second, one can compute individual kappa values for each category. This reveals the categories in which raters have trouble agreeing.

In addition to the AAA report, AAA graphics are also useful. They can be used to reflect the agreement clearly and directly. Generally, the percentages of assessment agreement within and between appraisers, kappa coefficient tables, and Kendall's coefficient (ordinal data only) tables are calculated. Moreover, a graph of the matched proportions for each appraiser can be displayed when the number of trials for each appraiser is more than one. Additionally, another graph of the matched proportions between the ratings of each appraiser and the attribute can be displayed only when the attribute is known and provided for each sample.

5.7 Conclusions and suggestions

Based on the results of the AAA, a judgement can be made about the adequacy of the attribute measurement process. Generally the disagreement within an appraiser shows the appraiser cannot make consistent measurement results (possibly because the appraiser did not follow the measurement procedure exactly at different trials). The disagreement between appraisers means the appraisers' procedures are not exactly the same or the appraisers' capabilities of measurement are different (possibly due to their different experiences or physical reasons, e.g. eyesight for visual inspection). Actions shall be taken after the root cause(s) is (are) found for the inadequate attribute measurement process.

After certain actions have been taken to improve the measurement system, e.g. effective training has been done for the operators, the AAA needs to be repeated to validate whether the improved measurement system is acceptable.

6 Description of Annexes A to E

Five distinct examples of AAA are illustrated in Annexes A to E, which have been summarized in Table 3 with the different aspects indicated.

Table 3 — Example summaries listed by annex

Annex	Example	AAA details
A	LCD manufacture	Three appraisers, randomly selected among the group of appraisers, judged LCD quality on 20 samples twice by visual inspection. The inspection results are binary. Minitab[a] software package was used to perform the analysis
B	Technical support triage of issues	Nominal response with 6 categories encountered in Service Sector; 4 appraisers, no repetition, 48 issues evaluated by each appraiser. SAS JMP[b] software package was used to perform the analysis. “Truth” on correct categorization of issue is known
C	Tasting differences in water	Nominal response with 4 categories; 3 testers, 3 repetitions, leading to 12 cups of water evaluated by each tester. SAS JMP[b] software package was used to perform the analysis. “Truth” on correct categorization of brand of water is known
D	Thermistor defects	Three appraisers, randomly selected among the group of appraisers, judged 20 thermistor samples twice by visual inspection. The inspection results are nominal data, falling into 8 categories and without natural ordering. Minitab[a] software package was used to perform the analysis
E	Assessment of level of disability following a stroke	Ordinal response with 5 ordered categories encountered in the medical sector; 2 appraisers, no repetition, 46 cases evaluated by each appraiser. SAS JMP[b] software package was used to perform the analysis. “Truth” on correct categorization of issue is known

[a] Minitab is the trade name of a product supplied by Minitab, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

[b] SAS JMP is the trade name of a product supplied by the SAS Institute, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

Annex A
(informative)

Liquid crystal display manufacture

A.1 General

In a liquid crystal display (LCD) manufacturer, the display feature is judged by operators through visual inspection. All the samples are tested under video graphics array (VGA) mode. The results can be either normal colour (marked as good) or deflected colour (bad). In the measurement phase, visual inspection, leading to subjective classification, is mainly employed by the appraisers to judge whether a sample is good or bad. Therefore, the experience of the appraisers and the training they have been given are of huge importance. The objective of this study is to evaluate the consistency and accuracy of the attribute measurement system.

A.2 Response variable

The response variable is binary data (two levels with no natural ordering).

A.3 Standard attribute

Standard attribute (the correct rating) is given in this case.

A.4 Possible reasons for incorrect judgement

Failing to follow work instructions could lead to incorrect judgement. Another factor could be the experience of the appraisers and the training they have been given.

A.5 Sampling plan

To assess the consistency and accuracy of ratings, three appraisers, Carol, Fiona, and Kaka, judged LCD quality on 20 samples (model: LCD40b66) twice by visual inspection. LCD samples were randomly presented to the three appraisers, who were randomly selected from a group with the same introductory training and similar experience.

The inspection results are binary.

A.6 Raw data

Table A.1 lists the raw data used in the AAA.

Table A.1 — Inspection results of LCD and standard attribute

Part	Standard	Carol		Fiona		Kaka	
		1st Trial	2nd Trial	1st Trial	2nd Trial	1st Trial	2nd Trial
1	Good	Good	Good	Good	Good	Good	Good
2	Good	Good	Good	Good	Good	Good	Good
3	Good	Good	Good	Good	Good	Good	Good
4	Bad	Bad	Bad	Bad	Bad	Bad	Bad
5	Bad	Bad	Bad	Good	Good	Bad	Bad
6	Bad	Bad	Bad	Bad	Bad	Bad	Bad
7	Good	Good	Good	Good	Good	Good	Good
8	Good	Good	Good	Good	Good	Good	Good
9	Good	Good	Good	Good	Good	Good	Good
10	Good	Good	Good	Good	Good	Good	Good
11	Good	Good	Good	Good	Good	Good	Good
12	Good	Good	Good	Good	Good	Good	Good
13	Good	Good	Good	Good	Good	Good	Good
14	Good	Good	Good	Good	Good	Bad	Bad
15	Good	Good	Good	Good	Good	Good	Good
16	Good	Good	Good	Good	Good	Good	Good
17	Good	Good	Good	Good	Good	Good	Good
18	Good	Good	Good	Good	Good	Good	Good
19	Bad	Bad	Bad	Bad	Bad	Bad	Bad
20	Bad	Bad	Bad	Bad	Bad	Bad	Bad

A.7 Attribute agreement analysis

AAA in Minitab 15[2)] is adopted to assess the consistency and accuracy of subjective classifications by examining the results within appraisers, between appraisers, and against the standard. AAA output consists of session window and graph window results.

The session window includes the following types of agreement:

a) within appraiser: it shows the consistency with which an appraiser rates the same sample across different trials;

b) between appraisers: it shows whether appraisers' ratings agree with each other, i.e. whether different appraisers give the same rating to the same sample.

Since the standard attribute (the correct rating) is given in this case, the session window output includes two additional types of agreement:

c) each appraiser vs. standard: it shows how well each appraiser's assessment of each sample matches with the standard, in other words, whether each rating of the same appraiser agrees with the standard rating;

d) all appraisers vs. standard: it shows how well responses of all appraisers agree with the known standard when they are combined.

For each type of agreement, the session window output includes assessment agreement and Fleiss's kappa statistics to assess the consistency and accuracy of the appraisers' responses.

2) Minitab is the trade name of a product supplied by Minitab, Inc. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.