TECHNICAL REPORT

ISO/TR 16705

First edition
2016-08-15

Statistical methods for implementation of Six Sigma — Selected illustrations of contingency table analysis

Méthodes statistiques pour l'implémentation de Six Sigma — Exemples sélectionnés d'application de l'analyse de tableau de contingence

Reference number
ISO/TR 16705:2016(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

Page

Foreword iv

Introduction v

1 Scope 1

2 Normative references 1

3 Terms and definitions 1

4 Symbols and abbreviated terms 2

5 General description of contingency table analysis 2

5.1 Overview of the structure of contingency table analysis 2

5.2 Overall objectives of contingency table analysis 3

5.3 List attributes of interest 3

5.4 State a null hypothesis 3

5.5 Sampling plan 3

5.6 Process and analyse data 4

5.6.1 Chi-squared test 4

5.6.2 Linear trend test 6

5.6.3 Correspondence analysis 6

5.7 Conclusions 7

6 Description of Annexes A through D 7

Annex A (informative) **Distribution of number of technical issues found after product release to the field** 8

Annex B (informative) **People's perception about contented life** 15

Annex C (informative) **Customer satisfaction research on a brand of beer** 20

Annex D (informative) **Proportions of nonconforming parts of production lines** 26

Bibliography 31

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

The committee responsible for this document is 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 and international statistical standards communities share a philosophy of continuous improvement and many analytical tools. The Six Sigma community tends to adopt a pragmatic approach driven by time and resource constraints. The statistical standards community arrives at rigorous documents through long-term international consensus. The disparities in time pressures, mathematical rigor, and statistical software usage have inhibited exchanges, synergy, and mutual appreciation between the two groups.

The present document takes one specific statistical tool (Contingency Table Analysis), develops the topic somewhat generically (in the spirit of International Standards), then illustrates it through the use of several detailed and distinct applications. The generic description focuses on the commonalities across studies designed to assess the association of categorical variables.

The Annexes containing illustrations do not only follow the basic framework, but also identify the nuances and peculiarities in the specific applications. Each example will offer at least one “winkle” to the problem, which is generally the case for real Six Sigma and other fields application.

Statistical methods for implementation of Six Sigma — Selected illustrations of contingency table analysis

1 Scope

This document describes the necessary steps for contingency table analysis and the method to analyse the relation between categorical variables (including nominal variables and ordinal variables).

This document provides examples of contingency table analysis. Several illustrations from different fields with different emphasis suggest the procedures of contingency table analysis using different software applications.

In this document, only two-dimensional contingency tables are considered.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

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

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at http://www.iso.org/obp

3.1
categorical variable
variable with the measurement scale consisting of a set of categories

3.2
nominal data
variable with a nominal scale of measurement

[SOURCE: ISO 3534-2:2006, 1.1.6]

3.3
ordinal data
variable with an ordinal scale of measurement

[SOURCE: ISO 3534-2:2006, 1.1.7]

3.4
contingency table
tabular representation of categorical data, which shows frequencies for particular combinations of values of two or more discrete random variables

Note 1 to entry: A table that cross-classifies two variables is called a “two-way contingency table;” the one that cross-classifies three variables is called a “three-way contingency table.” A two-way table with r rows and c columns is also named “r × c table.”

EXAMPLE Let n items be classified by categorical variables X and Y with levels X_1, X_2 and Y_1, Y_2, respectively. The number of items with both attribute X_i and Y_j is n_{ij}. Then, a 2 × 2 table is as follows.

Table 1 — 2 × 2 contingency table

Variable X	Variable Y	
	Y_1	Y_2
X_1	n_{11}	n_{12}
X_2	n_{21}	n_{22}

3.5
***p*-value**
probability of observing the observed test statistic value or any other value at least as unfavorable to the null hypothesis

[SOURCE: ISO 3534-1:2006, 1.49]

4 Symbols and abbreviated terms

H_0	null hypothesis
H_a	alternative hypothesis
χ^2	Chi-square statistic
G^2	likelihood-ratio statistic
n	total number of cell count
r × c table	contingency table with *r* rows and *c* columns
DF	degree of freedom

5 General description of contingency table analysis

5.1 Overview of the structure of contingency table analysis

This document provides general guidelines on the design, conduct, and analysis of contingency table analysis and illustrates the steps with distinct applications given in Annexes A through D. Each of these examples follows the basic structure given in Table 2.

Table 2 — Basic steps for contingency table analysis

1	State the overall objective
2	List attributes of interest
3	State a null hypothesis
4	Sampling plan
5	Process and analyse data
6	Accept or reject the null hypothesis (Conclusions)

Contingency table analysis is used to assess the association of two or more categorical variables. This document focuses on two-way contingency table analysis, which only considers the relation of two categorical variables. Particular methods for three or more categorical variables analysis are not included in this document. The steps given in Table 1 provide general techniques and procedures for contingency table analysis. Each of the six steps is explained in general in 5.2 to 5.7.

5.2 Overall objectives of contingency table analysis

Contingency table analysis can be employed in Six Sigma[1] projects in the “Analyse” phase of DMAIC methodologies, and often used in sampling survey, social science and medical research, etc. Apart from the usual statistical methods focusing on continuous variables, contingency table analysis mainly handles the categorical data, including nominal data and ordinal data. In the case that the observed value is the frequency of certain combinations of several objective conditions, but not the continuous value from the equipment, the contingency table analysis is needed.

The primary motivation of this method is to test the association of categorical variables, including the following situations:

a) to assess whether an observed frequency distribution differs from a theoretical distribution;

b) to assess the independence of two categorical variables;

c) to assess the homogeneity of several distributions of same type;

d) to assess the trend association of observations on ordinal variables;

e) to assess extensive association between levels of categorical variables.

5.3 List attributes of interest

This document considers the association of two categorical variables based on the observed frequency of the characteristic corresponding to combinations of different levels of attributes of interest.

If the association between quantitative variable and categorical variable is of interest (e.g. cup size versus surface decoration), it is necessary to divide quantitative data into ordinal classes (e.g. small, medium, large).

5.4 State a null hypothesis

This document is to determine whether row variable and column variable are independent. The null hypothesis for Chi-square test is

H_0: the row variable and column variable are independent;

and the alternative hypothesis is

H_a: the row variable and column variable are not independent.

5.5 Sampling plan

In the sampling plan for contingency table analysis, variables and the levels should be determined first. For two-way contingency tables, there are four possible sampling plans to generate the tables.

a) The total number of cell count n is not fixed.

b) The total number of cell count n is fixed, but none of the total rows or columns are fixed.

c) The total number of cell count n is fixed, and either the row marginal totals or the column marginal totals are fixed;

d) The total number of cell count n is fixed, and both row marginal totals and the column marginal totals are fixed.

[1] Six Sigma is the trademark of a product supplied by Motorola, 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. Equivalent products may be used if they can be shown to lead to the same results.

The aforementioned four sampling plans correspond to different purposes of categorical data analysis. Case a) is a random sampling, that all frequency numbers are independent. For example, the number of customers entering a supermarket during the day is a random variable. The customers are divided into four classes based on their gender and whether they are shopping or not (male/shopping, male/no shopping, female/shopping, female/no shopping). These four numbers form a contingency table. Case b) is applicable to a sampling survey where the sample size is fixed. Case c) is usually an analysis of a comparative analysis. For example, when conducting a research on the relationship of lung cancer and smoking, a group of patients with lung cancer and a group of healthy people with similar age, gender, and other physical condition are chosen for the research. The total number of people in each group is fixed. Case d) is another test of attribute agreement analysis, usually used to test whether the results from two measurement systems are consistent with each other. For attribute agreement analysis, one can refer to ISO 14468. The calculated statistics of the test of independence for the first three cases are the same.

Randomization is very important when sampling for experiments. The observations in each cell are made on a random sample. When it is inconvenient or difficult to attain adequate samples, one should pay close attention to any confounding factors that may affect the results of the analysis.

Table 3 shows a two-way contingency table with r levels of variable *X* and c levels of *Y*. The observed frequency of each combination of the two variables is n_{ij} (i =1,..., r, j=1,...,c).

Table 3 — Layout of a generic r × c contingency table analysis

Variable *X*	**Variable *Y***			
	Y_1	Y_2	...	Y_c
X_1	n_{11}	n_{12}	...	n_{1c}
X_2	n_{21}	n_{22}	...	n_{2c}
...	...	...	...	...
X_r	n_{r1}	n_{r2}	...	n_{rc}

5.6 Process and analyse data

5.6.1 Chi-squared test

Chi-square (χ^2) test is the most fundamental tool for contingency table analysis to test independence of variables. It is commonly used to compare observed data with some expected data according to a specific test purpose.

For a one-dimension contingency table, which has only one categorical variable with two or more levels, Chi-square test, usually called “goodness-of-fit test,” can be used to assess whether the observed data classified by levels follow an theoretical distribution.

For a two-dimensional contingency table, r × c table, Chi-square test can be used to evaluate whether two categorical variables are independent. It can test the homogeneity of distributions with same type, which is also called “homogeneity test.”

Chi-square test is defined to evaluate the distance of the observed data from the expected data. The formula for calculating Chi-square statistic is:

$$\chi^2 = \sum \frac{(o-e)^2}{e} \tag{1}$$

where

o is the observed frequency data;

e is the expected frequency data.

The formula is the sum of the squared difference between observed and expected frequency, divided by the expected frequency in all cells.

For r × c table, there are r levels for row variable X, c levels for column variable Y. With the null hypothesis H_0, X and Y are independent, and the alternative hypothesis H_α, X and Y are not independent, the Chi-square statistic is calculated as follows:

$$\chi^2 = \sum_{i=1}^{\mathrm{r}} \sum_{j=1}^{\mathrm{c}} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{2}$$

where

n_{ij} is the observed frequency in the ith level of row variable X and jth level of column variable Y;

m_{ij} is the expected value of n_{ij} assuming independence.

This statistic takes its minimum value of zero when all $n_{ij} = m_{ij}$. For a fixed sample size, greater differences between n_{ij} and m_{ij} produce larger χ^2 values and stronger evidence against H_0. The χ^2 statistic follows a asymptotic Chi-square distribution for large n, with degree of freedom (DF) = (r-1)(c-1). Reject the null hypothesis if the p-value is less than the pre-specified value, commonly taken at 0,05. The Chi-square approximation improves as m_{ij} increases. Note that when any cell expected value is less than 5, the Chi-square test is not appropriate.

An alternative method for independence test is using likelihood-ratio function through the ratio of two maximum functions. For r × c table, the likelihood functions are based on multinomial distribution, and the likelihood-ratio statistic is

$$G^2 = 2\sum_{i=1}^{\mathrm{r}} \sum_{j=1}^{\mathrm{c}} n_{ij} \log\left(\frac{n_{ij}}{m_{ij}}\right) \tag{3}$$

This statistic asymptotically follows the χ^2 distribution. The two methods usually have the same properties and provide same conclusions.

If n is small, the distribution of statistics χ^2 and $G^2$are less Chi-squared. Usually, if there exists at least one expected value less than 5 (in some statistical software, the criterion is slightly different), Fisher's exact test is used. In this case, the observed data follow the hypergeometric distribution, which is the exact distribution of the data. The calculation is much more complicated than χ^2 and G^2 where statistical software is often used. There is another way to handle the case when the expected value is too small. One can combine the row or column to the adjacent one to increase the expected values before the independence test. However, this method should be used with caution; combining columns/rows may reduce interpretability and also may create or destroy structure in the table. If there is no clear guideline, the combination method should be avoided.

It should be noted that Chi-square test is basically a test of significance. It can help decide if a relationship exists, but not how strong it is. Chi-square test is not a measure of strength of association.

In this document, Chi-squared test is used to assess three types of comparisons: test of goodness-of-fit, test of independence, and test of homogeneity.

a) Test of goodness-of-fit compares the observed values and theoretical distribution to determine whether an experimenter's prediction fit the data.

b) Test of independence determines whether there is a significant association between two categorical variables.

c) Test of homogeneity compares two sets of categories to determine whether the two groups are distributed differently among the categories.

5.6.2 Linear trend test

When both row and column variables are ordinal data, linear trend test can be used to test whether a trend exists when a variable changes. However, binary variable can also be treated as ordinal data (e.g. no heart disease as “low” risk condition versus having heart disease as “high” risk condition). There are two directions of the trend. Given two ordinal variables *X* and *Y*, as the level of *X* increase, response on *Y* tends to increase to higher levels (positive linear trend), or response on *Y* tends to decrease to lower levels (negative linear trend).

Correlation coefficients are sensitive to linear trends. Two common used correlation coefficients are Pearson’s r and Spearman’s rho. Pearson’s correlation is a measure of linear relationship between two variables. The calculation of Pearson’s r is based on the number of observations. Spearman’s rho is a nonparametric statistic, using rank orders instead of observations.

Other three correlation coefficients assessing the trend association are Goodman and Kruskal’s γ, Kendall’s τ (tau-b), and Somers’ D. These nonparametric methods depend on the scores of the data in the rows or columns, but not on the quantitative value of each cell.

The values of correlation coefficients range from −1 to 1. The closer the value to 1, the more positive the linear trend is; the closer the value is to −1, the more negative the linear trend is. If the value is 0, there is no relationship. Kendall’s τ, like Goodman and Kruskal’s γ, tests the significance of association between ordinal variables, and it includes tied pairs (identical pairs) in its calculation, which γ ignores. For Somers’ D*,* D R|C and D C|R measure the strength and direction of the relationship between pairs of variables, with row variable and column variable as the response variables, respectively.

These coefficients just give a result of the possibility of trend association, but the trend association test shows strength of the relationship. For each coefficient, calculate the test statistic *z* ($z = \gamma, \tau$, or *d)* and its standard error σ_Z. The statistic is

$$U = \frac{z}{\sigma_z} \qquad (4)$$

with the null hypothesis H_0 variables *X* and *Y* are independent. *U* follows the asymptotic standard normal distribution when H_0 is true. Reject the null hypothesis if the *p*-value is less than the pre-specified value, commonly taken to be 0,05.

Loglinear Model is another useful method for contingency table analysis, which is not included in this document. The models specify how expected cell counts depend on levels of categorical variables and allow for analysis of association and interaction patterns among variables. It often uses for high dimensional tables. Loglinear Model method contains quite a few calculations, which is usually with the aid of computer.

5.6.3 Correspondence analysis

Correspondence analysis (CA) is a statistical visualization method to analyse the association between levels of row and column variables in a two-way contingency table. This technique transforms the rows and columns as points in a two-dimensional space, such that the positions of the row and column point are consistent with their association in the table.

CA changes a data table into two sets of new variables called “factor scores” (obtained as linear combination of row and columns, respectively). The factor scores represent the similarity of the rows and columns structure. Plot the factor scores in a plane that optimally displays the information in the original table. This plot is the standard symmetric representation of CA.

CA is intimately related to the independence Chi-square test. The total variance (often called “inertia”) of the factor scores is proportional to independence χ^2 statistic and the factors scores in CA decompose this χ^2 into orthogonal components.

In a symmetric CA plot, the distance between two row (respectively column) points in the coordinate plane is a measure of similarity with regard to the pattern of row (respectively column) relative