TECHNICAL REPORT

ISO/TR 20693

First edition
2019-04

Statistical methods for implementation of Six Sigma — Selected illustrations of distribution identification studies

Méthodes statistiques pour la mise en œuvre du Six Sigma - Exemples choisis d'études d'identification de la distribution

Reference number
ISO/TR 20693:2019(E)

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

Published in Switzerland

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 Basic principles **3**
5.1 General 3
5.2 Exploratory data analysis (EDA) 4
5.3 Discrete data case 4
5.3.1 Graphical methods 4
5.3.2 Numerical methods 4
5.4 Continuous data case 5
5.4.1 Graphical methods 5
5.4.2 Numerical methods 5
5.4.3 Distribution family unknown and no prior information available 5
6 General description of distribution identification **6**
6.1 Overview of the structure of distribution identification 6
6.2 State overall objectives 6
6.3 Formulate a model theory 6
6.4 Collect, prepare and explore data 7
6.5 Select underlying probability distributions 8
6.6 Perform goodness of fit test 8
6.7 Draw conclusions 8
7 Examples **9**
Annex A (informative) **Test uniformity in the Super Lotto** **10**
Annex B (informative) **Distribution of the number of technical issues found after product release to the field** **13**
Annex C (informative) **Software development effort estimation** **18**
Annex D (informative) **Determining the warranty period of a product** **26**
Bibliography **33**

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 of the voluntary nature of standards, 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 www.iso.org/iso/foreword.html.

This document 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*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Many statistical techniques assume that the data to be analysed come from a given distribution (or population). Such assumptions are crucial to the effectiveness of subsequent statistical inference methods. In the Six Sigma community, when using such statistical methods, one needs to consider whether this assumption is reasonable. More generally, sometimes it is interesting and necessary to find the distribution which generated the data set (or sample) at hand. Identification of the distribution may provide some ways to answer this question. It consists of finding a distribution (or a family of distributions) which provides a good representation of a sample.

The distribution identification within Six Sigma projects should ideally be performed before the end of the Measure phase and can continue throughout the other phases of the DMAIC. From a Six Sigma perspective, the distribution identification can have multiple purposes based on the considered phase. It is used, for example, to characterise a baseline of the process performance, during the Measure or Analyse phase, to characterise the new process during the Improve phase, and to continuously monitor the process performance during the Control phase to ensure that the change is sustained. From a statistical perspective, distribution identification may be helpful to find appropriate statistical techniques for the related data, since many parametric statistical inference methods need certain distributional assumptions.

In general, distribution identification methods may be used as a tool to:

a) verify that a distribution used historically is still valid for the current data;

b) choose the appropriate distribution.

The choice of appropriate distribution should be guided by the knowledge of physical phenomena or the business process. It is recommended to start from a tentative theory to avoid just curve fitting.

In practice, there is always certain context or business background which can be used in determining the distribution. For example, under some circumstance, one can expect the measurement error is normally distributed. In reliability fields, the life distributions for certain products are exponential, lognormal, Weibull, or extreme distributions and so on. However, when such knowledge is not available, the possible underlying distribution for the data should also be identified if one wants to use parametric statistical methods. In this case, exploratory data analysis methods should be used to gain a better understanding. Through graphical visualisation methods, one could form a hypothesis on the possible distributions, stratification of the data or other aspects. Once the hypothesis is formed, hypothesis testing, including goodness of fit testing, can be applied to check one's guess. Finally, a suitable distribution may be found for the data.

In some commercial software packages including MINITAB[1)], SAS-JMP[1)] and Q-DAS[1)], although there are buttons for distribution identification, one should take knowledge of context and process related to data into consideration instead of simply relying on the software packages. Otherwise, misleading results can be given.

1) MINITAB is the trade name of a product supplied by Minitab Inc. JMP is the trade name of a product supplied by SAS Institute Inc. Q-DAS is the trade name of a product supplied by Q-DAS GmbH. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of these products.

Statistical methods for implementation of Six Sigma — Selected illustrations of distribution identification studies

1 Scope

This document provides guidelines for the identification of distributions related to the implementation of Six Sigma. Examples are given to illustrate the related graphical and numerical procedures.

It only considers one dimensional distribution with one mode. The underlying distribution is either continuous or discrete.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements 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 3534-1:2006, *Statistics — Vocabulary and symbols — Part 1: General statistical terms and terms used in probability*

3 Terms and definitions

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

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

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

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

3.1
population
totality of items under consideration

[SOURCE: ISO 3534-1:2006, 1.1, modified - Notes 1, 2, and 3 deleted.]

3.2
sample
subset of a *population* (3.1) made up of one or more sampling units

[SOURCE: ISO 3534-1:2006, 1.3, modified - Notes 1and 2 deleted.]

3.3
observed value
obtained value of a property associated with one member of a *sample* (3.2)

[SOURCE: ISO 3534-1:2006, 1.4, modified - Notes 1 and 2 deleted.]

3.4
family of distributions
distribution family
set of probability distributions

[SOURCE: ISO 3534-1:2006, 2.8, modified - Synonym "distribution family" added; Notes 1 and 2 deleted.]

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

[SOURCE: ISO 3534-1:2006, 1.49, modified - Example and Notes 1 and 2 deleted.]

3.6
descriptive statistics
summary statistics that capture information about the shape, centre or spread of a variable or a distribution

3.7
frequency distribution
empirical relationship between classes and their number of occurrences or *observed values* (3.3)

[SOURCE: ISO 3534-1:2006, 1.60]

3.8
histogram
graphical representation of the *frequency distribution* (3.7) of a data set

3.9
boxplot
horizontal or vertical graphical representation of the five-number summary

[SOURCE: ISO 16269-4:2010, 2.16]

3.10
Q-Q plot
scatter plot for theoretical quantiles and empirical quantiles

3.11
goodness of fit test
hypothesis testing on whether the *population* (3.1)distribution follows a given distribution or belongs to a *distribution family* (3.4)

3.12
normality test
hypothesis testing on whether the *population* (3.1) distribution belongs to a normal *distribution family* (3.4)

4 Symbols and abbreviated terms

$X_1, X_2, ..., X_n$	sample or observed values or data
χ^2	Chi-square distribution or statistics
ALT	accelerated life testing
BB	black belt
BTS	base transceiver station
CLT	central limit theorem
CRM	customer relationship management
DMAIC	Define, Measure, Analyse, Improve and Control

EDA	exploratory data analysis
EDF	empirical distribution function
MIS	management information systems
pdf	probability density function
TP	transaction processing
WEB	Web/online

5 Basic principles

5.1 General

The identification of distributions consists in finding a distribution (or a family of distributions) which best represents a sample (or a group of observed data X_1, X_2, ..., X_n). Based on a priori knowledge or the state of knowledge on the data-generating process, one may possibly know the distribution family for the data set. In that case, it is easy to verify (confirm or reject) it. Otherwise, it may be somewhat complicated to perform distribution identification. At that time, one must narrow down the possible distribution models to a few likely ones. Here are some general guidelines.

a) Apply basic knowledge about the process.

   — If theoretical models exist, they should be applied.

   — If the process generates discrete data, limit the test to discrete distributions.

   — If the process generates only positive number, limit the test only to positive distributions.

b) Apply the Occam's Razor — favour a simpler model unless evidence supports a more complex model.

   — The exponential distribution family is the simplest positive continuous distribution with one parameter.

   — The normal distribution should be favoured as many natural processes follow a normal distribution.

   — The Poisson distribution is among the simplest discrete distribution with one parameter.

In practice, the general flow chart of the procedures for identification of distributions is given in Figure 1.

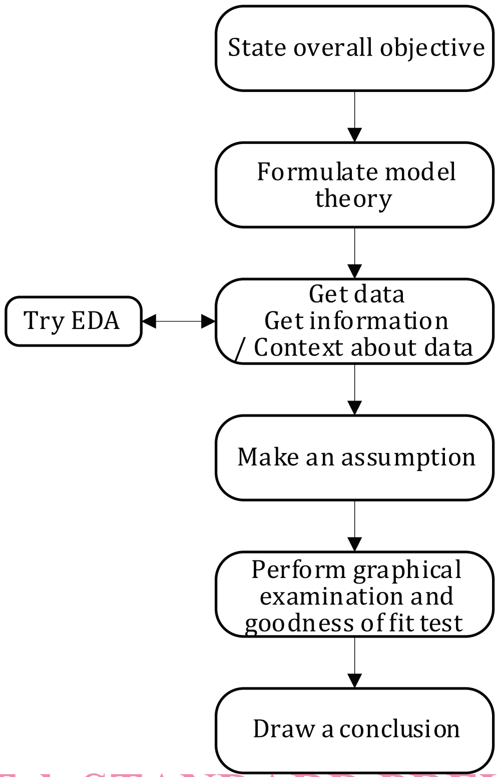

Figure 1 — General flow chart of the procedures for identification of distributions

5.2 Exploratory data analysis (EDA)

EDA is a collection of techniques for revealing information about the data and methods for visualising them. Its philosophy is that data should first be explored without assumptions about probabilistic models, distribution, etc. For one dimensional data, one can consider the following tools.

Descriptive statistics: Mean, standard deviation, skewness, kurtosis, median, max, min, quartile, inter-quantile range, and range are commonly used. Such statistics give the summary values of the data. Some information about the distribution can be derived from them. For example, whether the distribution is symmetric or not. It will be more clearly illustrated by visual tools such as a histogram and a boxplot. A histogram (or stem-and-leaf plot) is a way to graphically represent the frequency distribution of a data set. Though the graphical shape of a histogram may be affected by the different width of bins, the presence of multi-modal behaviour can always be seen from it. The boxplot is another way to display the distribution of a sample. It may provide insights on skewness, behaviour in the tails, and the presence of outliers. The Q-Q plot can be used to check normality, or more generally a location-scale distribution family, or whether two data sets come from the same distribution family.

Since there are some differences between methods of distribution identification for discrete data and for continuous data, the two cases are treated separately in 5.3 and 5.4.

5.3 Discrete data case

5.3.1 Graphical methods

The barplot and histogram can be used for data generated from discrete distribution.

5.3.2 Numerical methods

As a general goodness of fit test statistic, the Pearson χ^2 statistic can be used to test whether the data set comes from certain discrete distributions.

5.4 Continuous data case

5.4.1 Graphical methods

Besides the histogram and boxplot, the Q-Q plot can often be used to graphically check whether the population distribution belongs to a location-scale family. More generally, the Q-Q plot is also used to check whether two groups of data come from the same family of distributions.

5.4.2 Numerical methods

5.4.2.1 Regression method

If the Q-Q plot is nearly linear, the hypothesis about the population distribution can be accepted. To estimate the linearity of the Q-Q plot and evaluate the strength of this linear relationship, regression method may be used. This method is mainly used for testing the location-scale distribution family. Roughly speaking, by considering regression of order statistic on the corresponding population quantile or the expectation of standardised order statistic, correlation coefficient between the dependent variable and the predictor will be used to measure the strength of the linearity in some extent. More rigidly, generalised least squares estimation can be used. In this way, it can be used for testing the uniform distribution, normal distribution, exponential distribution, extreme distributions and logistic distribution. One can refer to [3] for more details.

5.4.2.2 Formal hypothesis testing methods

a) **χ^2-type test**

χ^2-type test statistics include the Pearson χ^2 statistic, likelihood ratio statistic, Neyman modified χ^2 statistic, Freeman-Tukey statistic, Class of power divergence statistics[1],[2] and so on.

b) **EDF-type test**

The Kolmogorov-Smirnov (K-S) test is one of test statistics based on empirical distribution function (EDF). There are still some other complicated test statistics such as supremum EDF type with power divergence weight statistics[6], Cramer-von Mises type statistics[3], etc.

c) **Special test for normality**

Because of its special importance in statistics, the test for normality is widely studied in literature. There are many test statistics for normality testing. Some of them can be found in ISO 5479:1997. The following list just names a few of them.

a) Testing on skewness or kurtosis (or both at the same time).

b) Shapiro-Wilk test (also known as Shapiro-Francia test).

c) Anderson-Darling test: a modification of Kolmogorov-Smirnov test.

d) Jarque-Bera test or Adjusted Jarque-Bera test.

e) Epps-Pulley test.

f) Cramer-Von Mises test.

g) Kolmogorov-Smirnov test.

5.4.3 Distribution family unknown and no prior information available

In the above, it is supposed that the possible distribution families for the data are known in some way. Graphical and numerical methods are provided to verify or disprove it. In some cases, there is no prior knowledge available about the distribution type of the data set. Except for the EDA method,

data transformation techniques such as the Box-Cox or Johnson transformations can be taken into consideration. Both graphical and statistical testing methods of identifying distribution may be used for the transformed data. If it is still hard to identify a good distribution for the data set, the density estimation methods may be invoked, which belongs to the nonparametric tools. When the kernel density estimation method, which is a generalisation of histogram estimate method, is chosen, the result is also affected by the different bandwidths used. The estimated probability density function (pdf) seldom agrees with a simple known distribution (such as normal, student *t* and so on). Thus it may not easy to perform subsequent data analysis in Six Sigma.

6 General description of distribution identification

6.1 Overview of the structure of distribution identification

This document provides general guidelines or principles on distribution identification and illustrates the steps with distinct applications given in Annexes A through D. Each of these examples follows the basic structure given in Table 1.

Table 1 — Basic steps for distribution identification

1	State overall objectives
2	Formulate a model theory
3	Collect, prepare and explore data
4	Select underlying probability distributions
5	Perform goodness of fit test
6	Draw conclusions

The steps given in Table 1 provide a general technique and procedures for distribution identification and how they dovetail with the Six Sigma roadmap (e.g. DMAIC). Each of the six steps of the procedures in Table 1 is explained in detail in 6.2 to 6.7.

6.2 State overall objectives

Distribution identification is implemented within the Six Sigma project ideally before the end of the Measure phase and can continue throughout the other phases of the DMAIC.

By the end of Define and Measure phases, the Six Sigma project team has a clear definition of the problem, the improvement objectives and description of the process structure under study and its scope. The problem is often related to the process performance which is described qualitatively and quantitatively by the end of Measure phase. This will lead to a set of measures. These measures may require identification of the probability distribution for performing further analyses (e.g. capability analysis).

The Six Sigma project team should link the project objectives and process structure, by which the data are generated or will be generated, to the motivation for performing the probability distribution identification. This may be refined or revisited during the following phases of DMAIC as required or appropriate.

6.3 Formulate a model theory

Starting from the objectives and the process by which the data are generated, will help form a tentative theory. This is motivated by W. E. Deming saying “Theory comes first”, so avoid simply curve fitting and instead use the understanding of the data generating process and its structure to identify the most appropriate distribution.

In a sense, there are some natural or physical phenomena that can be modelled by more appropriate probability distributions. Similarly, some probability distributions may be more convenient as they

make less assumptions in terms of parameters (process structure). In this case, use the parsimony argument: fewer parameters are generally better.

Other information, context or knowledge of data generating process such as data type (e.g. categorical, discrete or continuous) will have an impact on the selection of the possible potential distribution families.

One should not solely rely on the data for concluding its type or identifying the distribution, without referring to the context and the generating process, as this can be misleading. For example, one can conclude that a given data set is discrete, whilst in reality the values have been rounded due to the measurement system or by a transformation from one format to another.

Table 2 below lists some of the most common distributions and the motivational theories behind them.

Table 2 — Common distributions and their underlying motivations

Model theory	Candidate probability distributions	Justification
Physical wear out	Weibull	Flexible distribution
Aggregation	Normal	Central limit theorem (CLT), Simple law, appropriate for one dimensional physical measures (e.g. weight, length)
Multiplication	Lognormal	Log of product = sum, CLT
Minimum or maximum	Extreme value	Asymptotic
Random occurrence times	Exponential	Memoryless
Random occurrence counts	Poisson	Approximation for binomial and also suitable for (rare) defects per unit
Processes made up of sub-processes	Gamma Poisson, negative binomial	Processes made up of sub-processes
Process with natural lower boundary as zero, e.g. cost, failure times	Weibull, lognormal, gamma and log-logistic	Flexible distribution
	Normal	For practical reasons due to the ease and/or availability of statistical tools.
	Folded normal distribution, half-normal distribution	Truncated distribution
Process has a natural lower boundary which is not zero	Weibull, lognormal, gamma and log-logistic with a third parameter representing the threshold or minimum value are candidates	Distributions with location shift from zero, and the threshold is only used if physically relevant.
Process generating symmetric data	Normal or logistic	Central limit theorem (CLT)

NOTE The table is not comprehensive and further justification can be found in other publications.

6.4 Collect, prepare and explore data

This section describes the necessary steps for collecting, characterising, categorising, cleaning and contextualising the data to enable its analysis.

The data may be generated by the process, as defined during the Define and Measure phases or may be gathered from a designed experiment.

After collecting the data, it is highly recommended to check it for completeness (non-missing values), errors or outliers, stability since these types of anomalies may distort the identification of distributions.