Contents

Page

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.

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 16269-4 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods*.

ISO 16269 consists of the following parts, under the general title *Statistical interpretation of data*:

— *Part 4: Detection and treatment of outliers*

— *Part 6: Determination of statistical tolerance intervals*

— *Part 7: Median — Estimation and confidence intervals*

— *Part 8: Determination of prediction intervals*

Introduction

Identification of outliers is one of the oldest problems in interpreting data. Causes of outliers include measurement error, sampling error, intentional under- or over-reporting of sampling results, incorrect recording, incorrect distributional or model assumptions of the data set, and rare observations, etc.

Outliers can distort and reduce the information contained in the data source or generating mechanism. In the manufacturing industry, the existence of outliers will undermine the effectiveness of any process/product design and quality control procedures. Possible outliers are not necessarily *bad* or *erroneous*. In some situations, an outlier may carry essential information and thus it should be identified for further study.

The study and detection of outliers from measurement processes leads to better understanding of the processes and proper data analysis that subsequently results in improved inferences.

In view of the enormous volume of literature on the topic of outliers, it is of great importance for the international community to identify and standardize a sound subset of methods used in the identification and treatment of outliers. The implementation of this part of ISO 16269 enables business and industry to recognize the data analyses conducted across member countries or organizations.

Six annexes are provided. Annex A provides an algorithm for computing the test statistic and critical values of a procedure in detecting outliers in a data set taken from a normal distribution. Annexes B, D and E provide the tables needed to implement the recommended procedures. Annex C provides the tables and statistical theory that underlie the construction of modified box plots in outlier detection. Annex F provides a structured guide and flow chart to the procedures recommended in this part of ISO 16269.

Statistical interpretation of data —

Part 4:
Detection and treatment of outliers

1 Scope

This part of ISO 16269 provides detailed descriptions of sound statistical testing procedures and graphical data analysis methods for detecting outliers in data obtained from measurement processes. It recommends sound robust estimation and testing procedures to accommodate the presence of outliers.

This part of ISO 16269 is primarily designed for the detection and accommodation of outlier(s) from univariate data. Some guidance is provided for multivariate and regression data.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
sample
data set
subset of a population made up of one or more sampling units

NOTE 1 The sampling units could be items, numerical values or even abstract entities depending on the population of interest.

NOTE 2 A sample from a **normal** (2.22), a **gamma** (2.23), an **exponential** (2.24), a **Weibull** (2.25), a **lognormal** (2.26) or a **type I extreme value** (2.27) population will often be referred to as a normal, a gamma, an exponential, a Weibull, a lognormal or a type I extreme value sample, respectively.

2.2
outlier
member of a small subset of observations that appears to be inconsistent with the remainder of a given **sample** (2.1)

NOTE 1 The classification of an observation or a subset of observations as outlier(s) is relative to the chosen model for the population from which the data set originates. This or these observations are not to be considered as genuine members of the main population.

NOTE 2 An outlier may originate from a different underlying population, or be the result of incorrect recording or gross measurement error.

NOTE 3 The subset may contain one or more observations.

2.3
masking
presence of more than one **outlier** (2.2), making each outlier difficult to detect