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

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.

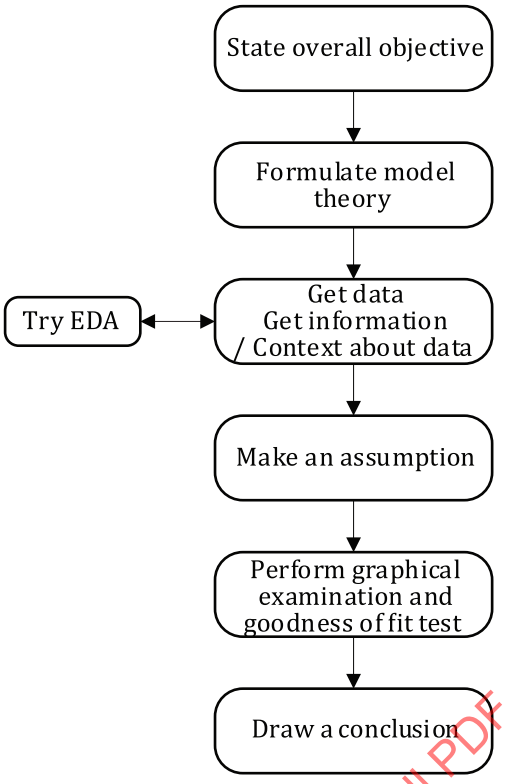


Figure 1 — General flow chart of the procedures for identification of distributions

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.

For missing data, one should decide whether to use imputation methods. The erroneous data must be removed or corrected, whilst for outlier detection and treatment one can refer to ISO 16269-4. Control charts are suitable for stability check.

One should use EDA methods for visualising data, exploring data patterns and suggesting hypotheses. It is always necessary to explore and to check the observed data against the model theory per step 2.

When EDA methods are combined with contextual information on the generating process, they are very helpful in selecting or narrowing down the convenient probability distribution, as well as confirming some of the distribution characteristics.

For example, information on distribution symmetry or unimodality can be found through EDA diagrams and then substantiated with the generating process (e.g., shift between two teams or machines).

6.5 Select underlying probability distributions

Based on the model theory developed in step 2 and the information resulting from the context and EDA in step 3, a set of candidate distributions can be obtained. This set can be refined for selecting the most convenient distribution by:

— refining the assumptions obtained from the model theory;

— narrowing down the set of distributions for practical choice due to the simplicity of the model and the analysis as well as the availability of statistical tools. Use the parsimony principle.

6.6 Perform goodness of fit test

Once a target distribution or a group of target distributions are identified, some specific methods both graphical and numerical are to be used for goodness of fit testing and distribution determination.

Graphical probability distribution plots enable assessing the distribution fitness by comparing the observed data distribution versus the model theory, including in the following situations:

— to assess whether an observed frequency distribution differs from a theorised model;

— to assess whether the observed values are not generated from a special distribution family;

— to fit certain proper distributions for the observed values.

As stated in Clause 5, for more accurate results, some proper statistical hypothesis testing methods for goodness of fit are necessary. These goodness of fit tests include the Pearson's χ^2 test, the Kolmogorov-Smirnov test, and many others.

6.7 Draw conclusions

The results obtained in step 5 may suggest a strong choice of data distribution or may not be convincingly conclusive. Based on the Six Sigma project objectives, the considered project phase and the information obtained from step 5 on the data distribution, the project team will need to make decisions such as:

a) select the most convenient distribution and move to the next activities within the phase or proceed to the next phase;

b) obtain more data or run further experiments;

c) review and analyse the process further;

d) refine and clarify the problem or objectives.

Options b), c) and d) would require performing again the above 6-step procedure totally or partially.

7 Examples

Some distinct examples of distribution identification are illustrated in the Annexes, which have been summarised in Table 3 with the different aspects indicated.

Table 3 — Example summaries listed by the Annexes

Annex	Example	Identification of distribution details
A	Test uniformity in the Super Lotto	Goodness-of-fit verification, R
B	Distribution of the number of technical issues found after product release to the field	Goodness-of-fit test, SAS-JMP
C	Software development effort estimation	Goodness-of-fit for response variable, Minitab
D	Determining the warranty period of a product	Goodness-of-fit for censored data, Minitab

Annex A
(informative)

Test uniformity in the Super Lotto

A.1 State overall objectives

Super Lotto is issued by a Lottery Management Centre. You either pick seven numbers by yourself from two separate pools of numbers: five different numbers from 1 to 35 (called the front zone), and two numbers from 1 to 12 (called the back zone), or let the computer pick all seven numbers for you. You win the jackpot by matching all seven winning numbers. The jackpot continues to grow until a ticket matches all seven numbers in the official drawing (regardless of the order in which the numbers are drawn) by machines with witness, then the player is a jackpot winner.

The Lottery Management Centre has a suspicion about the fairness of the Lotto Machine which has been reinforced by the compliance authority that regularly tests the machines and renews the yearly certificate. The Lottery Management Centre initiates an examination by implementing a Six Sigma project. During the Measure phase, this team reviewed historical data more precisely on the randomness (uniformity) about all the issued numbers. If the data are not uniform, the Lottery Management Centre may face serious consequences.

In what follows, only the front zone is considered.

A.2 Formulate a model theory

By design and by law, the machines are required to produce a uniform distribution of numbers. Any deviation from this is a violation. In each draw, 5 numbers are generated from 35 numbers (i.e. {1, 2, …,35}) at random without replacement. For randomness here, it is equivalent to say that all $\binom{35}{5} = 324\ 632$ combinations are equally likely, and the draws are independent.

A.3 Collect, prepare and explore data

There are 3 lottery draws per week. Each draw is uniquely identified by the issue number. Table A.1 illustrates an example of data for 3 draws corresponding to one week.

Table A.1 — Issued numbers for three consecutive terms (draws)

Issue No.	Date	Front zone
15070	06-20-2015	03 09 12 17 18
15069	06-17-2015	02 09 11 22 25
15068	06-15-2015	02 03 20 31 35

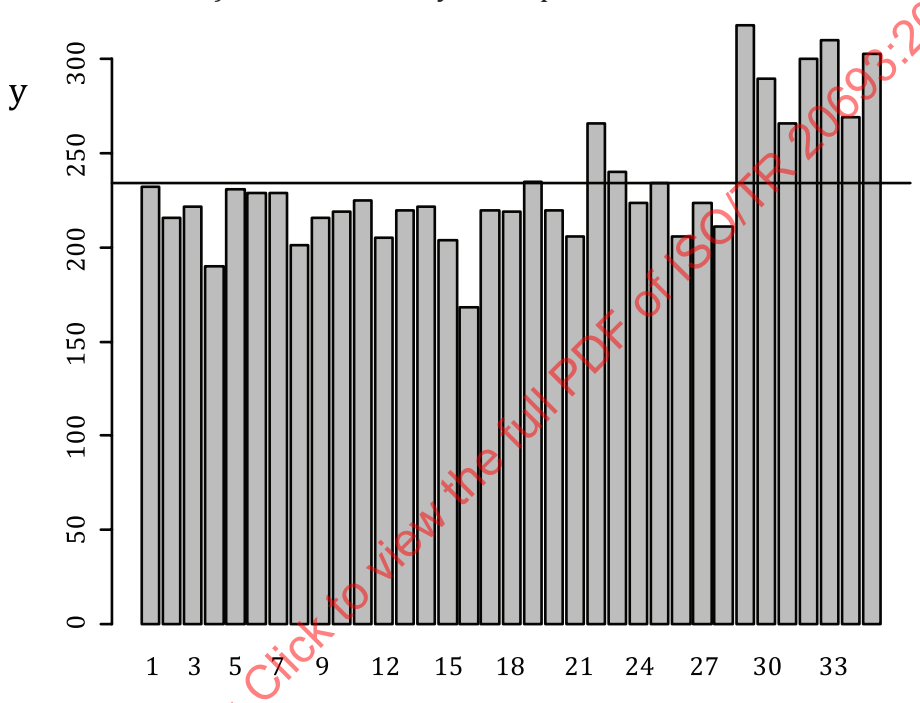
In this example, 1 638 draws were considered corresponding to the date 05/30/2007–01/22/2018.

Intuitively, one may plot the observed frequency of each number in 1 638 draws and compare it with its "expected" frequency (this is not accurate, see reason stated at the end of A.4). From the original data, observed frequencies can easily be obtained via R software. This is summarised in Table A.2.

Table A.2 — Observed frequency of each number

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Frequency	232	216	222	190	231	229	229	201	216	219	225	205	220	222	204	168	220
Number	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Frequency	219	235	220	206	266	240	224	234	206	224	211	318	290	266	300	310	269
Number	35																
Frequency	303																

The result can be visualised in the following barplot (Figure A.1) with its possible expected number (indicated by the horizontal line) under uniformity assumption.



Key

y frequency

Figure A.1 — Observed frequency and the possible expected frequency of all numbers

It seems that some of the frequencies appear to be far from their expected values, for example 16, 29, 33 and 35.

A.4 Select underlying probability distributions

From A.2, testing uniformity can be stated as the following hypothesis testing problem with

H_0: the probability of any set of five distinct numbers from {1,2, ..., 35} is $1/\binom{35}{5}$.

H_1: there is a set of five distinct numbers from {1,2, ..., 35} with probability not equal to $1/\binom{35}{5}$.

Since $\binom{35}{5}$ is too large compared to the total number of draws 1 638, it is impossible to check the above hypothesis directly.

However, as mentioned in [8] a uniform model has many implications and it can be checked whether these are as expected under a uniform model. For example, tests for uniformity of 1-combination, 2-combination or 3-combination (see detailed illustration in [8]) marginal distributions can be

performed. Because each draw is conducted without replacement, it should be noticed that the above null hypothesis H_0 is not equivalent to each number in {1,2, ..., 35} having the probability of occurrence 1/35. Thus, in A.3 quotation marks are added to "expected".

A.5 Perform goodness of fit test

For simplicity, only uniformity of 1-combination is considered. The corresponding random vector is $O = (O_1, O_2, ..., O_{35})$ where O_i stands for the observed frequency of number i. The adjusted χ^2 goodness of fit test statistic is

$$\chi^2_{\text{Adjust}} = \frac{35-1}{35-5}\frac{1}{E}\sum_{i=1}^{35}(O_i - E)^2$$

where E = 1 638 × 5/35. This statistic is asymptotically Chi-square distributed with 35 − 1 degrees of freedom. From this result, it is easy to find out the difference between draw with replacement and without replacement. χ^2_{Adjust} can be computed easily by R software. That is $\chi^2_{\text{Adjust}} = 203{,}089\ 5$. For significance level α = 0,05, the upper α-quantile of the Chi-square distribution with 34 degrees of freedom is 48,602 37. Thus one should reject uniformity of 1-combination with p-value less than 2.2e-16, and consequently the H_0 is rejected.

A.6 Draw conclusions

The null hypothesis is rejected. This means the results of the draws may not be as uniform as expected. One reason could be that the drawing machines didn't work well. Since the data is publicly available, one has the chance to select some more likely occurring numbers to become a winner. To check the latter belief, one would need to collect the data of all Lotto players.

Annex B
(informative)

Distribution of the number of technical issues found after product release to the field

B.1 State overall objectives

In the telecommunications industry, new platforms such as a new mobile device platform or a new base transceiver station (BTS) platform follow a very complex and lengthy pre-launch testing process before their "release" to optimise their success and their performance once introduced in the field. Many products are likely to be released for a given platform, with each one of these products going through a pre-launch process as well, somewhat simpler and shorter than the pre-launch process of a new platform, but still quite extensive. For a given product, many revisions consisting of small improvements to the product will also be launched according to a specified pre-launch process. This complex multi-level pre-launch process is very well documented and the corresponding products, be they a new platform, a new product within a platform, or a revision of hardware or software within a product, go through considerable testing (verification, validation, regression testing, etc.) before they are released.

The effectiveness and efficiency of the pre-launch process is assessed after release by the number of issues reported by customers to the technical support centre. Once a product is released, it is closely monitored during a period of one year to capture and fix as quickly as possible any issues that may have been overlooked during the pre-launch process. Thus, appropriate staffing needs to be allocated for that purpose. "Lessons learned" are developed as to why issues were missed — as well as why issues were introduced in the first place — and actions to prevent them in the future are taken.

At any point in time, there are around 200 products for a considered platform, all within one year from their release date, that are being monitored simultaneously by the technical support centre. Every technical issue found on these products is classified as a level A, B, or C, with level A being the highest priority and likely to cause a "stop shipment", level B being the next priority, and level C being the lowest priority as that category consists of issues not affecting the user directly. The number of technical issues at level A and B is tracked on an hourly basis and its distribution is modelled to provide information on the level of staffing required by the centre to ensure satisfaction of the customers.

The data of interest consists of the number of technical issues at level A and B received and logged in by the technical customer centre on an hourly basis. Only technical issues associated with products launched for less than a year are of interest in this study — it is rare to find such important technical issues, visible to the user, after the first year's release. This data is needed to estimate the staffing of the technical centre to ensure speedy and complete resolution of customer's issues.

B.2 Formulate a model theory

A platform is composed of multiple products similar to a process composed of sub-processes. Each product is behaving as a Poisson distribution, but as a platform due to the interaction among the products, it is not a Poisson distribution. It is rather a Gamma Poisson instead. For the sake of simplicity in this example, defect A and B are merged.

B.3 Collect, prepare and explore data

B.3.1 Collect data

The technical customer centre provided the log of all the technical issues that were reported during its 10 hour × 6 day weekly operation. For this study, the "sample" consists of all the technical issues of level A&B received over the last 2 weeks (120 h) and we assume that these 2 weeks are a "representative" sample of the past weeks and of the future weeks to come.

The data are shown in Table B.1.

Table B.1 — Number of first year post-release technical issues A&B logged in per hour

Number of first year post launch technical issues A&B per hour during validation week	Frequency
0	96
1	11
2	6
3	1
4	3
5	0
6	2
7	1
8	0
9	0
10	0

B.3.2 Prepare and explore data

The analysis of the data in Table B.1 was carried out with SAS JMP Pro Version 11.

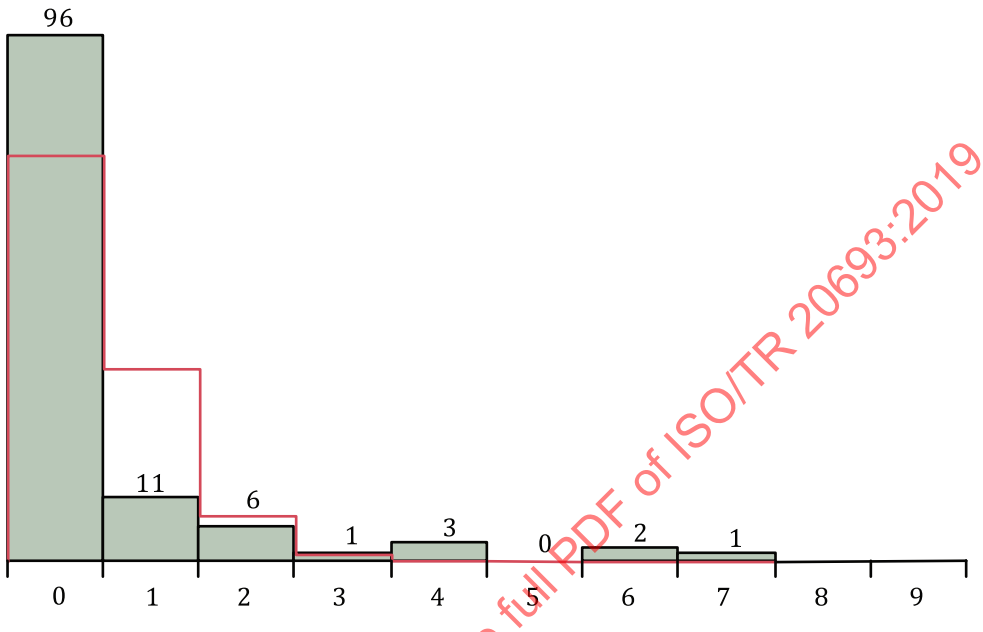
Summary statistics are provided below in Table B.2.

Table B.2 — Summary statistics

Mean	0,475
Std Dev	1,249 958
Std Err Mean	0,114 105
N	120

It is useful to look at the summary statistics, in particular at the mean (0,475) and standard deviation (1,25). A Poisson distribution has the characteristic that its mean and variance are equal. Here the sample variance is 1,562 4, somewhat larger than the mean, and this also raises a question as to whether the Poisson distribution will be a good fit for the data.

Figure B.1 shows the fitted Poisson distribution to the data, with mean set up to be equal to the sample mean. The red line is the fitted distribution. This reiterates the findings from summary statistics.



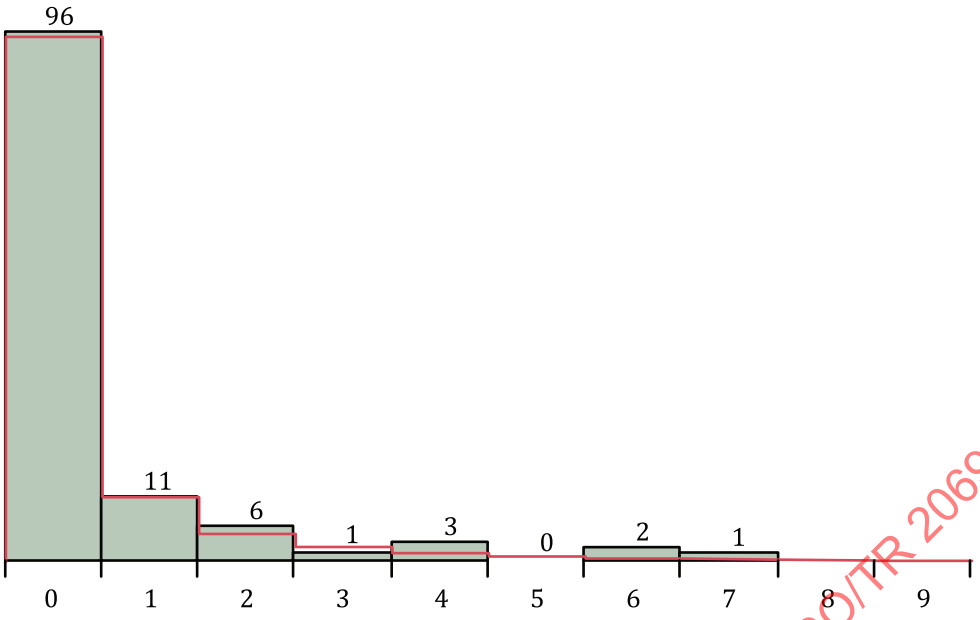
Key

Poisson (0,475)

Figure B.1 — Bar chart of number of technical issues A&B per hour and fitted Poisson distribution

Thus the Poisson distribution should not be used to make predictions about staffing level or improvement in the launch process. This confirms the model theory.

There are other distributions that can be fit to these data. In particular, in the case of count type of data and overdispersion (variance larger than the mean, which is the situation in this Annex), the Gamma Poisson distribution — also known as the negative binomial distribution—often is a better fit than the Poisson distribution as it arises from a mixture of Poisson distributions of which the means follow a Gamma distribution. If the number of technical A&B issues from each product follows a Poisson distribution but the means of the Poisson distributions for different products follow a Gamma distribution, then the resulting distribution of technical A&B issues which is being observed will follow a Gamma Poisson distribution.



Key

Gamma Poisson (0,475, 3,851 87)

Figure B.2 — Bar chart of number of technical issues A&B logged in per hour and fitted Gamma Poisson distribution

Figure B.2 shows the fitted Gamma Poisson distribution to the data. The red line is the fitted distribution. This gives much better fit than Figure B.1.

B.4 Select underlying probability distributions

From the model theory and the extensive exploration done for the data collected, the Gamma Poisson distribution is the preferred underlying probability distribution. To make it clear, this distribution identification problem can be summarised as the following formal hypothesis testing problem.

H_0: Number of technical issues A&B received per hour by the technical centre follows a Gamma Poisson distribution.

H_a: Number of technical issues A&B received per hour by the technical centre does not follow a Gamma Poisson distribution.

B.5 Perform goodness of fit test

There are a few cells in Table B.2 with expected frequency less than 5. To ensure the validity of the Pearson Chi-square test of goodness of fit for the Gamma Poisson distribution, we regrouped the cells with low cell count as shown in Table B.3.

Table B.3 — Regrouped number of technical issues A&B per hour and goodness-of-fit calculations for Gamma Poisson distribution

Regrouped number of first year post release technical issues A&B per hour	Frequency	Expected probability with regrouped cells	Expected frequencies for Gamma Poisson	Chi-square elements	Chi-square statistic	*p*-value
0	96	0,799	95,86	0,000		
1	11	0,099	11,82	0,057		
2	6	0,043	5,10	0,157		
3+	7	0,060	7,21	0,006		
					0,221	0,638

The large *p*-value 0,638 means that the null hypothesis — H_0: The data is from the Gamma Poisson distribution — is not rejected.

B.6 Draw conclusions

Both graphical visualisation and numerical testing show the Gamma Poisson distribution is acceptable and it should be used to make predictions about staffing level or improvement in the launch process.

Annex C
(informative)

Software development effort estimation

C.1 State overall objectives

A software development house is developing custom applications for customers. It has been struggling the last few years with estimating the software development effort, which had multiple impacts on the company's business. Many projects ran over budget, and/or slipped over their target delivery date leading to unhappy customers. Classically the development target delivery and cost have been decided in isolation by the sales team and the project plans were reversed by the software engineering team to meet the target. This has also led to frustration in the software engineering department and a very stressed and blame culture, leading to multiple resignations of senior software engineers and architects. Recently, the management took the issue seriously and hired a Six Sigma Black Belt (BB) with a strong software engineering background. The BB looked at multiple areas of improvement in their project management and estimation systems and provided an introduction to Six Sigma methodology.

The software development house has been recording the actual effort (time) for 34 different projects from various fields (MIS, WEB, TP and CRM), with their corresponding effort in person-hour and size in number of function points.

The aim of the project is to develop an accurate predictive model that will be used by the development team and bidding team alike for estimating the effort and then consequently deriving the cost, schedule, etc.

C.2 Formulate a model theory

The software engineering community has developed multiple model theories for software development effort estimation, e.g.:

— COCOMO model (empirical);

— Putnam-Norden-Raleigh model (empirical).

They all suggest an "exponential" relationship between effort (E) and size (S), which is supported by the Suze theorem $E = a \cdot S^b$, where a and b are parameters, which are a function of the process maturity, the technology, the team organization and skills (e.g., k1 and k2).

In addition, the software development process according to Table 2 is a process with a boundary different from zero, which suggests four candidate probability distributions: Weibull, lognormal, Gamma and logistic. Lognormal seems to be the most convenient choice. This is also consistent with the Suze theorem and the empirical models mentioned above.

C.3 Collect, prepare and explore data

C.3.1 Collect and prepare data

Table C.1 gives the necessary information about the variables involved in software development effort estimation.

Table C.1 — Name and definition of each variable

Field	Name	Definition
id	Project Id	Each project has a unique id
effort	Effort	Effort measured in hours
size	Application size	Measured in function points
app_type	Type of application	Management information system (MIS), transaction processing (TP), Web/online (WEB), customer relationship management (CRM)
case_tool	Case tool used	Case used for code generation Yes, No
k1	Staff application knowledge	1 = very low, 2 = low, 3 = nominal, 4 = high, 5 = very high
k2	Staff tools skills	1 = very low, 2 = low, 3 = nominal, 4 = high, 5 = very high

Data come from the software development house. Information about the total 34 projects is given in Table C.2, as follows.

Table C.2 — Raw data

id	effort	size	app_type	case_tool	k1	k2
2	7871	647	TP	No	4	4
3	845	130	TP	No	4	4
5	21272	1056	CRM	No	3	2
6	4224	383	TP	No	5	4
8	7320	209	TP	No	4	2
9	9125	366	TP	No	3	2
15	2565	249	WEB	No	2	4
16	4047	371	TP	No	3	3
17	1520	211	TP	No	3	3
18	25910	1849	TP	Yes	3	3
19	37286	2482	TP	Yes	3	1
21	11039	292	TP	No	4	2
25	10447	567	TP	Yes	2	2
26	5100	467	TP	Yes	2	3
27	63694	3368	TP	No	4	2
30	1745	185	WEB	No	4	5
31	1798	387	CRM	No	3	3
32	2957	430	MIS	No	3	4
33	963	204	TP	No	3	3
34	1233	71	TP	No	2	4
38	3850	548	CRM	No	4	3
40	5787	302	MIS	No	2	4
43	5578	227	TP	No	2	3
44	1060	59	TP	No	3	3
45	5279	299	WEB	Yes	3	2
46	8117	422	CRM	No	3	2
50	1755	193	TP	No	2	4
51	5931	1526	WEB	Yes	4	3

Table C.2 *(continued)*

id	effort	size	app_type	case_tool	k1	k2
53	3600	509	TP	No	4	2
54	4557	583	MIS	No	5	3
55	8752	315	CRM	No	3	3
56	3440	138	CRM	No	4	3
58	13700	423	TP	No	4	2
61	4620	204	WEB	Yes	3	2

C.3.2 Explore data

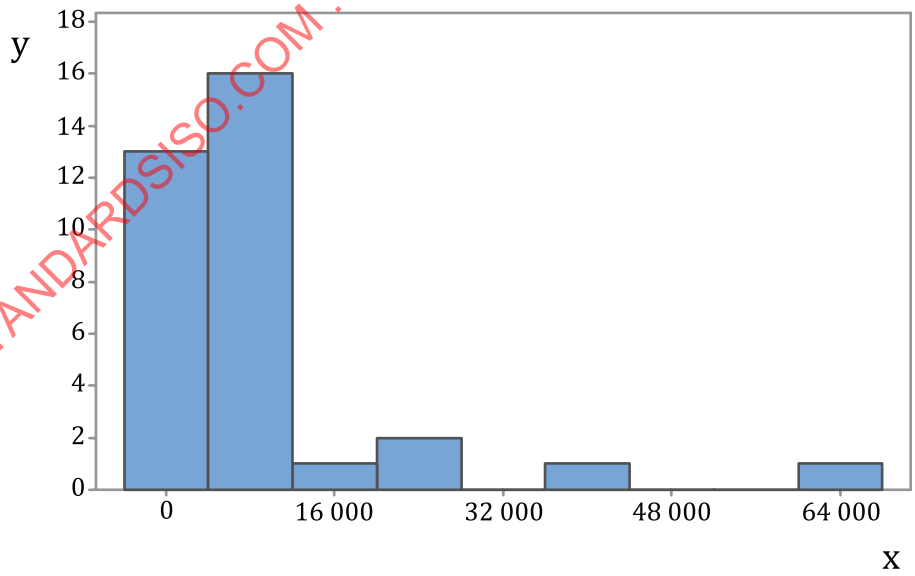
Firstly, completeness of the data is checked and the results are given in Table C.3.

Table C.3 — Data completeness check results

id	Count	effort	Count	size	Count	app_ type	Count	case_ tool	Count	k1	Count	k2	Count
2	1	845	1	59	1	CRM	6	No	27	2	7	1	1
3	1	963	1	71	1	MIS	3	Yes	7	3	14	2	11
5	1	1060	1	130	1	TP	20			4	11	3	13
6	1	1233	1	138	1	WEB	5			5	2	4	8
8	1	1520	1	185	1							5	1
...	...	...	...	...	...								
N=34		N=34		N=34		N=34		N=34		N=34		N=34	

The above result shows that there are no missing values.

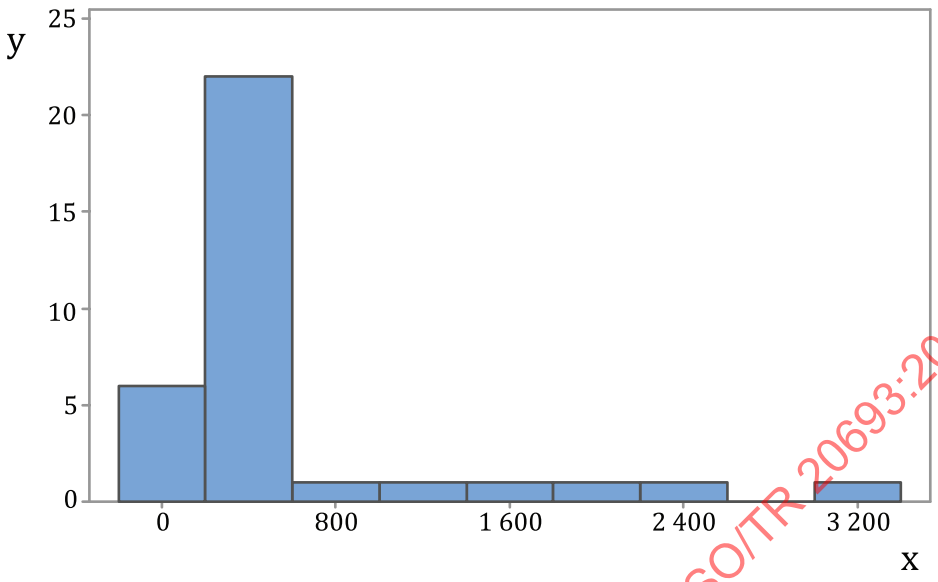
Secondly, histograms (see Figures C.1 and C.2) are given for the continuous variables, while boxplots (see Figures C.3 to C.6) of the response variable effort are conducted under each level of the categorical variables.



Key

x effort y frequency

Figure C.1 — Histogram of effort

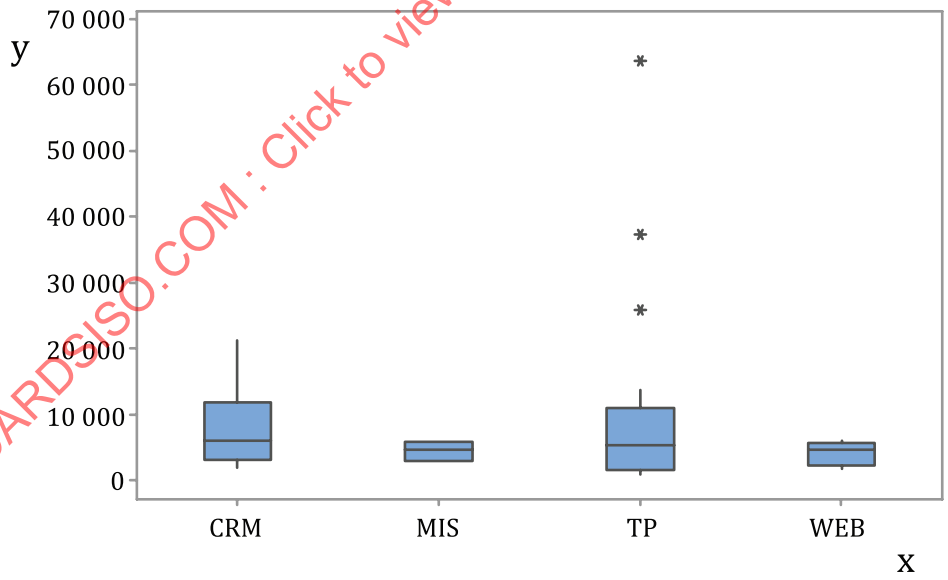


Key

x size y frequency

Figure C.2 — Histogram of size

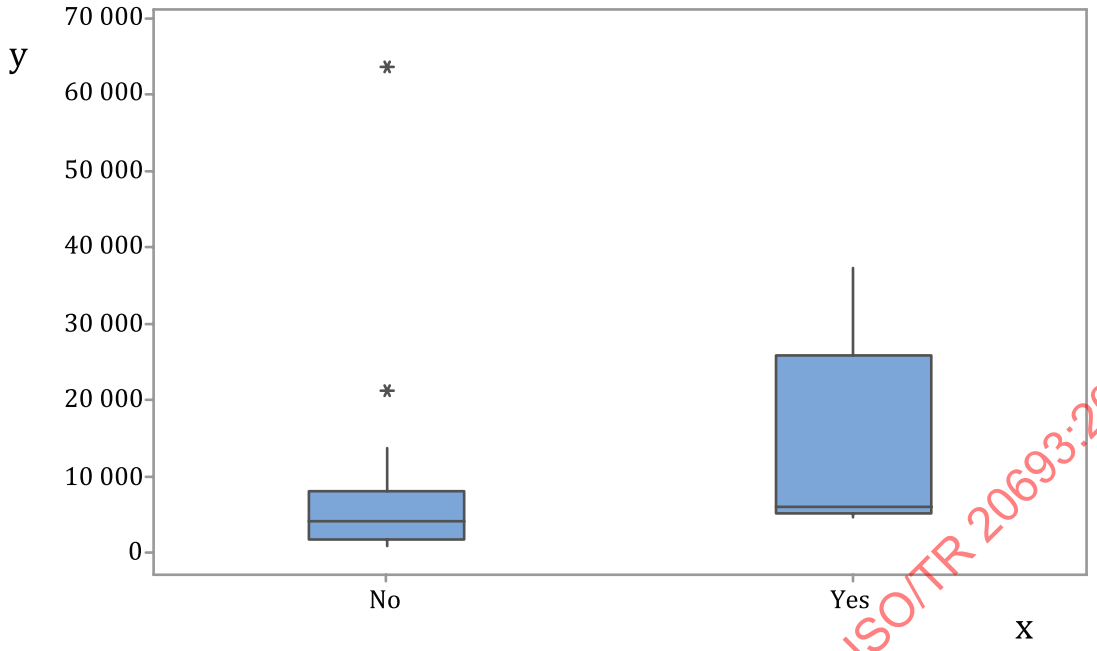
The distributions of effort and size show that they are not normally distributed. In fact, the database contains few projects with very high effort, or a very big size.



Key

x app_type y effort

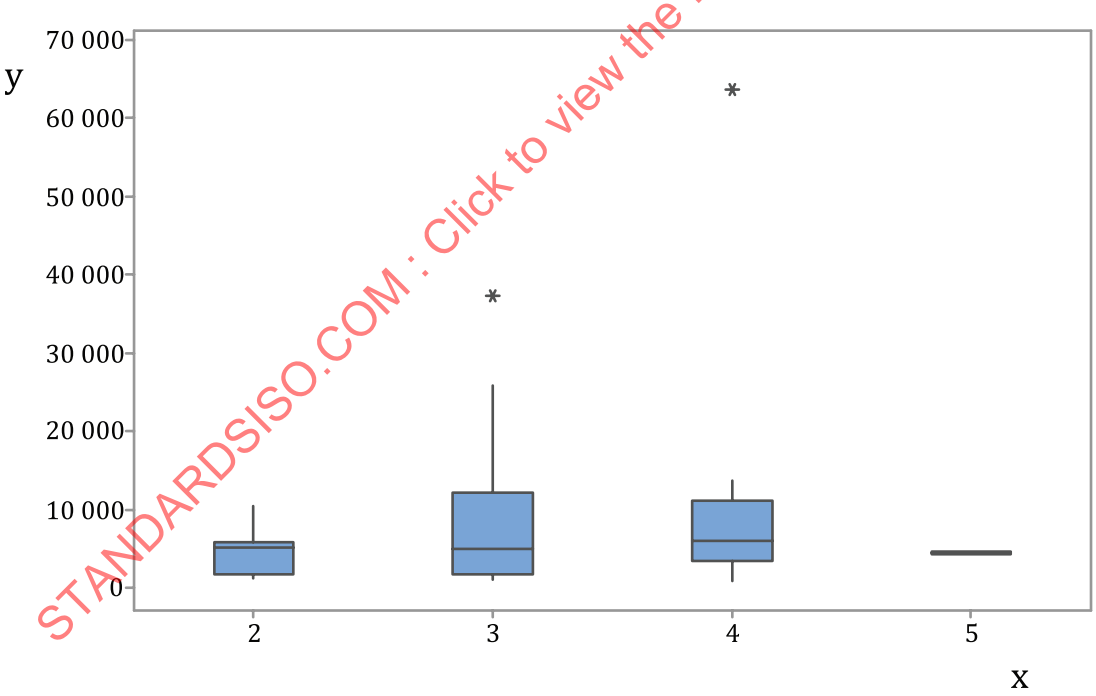
Figure C.3 — Boxplot of effort under each level of type of application



Key

x case_tool y effort

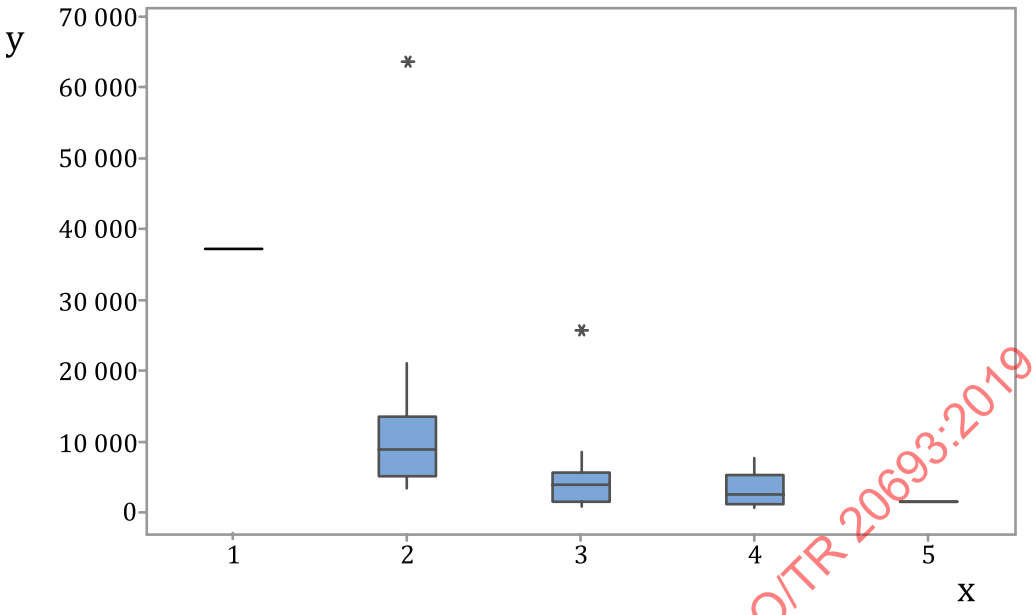
Figure C.4 — Boxplot of effort under each level of case tool used



Key

x k1 y effort

Figure C.5 — Boxplot of effort under each level of k1



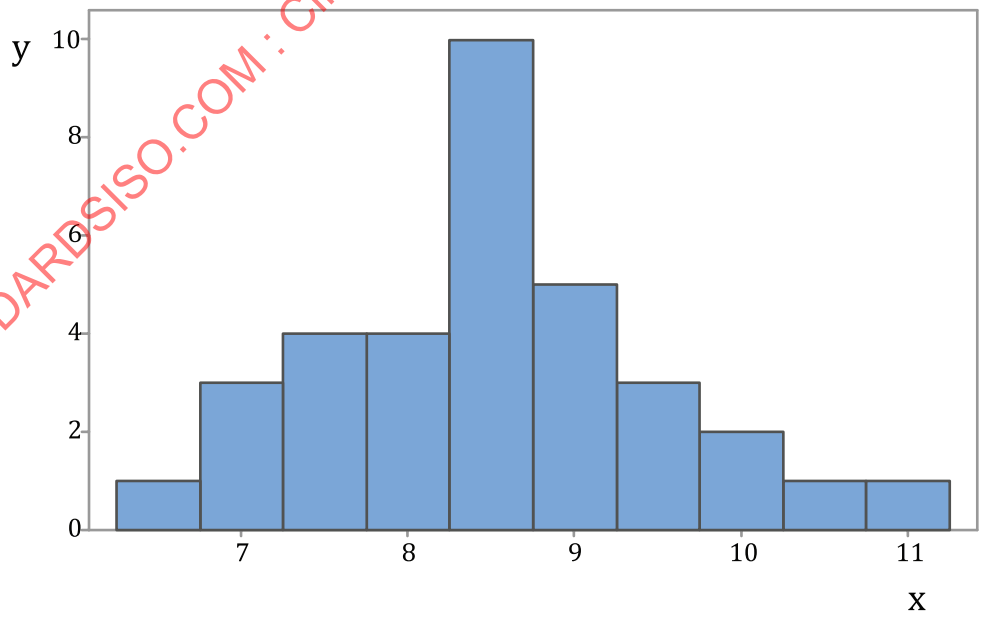
Key

x k2 y effort

Figure C.6 — Boxplot of effort under each level of k2

These boxplots show that effort is impacted by all these categorical variables. This in part provides support for the Suze theorem.

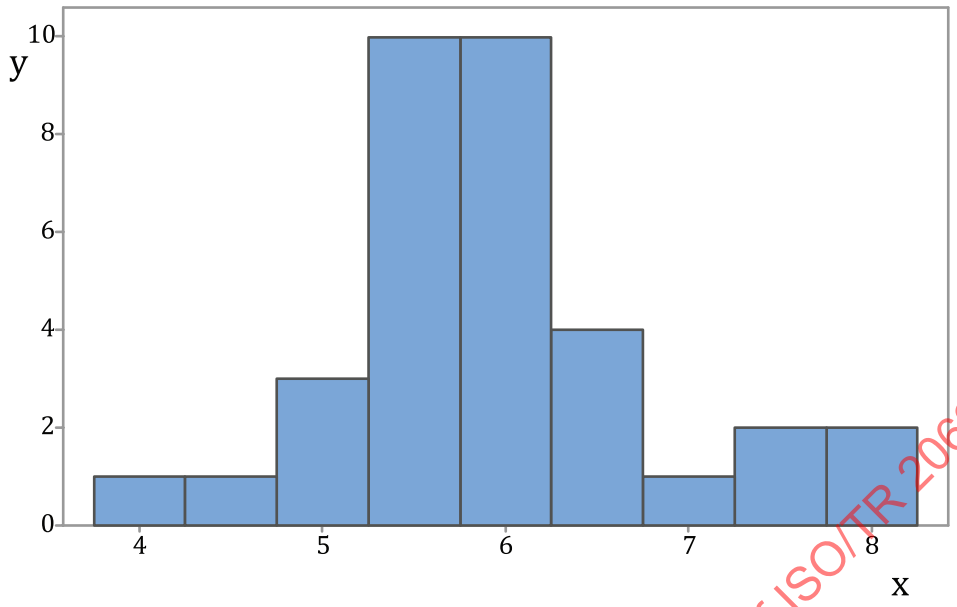
Finally, histograms of log transformation for effort and size are given (see Figures C.7 and C.8). From the shape of the histograms, it is easy to see that log effort is closer to the normal than log size. Because distribution of the dependent variable "effort" is more crucial to the estimation of the “exponential” model mentioned in C.2, only the distribution of effort is considered in the following two parts.



Key

x log effort y frequency

Figure C.7 — Histogram of log effort



Key

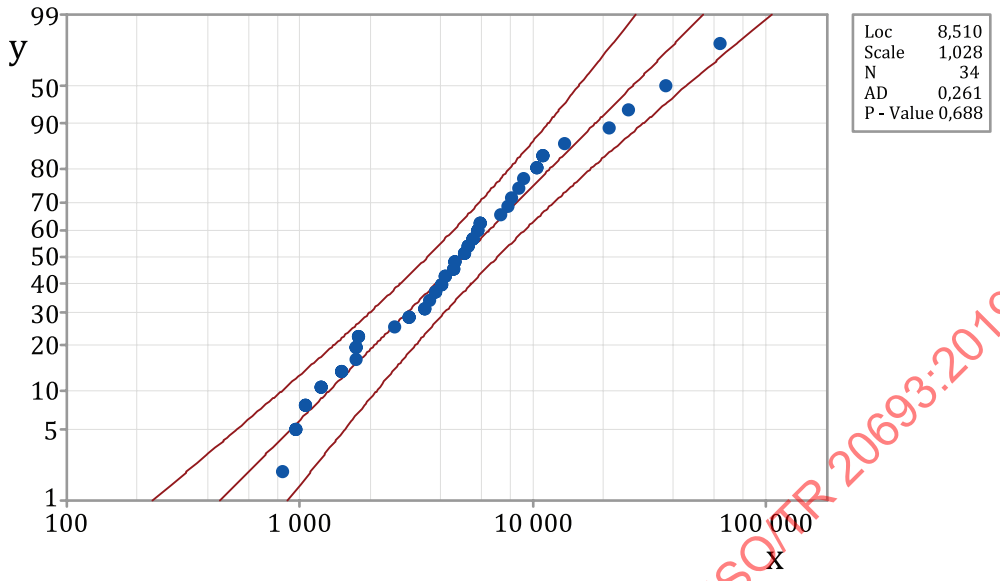
x log size y frequency

Figure C.8 — Histogram of log size

C.4 Select underlying probability distributions

Compared with gamma and logistic, lognormal is simple and the most convenient distribution for the "exponential" model. From Figure C.7, there is no evidence showing that the log effort deviates from normality. Thus lognormal is the first choice for effort variable. If lognormal is not accepted, then other distributions will be checked later.

C.5 Perform goodness of fit test



Key

x effort y percent

Figure C.9 — Lognormal probability plot of effort

Both the probability plot (see Figure 9) and the Anderson-Darling test show that the lognormal distribution is acceptable.

C.6 Draw conclusions

From C.5, for the effort prediction model $E = a \cdot S^b$, a log transformation is implemented to both sides and the least square method can be easily used to estimate parameters a and b. Substituting the estimates into the theoretical model, one gets the final empirical model. This final model will be used to predict the ongoing and upcoming software development projects. Such kinds of results can be used to optimise the resource allocation and improve the satisfaction of customers. Thus these type of models are important for the software industry. Understanding such models can provide competitive advantage.

Annex D
(informative)

Determining the warranty period of a product

D.1 State overall objectives

A consulting company is asked to deploy a Six Sigma project by a manufacturing company. This project intends to decrease the failure rate of its products and improve the quality of the product ultimately. The consulting company decides to choose a household appliance to do some pilot study. The technical manager hopes they can do some experiments and analyse the corresponding data. However, in general the life span of this product is quite long, thus accelerated life testing (ALT) is preferred to reduce the time to obtain sufficient data. Before this, they want to get some important information from the failure data collected from the market place by the customer service department. At this stage, one is interested in providing some support to the experiment design for the upcoming ALT. On the other hand, one also expects that such data will help product improvement and establish warranty periods for the product. For the first problem, the failure data will unveil some relationship between failure and environmental parameters such as temperature and pressure and so on. For the second problem, certain quantile estimation can be used to establish warranty periods. The process capability index can be used to evaluate the ability of the process producing the product. Here we focus on the second issue and prefer parametric methods, thus distribution identification for such data is a top priority.

For illustration purposes, only a part of the second problem, i.e. warranty period determination is considered.

The "LX" or "Bearing Life" refers to the time at which X % of items in a population will fail. It's particularly useful in establishing warranty periods for a product. The L10 life metric originated in the ball and roller bearing industry, but has now become a metric used across a variety of industries. L10 life is the time at which 10 % of units in a population will fail. Alternatively, one can think of it as the 90 % reliability of a population at a specific point in its lifetime — or the point in time when an item has a 90 % probability of survival. The L10 life metric became popular among ball and roller bearing makers due to the industry's strict requirement that no more than 10 % of bearings in a given batch fail by a specific time due to fatigue failure.

It's common to keep track of reliability field data in the form of number of items entered into market and number of items returned from a particular lot over time. When several lots are made at different dates and their corresponding returns noted, the recorded data are in the form of a triangular matrix.

D.2 Formulate a model theory

Since it is hard and in most time even impossible for customer service department to record the exact failure time of each product. In fact, the recorded data are in the form of a triangular matrix. That is to say, most of the exact failure times are unknown. What can be known is that the failure time belongs to an interval or right censored. There is little model theory for such censored data. But in reliability or survival analysis, empirical knowledge tells us that the exponential, lognormal, or Weibull, etc. are very common from physical phenomenon. Thus such kind of distributions will be considered.

D.3 Collect, prepare and explore data

The customer service department collected the data and stored them in the form of a triangular matrix. The raw data is given in Table D.1.