TECHNICAL REPORT

ISO/TR 18532

First edition
2009-04-15

Guidance on the application of statistical methods to quality and to industrial standardization

Lignes directrices pour l'application des méthodes statistiques à la qualité et à la normalisation industrielle

Reference number
ISO/TR 18532:2009(E)

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Foreword

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

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

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

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

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

ISO/TR 18532 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods*.

Introduction

This Technical Report demonstrates the advantages in the application of statistical methods in as simple and efficient a manner as possible so that they become accessible to the many rather than to the few.

As an introduction to the subject, three examples are given in Clause 4 to focus attention on some of the wider questions at issue. These examples suggest how statistical thinking coupled with the use of simple statistical tools and technical and operational knowledge of the process can help in improving designs, process efficiency and performance and product conformity to specification.

- Example 1, relating to the strength of wire, illustrates the role and value of division of data into so-called rational subgroups coupled with the use of cause and effect diagrams and line plots. It also shows how to exploit interrelationships between process parameters to achieve robust designs. The need is emphasized to treat numerical data not just as a set of figures but as potentially meaningful information on a process. It demonstrates clearly that an enquiring mind and sound judgement, coupled with an understanding of the actual process producing the numerical data, are required as distinct from a mere knowledge of statistical method. This indicates the need for non-statisticians to become more aware of the role of statistical method and to become more involved in their actual application to secure the maximum possible benefits to any organization.

- Example 2, on fabric mass, illustrates key aspects that need to be considered when sampling to establish conformance of an entity to specification. In this example, general conclusions are established by statistical theory and are turned to practical use.

- Example 3 concerns the mass fraction of ash (in %) in coal. Specifically, it demonstrates four principal concepts: how to handle apparent fluctuation of quality within a quantity of material; the need to determine, on a sound basis, the amount of sampling necessary to estimate the quality of a commodity; the necessity to establish, in advance, a well designed sampling procedure; and the value of progressive analysis of results, in a simple graphical manner, as they become available.

More generally, example 3 illustrates the importance of the application of statistical thinking and design method to a numerical study prior to it being undertaken. It also indicates that, to gain full benefit from such a study, persons familiar with the activity under scrutiny should be involved throughout.

Clause 5 introduces basic statistical terms and measures, and a wide range of simpler statistical tools used to present and analyse data. Emphasis has been placed on a pictorial approach that can most readily be communicated to, and understood by, the many.

Clause 6 describes the fundamentals of sampling on a statistical basis and distinguishes between statistical uniformity (stability of a process) and quality level (process capability). Clause 7 introduces sampling with reference to a product requirement. It draws out the two principal methods, *viz*. that of after the event acceptance sampling and that of the ongoing control of inherently capable processes. Clause 8 provides a detailed treatment of the statistical relationship between sample and batch. Clause 9 describes the methodology, terminology and rationale of acceptance sampling. Single, double, multiple, sequential, continuous, skip-lot, audit, parts per million, isolated lot and accept-zero plans for acceptance sampling by attributes are dealt with. Acceptance sampling by variables covers the following plans for individual quality characteristics: single sampling plans for known and for unknown standard deviation; double sampling plans; sequential sampling plans for known standard deviation and accept-zero plans. Multiple-quality characteristic plans are also described.

Clause 10 covers the fundamentals of statistical process control. It distinguishes between statistical process control and the use of statistical process control techniques for statistical product control. Over-control, under-control and control are discussed. The key steps in establishing and interpreting performance-based control charts that are intended primarily to differentiate between special and common causes of variation and provide a basis for capability and performance assessment are covered. The principal types of Shewhart-type control charts and the role and application of cumulative sum (CUSUM) charts are dealt with.

Clause 11 deals with performance benchmarking of stable processes under the heading of process capability assessment. Three very pertinent business questions are answered by a control chart: 1) is the process in control?; 2) what is the performance of the process?; and 3) is there evidence of significant improvement in process performance? Clause 11 focuses on answering the second question regarding process capability/performance of both measured data and attribute processes. It introduces the use of the internationally standardized capability indices, C_p, $C_{\text{pk}L}$ and $C_{\text{pk}U}$. It also discusses the business implications, in terms of aiming at preferred value and minimizing variation, with the quotation of minimum C_pm values, rather than the convention of tolerating maximum use of specified tolerances in determining whether or not an entity conforms to requirements.

Clause 12 begins by illustrating the role and value of simple economic experimental designs where the mathematical content is such that all the necessary calculations can readily be done manually. It then continues to exploit the development of computer software programs in the design and analysis of experiments. Nowadays the need for computational skills has become so minimal that the practitioner can concentrate his attention on choosing the right kind of design for a particular application, how to perform the experiment and how to interpret the computer outputs. In both cases, pictorial outputs are encouraged to facilitate understanding.

Clause 13 deals with the capability of measuring systems. Following a resumé of the basic statistical requisites of a measuring system that ensures the integrity of the data output, examples are given of the application of statistical method to the evaluation of resolution, bias and precision, uncertainty, repeatability and reproducibility.

Guidance on the application of statistical methods to quality and to industrial standardization

1 Scope

This Technical Report describes a broad range of statistical methods applicable to the management, control and improvement of processes.

2 Normative references

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

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

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

ISO 3534-3, *Statistics — Vocabulary and symbols — Part 3: Design of experiments*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 3534-1, ISO 3534-2 and ISO 3534-3 apply.

4 Illustration of value and role of statistical method through examples

4.1 Statistical method

The term “statistics” is commonly associated with an idea of lists of numbers, whether relating to output, costs, sales, prices or wages. It is thus advisable to make clear at the outset what in fact this statistical method is that may gainfully be applied in the field of quality and standardization. It is important to give some preliminary answers to certain questions. Why is statistical method needed at all? What does it consist of?

What kind of assistance can it give? Where can, when can, and should, it be applied? For this purpose, it has seemed best to deal first with the particular rather than the general, using specific examples to focus attention on the wider issues involved.

4.2 Example 1: Strength of wire

4.2.1 General

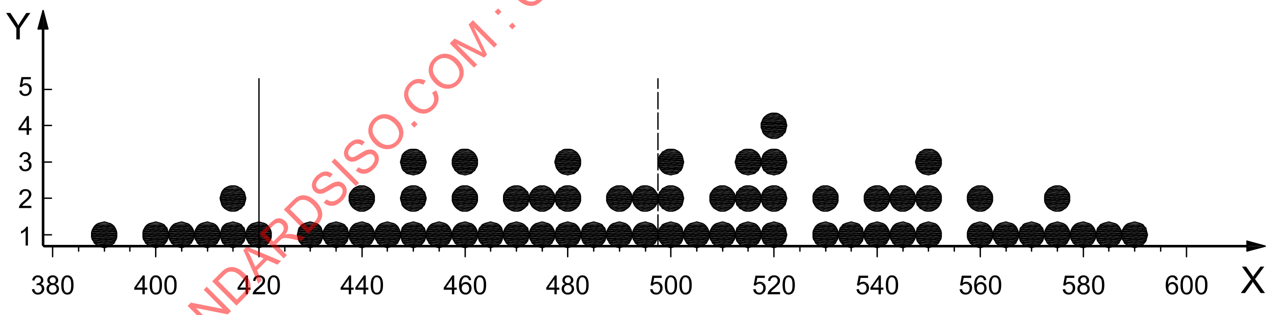
This example illustrates the role and value of the division of data into so-called *rational subgroups* coupled with the use of *cause and effect diagrams* and *dot plot*s. It shows their applicability to both problem solving and process and product enhancement. It also indicates the need to treat numerical data, not just as a set of figures, but as potentially meaningful information on a process. It demonstrates clearly that an enquiring mind and sound judgement, coupled with an *understanding of the actual process* producing the numerical data, are required, as distinct from a mere knowledge of statistical method. Hence, there is a need for technologists, technicians, and operational, administrative, marketing and management personnel to become more aware of the role of statistical method and become more involved in their actual application to secure the many benefits possible to any organization.

4.2.2 Overall test results and lower specification limit

Suppose 64 test results were obtained on the breaking strength of wire where the lower specification limit is set as 420 units. The results are shown in ascending order reading down the columns in Table 1 and as a dot plot in Figure 1.

Table 1 — 64 test results of wire breaking strength arranged in order from minimum to maximum (measurements were made to the nearest 5 units)

390	415	435	450	460	470	480	490	500	510	515	520	540	550	560	575
400	415	440	450	460	470	480	490	500	510	520	530	540	550	565	580
405	420	440	450	460	475	480	495	500	515	520	530	545	550	570	585
410	430	445	455	465	475	485	495	505	515	520	535	545	560	575	590
Mean = 495; median = 497,5; minimum = 390; maximum = 590.															



Key

X strength

Y number of observations

The dashed line at 497,5 represents the sample median and the full line at 420 represents the lower specification limit.

Figure 1 — Dot plot of breaking strength of 64 test specimens

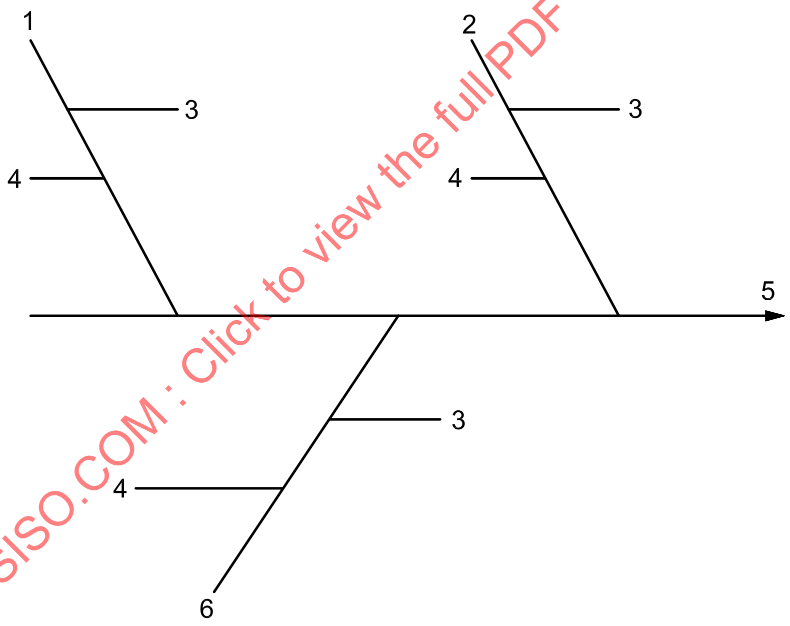
4.2.3 Initial analysis

It can be seen that 6 of the 64 test specimens have failed to achieve the 420 lower limit, although the mean is well above this at nearly 500; that is because there is a large amount of variation about the average. This is best indicated graphically in the form of the *dot plot* of Figure 1. (For the corresponding *line plot*, see Figure 12.)

It is obviously necessary to improve the quality of the wire, of which these are sample pieces, if the breaking strength is to be depended upon to always satisfy the minimum requirements of the specification. The pattern of variation is fairly symmetrical with a relatively large scatter. Whilst it may be possible to increase the mean strength, it is impossible to reduce excessive variation without some clue as to its main causes. If, on the other hand, some assignable (special) cause of variation can be located, it may be possible to take specific action both to increase the mean strength and reduce the overall variability. This will call for preliminary investigations into the causes to which the extreme variations may be assigned.

4.2.4 Preliminary investigation

This investigation would first require a consideration of such questions as the possible causes of variation in the wire strength. The outcome from a multi-disciplined team was the simple *cause and effect diagram* (see 5.3.13) as shown in Figure 2, which suggests a dependence of the wire strength on material composition and levels of steel and oil quench temperatures.

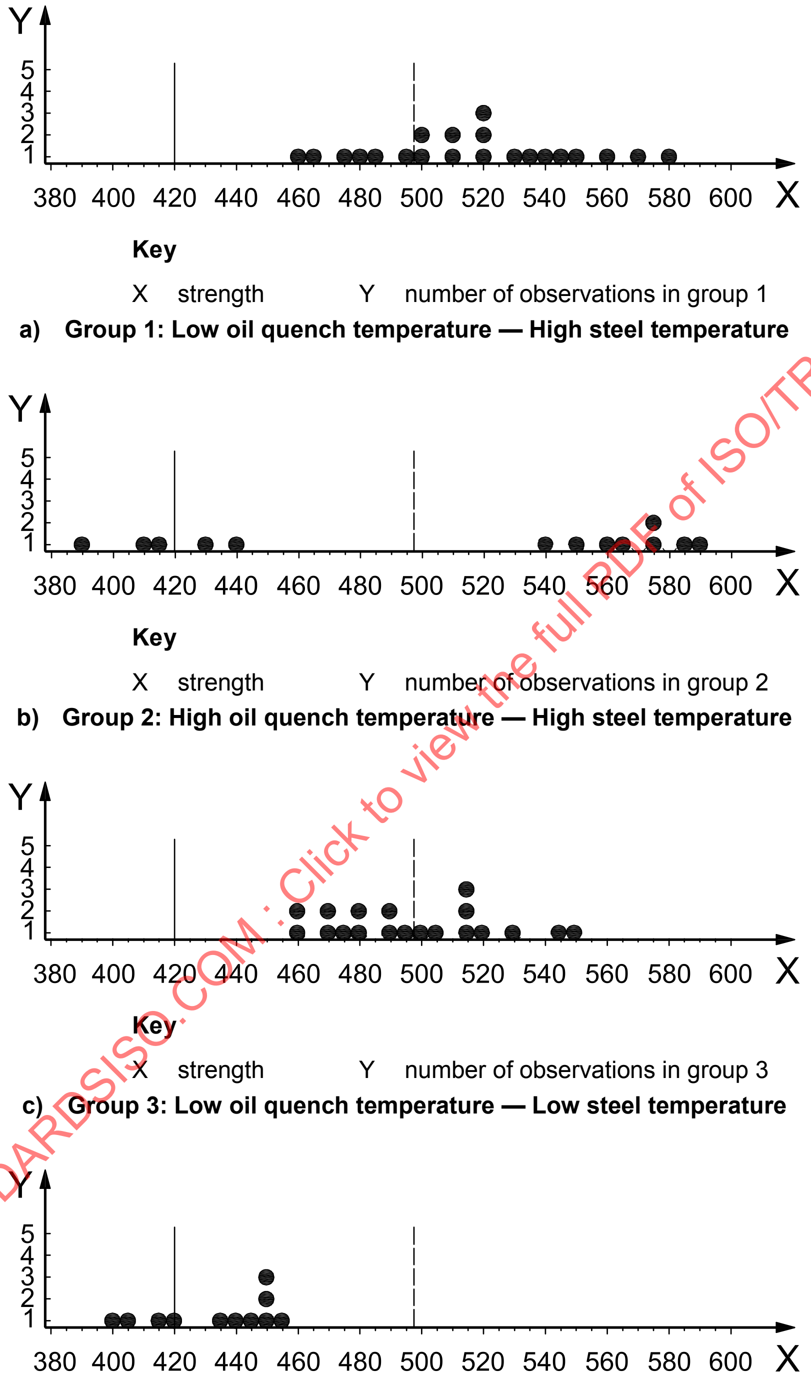


Key

1	steel temperature	4	low
2	oil quench temperature	5	strength
3	high	6	carbon content

Figure 2 — Basic cause and effect diagram for variation in wire strength (due to possible changes of material and process parameters within specified tolerances)

The next stage involves the division of the test records into a number of groups, within each of which all or some of these possible factors are roughly constant. This grouping, which is essential in any process of analysis, is described as division into *rational subgroups* (see 10.5.2). Suppose now that the 64 tests in the present example fall naturally into 4 subgroups, which is thought might be differentiated owing to changes in one or other of the factors suggested in Figure 2. The result is shown in the *dot plots* of Figure 3.



Key

X strength

Y number of observations in group 1

a) Group 1: Low oil quench temperature — High steel temperature

Key

X strength

Y number of observations in group 2

b) Group 2: High oil quench temperature — High steel temperature

Key

X strength

Y number of observations in group 3

c) Group 3: Low oil quench temperature — Low steel temperature

Key

X strength

Y number of observations in group 4

d) Group 4: High oil quench temperature — Low steel temperature

The dotted line in each subfigure at 497,5 represents the sample median of all 64 test results and the full line at 420 represents the lower specification limit.

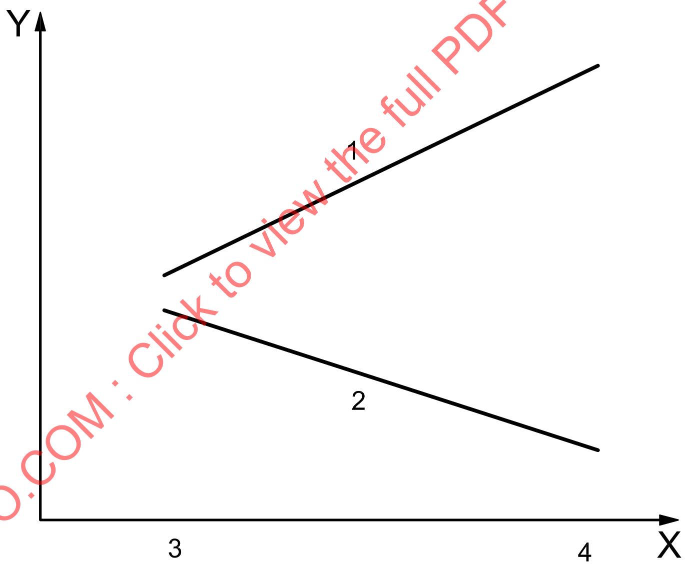
Figure 3 — Line plots showing patterns of results after division into rational groups

A study of Figure 3 indicates the following.

a) Group 1 results are similar to those of group 3. This suggests that strength does not appear to vary a lot at low oil quench temperatures even if the steel temperature varies. The technical expression for this is that the process is *robust* to steel temperature variation at low oil quench temperatures. The means are of the order of 500, or greater, and the minimum sample values about 460, compared with the minimum specification value of 420.

b) A study of group 2 and group 4 results appear to indicate a very different situation. Group 4 results are consistently low at, or around, the minimum specification limit. Group 2 results, on the other hand, are in two sets: one low set, with a mean below the specification limit, comparable with those of Group 4 and another contrasting set with an extremely high mean at about 570 with a relatively low variation.

A comparison of the records of these two sets indicated that the low set corresponded with operating conditions where the preset high steel temperature had inadvertently dropped to a low value for a short period. At a high quench temperature, the wire strength is extremely sensitive to variation in steel temperature and extremely low results, with a high proportion below the specification limit, may be expected at low steel temperatures, whereas at high steel temperatures the high quench temperature appears to yield a far superior strength performance with a mean of the order of 570 with relatively low scatter. The relationship (which was later confirmed by statistical experimentation, but is not reported in this Technical Report) is shown diagrammatically in Figure 4. Note that Figure 4 is not exclusively a summary of Figure 3 a) to d) but also draws on the results of extra experimentation.



Key

1 high steel temperature
2 low steel temperature
3 low
4 high

X oil quench temperature
Y strength

Figure 4 — Diagram indicating the effect of the interrelationship between oil quench temperature and steel temperature on wire strength

Now is decision time. How should this process be run to be likely to ensure uniform strengths of wire which do not contravene the lower specification limit? The choice depends on operational, economic, marketing and statistical considerations.

Option 1 is to run at low quench temperatures, which would be expected to give results similar to those of groups 1 and 3. Due to the predicted value of the mean and the pattern of variation, there would be some chance that occasionally the minimum specified strength may not be achieved. Variation in steel temperature between high and low would then be anticipated to have little impact on wire strength. Certain economies might be achieved using this option, by running with a lower steel temperature or a lower level of control of steel temperature.

Option 2 is to grasp the opportunity to achieve a relatively high mean wire strength with low variability by running the process with both a high steel and oil quench temperature. This may increase the process cost but it would ensure wire strength conformance to specification. It would also, perhaps, be appropriate to seek marketing advantage by improving the grade and increasing the price of the wire. However, the wire strength is seen to be particularly vulnerable to drops in steel temperature at high settings of oil quench temperatures. It is vitally important if this option is chosen to place strict controls on steel temperature.

4.2.5 General discussion on findings

This example has been used to suggest how simple statistical tools, coupled with technical and operational knowledge of the process, may help in improving process efficiency and performance and product conformity to specification. They provide powerful analytical and communicating tools and, at the same time, assist in determining, on a sound basis, simple routine checks on the efficiency of technical control.

Certain questions are posed. Should material of such great variation in strength be sold under the same specification? What is the relative cost to produce wire under some process parameter settings rather than others? Supposing that such variety is not desirable, what should be the best standard to aim at, having regard to the needs of the user and the obstacles to be overcome by the producer? Should the strength specification be modified, either downwards to encourage the attainment of the standard, or upwards following improvement in process settings and control, to increase the grade and price? To what extent are other product characteristics, such as hardness and brittleness, related to strength? Are trade-offs between one and the other involved?

In addition to these points, there is one more closely connected with statistical theory. The mere statement of means and minimum and maximum sample test strengths and the graphical display of the results, in the form of a dot or line plot, do not really provide measures of variation adequate for numerical prediction of the ability of the process to produce wire strengths conforming to standard. A number of other statistical aspects need to be considered, such as the stability of the process in relation to wire strength and the fitting of a probability distribution to the pattern of variation of the results (such as is shown later for a different example in Figure 26).

4.2.6 Explanation of statistical terms and tools used in this example

Rational subgroup: a group in which data is so organized through classifying, grouping or stratifying as to ensure the greatest similarity among the data within each subgroup and the largest difference between subgroups. The aim of rational subgroups is to include only *common causes of variation* within a subgroup with all *special causes of variation* occurring between subgroups. The object is to discriminate more readily between common and special cause variation in sets of data.

Knowledge and information, obtained through theory, experimentation or experience of the process, typically form the basis of the selection of rational subgroups. For example, in the administrative area, historical data on late payments could be grouped by account, account supervisor, product or by intervals of time. In a production process, the maximum homogeneity within a subgroup is frequently obtained by making up rational subgroups from consecutively produced parts taken from the same location or machine. For example, five consecutively produced parts from one machine may be taken every hour. It is then possible to segregate special causes of hour to hour variation, identified from subgroup to subgroup variation, from the inherent sources of common cause variation within a subgroup.

Common causes of variation: source of variation that is inherent in a process. It relates to those sources of *natural* variation in a particular process. For example, a turret capstan may produce to 0,25 mm, a grinder to 0,025 mm and a hand lapper to 0,002 5 mm; an investment casting may produce to 0,2 mm per metre and a sand casting to 0,8 mm per metre. Hence, only people responsible for the system can often reduce common cause variation. The variation is predictable in a process subject only to common cause variation.

Special causes of variation: source of intermittent variation in a process. A special cause arises because of specific circumstances that are not always present. For instance, it could be irregular (e.g. power surge), progressive (e.g. tool wear) or stepwise (e.g. change in datum of a gauge, or change in setting). As such, in a process subject to special causes, the magnitude of the variation from time to time is unpredictable. The presence of special cause variation is found using a statistical process control (SPC) chart by operational people, those who work in the system.

Dot/line plot: the frequency of readings at each measurement is shown by dots/lines built up vertically on a horizontal axis representing the scale of measurement. It can be used to compare or contrast, graphically, the pattern of variation of data both within a rational subgroup and between subgroups. It is particularly useful when working with limited sets of data.

Mean (arithmetic mean): sum of the values of the observations divided by the number of observations.

Median: value of a variable characteristic, which is greater than one half of the observations and less than the other half (the middle, or mid-value). In the case of an even number of observations, it is usual practice to set the sample median equal to the arithmetic mean of the central two observations after arranging the observations in ascending order.

Cause and effect diagram: frequently called a fishbone diagram (because of its shape) or an Ishikawa diagram (named after its creator, see Reference [101]). It applies where it is required to explore and display causes of a specific concern, problem or condition. The concern (effect) is shown on the right of a main horizontal spine. Possible categories of causes of the concern are shown on main branches from the spine. Subcategories are indicated on subbranches.

4.3 Example 2: Mass of fabric

4.3.1 General

This example illustrates a form of problem that arises in sampling anything, for example, a product or material, to determine whether or not it conforms to specification. It suggests the importance of establishing the relationship between size of sample and the precise rules to be laid down for acceptance or rejection, based on the resulting tests or measurements.

Specifications may be one-sided, with either a minimum (e.g. strength) or maximum value (e.g. eccentricity) quoted, or two-sided, with both a minimum and maximum given (e.g. assembly component dimension). When measurements are taken of successive results from however stable and precise a process, it cannot be supposed that the results will be identical. Some variation will be evident if the resolution of the measuring device is appropriate. Consequently, to obtain any adequate appreciation of the quality of the particular characteristic in question, a number of results need to be obtained. Furthermore, it is not only the resulting average that is of importance, but also the uniformity as measured by the variation about that average.

It follows that, in using a series of sets of measurements to check for conformity to a specification, it is helpful to take into account the following:

a) the relationship between average values, minimum values, range of variation, etc.; together with

b) the manner in which these are dependent upon the actual number of measured values taken.

4.3.2 Test results and specification limits

It is possible to illustrate the nature of the problem using the data given in Table 2. The figures represent the masses of specimens taken from a roll of fabric. They have been grouped for purposes of illustration into 32 samples of 4 specimens each.

Table 2 — Masses of 128 specimens from a roll of fabric — Minimum specification limit = 98 dg = 9,8 g

Sample number	Mass of specimens (1 dg)				Sample number	Mass of specimens (1 dg)			
	1	2	3	4		1	2	3	4
1	101	99	100	102	17	100	97	91	92
2	106	98	101	99	18	106	100	102	100
3	98	101	102	100	19	97	97	94	99
4	103	104	95	96	20	99	101	100	101
5	96	97	100	96	21	100	101	95	103
6	101	96	97	97	22	101	99	99	99
7	109	100	106	101	23	94	96	94	98
8	92	97	100	95	24	99	100	104	108
9	104	102	95	100	25	102	100	105	98
10	98	101	99	107	26	99	98	103	97
11	99	98	99	99	27	97	98	106	104
12	109	101	105	102	28	97	101	108	99
13	95	94	97	100	29	100	97	100	98
14	102	100	100	95	30	104	103	104	100
15	97	101	102	98	31	105	99	103	103
16	103	101	99	100	32	98	104	102	103

Figure 5 illustrates the data of Table 2 in 3 ways, in relation to sample means:

a) the first 32 samples each have 4 masses (as shown in Table 2);

b) samples 33 to 48 relate to the same data in Table 2, which has now been combined into 16 samples each of 8 masses with sample 33 being the union of samples 1 and 2, etc.; and

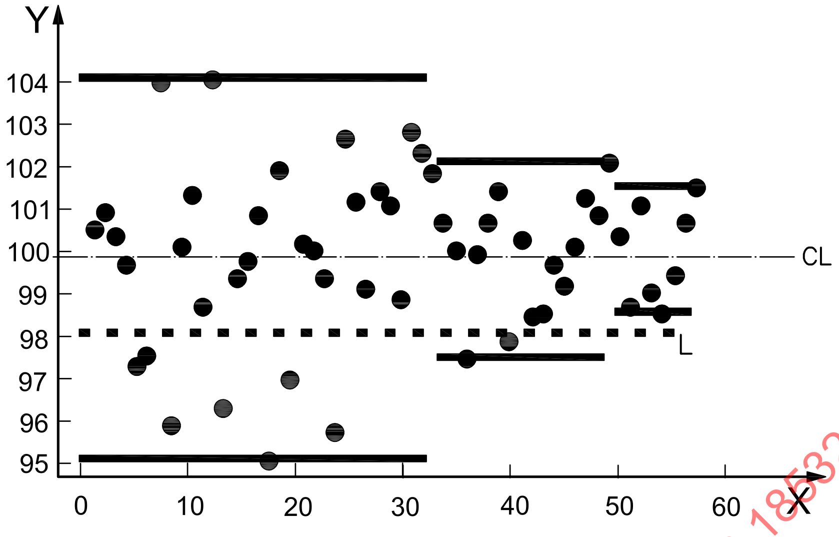
c) samples 49 to 56 relate to the same data in Table 2, which has now been combined into 8 samples each of 16 masses with sample 49 being the union of samples 1 to 4, etc.

Figure 6 illustrates the data of Table 2 in 3 ways, in relation to sample ranges (see 5.2 for definition of range) rather than sample means. As in Figure 5:

1) the first 32 samples each have 4 masses (as shown in Table 2);

2) samples 33 to 48 relate to the same data in Table 2, which has now been divided into 16 samples each of 8 masses with sample 33 being the union of samples 1 and 2, etc.; and

3) samples 49 to 56 relate to the same data in Table 2, which has now been divided into 8 samples each of 16 masses with sample 49 being the union of samples 1 to 4, etc.



Key

X sample number

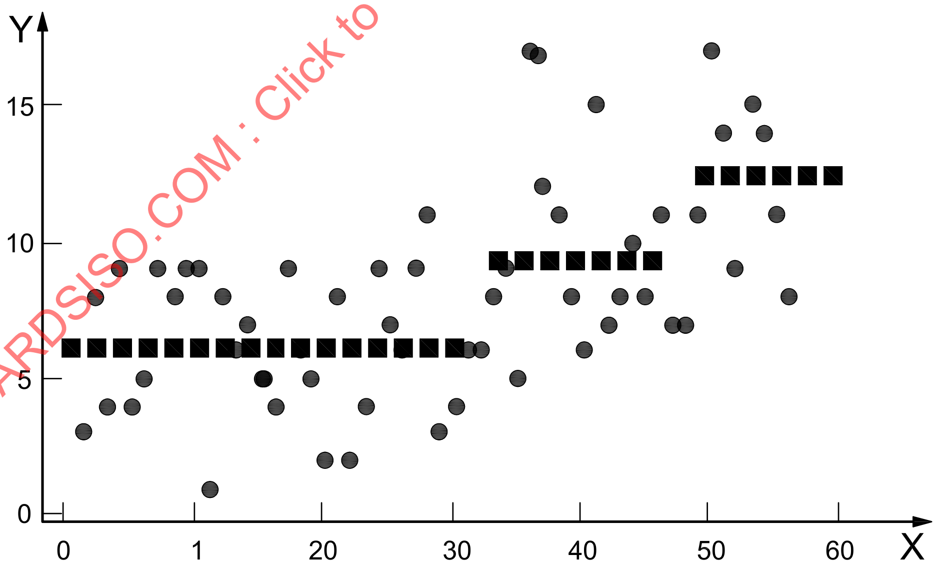
Y mass in units of 1 dg

NOTE 1 Bold horizontal lines indicate maximum ranges of means for sample sizes of 4, 8 and 16.

NOTE 2 L = lower specification limits

NOTE 3 CL = centreline

Figure 5 — Means of masses plotted against sample number (illustrating decreasing variation in the mean with the sample size increase)



Key

X sample number

Y range in units of 1 dg

NOTE Dotted lines show increase with sample size in mean range within a sample.

Figure 6 — Ranges of masses within each sample vs sample number [illustrating increasing (range) variation within a sample with sample size increase]

4.3.3 Discussion of specific results

Attention is drawn to the following points, brought out by examination of Figure 5.

a) The mean (Figure 5).

Sample size	Range of means (in units of 1 decigram)
4	95 to 104 = 9
8	97 to 102 = 7
16	99 to 101 = 2

The conclusion is that as the number of values upon which the mean is based becomes larger, the variation in the mean becomes smaller.

b) The range (Figure 6).

Sample size	Average range (in units of 1 decigram)
4	6
8	10
16	12

The conclusion is that the range of variation within a sample *increases* with the number of values in the sample.

c) Conformance to specification.

Suppose that the minimum mass, the lower specification limit (LSL), were to be set at 98.

1) If this criterion is applied to the *mean value* in a sample, then:

i) 7 of 32 samples of 4;

ii) 1 of 16 samples of 8;

iii) 0 of 8 samples of 16

would fail to meet the criterion.

2) On the other hand, if this criterion is applied to the *smallest value* in the sample, then:

i) 15 of 32 samples of 4;

ii) 12 of 16 samples of 8;

iii) 8 of 8 samples of 16

would fail to meet the criterion.

This example is discussed further in 8.3.

4.3.4 Discussion on general findings

Without placing undue emphasis on figures that are based on a single sample, the following general conclusions can be established by statistical theory and turned to practical account.

a) As the number of values or tests becomes larger, the (absolute value of the) expected difference between the mean of one set of tests and that of another becomes smaller.

b) As the number of values or tests becomes larger, their expected range of variation also becomes larger.

c) A statement by way of specification that the lower specification limit = 9,8 g, say, is inadequate unless supplemented by

   1) information concerning the number of specimens to be tested,

   2) whether or not something conforms to specification.

Without this information, it would not be possible to know whether something conforms or does not conform to specification.

4.4 Example 3: Mass fraction of ash (in %) in a cargo of coal

4.4.1 General

This example illustrates four principal concepts:

- the handling of apparent fluctuation of a quality characteristic within a quantity of material (or alternatively with time);
- the need to determine, on a sound basis, the extent of sampling necessary to estimate the quality characteristic of a commodity;
- the necessity to establish, in advance, a well designed sampling procedure based on sound, but basic, statistical principles;
- the value of progressive analysis of results, in a simple graphical manner, as they become available.

These concepts are explained by reference to bulk sampling. However, they are applicable generally to all kinds of sampling.

Sampling of bulk commodities, for example, particulates, liquids and gases, can be classified into two types:

a) sampling to make a decision on lot acceptance/rejection (see ISO 10725 [39]);

b) sampling to make an estimation of the average quality of a particular characteristic of the bulk commodity. (See ISO 11648[43].)

Illustrations of the application of b) include sampling of chemical products such as those in liquid state, cokes, ferroalloys and cements; agricultural products such as grains and flours; minerals and liquid state petroleum products. Sampling may take place on moving streams or in stationary situations such as stockpiles, silos, wagons and holds of ships and barges.

For this particular example, the quality characteristic chosen is the mass fraction of ash (in %) in a ship's cargo of coal. The aim is to estimate the average (arithmetic mean) value. The prime purpose of such sampling is, typically, to obtain an appreciation of the quality characteristic of the bulk of the fuel as a basis for determining the price to pay for the consignment.

4.4.2 Test results (reference ISO 11648-1: *Statistical aspects of sampling from bulk materials*)

Table 3 shows the test results from a series of 20 lots of coal being unloaded from a ship. For each lot, eight samples of coal were drawn and the mass fraction of ash (in %) measured.

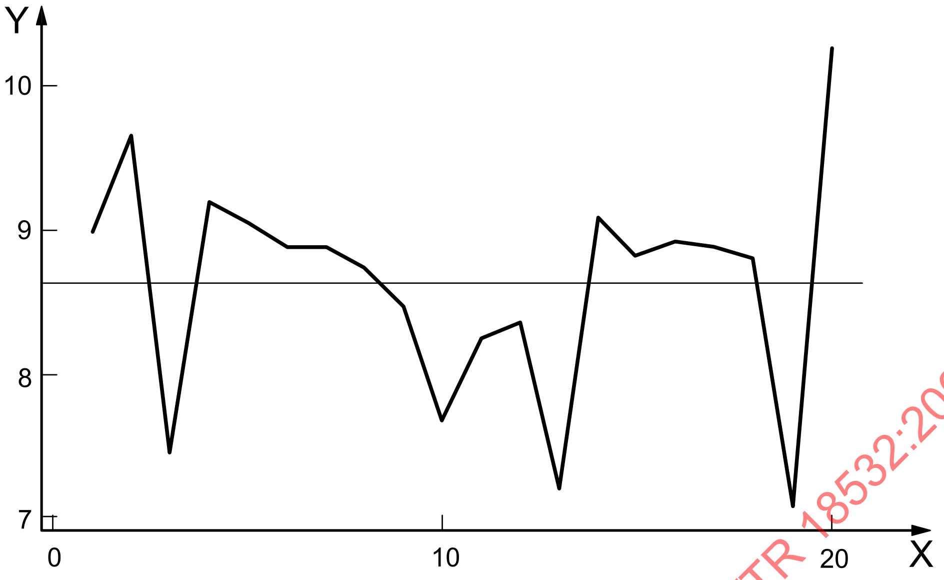
Table 3 — Mass fraction of ash (in %) measurement results by lot from ship's cargo on unloading

Lot no.	Result 1	Result 2	Result 3	Result 4	Result 5	Result 6	Result 7	Result 8
1	9,38	9,24	9,02	8,98	9,22	9,32	8,40	8,38
2	9,76	9,80	9,92	9,92	9,36	9,36	9,72	9,54
3	7,40	7,26	7,32	7,40	7,55	7,61	7,57	7,49
4	8,62	8,76	8,82	8,84	9,20	9,34	10,00	10,00
5	9,16	9,18	8,72	8,68	8,89	8,75	9,51	9,47
6	9,08	9,08	9,06	8,86	8,80	8,84	8,76	8,60
7	8,77	8,69	8,77	8,75	9,16	8,92	9,06	8,94
8	8,62	8,68	8,80	8,42	8,78	9,02	8,62	8,94
9	8,60	8,74	7,10	7,22	8,88	9,10	9,08	9,00
10	6,96	7,20	7,32	7,40	8,59	8,89	7,55	7,43
11	8,44	8,26	7,92	7,70	8,65	8,45	8,37	8,15
12	8,24	8,00	8,38	8,12	8,42	8,26	8,78	8,72
13	7,21	7,25	6,85	7,03	7,21	7,31	7,31	7,39
14	8,84	9,00	8,96	8,90	9,24	9,16	9,20	9,38
15	8,45	8,51	8,91	8,79	9,00	9,06	8,86	8,96
16	9,02	9,08	9,16	9,08	8,75	8,83	8,65	8,75
17	8,71	8,77	8,75	8,75	8,98	8,96	9,00	9,18
18	8,77	8,92	9,24	9,32	8,82	8,64	8,32	8,42
19	7,37	7,39	7,13	7,25	7,10	6,92	6,64	6,74
20	10,12	10,02	9,96	9,94	10,72	10,78	10,30	10,30

4.4.3 Initial graphical analysis of specific results

A plot of the averages of mass fraction of ash (in %) of the coal by lot is shown in Figure 7. The points have been joined by straight lines to aid visualization of the variation.

Fluctuation is observed about the overall average of 8,63 % ash content. Suppose 8,6 % is taken to represent the ash content in the whole consignment. A practical question would then be how many sets of tests need to be made before it would be reasonable to estimate this measure within, say, ± 1 % of its value (i.e. approximately 8,5 to 8,7). This question can be answered by reference to the progressive average plot in Figure 8.



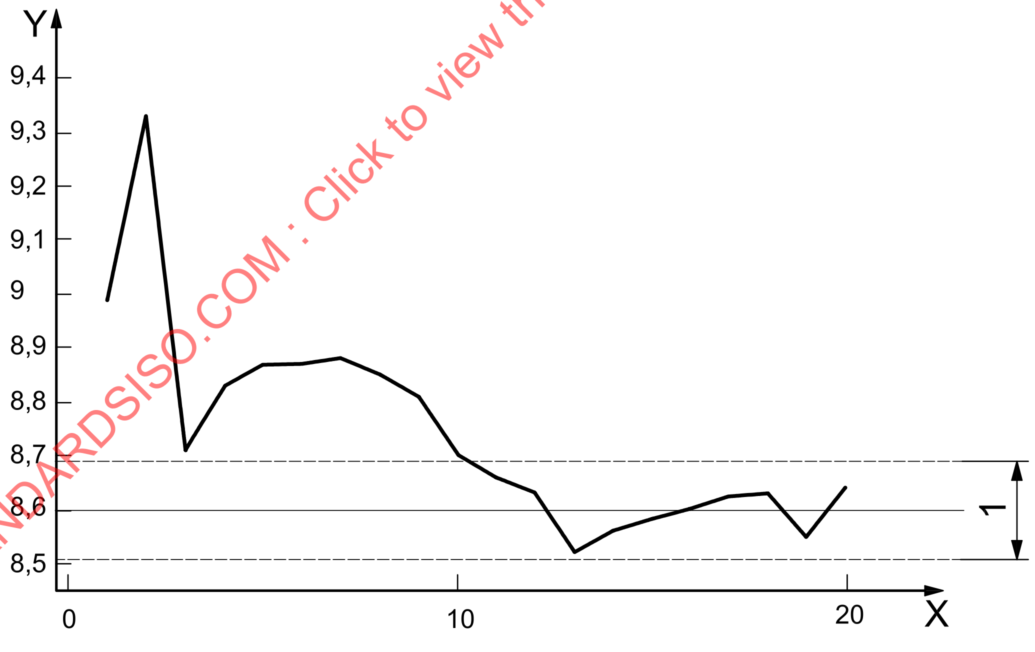
NOTE The value of mass fraction of ash (in %) in the figure for each lot represents the average of 8 results for the lot. Namely, from Table 3, the value for lot 1 = (9,38 + 9,24 + 9,02 + 8,98 + 9,22 + 9,32 + 8,40 + 8,38) /8 = 8,99, and so on.

Key

X lot number

Y average of mass fraction of ash (in %)

Figure 7 — Averages of mass fraction of ash (in %) of coal by lot from cargo



NOTE The value of mass fraction of ash (in %) plotted by lot number represent progressive, or cumulative, averages of all measured values up to that lot. For example, the averages of the first three lots are: 8,99, 9,67 and 7,45. The corresponding progressive means are 8,99, (8,99 + 9,67)/2 = 9,33 and (8,99 + 9,67 + 7,45)/3 = 8,70.

Key

1 ± 1 % band

X lot number

Y progressive average of mass fraction of ash (in %)

Figure 8 — Progressive averages of mass fraction of ash (in %) in terms of lot

Figure 8 shows that, as the number of sampling lots increases, the progressive average appears to approach a limiting value of approximately 8,6. From the 10th sampling lot, the fluctuation of the progressive average has stabilized to fall within the ± 1 % bounds. That is to say, in this particular case, some 10 sampling lots would be required before a stable estimate lying within ± 1 % of 8,6 % could have been made of the ash content. However, it is important that this is not taken as a general rule. Much will depend upon the homogeneity of the consignment, the mass of the sample, the sampling and sample preparation procedures, sampling plan design, instrument resolution, etc. It is in the determination of the relationships between these factors that the methods of statistical analysis are useful.

4.4.4 Benefits of a statistically sound sampling plan

The benefits of a statistically sound sampling design become evident on analysis and attempting to draw conclusions from the results. For example, it is noted from purely cursory observation of Table 3 that:

a) there is noticeable variation within each column of results;

b) the rows of results (i.e. within lots) for the main peaks and troughs of Figure 7, lots 2 (high), 3, 13 and 19 (low), and 20 (high), have similar patterns of variation;

c) there are adjacent column pairs of very low values in lot 9 (7,10 and 7,22 for results 3 and 4, respectively) compared with the six other values ranging from 8,6 to 9,1;

d) there are adjacent column pairs of very high values in lot 10 (8,59 and 8,89 for results 5 and 6 respectively).

What do these results indicate? To answer this question, it is necessary to refer to the plan for sampling percentage ash from the ship's cargo. This is shown in Figure 9.

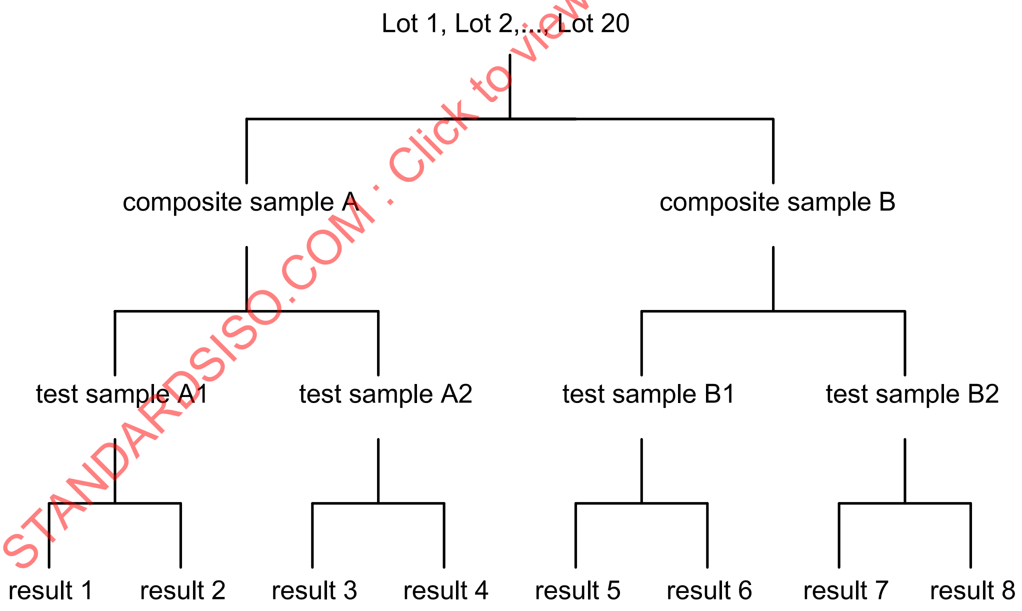
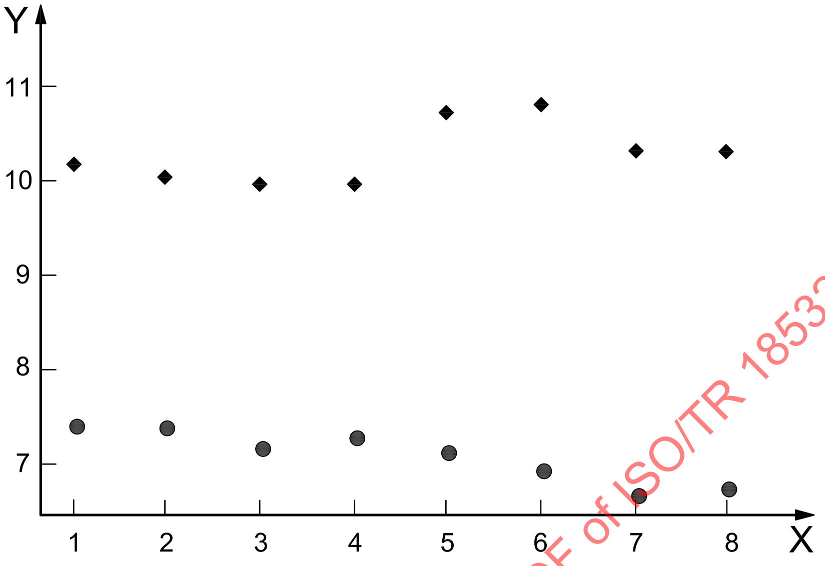


Figure 9 — Schematic diagram showing plan for sampling percentage ash from cargo of ship

This statistical design permits the isolation of lot to lot, composite sample to composite sample, test sample to test sample and measurement variation. This type of design is recommended (see ISO 11648-1) when there is no, or little, prior knowledge about the sampling situation.

In this particular case, the conveyor belt unloading coal from the ship was stopped at uniform time intervals.

A prespecified mass increment of coal was shovelled from the conveyor belt. Individual consecutive increments were placed alternately into two containers, A and B. Each of the two containers ultimately contains 30 such increments, which make up so-called composite samples. Two test samples are then prepared from each composite sample. Ash content is then analysed in duplicate on each test sample.



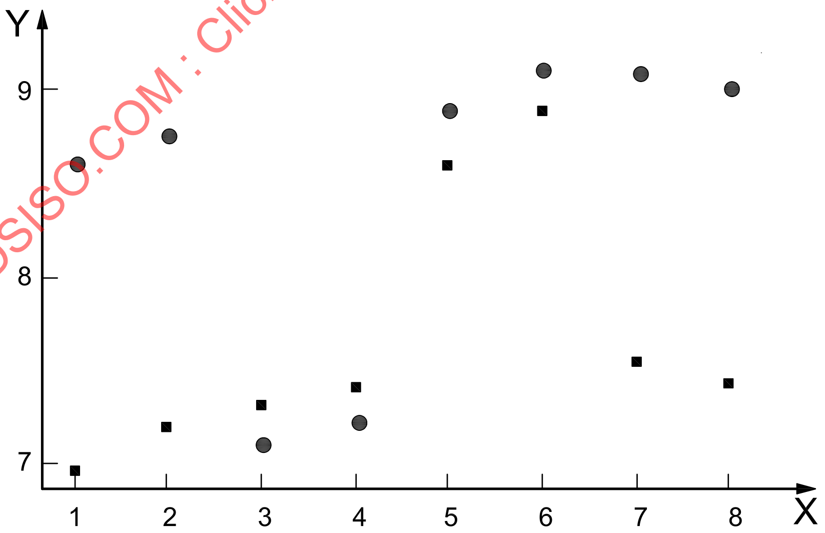
NOTE Circle: lot 19; Diamond: lot 20.

Key

X test number

Y mass fraction of ash (in %)

Figure 10 — Mass fraction of ash (in %) plotted against test number for lots 19 and 20 (illustrating relative consistency of percentage ash within each of these lots)



NOTE Circle: lot 9; Square: lot 10.

Key

X test number

Y mass fraction of ash (in %)

Figure 11 — Mass fraction of ash (in %) plotted against test number for lots 9 and 10 (illustrating rogue pairs in both lots)

Because of the design of the statistical sampling plan, certain conclusions may now be drawn, for example:

i) in general, there is more lot to lot variation (row to row variation in Table 3) than within lot variation (within row variation);

ii) Figure 10 confirms, with respect to lots 19 and 20, the relative consistency of test results within these lots;

iii) Figure 11 indicates two pairs of rogue values in lots 9 and 10. Reference to the sampling plan indicates that these pairs are associated with test sample A_2 in lot 9 and B_1 in lot 10. The latter phenomenon could be due to problems in sample preparation or, perhaps, an abrupt change in calibration level of the measurement system. In retrospect, it is not possible to assign this special cause specifically. However, if simple graphical analysis such as this is ongoing, as results become available, and is not left to be done retrospectively, it is more likely that the specific cause of events such as these could be identified, at the time and place of the sampling activity. Operational or technical personnel could then take action, to remove the cause and its effect by eliminating the rogue values or substituting more representative ones.

4.4.5 General conclusions

This example illustrates:

a) the importance of the deployment of statistical thinking and design method to a numerical study *prior* to its being undertaken;

b) the value of the progressive application of simple, mainly graphical, statistical tools to any numerical study at the time and place of the particular activity rather than *just* applying more sophisticated statistical methods retrospectively;

c) that to gain full benefit from b) it is essential that personnel who are technically and operationally familiar with the activity under scrutiny are involved in the progressive statistical analysis because it will facilitate the early identification and removal of any special cause variation that may be found to be present.

A more sophisticated retrospective statistical study of the results in this example included the use of analysis of variance (ANOVA). This confirmed that most of the overall variation in mass fraction of ash (in %) (84 %) was attributable to lot to lot variation, indicating variability in the ash content of the cargo. About 7 % to 8 % was attributed to each of composite and test sample variation and less than 1 % to measurement variation.

5 Introduction to basic statistical tools

5.1 General

The examples in Clause 4 give a general idea of the function of statistical methods in the analysis, control and reduction of variation and the usefulness of simple graphical presentation of data. Before developing in greater detail the application of these methods to *quality*, *specification* and *standardization*, it is necessary to describe more fully some of the basic tools.

Suppose that a single *quality characteristic* has been measured and recorded for each of a number of *objects*. The objects/characteristics of interest may be teeming temperature or vacuuming time in a steel mill; lateness of trains; time to pay invoices; length, diameter, surface finish or eccentricity of a component; hardness or silicon content of a material; time to answer a telephone; time to failure; noise level; emission level of engines. This partial listing gives some impression of the broad applicability of these tools. The measured values of characteristics will be termed values or observations.

5.2 Basic statistical terms and measures

If a group of units or quantities have been selected from a larger whole, it is defined in statistical terminology as a *sample*. It is also common to speak of the individual observations themselves as forming a sample. Thus, a sample may consist of 1, 2, 3,…, n units or observations.

A sample that is drawn without bias is termed a random sample. The larger whole of units, which is to be the subject of sampling (e.g. all students at a college at the time of a survey), is called a *population.* (Further discussion of this concept is provided in 7.1 and 8.1.1.) A sampling *frame*, on the other hand, is a list of sampling units from which the sample is taken (e.g. college register). In the subclauses that follow, it will be necessary to discuss various aspects of the relationship between the sample and the population. But it is first necessary to introduce certain statistical measures which, from the descriptive viewpoint, may be equally applied whether the group of units under consideration forms the sample or the population.

With a group of observations, three aspects are of prime importance. These are:

a) a measure of central tendency;

b) a measure of the magnitude of the variation;

c) the pattern of variation.

There are various methods for measuring the central tendency and the magnitude of the variation within a group of observations. An important feature that is, unfortunately, frequently not taken into consideration in the application of these measures is the *pattern of variation*. For example, too often normality (symmetrical bell-shaped distributions) is assumed in process capability studies, and constant failure rates in the specification of, and performance claims for, equipment reliability.

Central tendency is most commonly expressed in terms of:

1) *arithmetic mean* (or just *mean* or *averag*e): the total of the values divided by the number of values;

2) *median*: the central value when the data are ranked in order of size and when the number of observations is odd; if the number of observations is even, then the median is usually taken to be the average of the two central values;

3) *mode*: the most frequently occurring value.

The two most frequently used measures of *variability* are:

i) *range*: the difference between the smallest and largest values in the data;

ii) *standard deviation*: measures the variation of the data around the mean. The less the variation, the smaller the value. When derived from a sample, the value is given by the expressions:

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{n\sum x^2 - (\sum x)^2}{n(n-1)}}$$

where

$\sum$ is the sum of;

x is the individual value;

$\bar{x}$ is the arithmetic mean of the individual values;

n is the number of values;

s is the sample standard deviation.

A summary of the relative advantages and disadvantages of these measures is given in Table 4.

Table 4 — Advantages and disadvantages of various statistical measures

Measure	Advantages	Disadvantages
Mean	Easy to understand Commonly used	Affected by very high or low values Need all the data to calculate
Median	Unchanged by very high or very low values	Slow and tedious to calculate for large sample sizes unless a suitable calculator or computer facility is available
Mode	Unchanged by very high or low values	May be multi-modal
Range	Easy to calculate	Uses extreme values only
Standard deviation	More efficient than range	Less easy to calculate manually

To illustrate these terms, a set of five values is used: 7, 5, 10, 7 and 6. Using these values, the various statistical measures are as follows.

Arithmetic mean = (7 + 5 + 10 + 7 + 6)/5 = 7

Median = central value of ordered set, 5, 6, 7, 7, 10 = 7

Mode = most frequent value = 7

Range = maximum value − minimum value = 10 − 5 = 5

Standard deviation (manual method is shown below) = 1,87

Sample value	$(x-\bar{x})$	$(x-\bar{x})^2$
7	0	0
5	−2	4
10	3	9
7	0	0
6	−1	1

$$\sum(x-\bar{x})^2 = 14$$

and thus

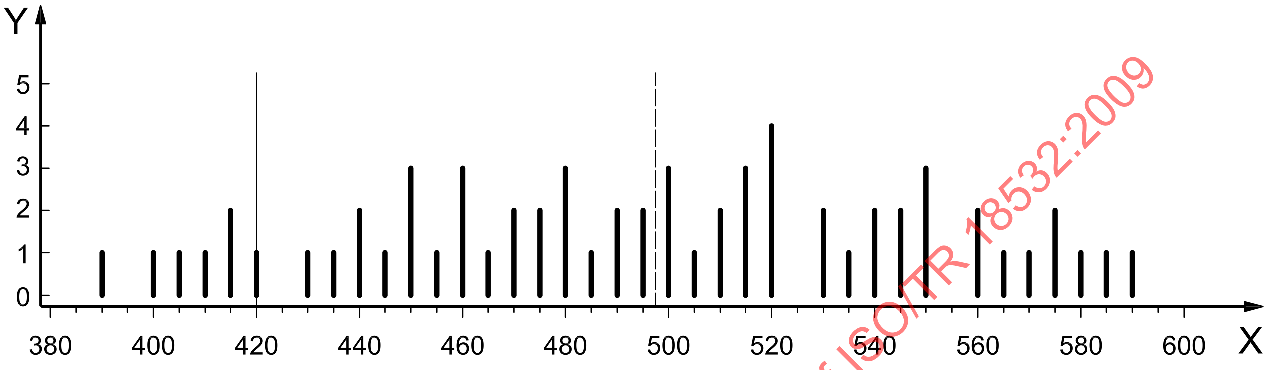
$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{14}{5-1}} = 1{,}87$$

Alternatively, the standard deviation may be obtained much more quickly and directly using a scientific calculator.

5.3 Presentation of data

5.3.1 Dot or line plot

Particularly when only a few observations are available, dot or line plots, such as those shown in Figure 1 and Figure 3, will often give a useful preliminary picture of the situation. Indeed for certain purposes, the consideration of such a diagram may be all that is needed. The corresponding line plot for the dot plot of Figure 1 is shown in Figure 12.



Key

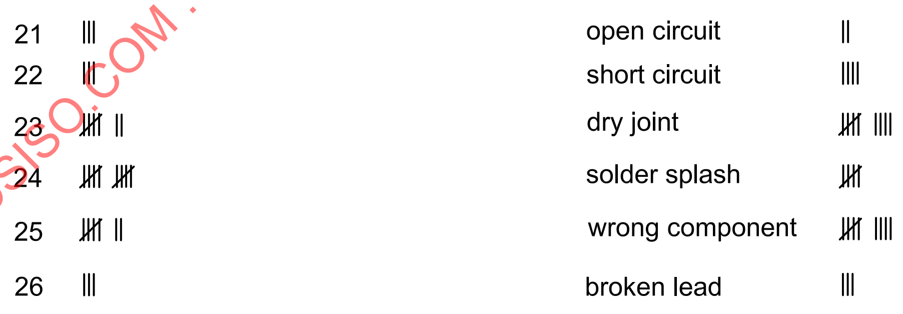
X strength

Y number of observations

Figure 12 — Line plot of breaking strength of wire (Table 1 data)

5.3.2 Tally chart

A tally chart may be applied to both measured and classified data. It is used to visually represent the frequency of a particular value, or a specific type of event, in a series. The five-bar gate notation is used. Examples are shown in Figure 13 for both measured data and classified events.



a) Tally chart for measurements

b) Tally charts for events/counts

Figure 13 — Typical tally charts

5.3.3 Stem and leaf plot

The stem and leaf plot displays the pattern of variation of measured data. It is an enhanced form of histogram or tally chart. In addition to showing the distribution of a set of data, it also shows individual values. Each value is split into two parts, the first part consisting of the leading digits, which are written downwards along the stem in ascending order, while the second part of the values, the remaining digits, are written in ascending order horizontally along the leaf on the line with their leading digits.

An example is shown for the following data in Figure 14.

Data: 29, 28, 41, 36, 36, 59, 50, 61, 44, 48, 35, 42, 53, 33, 31.

stem	leaf
2	8 9
3	1 3 5 6 6
4	1 2 4 8
5	0 3 9
6	1

Figure 14 — Stem and leaf plot for data

5.3.4 Box plot

The box plot is a very useful tool in exploratory data analysis. It is simple to construct and easy to interpret. Like the dot or line plot, it is used to depict the similarities within, or the differences between, different groupings of data. Tukey [130] defines

$$\text{Lower limit (LL)} = Q_1 - 1{,}5 \times \text{IQR}$$

and

$$\text{Upper limit (UL)} = Q_1 + 1{,}5 \times \text{IQR},$$

where IQR is the inter-quartile range defined, i.e. $\text{IQR} = Q_3 - Q_1$. Observations outside these limits may be outliers, and should be checked. If the data are from a normal distribution, the area outside these limits is approximately 0,7 %.

The box plot represents a five-number summary of a data distribution consisting of

$\tilde{x}$ median (mid) value;

Q_1 first quartile (value below which ¼ of values lie);

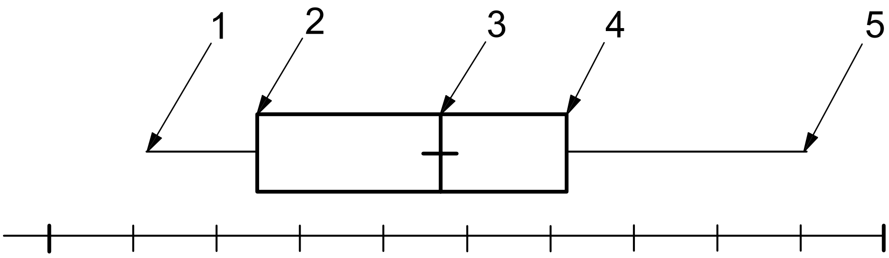
Q_3 third quartile (value above which ¼ of values lie);

A_1 lower adjacent value = smallest data value $\geqslant$ LL;

A_2 upper adjacent value = largest data value $\leqslant$ UL.

A basic box plot consists of a box, the length indicating the region where 50 % of the readings lie, a median line, and whiskers extending from the box to the lower and upper adjacent values. Each data value outside the whiskers is drawn as a point alongside the axis defined by the whiskers. A box plot is shown pictorially in Figure 15.

The box plot may be extended, for example, to include a display of statistical confidence limits around the median. An absence of overlap of these statistical bounds between groups would indicate statistically significant differences between the medians of these groups. Also, the width of the box may be varied to indicate changes in relative size of different groups. Outliers (apparent rogue values) may be shown by an asterisk.



Key

1 A_1 lower adjacent value

2 Q_1 first quartile

3 $\tilde{x}$ median,

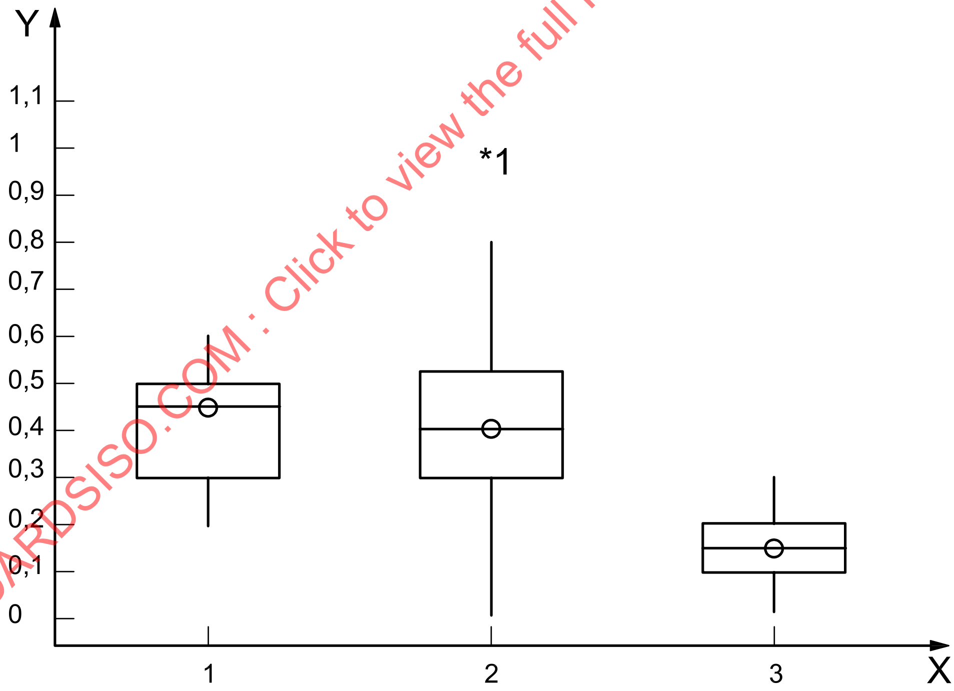
4 Q_3 third quartile

5 A_2 upper adjacent value

Figure 15 — Box plot

The box plot may be augmented by more formal statistical methods such as analysis of variance (ANOVA).

An example of the applicability and value of a box plot is shown in Figure 16. The shade variation of fabric of a particular colour was compared between adjacent panels on typical items of clothing sourced from three different suppliers. The results are shown in box plot form in Figure 16.



Key

1 two extreme values at 1,0 out of 30

X supplier

Y delta E value

NOTE 30 samples tested from each supplier.

Figure 16 — Box plot for Delta E panel shade variation between supply sources

The box plot indicates considerable variation in standards of performance between suppliers both in terms of process targeting (indicated by the relative positions of the median) and consistency about that target (indicated by the differences in lengths of the whiskers). The asterisk indicates two outlying values indicating lack of control of the dyeing process.

The box plot indicates that:

a) supplier 1 has approximately the same median Delta E value as supplier 2;

b) supplier 2 has a dyeing process with essentially the same variation as supplier 1; two very high values are also present (in a limited test sample of 30), which is likely to give rise to extreme customer dissatisfaction and a loss of quality reputation by the retailer; if typical of production, major recalls may be expected;

c) supplier 3 has a dyeing process distributed around a low (good) Delta E value with much smaller variation about that value than the other two suppliers.

The Delta E results on supplier 3 merchandise indicate what can be, and is being, achieved in terms of shading performance. This "current state of the art" or "best practice" result then becomes the benchmark or reference standard for all supply sources.

5.3.5 Multi-vari chart

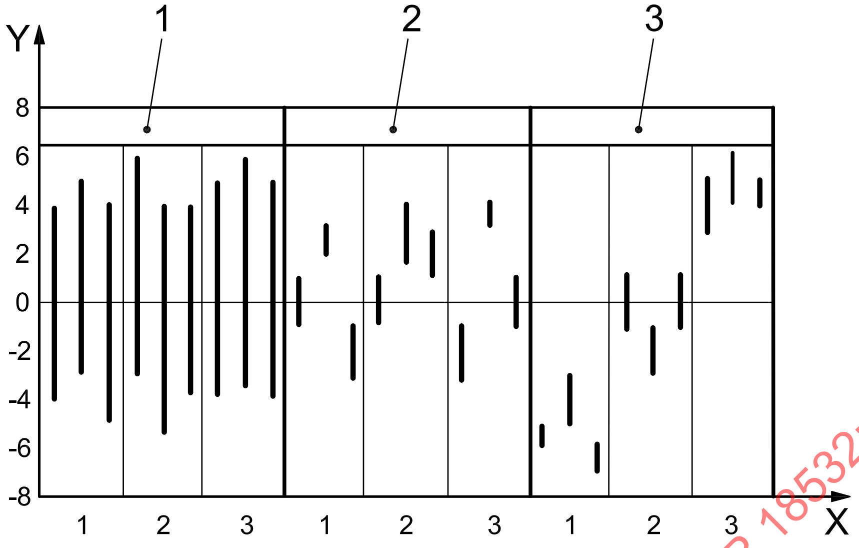
The multi-vari chart is a simple pictorial method of indicating and comparing the magnitude of different sources of variation. As such, it is very useful for diagnostic and investigation purposes rather than, and as a precursor to, ongoing process control. It consists essentially of vertical lines joining maximum and minimum values for a particular characteristic against a measurement scale: a max.-min. plot. A *dot* on nominal size represents an ideal value. The longer the line, the greater the variation.

Take a turned diameter where three consecutive components are taken from production and measured each hour. A clock gauge is used which records maximum and minimum values of the diameter of each component as it is rotated. The multi-vari chart in Figure 17 shows, for three quite different process performance scenarios, the dominant sources of variation prevailing, namely:

a) within part (geometric form) variation;

b) part to part variation;

c) time to time variation.



Key

1 scenario 1: Large within part variation

2 scenario 2: Large part to part variation

3 scenario 3: Large time to time variation

X time

Y value

NOTE Maximum and minimum of three consecutive parts every hour for three hours showing three quite different process performance scenarios.

Figure 17 — Multi-vari chart as a tool for process variation analysis

5.3.6 Position-Dimension (P-D) diagram

A P-D diagram can be looked upon as an extension to the multi-vari chart to handle more than one feature. A P-D diagram representing ideal values in relation to, say, ovality and taper of a cylinder, is a horizontal straight line on a vertical dimension scale. If this line is coincident with the nominal or targeted value of the overall mean of the diameter, then this represents the ideal situation.

An example illustrates its usefulness. The variation in the outside diameter of a cylinder is being investigated for nominal size, ovality and taper. Measurements are taken at right angles to one another at each end of the cylinder as shown in Figure 18. A, B, C and D, as shown in Figure 18, identified these positional measurement values. One cylinder from production was measured every shift for 4 shifts. The machine tool was then overhauled and another set of readings taken.

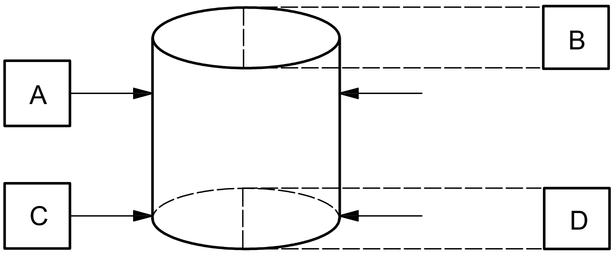
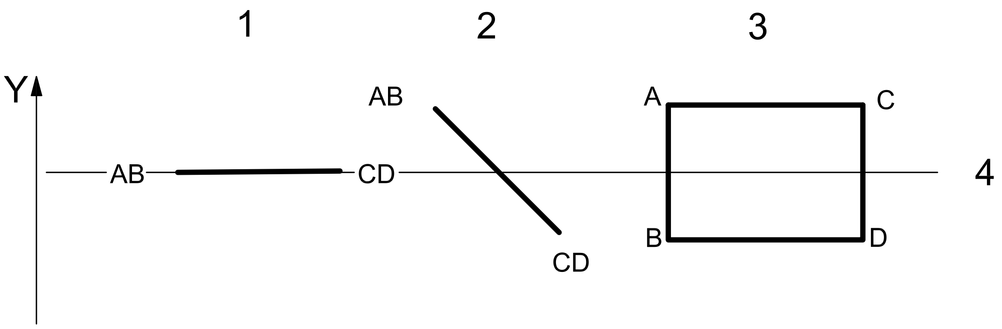


Figure 18 — Measurements on cylinder to determine nominal size, ovality and taper

Figure 19 provides the reference standard for this diameter and is used for judging the degree to which the diameter meets the preferred nominal value and the extent of geometric form variation present.



Key

1 ideal

2 pure taper

3 pure ovality

4 nominal value

Y diameter

NOTE AB indicates A is coincident with B dimensionally and CD indicates C is coincident with D dimensionally.

Figure 19 — Measurement on cylinder — P-D diagrams showing ideal diameter values, pure taper and pure ovality

Regarding Figure 20 and the factors under investigation:

a) ovality:

— A > B indicates ovality at the AB end and C > D indicates ovality at the CD end;

— this progressively increases with time until overhaul;

b) taper:

— the mean of A and B exceeding the mean of C and D indicates taper along the length of the cylinder;

— this, too, progressively increases with time until overhaul;

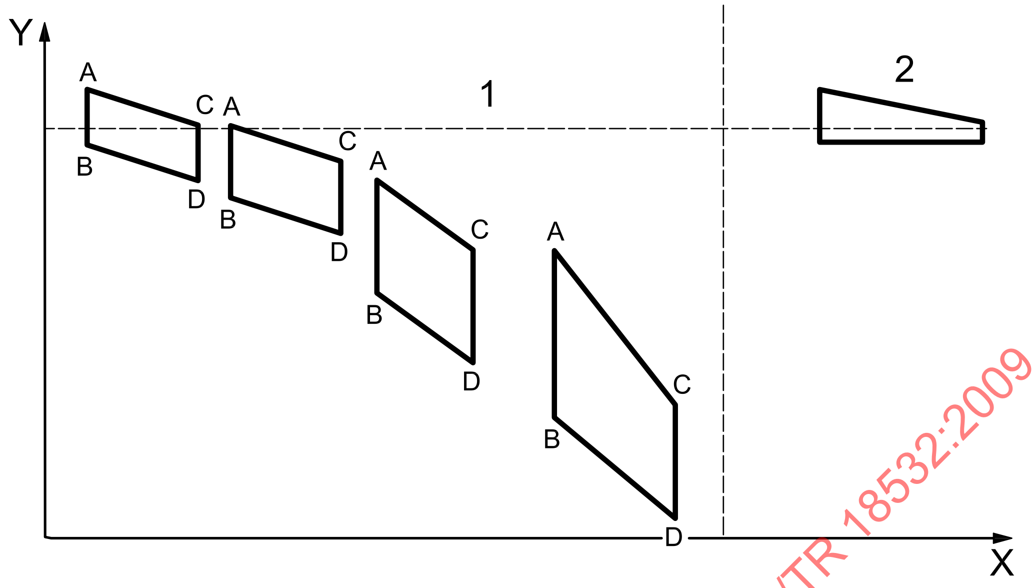
c) overall diameter size:

— the mean of A, B, C and D gives an estimate of the overall average diameter;

— this progressively decreases away from its nominal value until overhaul;

d) overhaul:

— improvements in overall diameter aim, and geometric form variation, with some ovality remaining at the AB end.



Key

1 before overhaul 2 after overhaul X time Y diameter

NOTE The dashed, vertical line indicates the time of the overhaul; on the left-hand side of the line, the situation before the overhaul is presented, and on the right-hand side, the situation after the overhaul. The dashed horizontal line represents the nominal value.

Figure 20 — Measurement on cylinder — P-D diagrams indicating progressive decrease of mean and increase in geometric form variation and the beneficial effects of overhaul

5.3.7 Graphical portrayal of frequency distributions

When only a few observations are available, *dot or line plots* as shown in Figures 3 and 12 or *stem and leaf plots* as in Figure 14 will often suffice. However, with a larger number of observations, and once the possibility of dependency has been discounted, it is generally found convenient firstly to arrange the data in numerical value order, the total observed range of variation in the measured characteristic is then divided into convenient equal intervals, and the number of observations falling into each interval is counted. This number is termed the frequency for that interval and the resulting tabulated series of numbers shows the frequency distribution. A simple method called Sturge's rule, i.e.

$$\text{Number of classes} = 1 + 3{,}3 \times \log_{10}(\text{number of observations})$$

provides guidance on the number of class intervals to select in terms of the total number of observations. This is shown, both in tabular and equation form, in Table 5. Sturge's rule should be taken purely as a starting point for experimenting with the number of classes to get the most informative view of the data. The number of class intervals actually chosen in a particular case should ultimately be chosen on the grounds of simplicity, clarity and ready understanding. This is illustrated in the example that follows.

Table 5 — Guidance on number of classes to select in terms of number of observations

Number of observations	20 to 45	46 to 90	91 to 190	191 to 370	371 to 750	751 to 1 500	1 501 to 3 000	3 001 to 6 000	6 001 to 12 000
Number of classes	6	7	8	9	10	11	12	13	14

A grouped frequency distribution may be represented in a number of ways. Typical ones are:

a) frequency table;

b) frequency tally chart;

c) histogram;

d) cumulative frequency table;

e) cumulative frequency plot.

The application of each of these methods is illustrated by example.

EXAMPLE Quality of zinc-coated item after galvanizing:

Selected test specimens typical of production are required to withstand a standard 4 min acid bath immersion test following galvanizing. Some 200 results that have accumulated over a period of time are used as the basis for this study. Measurements were taken to the nearest 0,1 min.

The results extended from 4,3 to 9,4. Sturge's rule suggests 9 class intervals for 200 results. This would give class intervals of (9,4 min − 4,3 min)/9 = 0,57 min. The results were arranged in ascending order and the actual class interval chosen was 0,5 min for greater simplicity and clarity. The resulting frequency and percentage frequency tabulation are shown in Table 6.

Table 6 — Frequency and percentage frequency table for immersion times withstood by test specimen

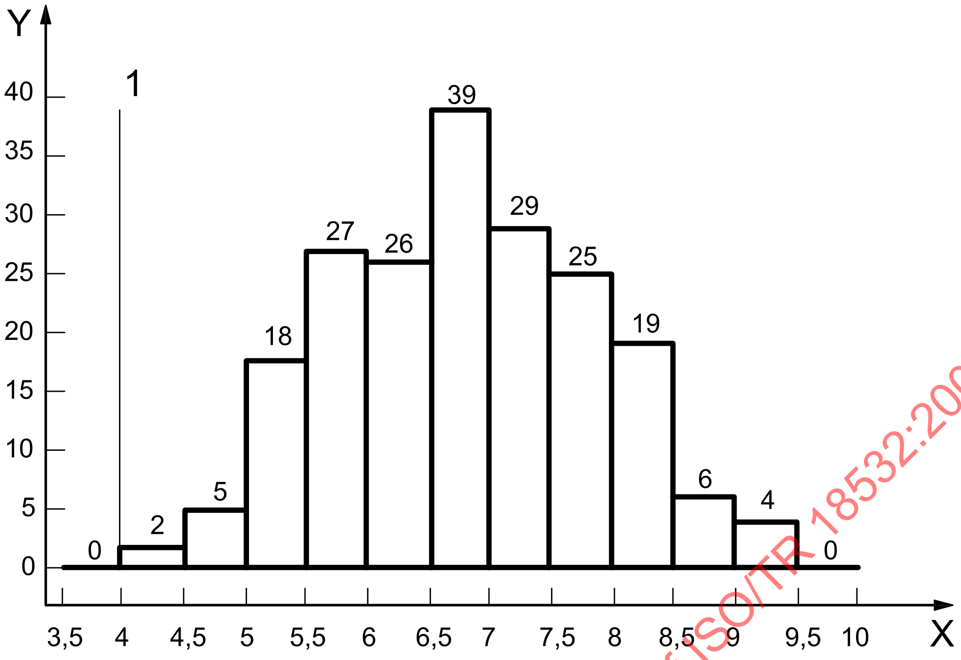
Immersion min	No. of observations (Frequency)	Frequency %
4,1 to 4,5	2	1
4,6 to 5,0	5	2,5
5,1 to 5,5	18	9
5,6 to 6,0	27	13,5
6,1 to 6,5	26	13
6,6 to 7,0	39	19,5
7,1 to 7,5	29	14,5
7,6 to 8,0	25	12,5
8,1 to 8,5	19	9,5
8,6 to 9,0	6	3
9,1 to 9,5	4	2

Table 6 illustrates that a frequency table summarizes a set of data by showing how often values within each class interval occur and that it may be enhanced by tabling the percentages that fall within each category.

This permits some feel for how the data, as a whole, are distributed.

Various pictorial representations of the test data of Table 6 are shown in Figure 21 to Figure 26. They further enhance perception of the shape and pattern of the distribution of the data and its relation to the lower specification limit of 4 min.

The horizontal axis of the histogram corresponds with the variable characteristic and the frequency of observations in a given interval is represented by a rectangle of height proportional to this frequency, standing on the appropriate base element. Using this method, *frequency* in the table corresponds with *area* in the histogram.



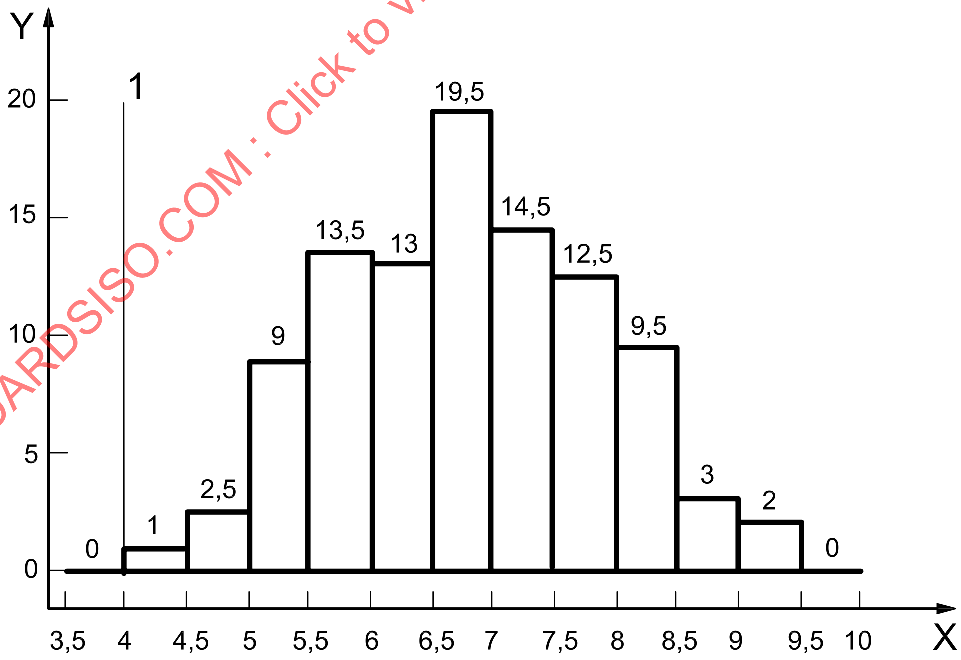
Key

1 lower specification limit

X time, in min

Y number of observations (frequency)

Figure 21 — Frequency histogram for immersion times in Table 6



Key

1 lower specification limit

X time, in min

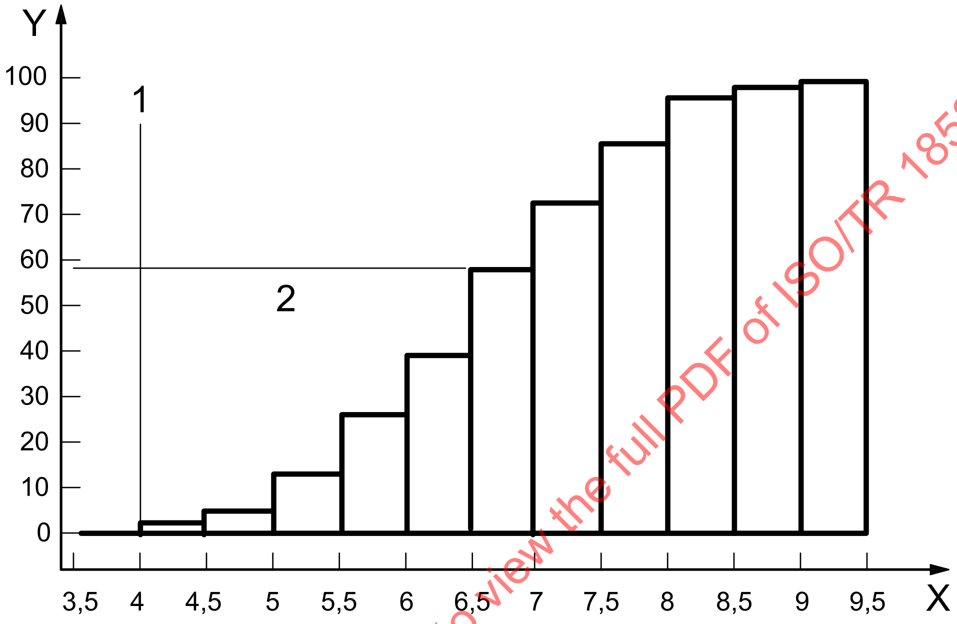
Y frequency, in percent

Figure 22 — Percentage frequency histogram for immersion times in Table 6

Sometimes, it adds to understanding if relative (percentage) frequencies rather than actual counts of frequencies are used to construct the histogram. It also demonstrates the intermediate step to be taken in constructing a cumulative percentage frequency diagram.

The cumulative relative frequency diagram shows the percentage of observations falling below (or above) particular values. By way of illustration, in Figure 23 it is seen that 58,5 % fall below 7,0 min. Hence, this is a very useful diagram for determining the situation in relation to specification limits (lower and upper).

It should be borne in mind that the cumulative percentage values relate to the *upper limit* of the class interval and not to the mid-value.



Key

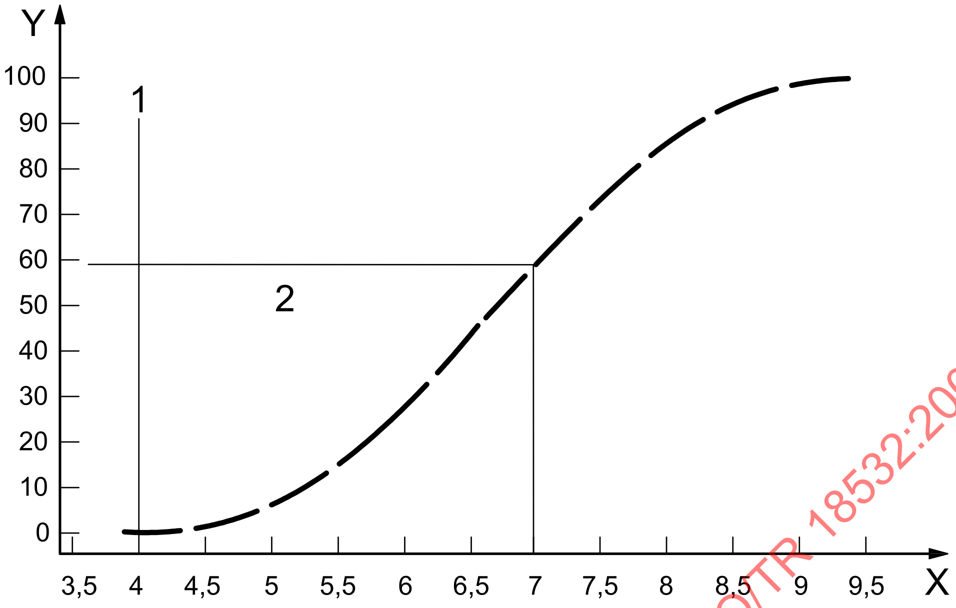
1 lower specification limit

2 58,5 % below 7,0 min

X time, in min

Y cumulative frequency, in percent

Figure 23 — Cumulative percentage frequency histogram for immersion times in Table 6



Key

1 lower specification limit

2 58,5 % below 7,0 min

X time, in min

Y cumulative frequency, in percent

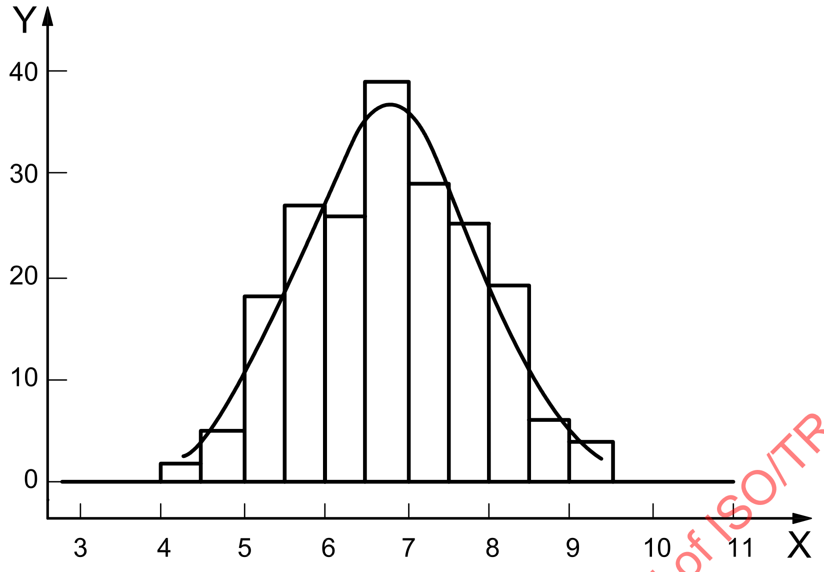
Figure 24 — Cumulative percentage frequency diagram for immersion times in Table 6

The cumulative histogram can alternatively be expressed in the form of a smooth curve as shown in Figure 24, or preferably as a straight line by transformation of the vertical scale. This latter concept will be developed later in this clause.

The diagrams of Figure 21 to Figure 26 portray the actual situation for a sample size of 200. What can be predicted about the galvanizing quality of production as a whole from this sample, assuming that the process is, and continues to be, stable about the present mean? This is where statistical modelling using the appropriate probability distribution can provide worthwhile quantitative information. A *best fit* probability distribution to match the actual frequency distribution is sought.

In the zinc plating case, the frequency histogram for immersion time (Figure 21) indicates a bell-shaped symmetrical pattern of variation about the mean. This is typical of the normal or Gaussian distribution. The normal curve has a definite mathematical equation that depends only on the values of the mean and standard deviation. Care should be taken not to interpret the word normal to mean that anything non-normal should be looked upon as being peculiar. Any characteristic that has a natural zero, for instance, such as taper, eccentricity and parallelism will naturally be skewed. Constant failure rates, looked upon as ideal in the reliability field, will naturally have a non-normal (negative exponential) frequency distribution. In fact, the existence of a normal failure frequency would indicate an undesirable increasing failure rate or wear-out regime.

However, a large number of the symmetrical frequency distributions met with in practice in the quality domain may be adequately represented by the normal curve. Does the normal distribution provide a reasonable fit to the immersion time data? The answer is given visually in Figure 25.



Key

X time, in min

Y number of observations (frequency)

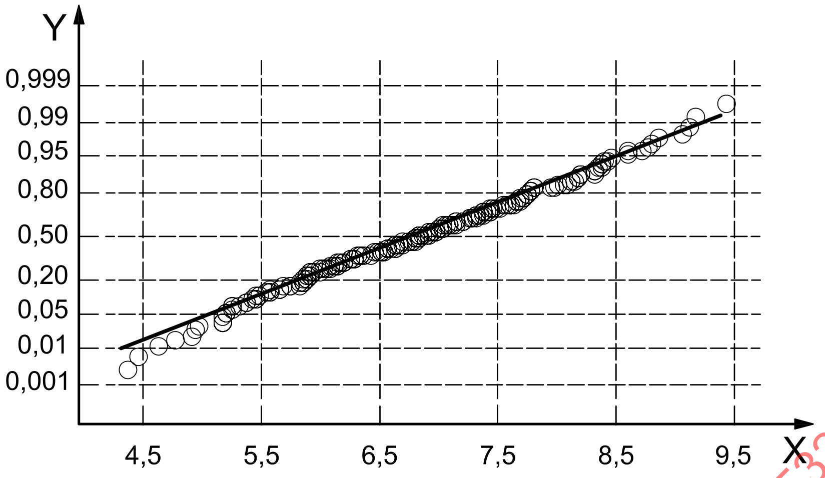
Figure 25 — Normal curve overlaid on the immersion time histogram (mean = 6,79; standard deviation = 1,08)

Based on a calculated mean and standard deviation of the 200 results, the normal curve has been fitted to the data in histogram form as shown in Figure 25. These show, by eye, a good correspondence between the two indicating that a normal distribution with a mean of 6,79 min and standard deviation of 1,08 min is a reasonable representation, or model, of the actual data. There are a number of formal statistical tests for departure from normality. These include the Shapiro-Wilk and Epps-Pulley tests (for more information on these tests see ISO 5479 [18]).

A simple, practical, effective and graphical method involves the plotting of cumulative percentage frequencies on normal probability paper. If such a plot follows a straight line, then the sample can reasonably be regarded as having come from a normal distribution. If the plot indicates a systematic departure from a straight line, then the shape of the plot often suggests the type of distribution it represents. For example, a plot that forms a concave curve (curving downwards in the middle) would result from a log-normal distribution, whereas a plot curving upwards at its centre would result from an exponential or a gamma distribution. The plot on normal probability paper is often called a quantile-quantile diagram or a Q-Q plot and is done automatically as a part of the analysis in many statistical software packages.

In addition to checking for normality, this method is used extensively in statistical process control for capability and performance measurement (see 6.1.7).

An example of this test applied to the galvanized item immersion data is shown in Figure 26. It is seen that the normal cumulative probability scale shown on the vertical axis of Figure 26 transforms the bell-shaped normal distribution into a straight line when plotted against immersion times.



Key

X time, in min

Y probability

NOTE Average: 6,778 5; standard deviation: 1,056 31; number of data points: 200.

Figure 26 — Straight line plot on normal probability paper indicating normality of data in Table 6

Whatever the distribution, it is desirable to work in terms of a straight line reference standard for a number of reasons:

1) it permits a simple immediate visual test of fit against the underlying distribution (normal here);

2) it makes for ease of extrapolation and so facilitates numerical prediction of the likelihood of having values in a larger lot or consignment outside of those experienced in the sample;

3) it facilitates the correction of individual measurements and other errors;

4) it gives an immediate visual appraisal of the relationship of the data to any specification limits or reference standards in terms of both targeting and variability;

5) for an incapable process, it immediately provides an estimate of the proportion of values likely to be above and/or below specification limits;

6) it serves as a diagnostic tool to detect divergences from the model; for instance, a smooth concave or convex plot on a normal probability plot indicates skewness of the data.

Using probability paper based on the normal or other standard statistical distributions to represent data thus offers many practical advantages in the interpretation of results from samples. These other standard distributions include the fixed shape log-normal and extreme value distributions for moderately skewed data and the versatile multi-shaped Weibull distribution. The Weibull distribution is used extensively in the reliability field to model the various regimes of failure: infant mortality (decreasing failure rate), prime of life (constant failure rate) and wear-out (increasing failure rate). Confidence bands may readily be plotted around the best estimate straight line plots. These would be represented by curves.

5.3.8 The normal distribution

The previous example suggests the important descriptive part that the normal curve can play, always provided that its suitability to represent the type of variation in question has first been established. It would seem appropriate to take this opportunity to discourage a common belief. There is no magic about the normal curve that if a distribution follows this law then that is proof that the process giving rise to the product or service is, in fact, in control (i.e. stable). To arrive at reasonable judgements on past performance and to make rational predictions of future performance based on accumulated data, it is necessary to have prior knowledge that no special causes of variation were present over the period in which the data were gathered.

The name *normal* distribution originates from its use in pioneering statistical studies of human populations.

For a process not in control, the variation in its output is unpredictable. A primary role of statistical process control is to ensure, and assure, process stability. However, the normal distribution is the one most frequently encountered in many processes.

Moreover, the distribution of means of samples, or subgroups, will be very close to normal, even when the sample size is as low as 4 or 5, in cases where the distribution of individuals is distinctly non-normal.

The normal distribution is a two-parameter distribution uniquely described by its mean and standard deviation. Consequently, its characteristics can be made available in a convenient, practical format for users. This will initially be illustrated generally in a graphical manner and, secondly, in the form of a table (Table 7). Figure 27 shows a standardized symmetrical bell-shaped curve that characterizes this distribution. Additionally, some key percentages are included in relation to distances from the mean in terms of standard deviations. Whilst, in Figure 27 and Figure 28, the normal curve appears to end at some finite value about ± 3 to 4 standard deviations from the mean, mathematically it extends to infinity in both directions.

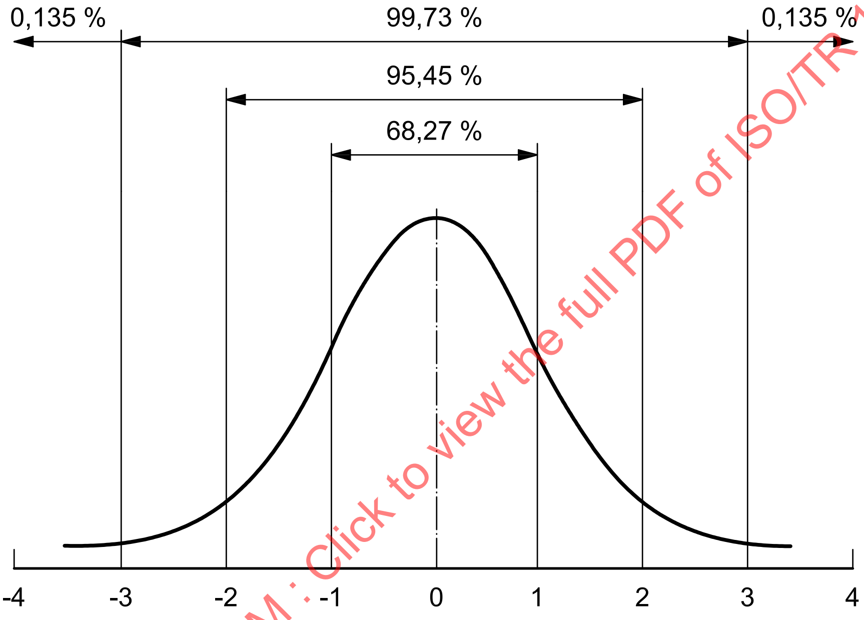


Figure 27 — Percentages of normal distribution in relation to distances from the mean in terms of standard deviations

Figure 27 illustrates that for a normal distribution

a) 99,73 % of values lie within the limits: mean ± 3 standard deviations;

b) of the remaining 0,27 %, 0,135 % lie below the mean −3 standard deviations and 0,135 % above the mean +3 standard deviations;

c) 95,45 % of the values lie within the limits: mean ± 2 standard deviations;

d) just over two thirds (68,27 %) of the values lie within the limits: mean ± 1 standard deviation.

This demonstrates a simple but useful property of the mean and standard deviation. Such a diagram shows the effectiveness of the normal distribution in predicting, from a sample, the proportion of the population lying within a specified range or above, or below particular limits. Whilst it is helpful in conveying certain principles, it is, however, not of sufficient resolution to be of real value in practice. Table 7 provides this by giving the probabilities of exceeding z in a standard normal distribution, i.e. a normal distribution with mean 0 and standard deviation 1. Because of the symmetry of the normal distribution, Table 7 also provides the probabilities of falling below $-z$ in a standard normal distribution. The probability of exceeding z in a standard normal distribution is the same as exceeding a point U that is z standard deviations above the mean in any normal distribution. Similarly the probability of falling below $-z$ in a standard normal distribution is the same as falling below a point L that is z standard deviations below the mean in any normal distribution. Those probabilities are illustrated in Figure 28 and they are also covered by Table 7.

NOTE 1 Example 2 following Table 7 shows how the percentage above or below a selected value can be determined when the mean and standard deviation are known. An alternative to the use of Table 7 is the application of the straight line probability plot shown in Figure 26 for making similar predictions.

EXAMPLE 1 Clothing size survey:

A size survey conducted on a random sample of a prospective customer base indicated that one particular characteristic, height, was normal with a mean equal to 175 cm, and a standard deviation equal to 7,5 cm.

By using the results of the survey, it was predicted, for instance, that:

— 16 % (15,87 %) of the target customer population are taller than 182,5 cm (mean + 1 standard deviation);

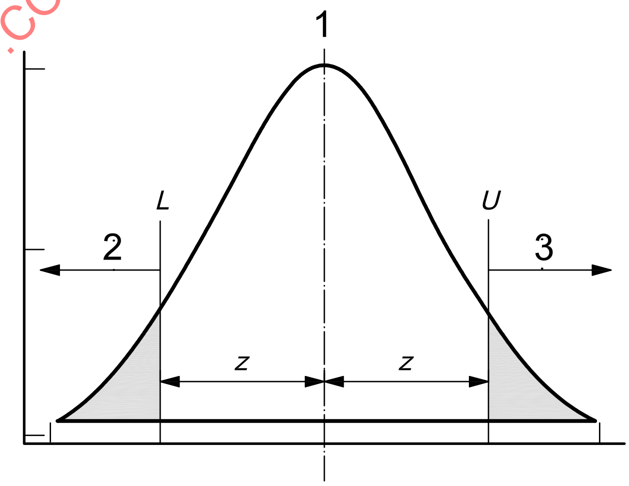
— 25 % (25,14 %) of the target customer population are shorter than 170 cm (mean $-b$ × standard deviation);

NOTE The b arises as 175 minus 170 expressed as a fraction of the standard deviation.

— 59 % (58,99 %) of the target customer population are between 170 cm and 182,5 cm.

Such estimates enable appropriate size ratios of garments to be ordered.

NOTE 2 Figure 27 and Figure 28 and Table 7 relate to a theoretical distribution representing a whole population. By convention, population parameters are symbolized by lower case italic Greek letters (e.g. population mean = μ and population standard deviation = σ).



Key

1 mean

2 percent below L

3 percent above U

Figure 28 — Standard normal probability density with indications of percentage expected beyond a value, U or L, that is z standard deviation units from the mean

In the real life examples shown, sample statistics are used to provide estimates. By convention, sample statistics are distinguished from population parameters by italic Roman letters (e.g. sample mean $= \overline{x}$ or $\overline{X}$ and sample standard deviation $= s$ or S). Sometimes such sample statistics are limited to upper case to distinguish from their actual realization, which are then shown in lower case. In this standard, this latter distinction is not used because of common usage considerations in the application areas concerned.

Clause 8 deals with the statistical relationship between sample and population.

Table 7 — Probabilities (in percent for $z < 4$ and in parts per million for $z \geqslant 4{,}0$) of exceeding z in a standard normal distribution

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
6,0	0,001	0,000 9	0,000 9	0,000 8	0,000 8	0,000 7	0,000 7	0,000 6	0,000 6	0,000 6
5,0	0,286 7	0,272 2	0,258 4	0,245 2	0,232 8	0,220 9	0,209 6	0,198 9	0,188 7	0,179
4,0	31,67	30,36	29,10	27,89	26,73	25,61	24,54	23,51	22,52	21,57
NOTE For $z \geqslant 4{,}0$, values are given in parts per million.										
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
3,5	0,023 3	0,022 4	0,021 6	0,020 8	0,020 0	0,019 3	0,018 5	0,017 8	0,017 2	0,016 5
3,4	0,033 7	0,032 5	0,031 3	0,030 2	0,029 1	0,028 0	0,027 0	0,026 0	0,025 1	0,024 2
3,3	0,048 3	0,046 6	0,045 0	0,043 4	0,041 9	0,040 4	0,039 0	0,037 6	0,036 2	0,034 9
3,2	0,068 7	0,066 4	0,064 1	0,061 9	0,059 8	0,057 7	0,055 7	0,053 8	0,051 9	0,050 1
3,1	0,096 8	0,093 5	0,090 4	0,087 4	0,084 5	0,081 6	0,078 9	0,076 2	0,073 6	0,071 1
3,0	0,135 0	0,130 6	0,126 4	0,122 3	0,118 3	0,114 4	0,110 7	0,107 0	0,103 5	0,100 1
2,9	0,186 6	0,180 7	0,175 0	0,169 5	0,164 1	0,158 9	0,153 8	0,148 9	0,144 1	0,139 5
2,8	0,255 5	0,247 7	0,240 1	0,232 7	0,225 6	0,218 6	0,211 8	0,205 2	0,198 8	0,192 6
2,7	0,346 7	0,336 4	0,326 4	0,316 7	0,307 2	0,298 0	0,289 0	0,280 3	0,271 8	0,263 5
2,6	0,466 1	0,452 7	0,439 6	0,426 9	0,414 5	0,402 5	0,390 7	0,379 3	0,368 1	0,357 3
2,5	0,621 0	0,603 7	0,586 8	0,570 3	0,554 3	0,538 6	0,523 4	0,508 5	0,494 0	0,479 9
2,4	0,819 8	0,797 6	0,776 0	0,754 9	0,734 4	0,714 3	0,694 7	0,675 6	0,656 9	0,638 7
2,3	1,072	1,044	1,017	0,990 3	0,964 2	0,938 7	0,913 7	0,889 4	0,865 6	0,842 4
2,2	1,390	1,355	1,321	1,287	1,255	1,222	1,191	1,160	1,130	1,101
2,1	1,786	1,743	1,700	1,659	1,618	1,578	1,539	1,500	1,463	1,426
2,0	2,275	2,222	2,169	2,118	2,068	2,018	1,970	1,923	1,876	1,831
1,9	2,872	2,807	2,743	2,680	2,619	2,559	2,500	2,442	2,385	2,330
1,8	3,593	3,515	3,438	3,363	3,288	3,216	3,144	3,074	3,005	2,938
1,7	4,457	4,363	4,272	4,182	4,093	4,006	3,920	3,836	3,754	3,673
1,6	5,480	5,370	5,262	5,155	5,050	4,947	4,846	4,746	4,648	4,551
1,5	6,681	6,552	6,426	6,301	6,178	6,057	5,938	5,821	5,705	5,592
1,4	8,076	7,927	7,780	7,636	7,493	7,353	7,215	7,078	6,944	6,811
1,3	9,680	9,510	9,342	9,176	9,012	8,851	8,692	8,534	8,379	8,226
1,2	11,51	11,31	11,12	10,93	10,75	10,57	10,38	10,20	10,03	9,853
1,1	13,57	13,35	13,14	12,92	12,71	12,51	12,30	12,10	11,90	11,70
1,0	15,87	15,62	15,39	15,15	14,92	14,69	14,46	14,23	14,01	13,79
0,9	18,41	18,14	17,88	17,62	17,36	17,11	16,85	16,60	16,35	16,11
0,8	21,19	20,90	20,61	20,33	20,05	19,77	19,49	19,22	18,94	18,67
0,7	24,20	23,89	23,58	23,27	22,97	22,66	22,36	22,07	21,77	21,48
0,6	27,43	27,09	26,76	26,43	26,11	25,78	25,46	25,14	24,83	24,51
0,5	30,85	30,50	30,15	29,81	29,46	29,12	28,77	28,43	28,10	27,76
0,4	34,46	34,09	33,72	33,36	33,00	32,64	32,28	31,92	31,56	31,21
0,3	38,21	37,83	37,45	37,07	36,69	36,32	35,94	35,57	35,20	34,83
0,2	42,07	41,68	41,29	40,90	40,52	40,13	39,74	39,36	38,97	38,59
0,1	46,02	45,62	45,22	44,83	44,43	44,04	43,64	43,25	42,86	42,47
0,0	50,00	49,60	49,20	48,80	48,40	48,01	47,61	47,21	46,81	46,41

EXAMPLE 2 Use of Table 7:

Specified tolerance $= 42 \pm 4$

Mean $= 40$

Standard deviation $= 2{,}2$

Process is in statistical control with an output that is normal.

What percentage is expected outside the specification limits?

To find the percentage above the upper specification limit:

$$Z_{\text{upper}} = \frac{\text{upper specification limit } (U) - \text{mean}}{\text{standard deviation}} = \frac{46-40}{2{,}2} = 2{,}73$$

Enter Table 7 at 2,73 (2,7 from the left and 0,03 from the top as indicated by the arrows) to give 0,32 % above the upper specification limit.

To find the percentage below the lower specification limit:

$$Z_{\text{lower}} = \frac{\text{mean} - \text{lower specification limit } (L)}{\text{standard deviation}} = \frac{40-38}{2{,}2} = 0{,}91$$

Enter Table 7 at 0,91 (0,9 from left and 0,01 from the top) to give 18,41 % below the lower specification limit.

Hence, the expected total fraction nonconforming is 0,32 % + 18,41 % = 18,73 %.

5.3.9 The Weibull distribution

5.3.9.1 General

Most of the earlier discussion has been based on the assumption that the population or populations under consideration are normally distributed, at least approximately. This assumption is found in practice to be valid for a very wide range of situations. However, it is not appropriate for distributions that are typically skewed, and the Weibull distribution provides a better approximation to the kind of skewed distributions often arising in time-to-failure or breaking-strength data. For the purposes of discussion, we shall suppose that the characteristic in question is a failure time, and denote it by t. Here we shall briefly consider the simplest form of the Weibull distribution, with two parameters α and β where α controls the scale and β the shape. In the three-parameter Weibull distribution, the third parameter is the threshold which is the smallest observation possible. The smallest observation possible is 0 for the two-parameter Weibull distribution. The three-parameter Weibull distribution is briefly mentioned in 5.3.9.5.

The probability density of the two-parameter Weibull distribution is as follows:

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \mathrm{e}^{-\left(\frac{t}{\alpha}\right)^{\beta}}, \text{ for } t > 0 \qquad (1)$$

For $\beta = 1$, the expression [Equation (1)] simplifies to

$$f(t) = \frac{1}{\alpha}\mathrm{e}^{-\frac{t}{\alpha}} \text{ for } t > 0 \qquad (2)$$

which is the density of the exponential distribution with mean equal to α. Figure 29 shows the way the Weibull density function changes shape for the case $\alpha = 1$ as β increases from ½ to 4. Four different situations can be distinguished and they are described as follows:

1) $\beta < 1$ represents a range of hyper-exponential distributions;

2) $\beta = 1$ represents an exponential distribution;

3) $1 < \beta < 3{,}5$ represents a range of skew distributions with the skewness decreasing as beta increases until at about 3,5 the distribution looks roughly symmetrically normal;

4) $\beta > 3{,}5$ represents a distribution that stays largely symmetrical (slight skewness) and becomes progressively more peaky as beta increases.

Increasing or decreasing α has the effect of simply stretching or compressing the horizontal scale.

The probability that the failure time is less than t is given by Equation (3):

$$F(t) = 1 - \mathrm{e}^{-\left(\frac{t}{\alpha}\right)^{\beta}} \quad \text{for } t > 0 \tag{3}$$

The *reliability function*, sometimes called the *survival function*, $R(t)$, is the probability that an item is still functioning at time t, so it is the complement of $F(t)$, i.e.:

$$R(t) = 1 - F(t) = \mathrm{e}^{-\left(\frac{t}{\alpha}\right)^{\beta}} \quad \text{for } t > 0 \tag{4}$$

It is often impracticable to continue a trial until all the members of the sample reach the end of their lives. For example, twenty light bulbs may be switched on and left burning in order to provide information about their lifetime distribution. To prevent the trial going on indefinitely, a time limit may be set, say at 1 500 h, at which time the trial will be stopped. Alternatively, it may be decided in advance that the trial will be stopped when a specified number, say 15, of the light bulbs have burnt out. Both lead to what is called censored data, the former with respect to time and the latter with respect to numbers of failures. For small samples, it can be important when estimating parameters and calculating confidence intervals to take account of which type of censoring was used.

5.3.9.2 The failure rates of the Weibull distributions

Because of the application of the Weibull distribution to reliability or survival data, it is convenient to consider the (*instantaneous) failure rate* (also known as the hazard rate) of the Weibull distribution. It is the ratio of the density function [Equation (1)] to the reliability function [Equation (4)], so it is

$$r(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \quad \text{for } t > 0 \tag{5}$$

The probability that an item will fail in the small interval between time t and $t + \mathrm{d}t$ given that it has been functioning up to time t is given as $r(t)\mathrm{d}t$, where $\mathrm{d}t$ is the length of the short interval.

The three possible failure regimes are often represented by the generic *bathtub* curve, that of infant mortality (decreasing failure rate), prime of life (constant failure rate) and wear-out (increasing failure rate).

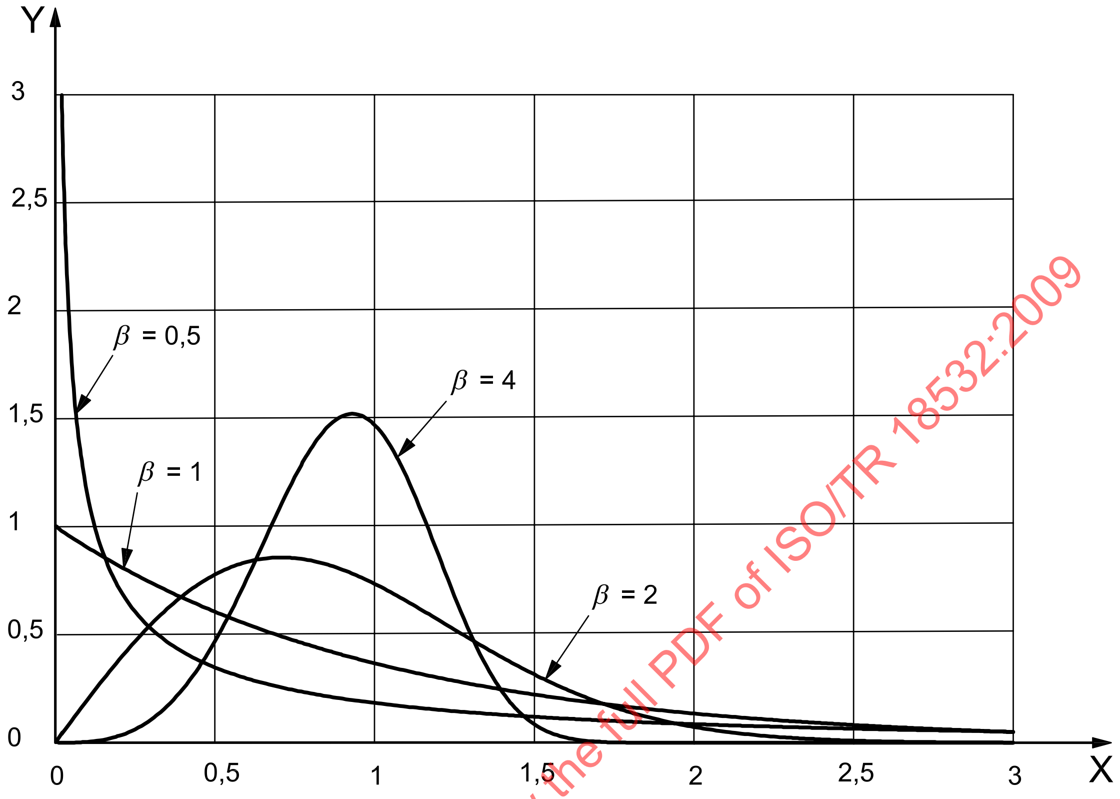
The Weibull β parameter distinguishes between these regimes, thus:

a) $\beta < 1$ represents a decreasing failure rate regime (colloquially called infant mortality);

b) $\beta = 1$ represents a constant failure rate regime (often called prime of life);

c) $\beta > 1$ represents an increasing failure rate regime (frequently termed wear-out).

So the failure rate of a Weibull distribution does not look like a bathtub curve. It is either decreasing, constant, or increasing.



Key

X t

Y probability density

Figure 29 — Comparison with Weibull distributions, all with $\alpha = 1$

5.3.9.3 Does the Weibull distribution fit the data?

Although it is somewhat subjective, the easiest way to check if the two-parameter Weibull distribution will provide a reasonable fit to a given set of data is to use a graphical method. It is based on the fact, which follows from the expression for the reliability function [Equation (4)], that:

$$\ln\left[-\ln(1-F(t))\right] = \ln\left[\left(\frac{t}{\alpha}\right)^{\beta}\right] = \beta\ln(t) - \beta\ln(\alpha) \qquad (6)$$

which is a straight line relationship between ln[–ln(1 – $F(t)$)] and ln(t).

The plotting procedure is as follows. The n sample values are first arranged in ascending order to give the *order statistics* $t_{[1]}$, $t_{[2]}$, ..., $t_{[n]}$, i.e. such that $t_{[1]} \leqslant t_{[2]} \leqslant \ldots \leqslant t_{[n]}$. For each $t_{[i]}$, the value of $F(t_{[i]})$ is estimated by $(i - 0{,}5)/n$, and the point with co-ordinates (ln($t_{[i]}$), ln{–ln[1–(i – 0,5)/n]} is plotted. If the data is a sample from a Weibull distribution, the plotted points will lie approximately on a straight line. The reason that the plotted points will only approximately lie on a straight line even if the data is a sample from the Weibull distribution is that $(i - 0{,}5)/n$ is just an estimate of $F(t_{[i]})$, and all estimates are subject to random variation. So this graphical method consists of making the plot as described and deciding whether or not the points lie approximately on a straight line. It is the last part that is described as somewhat subjective in the beginning of this clause.

The graphical method is illustrated with the data in the example in 5.3.9.4, and the plot is shown in Figure 30.

5.3.9.4 Example: Days between accidents in a company

The following data are the times in days between accidents in a large industrial plant. The data are recorded in days except for one case, where two accidents occurred within a few hours on the same day. The data are given ascending order in column 1 of Table 8. The calculations needed to make the plot to check if the data can be described by a Weibull distribution are given in columns 2 to 5. Column 2 contains the natural logarithm of the observations, which will be plotted along the first axis. The remaining three columns go through the calculations needed to find the quantity that will be plotted along the second axis. First, the observations are numbered from 1 to n in column 3. This number is used to calculate $(i - 0,5)/n$ which is shown in column 4. Finally, the logarithm of minus the logarithm of the values of column 4 is shown in the fifth column. The Weibull plot of Figure 30 is obtained by plotting the second column against the fifth column.

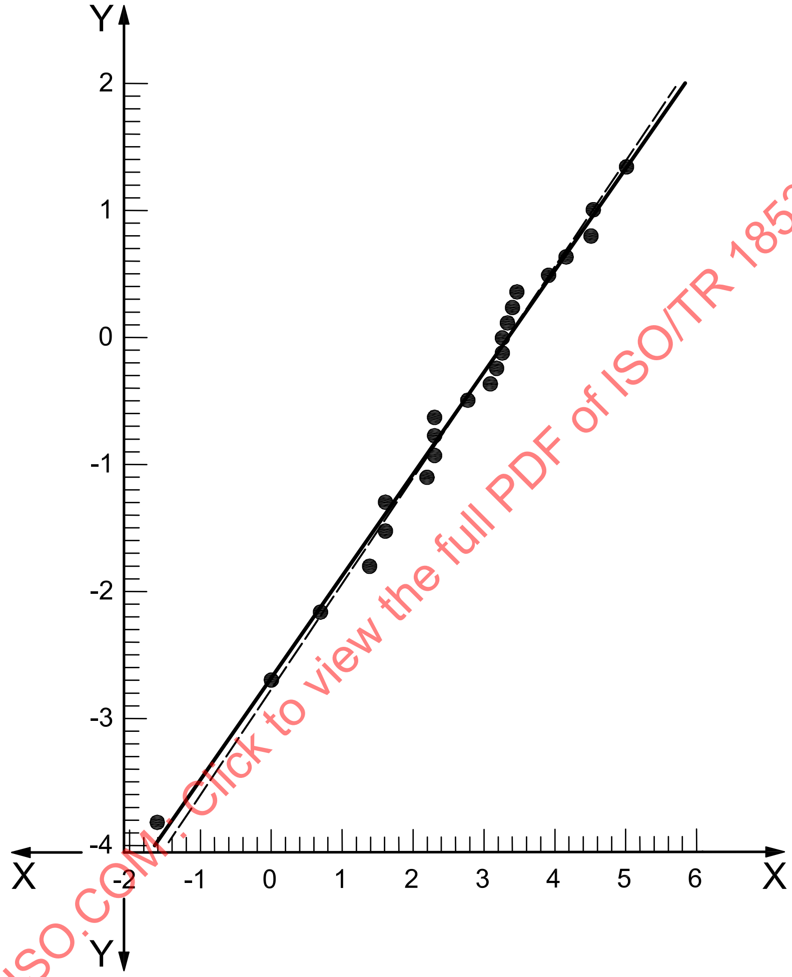
The points lie sufficiently close to a straight line, so the plot confirms that the data can be described by a Weibull distribution. The full line on the plot is a line fitted by eye, or rather by least squares. The dashed line corresponds to the Weibull distribution fitted by the maximum likelihood method.

Table 8 — Days between accidents in the first column sorted in ascending order

Days between accidents in ascending order	**Logarithm** (ln) **of days between accidents**	**Number of observation after ordering** i	$(i - 0,5)/n$	ln{−ln[1−(i−0,5)/n]}
0,2	−1,61	1	0,022	−3,81
1	0,00	2	0,065	−2,70
2	0,69	3	0,109	−2,16
4	1,39	4	0,152	−1,80
5	1,61	5	0,196	−1,52
5	1,61	6	0,239	−1,30
9	2,20	7	0,283	−1,10
10	2,30	8	0,326	−0,93
10	2,30	9	0,370	−0,77
10	2,30	10	0,413	−0,63
16	2,77	11	0,457	−0,49
22	3,09	12	0,500	−0,37
24	3,18	13	0,543	−0,24
26	3,26	14	0,587	−0,12
26	3,26	15	0,630	−0,01
28	3,22	16	0,674	0,11
30	3,40	17	0,717	0,23
32	3,47	18	0,761	0,36
50	3,91	19	0,804	0,49
64	4,16	20	0,848	0,63
91	4,51	21	0,891	0,80
94	4,54	22	0,935	1,01
150	5,01	23	0,978	1,34

NOTE The second column is plotted against the fifth column to produce the Weibull plot of Figure 30.

Before the pocket calculators and computers became common, it was customary to use Weibull plotting paper, where the transformations given in Table 8 have been applied to the axes, so the ordered observations can be plotted directly against the estimates of the distribution function $(i - 0{,}5)/n$ expressed in percent, i.e. against $100(i - 0{,}5)/n$ percent. This tradition is carried forward in today's statistical software where a Weibull probability plot is exactly as described. An example is shown in Figure 31, where the original observations (column 1 of Table 8) are plotted against the estimates of the distribution function (column 4 of Table 8), expressed in percent.



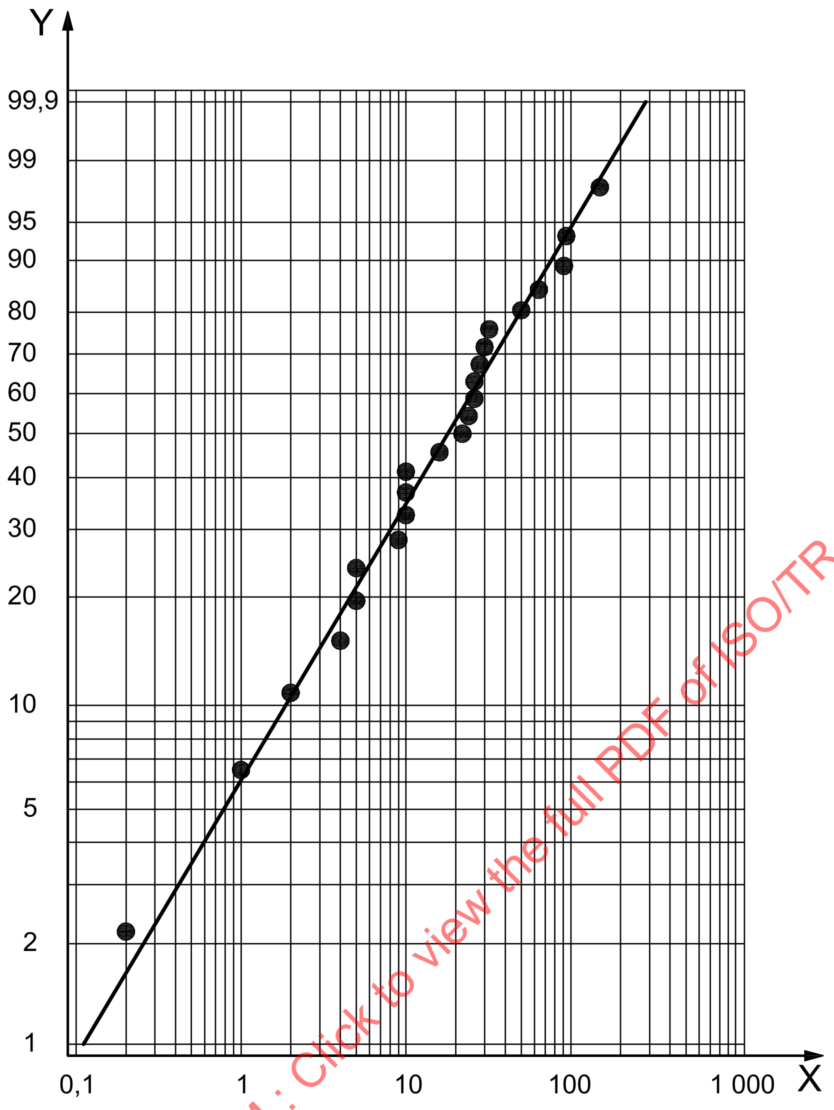
Key

X logarithm of ordered observations, $\ln t_{(i)}$

Y $\ln\{-\ln[1 - (i - 0{,}5)/n]\}$

The full line on the plot is a line fitted by eye, or rather by least squares. The dashed line corresponds to the Weibull distribution fitted by the maximum likelihood method. The plot is equivalent to the probability plot of Figure 31.

Figure 30 — Q-Q plot to assess the fit of days between accidents (data in column one of Table 8) to a Weibull distribution



Key

X ordered observations

Y estimated distribution function in percent: 100(i – 0,5)/n %

Figure 31 — Weibull probability plot of days between accidents (data in column one of Table 8)

5.3.9.5 The three-parameter Weibull distributions

On those occasions when the possibility of failure (or the origin of the distribution) does not start at time zero, the third Weibull parameter, γ, comes into play as the point in time where the possibility of failure begins. The formulas for the density function, the reliability function, and the instantaneous failure rate are easily obtained by substituting $t - \gamma$ for t everywhere in the formulas.

The formula for the instantaneous failure rate for the three-parameter Weibull distribution, for example, is

$$r(t) = \frac{\beta}{\alpha}\left(\frac{t-\gamma}{\alpha}\right)^{\beta-1}, \text{ for } t-\gamma > 0 \qquad (7)$$

Reliability illustrations are when it is feasible for an entity to fail before operation. For example, it fails on the shelf or is found dead on delivery. When γ is not zero, such a situation is recognized in a Weibull probability plot by the data points lying on a smooth curve rather than on a straight line. In such a case, the estimated value of γ is subtracted from each of the plotted points and the new values re-plotted. It may take a few iterations to arrive at a best estimate straight line and hence at a reliable value for γ.

5.3.10 Graphs

A graph is essentially a representation of data by a continuous curve (a line on a graph may be referred to as a curve even though it may be straight). Graphs are constructed to provide visual communication of information with clarity and precision. There are various forms of graphs, such as the following:

a) arithmetic (linear): these are the most familiar and are easily identified by the fact that both horizontal and vertical scales are arithmetic (linear) (see Figure 32);

b) log-linear: semi-logarithmic graphs have a linear horizontal scale and a logarithmic vertical scale and are used to display rates of change; a constant rate of change will appear as a straight line;

c) log-log: these have both scales logarithmic and are used to express learning curves in straight line form;

d) Q-Q plot or probability plot: these transform a regular pattern of variation into a straight line (see examples in Figure 26, Figure 30 and Figure 31);

e) nomograph: these provide graphical solutions to formulae.

5.3.11 Scatter diagram and regression

A scatter diagram or scatter plot is used to display possible relationships between one variable and another. Often a sample correlation coefficient is calculated when the covariance of two variables is of interest. The expression for the sample correlation is

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

The data in the table below is plotted as data set A in Figure 32.

	Observation										
	1	2	3	4	5	6	7	8	9	10	11
X	10	8	13	9	11	14	6	4	12	7	5
Y	8,04	6,96	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68

Taking this data set as an example, we calculate the numerator as

$$\sum(x-\bar{x})(y-\bar{y}) = \sum xy - \left(\sum x\right)\left(\sum y\right)\Big/n = 797{,}600\,0 - 99 \cdot 82{,}51/11 = 55{,}01$$

and the two sums of squares in the denominator as

$$\sum(x-\bar{x})^2 = \sum x^2 - \left(\sum x\right)^2\Big/n = 1001 - 99 \cdot 99/11 = 110$$

and

$$\sum(y-\bar{y})^2 = \sum y^2 - \left(\sum y\right)^2\Big/n = 660{,}172\,7 - 82{,}51 \cdot 82{,}51/11 = 46{,}661\,6$$

Finally, the sample correlation coefficient is calculated as

$$r = \frac{55{,}01}{\sqrt{110 \cdot 46{,}6616}} = 0{,}816$$

The sample correlation coefficient is always between −1 and 1, and it is 1 if the data points is on a straight line with a positive slope. Similarly, the sample correlation is −1 if the data points lie on straight line with a negative slope. This occasionally leads to the misconception that the sample correlation coefficient measures the degree of linearity of two variables in a sample. This is not the cases as is illustrated by the four samples in Figure 32.

Anscombe [65] has given four data sets of observations on pairs of variables. The scatter plots are given in Figure 32. All four cases give rise to the same fitted regression line and the same sample correlation coefficient, but it is only data set A for which it might make sense to fit a regression line or to report a sample correlation coefficient.

This emphasizes that when the relationship between two variables is of interest, it is very important to make the scatter diagram before any assumptions are made about the nature of the relationship. Indeed, Figure 32 B indicates a possible quadratic relationship between Y and X, Figure 32 C indicates a possible linear relationship together with a single outlier, while Figure 32 D simply shows variation in Y at X = 8 together with a single value of Y at X = 19, which is insufficient to indicate the form of any underlying relationship between Y and X.

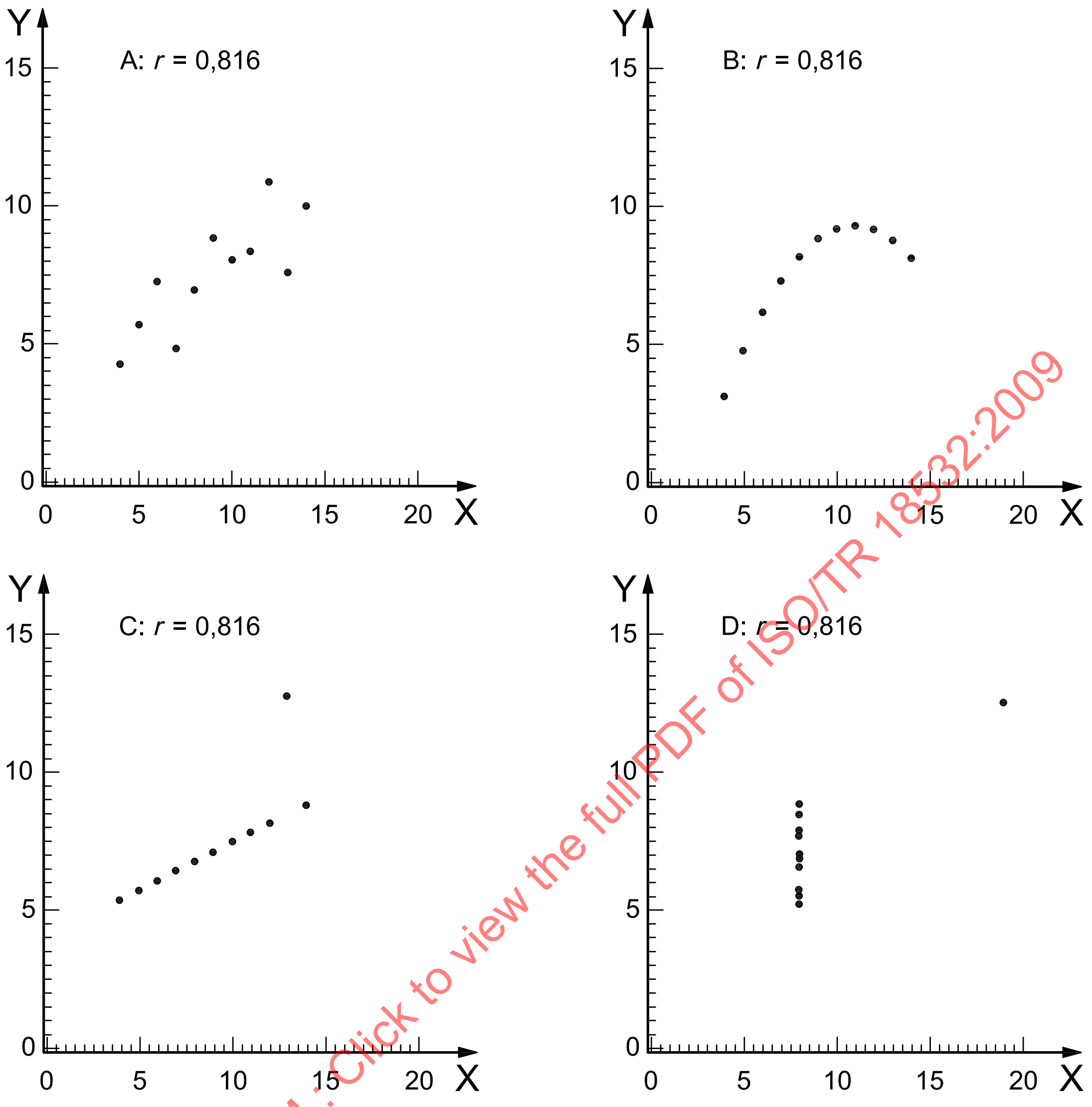
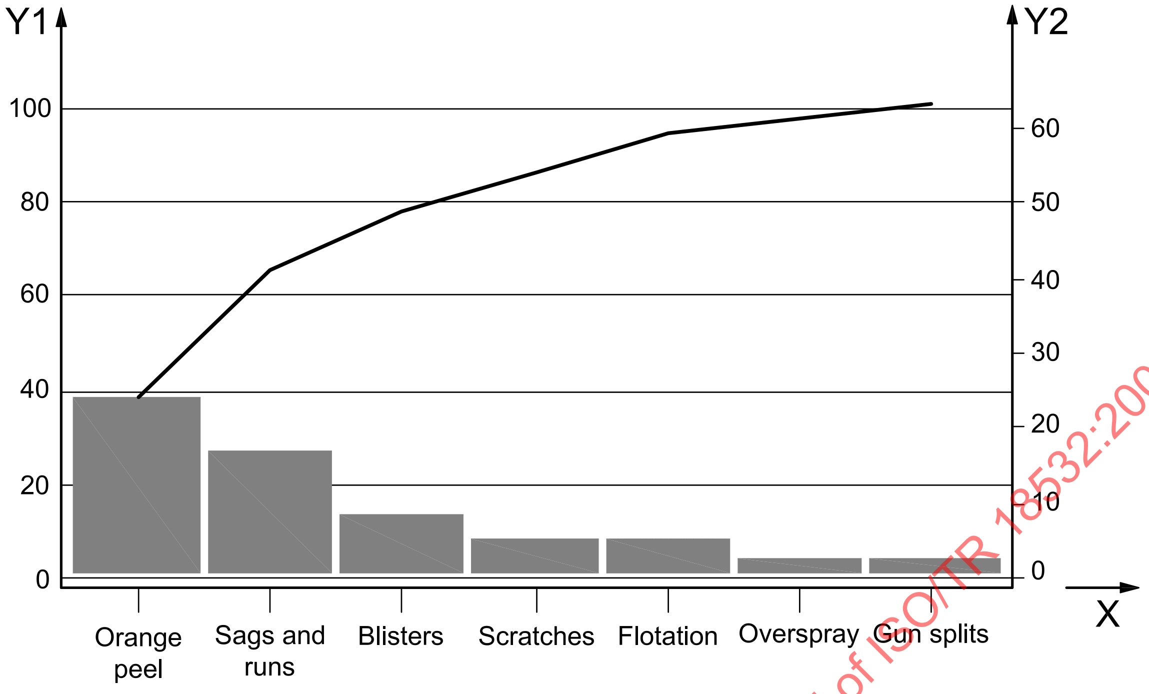


Figure 32 — Scatter diagrams of four data sets that all have the same correlation coefficients (*r*) and fitted regression lines

5.3.12 Pareto (or Lorenz) diagram

A Pareto diagram is a simple graphical technique for displaying the relative importance of features, problems or causes of problems as a basis for establishing priorities. It distinguishes between the vital few and the trivial many and hence focuses attention on issues where maximum quality improvement may be secured most quickly.

It displays, in decreasing order, the relative contribution of each element (or cause) to the total situation (problem). Relative contribution may be based on relative frequency, relative cost or some other measure of impact. Contributions are shown in bar chart form. Sometimes a cumulative line may be added to show the accumulated contribution. An example is shown in Figure 33, which shows that orange peel and sags and runs make up some 65 % of total paint faults in a particular paint shop. These were selected for priority attention in a quality improvement drive.



Count	24	17	8	5	5	2	2
Per cent	38,1	27,0	12,7	7,9	7,9	3,2	3,2
Cum. %	38,1	65,1	77,8	85,7	93,7	96,8	100,0

Key

X type of in-process fault

Y1 percent

Y2 number of occurrences

NOTE Pareto chart for defects — Relative contribution of different types of in-process faults with specified action priorities in process paint faults.

Figure 33 — Relative contribution of different types of in-process paint faults

5.3.13 Cause and effect diagram

A cause and effect diagram is frequently referred to as a fishbone diagram (because of its shape) or an Ishikawa diagram (after its creator). It applies where it is required to show, pictorially, cause and effect relationships. There are several types, based on the formation of the main branches (categories), including:

a) general 4M (manpower, machines, materials, methods);

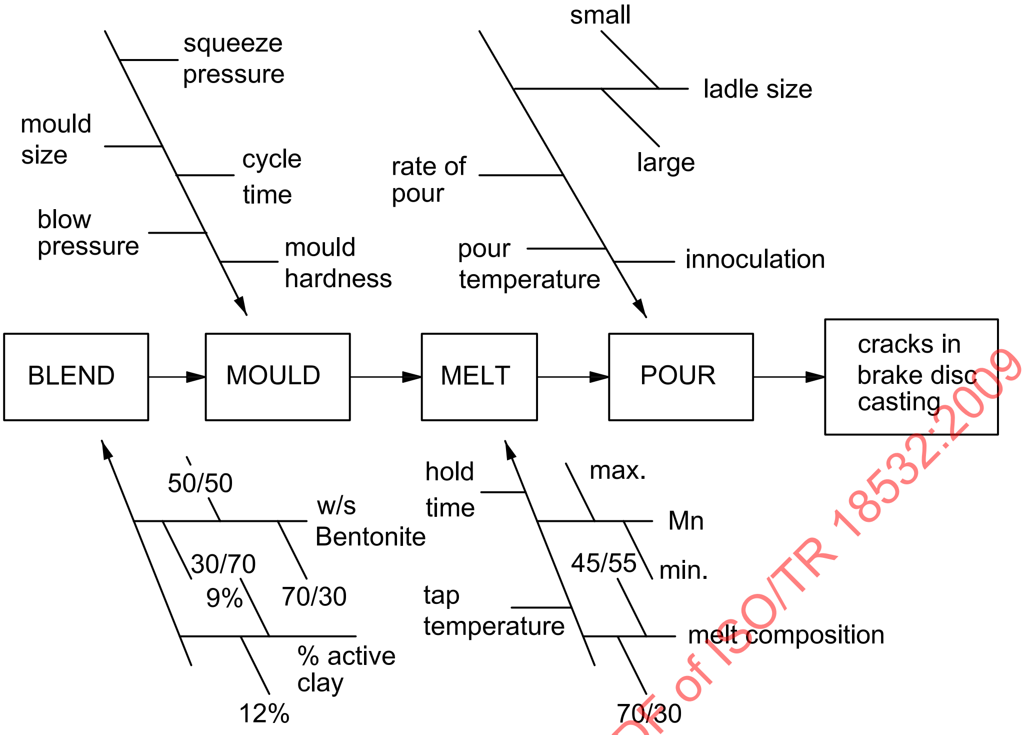
b) general 4P (people, procedures, plant, process);

c) process (by process steps and sequence);

d) assembly (by subassemblies);

e) specific (by technical consideration).

A process cause and effect diagram for a foundry process is shown in Figure 34. An earlier example is given in Figure 2.



NOTE w/s identifies the origin of the bentonite, namely western/southern.

Figure 34 — Process cause and effect diagram for cracks in a casting

6 Variation and sampling considerations

6.1 Statistical control and process capability

6.1.1 Statistical control

Post-process 100 % inspection is often neither practicable, relevant nor timely enough to meet today's needs. Monitoring in real time is required to enable processes to be steered and managed in an effective manner.

From economic considerations, amongst others, monitoring usually involves the assessment of process parameters and resulting product characteristics from a limited number of observations or items, which are defined in statistical terms as a sample.

It is essential to take the variation between similar items or observations into account when considering the relation of the sample to the totality of objects under consideration. This is true whether the object is a process parameter such as teeming temperature, a constituent of a material such as fraction silicon (in %), or a characteristic of a product such as a the diameter of a rod. In all cases where sampling is undertaken, estimates for the totality of objects under consideration can only be answered satisfactorily with the aid of statistical treatment.

Two statistical definitions, relating to population and lot, are relevant to an understanding of the following text. A population is defined as the totality of objects under consideration. A lot is defined as a definite part of a population constituted under essentially the same conditions as the population with respect to the sampling purpose.

The relationship between sample and lot is the kernel of the problem. In discussing it, it is necessary to introduce certain ideas that may appear difficult because they are unfamiliar. The following paragraphs, if studied in conjunction with Figure 35, should convey the essential features of the statistician's method of approach.

Suppose that Figure 35 represents the results of tests made at a supplier on similar articles or component parts sampled five at a time, at regular intervals during production. For example, the data may relate to the number of millilitres of battery acid per bottle, or to the minimum temperature of operation of a certain device, etc. The results of each sample of five tests are shown in the figure as dots, with the measurements displayed on a horizontal scale. Six cases are presented, numbered 1 to 6, in each of which the results of 12 samples from consecutive lots are shown.

Beneath the dots for each case, curves have been drawn. These represent the hypothetical distribution of the measured characteristic that would be found were it possible to test all the items in the 12 lots that have been sampled.

Consider case 1. There are, of course, considerable differences between the dot patterns of the 12 samples. Yet a certain stability or uniformity in the variation from sample to sample is evident, which is clearly not so in case 5. In case 1, there is no indication that the production process is anything but stable through time, as the samples could quite easily be imagined to have been drawn from the same lot.

If the pattern of variation is stable, then:

a) when all that are available are the measurements of the characteristic in a random sample, it is possible to use these measurements to estimate the distribution curve of the characteristic for the process;

b) when the distribution curve is known from experience, it is possible to predict the nature of the variation to be expected from one random sample to another.

Notice that situations a) and b) are the inverse of one another.

In forming the estimate described in a), the larger the size of the sample, the more reliable the estimate. For example, consider the distribution curve shown for case 1. Its mean and standard deviation would be estimated more reliably from 12 samples of size 5 than from a single sample of size 5. This is simply common sense but, when it is required to draw inferences about the distribution curve that depend on the *extent* of this reliability, the assistance of statistical theory is necessary. Much of what follows in this Technical Report is concerned, directly or indirectly, with the problems of drawing inferences and making decisions related to the process distribution curves, based on limited information. The construction of confidence intervals, prediction intervals and statistical tolerance intervals, addressed in Clause 8, are but three examples of these problems.

The following example illustrates situation b):

EXAMPLE For safety and ease of transportation, car batteries are supplied dry, together with plastic bottles of acid. With too little acid per bottle, the battery electrodes will not be fully covered, while if there is too much, the cell of the battery could overflow or present the user with the problem of disposing of the surplus acid. Suppose it is known, based on extensive experience, that the extent to which the bottles are filled varies according to the normal distribution curve. Suppose also that the mean and standard deviation of this normal curve are known to be 729,0 ml and 2,0 ml respectively. Then statistical theory enables such statements as the following to be made.

1) The chance that the mean contents in a random sample of six bottles will fall below 726,5 ml is 0,001 1, or such a result may be expected only once in about 900 samples.

2) The chance that *at least one bottle* in a sample of six bottles will contain less than 726,5 ml is 0,488 3, i.e. this result is over 400 times as likely as the previous result.

Statistical theory leads one to *expect* these two chances to be entirely different; moreover, statistical theory is able to give precision to the expected.

When this characteristic of stability of distribution is obtained, as represented by case 1 of Figure 35, the process will be described as *stabl*e, or *under statistical control*. It is then possible to make use of the methods of statistical theory described in the following clauses for such purposes of inference or prediction as were referred to under a) and b) above.

Although it is not easy to give a precise non-mathematical definition of what is meant by saying that a process is under statistical control, the concept is not difficult to grasp. It will be illustrated below using Figure 35, by contrasting it with some cases where the pattern of variation from lot to lot is not in statistical control.

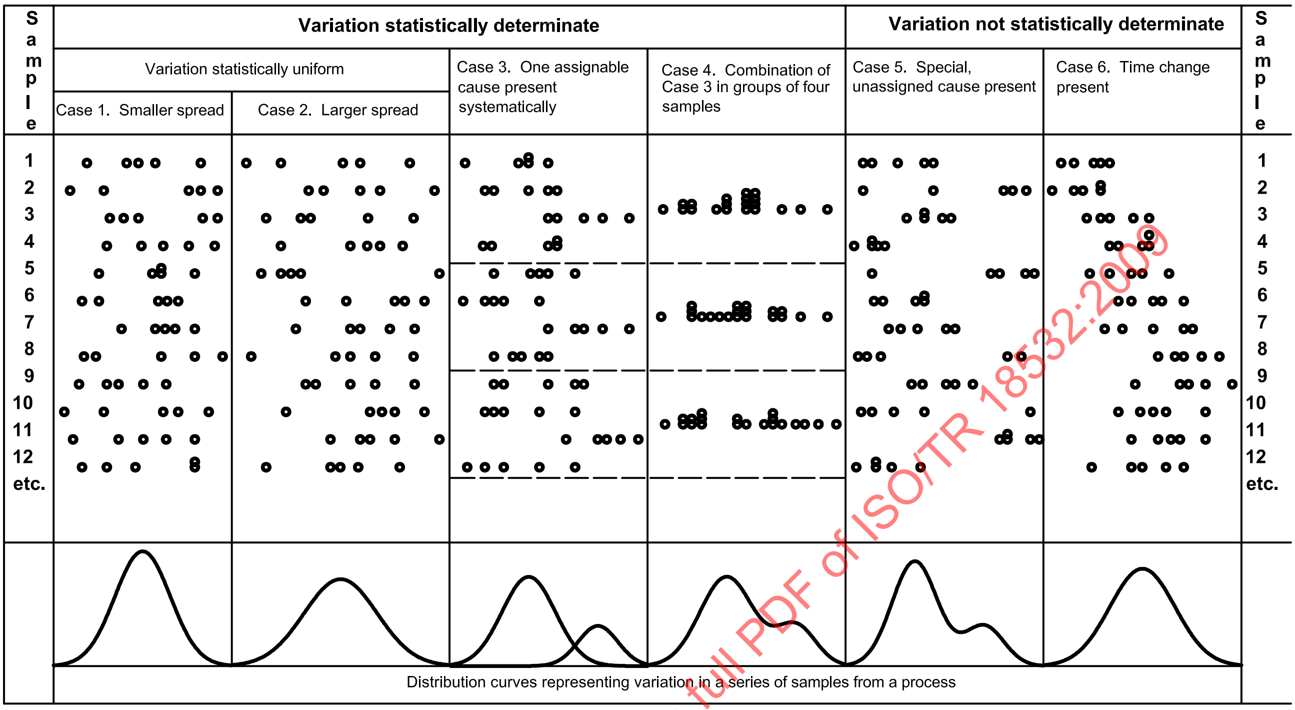

Sample
Variation statistically determinate
Variation not statistically determinate
Variation statistically uniform
Case 1. Smaller spread
Case 2. Larger spread
Case 3. One assignable cause present systematically
Case 4. Combination of Case 3 in groups of four samples
Case 5. Special, unassigned cause present
Case 6. Time change present
1
2
3
4
5
6
7
8
9
10
11
12
etc.
Distribution curves representing variation in a series of samples from a process

Figure 35 — Diagram indicating types of variation in samples

6.1.2 Erratic variation

While the variation in case 1 appears to be in statistical control, the variation for case 5 most certainly does not. A production process that leads without assignable cause to both sample No. 4 and sample No. 11 in case 5 can hardly be considered to be in statistical control. Indeed, the dot patterns suggest that there may be factors at work leading to two centres of variation with sometimes one operating, sometimes the other and sometimes both at the same time. This lack of homogeneity is suggested by the distribution curve for the 12 lots combined, shown beneath the dot patterns. The variation in case 5 may best be described as out of statistical control. Without any further understanding of the factors affecting the lot to lot variation, any attempt to predict the variation in subsequent products from the process is futile.

6.1.3 Systematic variation

Another situation is presented by case 3, where samples 3, 7 and 11 appear different from the others.

However, here there is a systematic repetition in the irregularities, which was not evident in case 5, and it may be the case that it is possible to determine a cause for these differences. If this were so, then the preferred procedure would be to eliminate the cause. However, if this were not possible, the series of lots could be divided into two homogeneous subseries, within each of which there is statistical control, and to each of which statistical methods could usefully be applied.

Suppose, in case 3, that samples 1, 5, 9, etc. were from sublots of material from one machine, samples 2, 6, 10, etc. from sublots from a second machine, 3, 7, 11, etc. from a third and 4, 8, 12, etc. from a fourth. If lots were formed by combining one sublot from each machine, then the samples of size 20 shown as case 4 could be described as a representative *stratified sample* of the total output, where each stratum (plural: strata) is the output from one machine. This illustrates the difference between simple random sampling and representative stratified random sampling. Under representative stratified random sampling, the output is divided into homogeneous parts (strata) from each of which a random subsample is drawn of size proportional to the part, and the subsamples are then combined into a representative stratified random sample. Thus, in the example just described, a quarter of each sample of size 20, i.e. five items, are selected from a sublot from each machine. Contrast this with a simple random sampling of the same size from the lot, under which every possible sample of the same size from the lot would have exactly the same chance of being selected. Thus, for example, under simple random sampling it would be possible for the sample of 20 items to contain no items at all from the third machine. Conversely, it would be possible for half or more of the items in the sample to come from the third machine. Clearly, such events would be undesirable in the present example. The aim of such stratified sampling is to secure as far as possible that the sample is a miniature of the whole lot, with similar distributions.

The samples taken from the sublots of material from the different machines all had the same size. Assuming that the sublots of material from each of the four machines also had the same size, the stratified random sampling we have undertaken, is a *proportional stratified random sampling* or, briefly, *proportional sampling*, in which the sizes of subsamples are proportionate to the sizes of the strata. This is the most common type of stratified random sampling.

To illustrate the advantage of representative stratified random sampling, consider the manufacture of sheet brass, the thickness of which at the edges is less than at the centre owing to the nature of the rolling process. If sheets are cut into narrower widths, the thickness will vary according to the position from which the strip has been cut. If this variation is recognized, the product will be divided accordingly into parts that will be homogeneous for sampling purposes. However, if it is not recognized, it is then possible that samples would sometimes consist of test pieces all taken from the edges, sometimes all from the centre, and sometimes from both in various proportions. The situation will then be as case 5, with no reliable inference possible from the sample measurements.

The distinction typified by the differences between case 1 and case 3 is an important one. In the former case, any one of the samples may be used to give information regarding the total output from the manufacturing process; in the latter, care needs to be exercised in choosing representative samples in the construction of the different strata. In the one, the variation within and between the samples from individual lots is no different from what might have been expected if a series of random samples had been selected from a consignment formed by first combining and mixing the items from the separate lots. In the other, when constructing the homogeneous strata, we have to identify the different sources and combine the material from the same source into a single stratum before we select any sample. In our example, we assume that the proportion taken from each source was proportional to the total quantity from that source, and that the drawing from within the source was at random.

Proportional stratified sampling is the special case of stratified sampling for which the sample sizes from each stratum are in proportion to the stratum size. Another special case is *optimal stratified sampling*, where the choice of sample size from each stratum under optimal stratified sampling takes into account the cost of sampling an item from each of the strata and prior estimates of the variability within each stratum. The choice is made with the objective of minimizing the cost of achieving a given precision in estimating the average value of the characteristic of interest, or maximizing the precision for a given cost.

Representative proportional stratified sampling is, therefore, a special case of optimal stratified sampling in which the strata have known and equal variability, and the cost of sampling an item from each stratum is the same.

Further discussion of stratified sampling is provided in 8.2.1.

6.1.4 Systematic changes with time

Case 6 represents another situation, where there is a systematic change with time taking place in the quality of material produced. If causes can be found for these fluctuations then, if these causes cannot be eliminated, it may alternatively be possible to divide the process output into streams that may separately be considered to be under statistical control. Examples are fluctuations due to known changes in temperature or humidity (perhaps in some textile process) or to differences between operators or shifts. However, if there are irregular fluctuations in time *without* known cause, prediction of characteristics of the process based on sample measurements from a few lots will be impossible. Thus, samples numbers 1 and 2, or again numbers 8 or 9, would not be representative of the typical process distribution shown below the dot patterns.

6.1.5 Statistical indeterminacy

If the total output of a particular article is made up from a number of sources, where each source is under statistical control and the proportion of the total coming from each is known, we have seen that statistical methods can be used to estimate the quality of the total from properly drawn samples. If, on the other hand, there is insufficient information to enable properly representative samples to be drawn in such a way, the variation is statistically indeterminate. This indeterminacy may be due to changes in space, e.g. from one machine or supplier to another; or it may be due to changes in quality with time, e.g. changes in the product from one supplier due to seasonal influences or changes in raw material.

6.1.6 Non-normal variation

It is important to realize that variation that is under statistical control is not necessarily represented by the normal distribution curve. It is true that the underlying distribution is *approximately* normal in a very large proportion of cases met with in industrial experience. Indeed, most of the methods described in Clauses 8 and 9 rely on the variation being approximately of the normal form. Nevertheless, it should be recognized that examples also abound for which the distribution curve of the measured characteristic is far from being symmetrical, e.g. the distributions of lifetimes and breaking loads, which typically have a long tail to the right. Yet provided the distribution remains stable from lot to lot, the concept of a process being in statistical control remains appropriate.

Distribution-free methods are discussed in 8.8.2 and 8.10.

6.1.7 Quality level and process capability

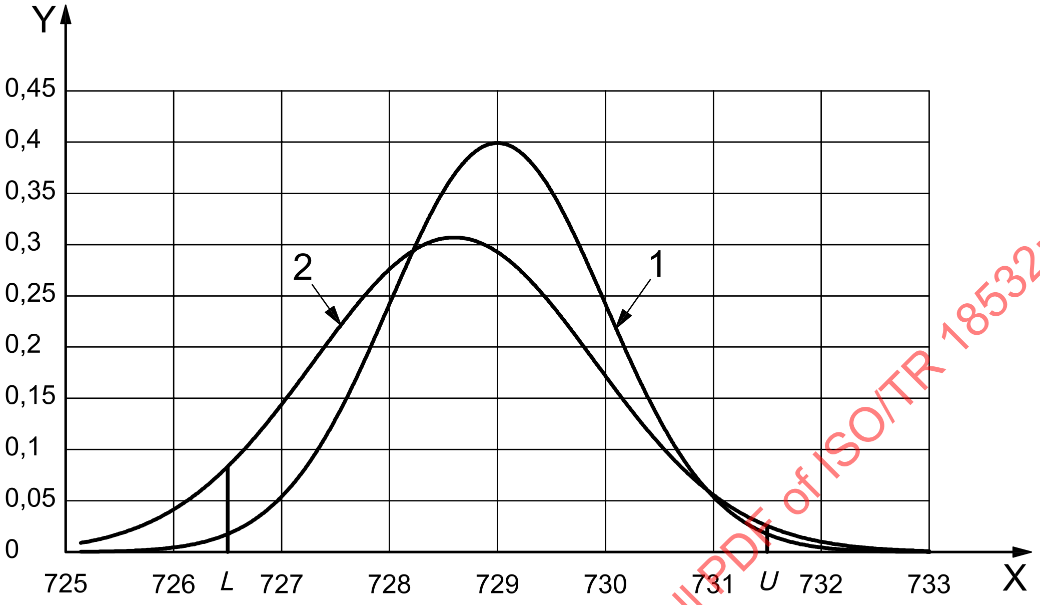
There is one further important concept, *process capability*, that needs to be introduced before dealing with the significance of these ideas to the supplier and the customer. Consider case 2, which has not yet been discussed. As in case 1, the variation in case 2 appears to be statistically uniform from lot to lot, i.e. under statistical control. In practice, evidence of stability is not enough. Statistical uniformity does not of itself indicate whether a process is operating at a high or at a low quality level. In order to be able to assess the quality level, information is also required concerning the process mean and the process variation. This information is provided by the sample mean and the sample standard deviation. (Incidentally, with this information, it will also be possible to detect a departure from uniformity during production, which will often enable adjustments to the process to be made to maintain a good quality level. Control chart methods that may be used for this purpose are discussed in Clause 10.) See also comments in 5.3.7.

The process variation for both case 1 and case 2 has been represented by normal distribution curves beneath their respective dot patterns in Figure 35. However, the cases differ in that case 2 has greater variation of individual items within samples than case 1. It has been previously pointed out that a normal distribution curve is completely defined by its mean, μ, and its standard deviation, σ. Denoting the process mean and process standard deviation for case 1 by μ_1 and σ_1, and for case 2 by μ_2 and σ_2, it is evident that the difference between case 2 and case 1 is that σ_2 is greater than σ_1.

Suppose that case 1 and case 2 represent the variation in the amount of car battery acid per bottle from two different filling machines with $\sigma_1 = 1{,}0$ ml and $\sigma_2 = 1{,}3$ ml. Suppose also that the mean contents also differ, say $\mu_1 = 729{,}0$ ml and $\mu_2 = 728{,}6$ ml, although this is not possible to see from Figure 35 as no scale is given. In both cases, statistical theory could be used to predict the proportion of bottles whose contents lie within *any* given limits. Suppose the specification is for a minimum of $L = 726{,}5$ ml and a maximum of $U = 731{,}5$ ml. The situation is shown in Figure 36.

Then, evidently, the capabilities of the filling machines to satisfy the requirements are different, with the first machine turning out a more homogeneous and more acceptable product. Indeed, it can be seen that the fraction of bottles that violate the lower limit for case 2 is many times that for case 1. Case 1 is therefore said to have greater *process capability*, i.e. the quality level of its output will be better than that of case 2. This distinction between the concepts of the statistical uniformity and the capability of a process is important.

One final remark about statistical uniformity, or statistical control, may be appropriate. They are terms used to describe the variation when the distribution curve *appears* to be stable from sample to sample. This stability is relative to the sampling technique employed, and is sometimes more apparent than real. For example, in a product that is being continuously produced, sampling at short intervals may identify a lack of statistical uniformity, e.g. a high-frequency cyclical effect, which sampling at longer intervals could fail to detect.



Key

1 case 1

2 case 2

X bottle contents, in millilitres

Y probability density

NOTE Comparison of capabilities of a normal distribution having $\mu_1 = 729{,}0$ and $\sigma_1 = 1{,}0$ with a normal distribution having $\mu_2 = 728{,}6$ and $\sigma_2 = 1{,}3$ when the specification limits are $L = 726{,}5$ and $U = 731{,}5$.

Figure 36 — Contrast of the capabilities of two filling machines

6.2 Sampling considerations

Consider now the way in which the principles discussed above bear upon the problems of sampling in practice. In general, to what extent are samples drawn to enable statistical theory to be profitably applied?

The question is too wide to give a single answer, as the methods of sampling that are practicable can vary enormously from one type of product to another. This notwithstanding, certain illustrations may profitably be presented to show some of the inherent difficulties and to indicate how they may be overcome.

Consider first the situation where the material sampled consists of a number of similar units, either component parts or finished articles. In some instances, it will be relatively straightforward to secure a random sample from a single well-mixed source of supply, for example in sampling small engineering parts such as ball bearings, bolts, screws, etc. A supplier who is confident that his process is in statistical control can adopt a simple procedure such as setting aside every 500th or 1 000th item (or whatever the need may be) to form samples for inspection purposes. The danger in such a procedure is of the time interval between the selection of sample items for inspection being in step with any periodic fluctuation in quality that may exist.

Were this to occur, the sample may well be biased, in which case the conclusions drawn from it would be misleading. Examples of possible reasons for such fluctuations are diurnal changes in temperature, increasing fatigue or inattention of operators during the course of shifts, or periodic replenishment of the raw material from which the product is made.

More often, however, the problem is not so simple. This is usually the case when sampling needs to be carried out not just to determine the acceptability of a lot but also to determine the grade, and therefore the price, of the product before acceptance by the user. As there is often no evidence available that the supplier's quality level has remained constant, it is important to plan a sampling procedure that will provide a reliable estimate of the quality of each lot, even if each lot is inhomogeneous.

In order to make valid, non-trivial generalizations from samples about characteristics of the populations from which they came, the samples must have been obtained by a sampling scheme which satisfies two conditions:

i) there must be a known relation between relevant characteristics in the population from which the samples are drawn and corresponding characteristics in the samples obtained by the sampling scheme;

ii) generalizations may be drawn from such samples in accordance with rules based on probability theory.

In order to satisfy these demands on the sampling scheme, the selection of the samples has to be done by some sort of *random selection*, which means that each possible sample has a fixed and determinable probability of being selected.

The most widely used type of random selection is *simple random sampling*. By this type of sampling, each unit in the population has the same probability of being the first unit to be selected for the sample; after the first member of the sample has been selected, each of the remaining units in the population has the same probability of being selected as a member of the sample and so on. The sampling scheme, simple random sampling, does not only demand that each item in the population has the same probability of being selected, but also that all possible samples of the same size have the same probability of being selected.

It is important to state that the *randomness* of a sample is inherent in the sample scheme used to obtain the sample and not an intrinsic property of the sample itself. Experience shows that it is not safe to assume that a sample selected haphazardly, without any conscious plan, can be regarded as if it had been selected by simple random sampling. Nor does it seem to be possible to consciously draw a sample *at random*.

If someone just “grabs a handful”, the items in the handful almost always resemble one another (on average) more than do the members of a simple random sample. Even if the “grabs” are randomly spread around so that every individual has an equal chance of entering the sample, there are difficulties. Since the individuals of grab samples resemble one another *more* than do individuals of random samples, it follows (by a simple mathematical argument) that the means of grab samples resemble one another *less* than the means of random samples of the same size. From a grab sample, therefore, we tend to *under*estimate the variability in the population, although we should *over*estimate it, in order to obtain valid estimates of grab sample means, by substituting such an estimate into the formula for the variability of means of simple random samples. Thus, using simple random sample formulae for grab sample means introduces a double bias, both parts of which lead to an unwarranted appearance of higher stability.

In order to make a random sample, we have to define a *target population* from which the sample is to be drawn, e.g. the units produced within the last hour, a *sampling frame*, which is a list of the items in the population, e.g. each unit has a different number, a *sampling design*, which is a pattern, arrangement or method for selecting a sample or sampling units from the target population, and a *sampling plan*, which is an operational plan, including the sampling design, for actually obtaining the sampling units for the sample.

Drawing a simple random sample may be done as follows:

a) assign to each unit in the target population a different number;

b) put each number on an individual slip of paper and put all the slips in a hat;

c) draw one slip at a time until there are as many slips drawn as units required in the sample.

The simple random sample consists of those units in the population corresponding to the drawn numbers.

In practice, sample selection is rarely done by means of slips and a hat. Instead, the units may be selected by means of a table of random digits or by means of a computer.

A table of random digits can be a table consisting of numbers with, for example, 6 digits. The table is constructed in such a way that all of the possible 6 digits have the same probability of occurring at a given entry in the table. If the target population consists of, for example, more than 100 and less than 1 000 units and one wants to select a sample of say $n = 10$, one starts at a random place in the table, e.g. at the top of the third column and considers the following numbers in a chosen direction from this one, e.g. downwards in the column. The first sampling unit to be selected from the target population is that one having as its population number the three first digits in the number chosen in the table, if such an item exists in the population, otherwise this number in the table is skipped. Then one considers the next number in the table in the chosen direction and selects as the next (or first) item that one having the three first digits as its number. Otherwise, this number is skipped. If the same population number comes up twice, it is skipped the second time. One continues in this way until the determined number n members of the sample have been selected.

By the way the table of random digits has been constructed, the procedure above will ensure that the sample is a simple random sample.

By computer, a simple random sample of size n from a target population of size N is drawn as follows.

1) Assign each sampling unit in the target population a number from 1 to N.

2) Generate N random numbers using a random-number generator. Assign the first random number to sampling unit number 1, the second to sampling unit number 2, and so on.

3) Sort the random numbers from smallest to largest (or vice versa).

4) Take as the sample the sampling units associated with the first n sorted random numbers.

Different computer packages give detailed instructions about how to do this.

The following illustration of random selection comes from the sampling subclause of EN 12326-1 [63]:

> Sampling shall be carried out by selecting slates from each lot separately in a random way so that every slate has an equal chance of being selected. Selected slates shall be marked so as to identify which lot they came from.
>
> When there is a possibility that the slates being tested may contain localized harmful inclusions such as calcite veins or oxidizable pyrite, the preparation of the test pieces shall be modified to ensure sufficient inclusions are contained in the specimen to provide a representative result.
>
> The acceptance procedure itself is not simple:
>
> > Where one or more of the tests do not satisfy the requirements of this standard, the unsatisfactory tests are repeated. If the results of the unsatisfactory test are confirmed, the lot is rejected or re-designated depending on the results.
> >
> > If the repeated test is satisfactory, a second check is carried out and if the result is satisfactory, the lot is accepted. If the repeated test is unsatisfactory, the lot shall be rejected or re-designated.

Different problems arise in sampling where material does not consist of discrete items, but is delivered in bulk, which for one reason or another may not be homogeneous. It is then necessary to withdraw small equal portions of material from a number of different parts of the whole mass. Alternatively, if the material is in movement on conveyors or in barrows, similar portions may be taken at regular intervals during the whole period of movement. The usual practice is then to combine these portions to form initial samples, which are then reduced after thorough mixing to form small quantities of material suitable for analysis in the laboratory. The object of these activities is to obtain final samples that mirror the distributions in the material as well as possible. ISO 3082 [9] illustrates the precautions that are necessary when sampling from bulk materials such as iron ore.

ISO 3082 contains diagrams of many types of sampling and dividing devices, illustrating the difficulties of obtaining samples from which sound inferences about the lot or consignment may be made. The following extracts from ISO 3082:2000 indicate the general considerations for sampling and sample preparation.

> The basic requirement for a correct sampling scheme is that all parts of the ore in the lot have an equal opportunity of being selected and becoming part of the partial sample or gross sample for analysis. Any deviation from this basic requirement can result in an unacceptable loss of accuracy and precision. Incorrect sampling schemes cannot be relied upon to provide samples that have similar distributions to those present in the whole material.
>
> The best sampling location to satisfy the above requirement depends on the type of sampling machine. Falling stream samplers require a transfer point between conveyor belts where the full cross-section of the ore stream can be conveniently intercepted at regular intervals. Sweep arm samplers that remove a cross-section of the ore stream directly off the conveyor belt are rapidly becoming the choice of preference because they offer a wide choice of sampling locations imposing minimal restrictions in regard to space and support. Here, the full cross-section of the ore stream can be conveniently intercepted at regular intervals, enabling samples to be drawn that reflect the variation in the material. However, precautions have to be taken to ensure that the regular intervals do not coincide with any regular fluctuations in the material.
>
> *In situ* sampling of ships, stockpiles, containers and bunkers is not permitted, because it is impossible to drive the sampling device down to the bottom and extract the full column of ore. Consequently, all parts of the lot do not have an equal opportunity of being sampled. The only effective procedure is sampling from a conveyor belt when ore is being conveyed to or from the ship, stockpile, container or bunker.
>
> *In situ* sampling from stationary situations such as wagons is permitted only for fine ore concentrates, provided the sampling device, e.g. a spear or auger, penetrates to the full depth of the concentrate at the point selected for sampling and the full column of concentrate is extracted.
>
> Moisture samples shall be processed as soon as possible, and test portions weighed immediately. If this is not possible, samples shall be stored in impervious airtight containers with a minimum of free air space to minimize any change in moisture content, but should be prepared without delay.
>
> Minimization of bias in sampling and sample preparation is vitally important. Unlike precision, which can be improved by collecting more increments or repeating measurements, bias cannot be reduced by replicating measurements. Consequently, the minimization or preferably elimination of possible biases should be regarded as more important than improvement of precision. Sources of bias that can be completely eliminated at the outset by correct design of the sampling and sample preparation system include sample spillage, sample contamination and incorrect extraction of increments, while sources that can be minimized but not completely eliminated include change in moisture content, loss of dust and particle degradation (for particle size determination).

In this example, there is no question of lots of ore being rejected; the sampling is solely to determine the grade and price.

If, from each lot, only one final sample were produced for laboratory analysis, there would be no way of assessing the reliability of the estimates of the lot characteristics. ISO 3084 [10] provides details of how this problem may be handled by the use of interleaved samples. These are "samples constituted by placing consecutive primary increments alternately into two sample containers" where the primary increments are the quantities of ore collected in a single operation of the sampling device. ISO 3084 provides for four scenarios.

I) When lots are frequently delivered, the quality variation may be determined from a large number of lots of almost equal mass by treating each lot separately and making up a pair of interleaved samples for each lot.

II) When large lots are infrequently delivered, the quality variation may be determined from a single lot by splitting the lot into at least 10 parts of almost equal mass and making up a pair of interleaved samples for each part.

III) When small lots are frequently delivered, the quality variation may be determined from several lots of almost equal mass by splitting all the lots involved into a total of at least 10 parts of almost equal mass and making up a pair of interleaved samples for each part.

IV) When sampling a wagon-borne lot where increments are taken from all wagons comprising the lot, the quality variation may be determined by treating each lot separately and making up a pair of interleaved samples for each lot.

Instructions are given in ISO 3084 for utilizing the information thus obtained under each scenario.

7 Methods of conformity assessment

7.1 The statistical concept of a population

To some extent, the customer and supplier have different viewpoints when it comes to the question of setting specifications. Broadly speaking, the customer is interested in the whole range of quality of individual items on the market from which he can purchase. The supplier has one eye on the competition, but is also concerned with the statistical control and capability of the production process, which can or needs to be maintained in a particular organization or organizations, having regard to technical and economic constraints. In both cases, however, the form of variation in the characteristics of a large collection of individual items is a matter of concern.

In discussing the concept of statistical uniformity, frequent reference has been made to the distribution curves shown at the bottom of the charts in Figure 35. These curves were drawn to represent the frequency distribution of a characteristic that would be obtained if measurements were made on a large collection or aggregation of items. In statistical terminology, the word *population* has been used to describe such a large collection of individual items, each possessing perhaps a number of different variable characteristics.

The use of this term arose because the early development of statistical method was associated with the study of human populations, formed of individuals, which were variable and many-charactered. In such a case, it is easy to grasp the concept of populations that are homogeneous or heterogeneous, stable or changing, the necessity of sampling, and different ways that this can be done by probability sampling or by purposely constructing a sample, and different types of samples: a simple random sample or a stratified random sample, a sample that properly reflects the variation in the data, or a biased sample, which does not, and so forth.

Deriving their origin from this special field of application, the terms *sample* and *population* have had associated with them very definite meanings in statistical theory. In the field of industrial production, the meaning of a sample is clear, but the concept of a population will perhaps be more readily understood if a different terminology is employed. It is sensible for the larger collection of items from which the sample is drawn (the statistician's *population*) to be described differently according to the particular situation under consideration. The terms output, consignment, batch or lot may each be used in their respective places, and no confusion would appear likely to arise since each of the terms will be found to be self-explanatory in its use.

The parallel with the human case can still be drawn. The different suppliers are the sources from which the output or consignments of manufactured items (corresponding to some extent to the different ethnic groups) are supplied.

The customer's interest in the populations, i.e. the outputs of the various suppliers who provide products of the type they desire to purchase, will depend on a number of considerations.

a) In certain cases, it will be essential for the variation in a characteristic to lie within narrowly defined limits. This will be so for the dimensions of component parts that need to be fitted together, for the analytical properties of certain chemical products, etc. The ideal, from the user's point of view, would probably be attained if the variation in items from all sources could be described by a normal curve with its mean on target and its standard deviation no greater than a certain value.

b) In other cases, wider latitude is permissible, so long as a minimum level is reached by virtually all the items; this is true, for example, when the qualities tested relate to strength or durability. For example, the average and standard deviation of the breaking strength of wire may differ considerably between suppliers, yet still meet the user's requirement.

c) Sometimes what is important to the user is not a particular mean value of a quality in a product, but a limit to the amount of variation about some mean that remains constant from one consignment to another. An example is products requiring craftsmanship in the finishing process, such as plasters or paints. The total material on the market may well be heterogeneous, consisting of several suppliers' outputs, all of which are of differing quality. However, the customer needs to be able to draw continually from one stable source of supply, i.e. to use material for which particular characteristics have a constant mean and low variation.

In all of the above cases, a statistical methodology is required that will indicate how best to determine from samples, whether the output or consignment does in fact conform to the desired standard.

The value of the guarantee that the specifications defined are satisfied depends on the statistical test for conformity applied. When the statement of conformity with the specifications shall be true beyond any reasonable doubt, the test for conformity shall be performed according to the principles given in ISO 10576-1 [38].

7.2 The basis of securing conformity to specification

7.2.1 The two principal methods

The provisions of a specification, the limits for the various quality characteristics, and the sampling technique to be adopted should be designed so as to provide assurance to the customer that each consignment or batch of material that he purchases is up to the stipulated standard. At the same time, the supplier will be required to know that the standard prescribed is one that is consistent with the capability of their production processes, and which is economically feasible for them to maintain. There are two principal methods of securing conformity to a specification:

a) by a system of tests of samples taken from batches of finished material. In certain cases, these samples may be drawn at random from the whole bulk of material; in other cases, it may be necessary to take special precautions to ensure that representative samples are obtained. In either event, this method is called *acceptance sampling*;

b) by requiring that records be kept that will provide statistical evidence of both the control and the capability of the manufacturing processes. Such a procedure could form the basis of a guarantee system of specification, so long as occasional audits, independent of the supplier, are made in order to satisfy the certifying authority that the routine tests are actually being carried out in the production facility.

Both these methods can form the basis of a system of quality marking or guarantee to show conformance with a specification. Statistical theory can assist by providing the user with assurance as to the adequacy of the sampling, and the supplier with confidence that no unsuspected variations in his processes are affecting the quality of the product. In many cases, however, it will be found on statistical analysis that acceptance sampling on an adequate scale will be too cumbersome or expensive, or even quite impracticable, while the second method would appear likely to provide effectively for a guarantee system.

It is not appropriate to make too sharp a distinction between supplier and customer, as the supplier will not only be a user of raw materials but will also be interested in the range of quality on the market of the commodities that the supplier is providing. Nevertheless, in making comparison of the two methods, it will be convenient to distinguish between questions of special importance to the customer and to the supplier.

7.2.2 Considerations of importance to the customer

The following considerations are of importance to the customer.

a) It has already been pointed out that in some cases it may be extremely difficult to obtain a sample that properly reflects the distributions within a consignment. This is well illustrated by the following example taken from Shewhart [123].

Given a consignment consisting of 10 truckloads of boxed material, there being 12 items in a box and roughly 1 000 boxes in a truck, how would it be possible to obtain a sample that mirrors the variation among these 120 000 items? Clearly, if the output were not homogeneous, certain boxes may contain articles of significantly different quality from others. According to the method of packing, these differences may be associated with certain trucks, or parts of a truck, or they may be scattered at random in the process of loading. Again, it may be the case that the articles at the bottom of each box are different from those at the top.

b) Statistical theory may show that the number of items that need to be tested to give the desired degree of information about the lot is prohibitive from an economic standpoint. This is particularly likely to be true where the test required is destructive and where, at the same time, there is considerable variation in the quality characteristic from item to item. For example, to burn out sufficient electric light bulbs to obtain a valid test of the difference in quality between the products of two manufacturers may not be beneficially viable in some cases.

c) For many processes, the process variation can be controlled successfully and the problem then becomes one of providing assurance that the process mean has not moved too far from the target value.

For a continuing series of lots from the same source, the amount of random sampling necessary can be very much reduced if the process variation remains demonstrably constant over time (i.e. from hour to hour and day to day). As soon as it is established that the process standard deviation is constant at a given value, smaller sample sizes can be used on subsequent lots. The process variation would still need to be checked and more intensive sampling resumed if evidence emerges that the process variation is no longer stable.

To reduce their intensity of sampling inspection with safety, it is necessary for the user to know that effective quality control procedures are in place. If the supplier maintains quality control records, what is required is an agreed means of making this information available to the customer. If access to such records is available, the question then naturally arises whether the guarantee system of securing conformity to specification would not be far more satisfactory than the method of testing samples from consignments. This is especially true for those materials for which inspection of the final product entails elaborate and costly procedures.

7.2.3 Considerations of importance to the supplier

The supplier is concerned with the day-to-day routine problem of turning out goods that will satisfy the requirements of a specification. As a more distant objective, probably involving research and experimentation, the supplier aims to reduce variation and increase the efficiency of the production process. Points they will consider are as follows.

a) If acceptance sampling is specified, the supplier who does not realise the waywardness of chance when dealing with variable material may find a sample from his product unexpectedly failing to pass specification. If, however, the supplier has studied and measured this variability, he may judge at what level quality should be maintained in order to reduce the risk of rejection to an acceptable level. Without this knowledge, the quality level he is maintaining for safety may in fact be uneconomically high.

b) The form of routine control required under a guarantee system of specification, depending on tests analysed on a statistical basis, is no different from that which is necessary to assure the same level of safety under an acceptance sampling system.

c) Stability in the quality of a manufactured product has a number of advantages to the supplier. Besides its relation to sales owing to user confidence, it may have an important bearing on the economy of management. The following example of a problem that might arise in the production of high-grade cotton fabric illustrates this point.

Owing to uncontrollable faults, a certain percentage of the lengths turned out by looms always needs to be put into a lower quality grade. If this percentage reaches a high figure, the manufacturer is faced with the necessity of disposing of this unwanted burden of low quality material, which is attached to his high-grade produce as an awkward but unavoidable shadow. Clearly, fluctuations in the magnitude of this percentage figure will upset his costing forecasts.

d) The concept of statistical uniformity has so far in this publication mainly been associated with the stability of variation in time. But the methods of statistical analysis that need to be used to decide whether variation is statistically uniform will also be invaluable to the supplier in research and development when he is attempting to reduce variability and to detect and eliminate sources of trouble.

For these aspects of the problem, reference should be made to textbooks, journal articles and standards on statistical process control, a brief selection of which may be found in the Bibliography. This is not strictly a problem of securing conformity with a specification, but some indication of its treatment is given in Clause 10.

From these considerations, it will be seen that there are clear advantages in the second method of securing conformity to a specification, namely by requiring that definite evidence be furnished of effective process control during manufacture. For this purpose, statistical theory can suggest systems for routine tests in the workplace. These will go far to arm a certifying authority with competence to assess the quality level of a product that is sold under a quality mark or guarantee.

In conclusion, the advantages of this method may be summarized as follows.

- It avoids the difficulty that often arises of determining how to draw a representative sample from a lot or consignment.
- It saves the cost of sampling on the large scale often necessary to give adequate assurance.
- The amount of sampling necessary will be far less than that required to provide definite protection in the face of erratic quality levels. This is generally true even where it is desirable to carry out occasional tests on samples from lots to gain assurance that the process control remains effective.
- The form of routine statistical analysis necessary to provide the basis of a system of quality certification is that which a supplier would employ anyway in attempting to increase the efficiency of his production process.

8 The statistical relationship between sample and population

8.1 The variation of the mean and the standard deviation in samples

8.1.1 General

In the preceding clauses, it has been explained why many of the problems that arise in attempting to achieve effective standardization of production and conformity to specification are essentially statistical in nature. A detailed development here of the relevant statistical theory would be inappropriate, but it is necessary to outline sufficient elements of this theory to clarify the treatment of some typical problems.

Suppose that the variation in initial efficiency of a specified type of electric light bulb is under consideration. If all the light bulbs in a lot of several thousand were tested, it would be possible to calculate the mean efficiency and the standard deviation of efficiency, measured in lumens per watt, for the whole lot. If, however, tests were only made on several samples each consisting of 10 lamps, then a different mean and a different standard deviation would be obtained for each sample. Not only would these means and standard deviations differ among themselves from sample to sample, but also they would not correspond exactly to the values that could, in theory, be determined from the whole lot. It is clearly important to have some means of defining the extent of the differences that can arise through the chance fluctuations of sampling.

Some further mathematical results need to be introduced in order to be able to do this in a precise manner. To avoid ambiguity, it is essential to make a clear distinction in the notation used for the characteristics of the population (lot, consignment) and that used for the characteristics of a sample drawn from this population. The most common notation for this is as follows. For the population, containing N items, the mean is denoted by μ and the standard deviation by σ. For a sample, containing n items, the sample mean is denoted by $\overline{x}$ and the sample standard deviation by s. Values of the sample characteristics for different samples are identified by the use of subscripts, for example:

— 1st sample, size n_1, sample mean $= \overline{x}_1$, sample standard deviation $= s_1$;

— 2nd sample, size n_2, sample mean $= \overline{x}_2$, sample standard deviation $= s_2$;

— 3rd sample, size n_3, sample mean $= \overline{x}_3$, sample standard deviation $= s_3$;

— 4th sample, size n_4, sample mean $= \overline{x}_4$, sample standard deviation $= s_4$.

If a number of random samples of the same size n were drawn from the population, the standard deviation of the resulting values of $\overline{x}$ would be a measure of the magnitude of the error likely to be involved in using the mean of just one sample of n items as an estimate of the population mean μ. This standard deviation of the sample means $\overline{x}_1$, $\overline{x}_2$, …, etc. is often called the standard error of the mean. Similarly, the standard deviation of the sample standard deviations s_1, s_2,…, etc. is often called the standard error of the standard deviation, since it measures the error involved in using s as an estimate of σ.

NOTE The standard deviation of an estimator is called the standard error. In order to avoid confusing the reader with terminology, the words “standard deviation” will be used instead of “standard error” throughout this Technical Report.

In general, it will not be practicable to take more than one random sample from the population. Fortunately, statistical theory comes to our aid by providing a means of estimating both of these standard deviations from the results of one sample. It will be assumed in 8.1.2 and 8.1.3 that the sample size n is small compared to the population size N, say less than one twentieth of N.

8.1.2 Variation of means

The average of the sample means from all possible samples of size n from the population equals the population mean, i.e.

$$\text{Mean of } \overline{x} = \mu \tag{8}$$

In fact, to make it clear that the left-hand side of this equation is a population mean rather than a sample mean, a better notation is:

$$\mu_{\overline{x}} = \mu \tag{9}$$

The variability of the sample mean varies directly as the standard deviation of the characteristic, x, in the population and inversely as the square root of the sample size, i.e.

$$\text{Standard deviation of } \overline{x} = \sigma / \sqrt{n} \tag{10}$$

Again, to make it clear that the left-hand side represents a population standard deviation rather than one based on a sample, a better notation is:

$$\sigma_{\overline{x}} = \sigma / \sqrt{n} \tag{11}$$

The variation in the sample means will approximate to a normal distribution except in cases of extremely asymmetrical variation in x in the population.

In simple terms, the standard deviation may therefore be interpreted as follows, provided that the sample contains some 20 or more items. Since $\bar{x}$ varies approximately in accordance with a normal distribution about μ with a standard deviation of $\sigma / \sqrt{n}$, it follows that it is rather unlikely that, in any particular random sample, the magnitude of the difference $(\bar{x} - \mu)$ will be greater than $2\sigma / \sqrt{n}$, and very unlikely that it will be greater than $3\sigma / \sqrt{n}$. Consequently, when only the data obtained from a sample are available, there is reasonable assurance that the population, lot or consignment mean will not differ from the sample mean $\bar{x}$ by more than ± 2 to ± 3 times $\sigma / \sqrt{n}$. If the standard deviation, σ, is not known from past experience, an estimate of σ obtained from the sample needs to be used. More precisely, tables of multipliers can be derived from statistical theory, such as Table 9 and Table 10, whose uses are described in 8.1.3.

Table 9 — Factors for confidence limits for the population mean and population standard deviation

Sample size n	Mean				Standard deviation							
	Limits $\mu_1 = \bar{x} - as,\ \mu_2 = \bar{x} + as$				Limits $\sigma_1 = b_1 s,\ \sigma_2 = b_2 s$							
	Chance of error				Chance of error							
	10 %	5 %	2 %	1 %	10 %		5 %		2 %		1 %	
	a	a	a	a	b_1	b_2	b_1	b_2	b_1	b_2	b_1	b_2
5	0,954	1,242	1,676	2,060	0,649	2,373	0,599	2,874	0,548	3,670	0,518	4,396
6	0,823	1,050	1,374	1,647	0,672	2,090	0,624	2,453	0,575	3,004	0,546	3,485
7	0,735	0,925	1,188	1,402	0,690	1,916	0,644	2,203	0,597	2,623	0,568	2,980
8	0,670	0,837	1,060	1,238	0,705	1,798	0,661	2,036	0,615	2,377	0,587	2,661
9	0,620	0,769	0,966	1,119	0,718	1,712	0,675	1,916	0,631	2,205	0,603	2,440
10	0,580	0,716	0,893	1,028	0,729	1,646	0,687	1,826	0,644	2,077	0,617	2,278
11	0,547	0,672	0,834	0,956	0,739	1,594	0,698	1,755	0,656	1,978	0,630	2,154
12	0,519	0,636	0,785	0,897	0,747	1,551	0,708	1,698	0,667	1,899	0,641	2,056
13	0,495	0,605	0,744	0,848	0,755	1,516	0,717	1,651	0,676	1,834	0,651	1,976
14	0,474	0,578	0,709	0,806	0,762	1,486	0,724	1,612	0,685	1,780	0,660	1,910
15	0,455	0,554	0,678	0,769	0,768	1,460	0,732	1,578	0,693	1,734	0,668	1,854
16	0,439	0,533	0,651	0,737	0,774	1,438	0,738	1,548	0,700	1,694	0,676	1,806
17	0,424	0,515	0,627	0,709	0,780	1,418	0,744	1,522	0,707	1,660	0,683	1,764
18	0,411	0,498	0,606	0,684	0,785	1,401	0,750	1,500	0,713	1,629	0,689	1,728
19	0,398	0,482	0,586	0,661	0,789	1,385	0,755	1,479	0,719	1,602	0,696	1,696
20	0,387	0,469	0,568	0,640	0,793	1,371	0,760	1,461	0,724	1,578	0,701	1,667
21	0,377	0,456	0,552	0,621	0,797	1,358	0,765	1,445	0,729	1,557	0,707	1,641
22	0,367	0,444	0,537	0,604	0,801	1,346	0,769	1,430	0,734	1,537	0,712	1,617
23	0,359	0,433	0,524	0,588	0,805	1,336	0,773	1,416	0,738	1,519	0,716	1,596
24	0,350	0,423	0,511	0,574	0,808	1,326	0,777	1,403	0,743	1,502	0,721	1,576
25	0,343	0,413	0,499	0,560	0,811	1,317	0,780	1,392	0,747	1,487	0,725	1,559
26	0,335	0,404	0,488	0,547	0,814	1,309	0,784	1,381	0,751	1,473	0,729	1,542
27	0,329	0,396	0,478	0,535	0,817	1,301	0,787	1,371	0,754	1,460	0,733	1,527
28	0,322	0,388	0,468	0,524	0,820	1,293	0,790	1,362	0,758	1,448	0,737	1,513
29	0,316	0,381	0,459	0,514	0,823	1,287	0,793	1,353	0,761	1,437	0,741	1,499
30	0,311	0,374	0,450	0,504	0,825	1,280	0,796	1,345	0,764	1,427	0,744	1,487

8.1.3 Variation of standard deviations

The square of a standard deviation is called a variance, i.e. s^2 is the sample variance and σ^2 is the population (lot or consignment) variance. The average of the sample variances over all possible samples of size n from the population equals the population variance, i.e.

$$\mu_{s^2} = \sigma^2 \tag{12}$$

Unfortunately, there is no such simple result for the average of all possible sample standard deviations for samples of size n. In fact, the average value of s is less than σ. This effect is described as a bias, a negative bias in this case. The bias depends on the sample size, and gets smaller as the sample size increases.

NOTE The fact that s is a biased estimator of σ although s^2 is an unbiased estimator of σ^2 may seem puzzling at first, but it should be remembered that the mean value of a set of numbers is not the same as the square root of the mean of their squares. For example:

$$\tfrac{1}{5}(1+3+3+5+6) = 3{,}6 \text{ whereas } \sqrt{\tfrac{1}{5}(1+9+9+25+36)} = 4{,}0$$

The average of the sample standard deviations over all possible samples of size n from the population is given by the following equation:

$$\mu_s = c_4\sigma \tag{13}$$

where c_4 is the bias factor and depends on the value of n.

NOTE This relation [Equation (13)] is only true if the variation among the observations is of the normal form.

Values of c_4 are given for sample sizes from 2 to 30 in Table 10; note that c_4 is approximately equal to $4(n-1)/(4n-3)$.

Table 10 — Factors for removing bias from sample standard deviations

Sample size n	c_4	$1/c_4$	**Sample size** n	c_4	$1/c_4$	**Sample size** n	c_4	$1/c_4$
—	—	—	11	0,975 4	1,025 3	21	0,987 6	1,012 6
2	0,797 9	1,253 3	12	0,977 6	1,023 0	22	0,988 2	1,012 0
3	0,886 2	1,128 4	13	0,979 4	1,021 0	23	0,988 7	1,011 4
4	0,921 3	1,085 4	14	0,981 0	1,019 4	24	0,989 2	1,010 9
5	0,940 0	1,063 8	15	0,982 3	1,018 0	25	0,989 6	1,010 5
6	0,951 5	1,050 9	16	0,983 5	1,016 8	26	0,910 1	1,010 0
7	0,959 4	1,042 4	17	0,984 5	1,015 7	27	0,990 4	1,009 7
8	0,965 0	1,036 2	18	0,985 4	1,014 8	28	0,990 8	1,009 3
9	0,969 3	1,031 7	19	0,986 2	1,014 0	29	0,991 1	1,009 0
10	0,972 7	1,028 1	20	0,986 9	1,013 2	30	0,991 4	1,008 7

If the sample standard deviation is to be used as an estimate of the population value, for some applications it is desirable or customary to eliminate the bias. Following Shewhart [124], this is done by taking s/c_4 as the estimate of σ.

With regard to these corrections, the following points should be noted.

a) Unless the sample contains very few items (i.e. unless n is very small), the bias is inconsequential.

b) If n is small, no single estimate of σ can be regarded as satisfactory; what is required is a pair of lower and upper limits σ_1 and σ_2 within which we may feel confident that σ lies. These are so-called "confidence limits" (see 8.4.1), for the calculation of which Table 9 has been given. This table shows very clearly the extent of the uncertainty that remains when σ is estimated from a few observations only.

c) Corrections in the case where σ is estimated from a *number* of small samples are, however, important, for in this case a really reliable unbiased estimate is possible. These corrections are discussed further in 10.7.

Corresponding to Equation (10), there is an approximate theoretical expression for the standard deviation of a standard deviation, namely:

$$\text{standard deviation of } s = \sigma / \sqrt{2(n-1)} \qquad (14)$$

i.e.

$$\sigma_s = \frac{\sigma}{\sqrt{2(n-1)}} \qquad (15)$$

The accuracy of this approximation is subject to several limitations:

1) the variation among the observations, x, has to approximately follow the normal distribution;

2) as σ will generally not be known, it will be necessary to substitute s in the right-hand side of Equation (15), so the sample should consist of at least 30 observations.

Subject to these restrictions, the result is nevertheless helpful in giving some idea of the reliability of s as an estimate of σ.

The data in Table 11 illustrates the variation in $\bar{x}$ and s among samples from the same population. A can of tomatoes is taken from a production line once every two hours, and its contents weighed. Four observations therefore become available every 8-hour shift. Observations over a period of 40 shifts are given in this table.

Statistical analysis of the type described in 4.4 shows that the variation from item to item is under statistical control during the period covered by the first 40 shifts.

Table 11 — Mass of tomato can contents

Shift	Mass of can contents g			
1	401,5	401,5	404,8	402,8
2	404,4	403,4	406,3	403,1
3	405,7	405,5	406,1	404,8
4	405,0	402,6	406,6	402,9
5	402,6	404,0	404,4	404,0
6	404,2	403,6	403,7	407,9
7	404,4	405,2	402,5	403,5
8	407,7	403,9	403,8	407,1
9	409,7	400,7	405,0	405,5
10	405,7	400,4	402,3	405,4
11	402,8	403,2	402,3	402,0
12	400,8	406,3	403,6	402,6
13	401,0	403,9	403,0	403,4
14	402,3	405,6	402,5	404,8
15	403,7	404,7	405,8	403,9
16	403,9	402,2	403,7	402,7
17	404,2	404,9	406,3	401,4
18	403,6	404,0	401,0	400,9
19	405,9	403,8	405,6	398,4
20	401,5	401,7	404,0	403,8
21	405,0	405,7	404,1	404,4
22	403,8	405,4	406,5	401,3
23	401,8	404,4	407,6	405,6
24	402,4	401,8	403,8	401,6
25	407,3	404,1	406,3	403,1
26	401,4	407,4	402,1	404,4
27	402,8	403,7	405,5	402,4
28	401,6	406,5	400,8	404,1
29	407,3	401,3	406,1	405,9
30	401,2	405,3	405,2	403,2
31	408,4	403,3	404,1	402,9
32	404,8	404,9	406,0	404,5
33	403,2	402,0	403,4	404,0
34	404,9	400,9	400,9	400,4
35	405,0	402,1	405,6	402,0
36	402,1	403,1	403,8	404,2
37	405,3	403,9	404,7	404,3
38	404,5	401,5	404,7	402,7
39	405,2	399,7	405,1	406,2
40	402,0	400,7	402,6	404,9

Table 12 — Canned tomatoes data — Mean and standard deviation of four masses per shift

Shift	Mean	Standard deviation
1	402,65	1,559
2	404,30	1,445
3	405,52	0,544
4	404,28	1,882
5	403,75	0,790
6	404,85	2,050
7	403,90	1,163
8	405,62	2,065
9	405,22	3,680
10	403,45	2,549
11	402,58	0,532
12	403,32	2,297
13	402,82	1,271
14	403,80	1,651
15	404,52	0,954
16	403,12	0,810
17	404,20	2,061
18	402,38	1,654
19	403,42	3,476
20	402,75	1,333
21	404,80	0,707
22	404,25	2,258
23	404,85	2,424
24	402,40	0,993
25	405,20	1,936
26	403,82	2,706
27	403,60	1,378
28	403,25	2,583
29	405,15	2,640
30	403,72	1,941
31	404,68	2,533
32	405,05	0,656
33	403,15	0,839
34	401,78	2,097
35	403,68	1,893
36	403,30	0,920
37	404,55	0,597
38	403,35	1,526
39	404,05	2,942
40	402,55	1,756

Table 13 — Canned tomatoes data — Frequency distribution of individual observations and of means and standard deviations of 4 tests

Mass g	Frequency	Mean, $\bar{x}$ g	Frequency	Standard deviation, s g	Frequency
398,00 to 398,99	1	400,50 to 400,99	1	0,500 to 0,999	11
399,00 to 399,99	1	401,00 to 401,49	2	1,000 to 1,499	5
400,00 to 400,99	9	401,50 to 401,99	5	1,500 to 1,999	9
401,00 to 401,99	16	402,00 to 402,49	8	2,000 to 2,499	7
402,00 to 402,99	26	402,50 to 402,99	7	2,500 to 2,999	6
403,00 to 403,99	30	403,00 to 403,49	5	3,000 to 3,499	1
404,00 to 404,99	32	403,50 to 403,99	6	3,500 to 3,999	1
405,00 to 405,99	25	404,00 to 404,49	4		
406,00 to 406,99	11	404,50 to 404,99	2		
407,00 to 407,99	7				
408,00 to 408,99	1				
409,00 to 409,99	1				
Total	160	Total	40	Total	40

The mean and standard deviation for the whole population of 160 masses are $\mu = 403{,}84$ g and $\sigma = 1{,}909$ g.

The means and standard deviations, $\bar{x}$ and s, of the 40 samples each consisting of 4 test results, are shown in Table 12. They have been grouped together in Table 13 where they form frequency distributions analogous to that in Table 6.

If the mean and the standard deviation of these distributions are calculated, they may be compared with the theoretical values obtained by inserting the values of n and s in Equations (9), (11), (13) and (14), as shown in Table 14.

Table 14 — Canned tomatoes data — Comparison of the sample mean and sample standard deviations of the means and standard deviations in groups of 4 tests with theoretical results

Measure	Results from Table 12	Results from theoretical formulae
Mean of $\bar{x}$, i.e. μ_x	403,84 g [a]	403,84 g [a]
Standard deviation of $\bar{x}$, i.e. $\sigma_{\bar{x}}$	0,965 g	
Mean of s, i.e. μ_s	1,727 g	$4(n-1)\sigma/(4n-3) = 12 \times 1{,}909/13 = 1{,}762$ g
Standard deviation of s, i.e. σ_s	0,808 g	$\sigma/\sqrt{2(n-1)} = 1{,}909/\sqrt{6} = 0{,}779$ g

[a] Note that these two values necessarily agree, as the mean of the 40 sample means has to equal the mean of the population of 160 tests.

Corresponding figures are seen to be in quite close agreement.

8.2 The reliability of a mean estimated from stratified and duplicate sampling

8.2.1 Stratified sampling

An important extension to some of the previous results is necessary when the items on which the observations or tests have been made are drawn from a number of sources, within each of which there is statistical control, but between which there may be differences in the process averages and variances.

Suppose a population (batch, consignment) consists of a large number of items, of which a proportion p_1 has come from one source, p_2 from a second, etc., and finally p_k from a kth source. Suppose that the mean and standard deviation of the quality characteristic under consideration in material from the first source are μ_1 and σ_1, for the second source μ_2 and σ_2, etc. The mean for the whole population will then be:

$$\mu = p_1\mu_1 + p_2\mu_2 + \cdots + p_k\mu_k \qquad (16)$$

The different sources can be considered as strata. And if the means within the different strata vary from stratum to stratum, we will get a more accurate estimate of the mean of the whole population if we select independent random samples from each separate stratum or source and combine the results from the different strata in an appropriate way. The simplest situation will be that a sample of n items is drawn by taking n_1 items at random from the material coming from source 1, n_2 items from the material from source 2, and so on, where:

$$n_1 = np_1,\ \ n_2 = np_2, \ldots,\ \ n_k = np_k \qquad (17)$$

and $n_1 + n_2 + \cdots + n_k = n$ since the proportions sum to unity, i.e. $p_1 + p_2 + \cdots + p_k = 1$. This is an example of a proportional stratified sample, or simply a proportional sample, where the number of items taken from each stratum is proportional to the size of each stratum or, in other words, that the sampling fraction from each stratum is equal to its relative size. (Stratified sampling has also been discussed in 6.1.3 to 6.1.5 and 6.2.)

If $\overline{x}$ is the mean value of the characteristic for the n items of this proportional sample, then it may be used as an estimate of the true population mean μ. For a proportional sample we simply pool the subsamples for all the strata and take the mean of this sample. It is known that the standard deviation (i.e. the standard deviation in repeated samples) of $\overline{x}$ is given by the relation:

$$\text{standard deviation of } \overline{x} = \frac{1}{n}\sqrt{n_1\sigma_1^2 + n_2\sigma_2^2 + \cdots + n_k\sigma_k^2} \qquad (18)$$

In practice, it will commonly happen that the standard deviations within the different sources of supply will be approximately the same, i.e. $\sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma$. This relation [Equation (18)] then simplifies to:

$$\text{standard deviation of } \overline{x} = \frac{\sigma}{\sqrt{n}} \qquad (19)$$

Results [Equations (18) and (19), like Equations (8) to (12)] are independent of any assumption of normality in the variation. A number of points of practical importance may be deduced.

a) Provided that the sample can be made properly proportionate by drawing subsamples that satisfy the condition [Equation (10)], the reliability of the estimate $\overline{x}$ depends only on the variation *within* each source of supply, and not upon the differences between the mean values μ_1, $\mu_2, \ldots, \mu_k$.

b) If the values of σ_1, $\sigma_2, \ldots, \sigma_k$ are not known from previous experience, they may be estimated from the standard deviations of the subsamples and by replacing the σ_i's in Equation (18) by the corresponding s_i's. An estimate of the standard deviation calculated from the whole sample, disregarding the fact that the subsamples originate from different strata (as we did above by estimating the mean), might be quite misleading. [See comments to Equation (22) below.]

c) If the samples from the different strata are not proportional to the relative sizes of the strata, then the estimate of the mean of the whole population will be a weighted sum of the sample means from the different strata with weights equal to the p_i's, the relative sizes of the strata. For stratified but not proportional samples, the standard deviation and the variance of the estimate of that estimate will also be more complicated because the variance has to be calculated for each subsample separately, even if the population variances within the strata are equal. The variance of the estimate of the total mean will be a weighted sum of the variances for the different strata and with weights equal to the squares of the p_i's."

d) If no attempt is made to take into consideration the differences between the mean values of the different strata, and a simple random sample is drawn instead of a stratified random sample, then the mean μ of the population will be estimated with less precision. The whole population of items then needs to be regarded as a single group having a standard deviation σ'. The mean $\bar{x}$ of the n observations in the sample drawn at random from the whole population will have a standard deviation given by Equation (10) in 8.1.2, i.e.:

$$\text{standard deviation of } \bar{x} = \frac{\sigma'}{\sqrt{n}} \tag{20}$$

In the case of proportional stratified sampling, in which the conditions [Equation (17)] are satisfied, it can be shown that:

$$\sigma' = \sqrt{p_1\sigma_1^2 + p_2\sigma_2^2 + \cdots + p_k\sigma_k^2 + p_1(\mu_1 - \mu)^2 + p_2(\mu_2 - \mu)^2 + \cdots + p_k(\mu_k - \mu)^2} \tag{21}$$

which, when $\sigma_1 = \sigma_2 = \cdots = \sigma_k = \sigma$, simplifies to:

$$\sigma' = \sqrt{\sigma^2 + p_1(\mu_1 - \mu)^2 + p_2(\mu_2 - \mu)^2 + \cdots + p_k(\mu_k - \mu)^2} \tag{22}$$

a quantity clearly at least equal to σ. Hence, if $\mu_1, \mu_2, \ldots, \mu_k$ are not all equal to μ, i.e. the process average changes from one source of supply to another, the mean of the batch will be estimated with less precision.

e) The results discussed above hold only when the sample fractions are small. If the sizes of the subsamples are larger than 0,1 of the stratum sizes, it is necessary to make so-called finite population corrections in the calculations of the variances and the standard deviations. This will cause a reduction in the variance. The larger the sample fractions are, the smaller the variance will be. However, the estimate of the overall mean is not affected by the finite population corrections.

f) Even when no exact attempt is made to estimate the values of $\sigma_1, \sigma_2, \ldots, \sigma_k$, stratified sampling is often employed to ensure that the resulting estimate of the population mean is as reliable as possible. In other words, although no calculations of reliability are made, sampling is in fact planned so that the standard deviation is given by Equations (18) or (19) rather than Equations (20) with (21) or (22).These points may be illustrated by the following numerical example. The strength and other properties of bricks depend to some extent on the position of the bricks during firing in the kiln. An investigation has shown that in a particular case the standard deviation of dry strength for the whole batch of bricks from a single firing of a kiln was given by:

$$\sigma' = 8\ 846 \text{ kPa}$$

If, however, the cross section of the kiln was divided into nine areas, the averaged standard deviation of brick strengths in a single area was the following:

$$\sigma = 5\ 081 \text{ kPa}$$

It follows that the producer, taking say four bricks at random from each of the areas, could obtain from the mean of the 36 test results an estimate of the mean brick strength of the kiln, having a standard deviation of the following:

$$\frac{\sigma}{\sqrt{36}} = 847 \text{ kPa}$$

NOTE It may be helpful to see this result obtained in two steps, as follows. The standard deviation for the mean result of tests made on four bricks from any one position in the kiln is $5\,081/\sqrt{4} = 2\,540$ kPa. As nine averages of similar tests are then themselves averaged, the standard deviation for the mean of the 36 tests will be $2\,540/\sqrt{9} = 847$ kPa.

The user, on the other hand, does not have the opportunity to obtain in this way a sample that reflects the variations in the data. It is likely, but by no means certain, that neighbouring bricks in the consignment he receives will have come from the same parts of the kiln. The user therefore needs to take a sample from the whole consignment and associate a standard deviation of $\sigma'/\sqrt{n}$ with the resulting estimate of the mean strength. To obtain an estimate about as reliable as that of the producer, which was based on 36 bricks, the user needs to test about $n = 110$ bricks because, approximately:

$$\frac{\sigma'}{\sqrt{110}} = 843 \text{ kPa}$$

It needs to be understood that the above argument holds only when principally it is the *mean* value of a characteristic that it is desired to control. In the example taken for illustration, the mean strength of a consignment of bricks is not in fact the best criterion for assessing quality. Both the mean strength and the standard deviation of strength require control, as they both relate to the proportion of bricks that are below a given strength (see 8.5 and 9.5.2). The question of how to plan sampling appropriate to the simultaneous estimation of the mean and the standard deviation involves other considerations, which cannot be entered into here. The comparison given, however, expresses in numerical form the advantage that follows if process data is obtained at the time of production rather than by the sampling of consignments.

8.2.2 Duplicate sampling

For certain products, it is the practice not to measure the characteristics of each individual item in the sample but to record a single value that is the grand total of the individual values. For example, the total mass of a sample of n articles may be taken, but not the n separate masses. For sampling bulked materials such as coal, cement and oil, the sampling methods may aim only at obtaining a single total measure as an estimate of quality. However, without additional information, it is impossible to determine the reliability of this single measure. This is because even if results from a number of consignments are collected and compared, the variation between them may be due to changes in the process mean value and not to sampling error.

The only satisfactory method of determining the reliability of a sampling procedure is for the same sampling procedure to be carried out independently several times on the same consignment or batch, and for the standard deviation of these independent results to be obtained. If the process variation remains approximately stable, then the reliability of the sampling procedure may be examined initially and rechecked only occasionally. An economic method of maintaining assurance of the continued reliability is to arrange that independent *duplicate* samples be taken. For example, as described in 6.2 for iron ore, for some products a number of small portions may be taken at regular intervals from a conveyor, alternate portions put into two separate receptacles, and the process of mixing, quartering, etc. and final analysis performed independently in duplicate.

Suppose that x_1 and x_2 are the two test results that are to be used as an estimate of the real quality of the consignment. Their difference may be expressed as $d = x_1 - x_2$. From statistical theory, it is known that the standard deviation of d in the consignment may be expressed in terms of the standard deviation of x in the consignment by the formula:

$$\sigma_d = \sqrt{2\sigma_x} \tag{23}$$

This result remains true even if the process mean value changes from one consignment to another, provided

a) that the variation about the mean in the parts of the sampled bulk is approximately the same in all consignments, and

b) that the duplicate samples are independent, e.g. if x_1 is above the real consignment value, then x_2 is as likely to be below as to be above.

It follows from Equation (23) that, if these conditions are satisfied, $\sigma_d / \sqrt{2}$ may be used as a measure of σ_x, the standard deviation of either of the estimates x_1 and x_2. As it is known that the standard deviation of the mean of x_1 and x_2 is $\sigma_x / \sqrt{2}$, it follows that $\sigma_d / \sqrt{2}$ may be used for the standard deviation of $\overline{x} = (x_1 + x_2)/2$. In order to keep a check on the continued reliability of the sampling and analytical processes, the successive values of $d = x_1 - x_2$ may be plotted on a control chart (see 10.7).

8.3 Illustration of the use of the mean mass, and the lowest mass, in a sample of prescribed size of specimens of fabric

The principles considered in 8.1 and 8.2 are of equal importance whether the assistance of statistical theory be required in connection with the method of consignment sampling or in assessing how well the process mean and variation are being controlled.

The following is an illustration of the use of the standard deviation of the mean, i.e. $\sigma / \sqrt{n}$, in consignment sampling. Example 2 in 4.3 gave some figures for the masses of 128 specimens from a roll of fabric. A potentially large-scale user first investigates the quality of such fabrics in the marketplace. He then decides that the specification and method of sampling should be as follows: while it will penalize occasionally the producer whose fabric has an average mass of $\mu = 100$ dg and a standard deviation of individual test specimens of $\sigma = 3{,}5$ dg, it will penalize less and less frequently as quality improves above this level. He intends to do this by introducing a test clause into the specification such that if the tests on the sample material fail to pass the standard, the whole roll of fabric from which the sample has been taken is to be rejected.

Suppose that it is under discussion whether the tests on each roll should be made on $n = 4$, 8 or 16 specimens, and that it is proposed that the clause should specify a *minimum mass that the mean of n tested specimens has to exceed*. The problem is how to determine what this minimum average mass should be in each case.

If the production process is under statistical control, we know that the means of samples of n pieces will be closely represented by a normal curve with mean μ and standard deviation $\sigma / \sqrt{n}$.

Table 15 — Fractiles of the normal distribution corresponding to selected confidence levels

Confidence	Chance of error		One-sided interval		Two-sided interval
%	%	α	$u_{1-\alpha}$	$\frac{\alpha}{2}$	$u_{1-\frac{\alpha}{2}}$
90	10	0,10	1,281 6	0,05	1,644 9
95	5	0,05	1,644 9	0,025	1,960 0
98	2	0,02	2,053 7	0,01	2,326 3
99	1	0,01	2,326 3	0,005	2,575 8

From Table 15, which gives details of the fractiles of the standard normal probability curve, it may be expected in the long run that, for example:

10 percent of means (i.e. 1 in 10) will fall below $\mu - 1{,}281\ 6\sigma / \sqrt{n}$;

5 percent of means (i.e. 1 in 20) will fall below $\mu - 1{,}644\ 9\sigma / \sqrt{n}$;

1 percent of means (i.e. 1 in 100) will fall below $\mu - 2{,}326\ 3\sigma / \sqrt{n}$;

0,5 percent of means (i.e. 1 in 200) will fall below $\mu - 2{,}575\ 8\sigma / \sqrt{n}$.

He decides to fix the minimum so that a producer whose standard of quality is represented by $\mu = 100$ dg, $\sigma = 3{,}5$ dg will be liable to have only one sample in 20 rejected. The limits were therefore set as follows.

For samples of size 4, $L_4 = 100 - 1{,}644\ 9 \times 3{,}5 / \sqrt{4} = 97{,}12$.

For samples of size 8, $L_8 = 100 - 1{,}644\ 9 \times 3{,}5 / \sqrt{8} = 97{,}96$.

For samples of size 16, $L_4 = 100 - 1{,}644\ 9 \times 3{,}5 / \sqrt{16} = 98{,}56$.

These limits are shown in the left hand side of Table 16 together with the number of samples, represented in Figure 5 in 4.3.2, which would be *rejected* if these specification limits were imposed.

Suppose that it was decided to demand a higher quality of material having a mean mass per specimen of at least 103 dg, and as before a standard deviation not greater than $\sigma = 3{,}5$ dg. The specification limits, which are now 3 higher than before, are shown on the right-hand side of Table 16, together with the number of samples (from Figure 5) that would now be *accepted*.

Table 16 — Fabric specimens — Testing rule based on sample mean

Size of Sample	Case $\mu = 100$ dg		Case $\mu = 103$ dg	
	Limit in dg	Number of samples in Figure 5 rejected	Limit in dg	Number of samples in Figure 5 accepted
4	97,12	5 out of 32	100,12	17 out of 32
8	97,96	2 out of 16	100,96	3 out of 16
16	98,56	1 out of 8	101,56	0 out of 8

The advantages of adjusting the limits to suit the sample size are now evident. Note that the mean and standard deviation of the 128 test results are 99,91 dg and 3,49 dg respectively. The first case, with $\mu = 100$ dg, illustrates the way in which a producer whose material lies on the borderline will receive broadly similar treatment whatever the sample size. The second case vividly demonstrates how the user may protect himself against receiving material of quality inferior to the standard at which he aims by increasing the number of specimens to be subjected to the test.

In the earlier use of these data in 4.3, it was suggested that the minimum criterion might be applied to the lowest mass in a sample of n test specimens instead of to the mean. If the masses vary according to a normal distribution with a mean μ and a standard deviation σ, then such limits can be determined; the limits are of the following form:

$$L = \mu - k\sigma$$

where the value of k depends upon the number, n, of items in the sample and the chance, α, of rejection or of acceptance that it is decided to adopt. In fact, k is the $(1 - \alpha)^{1/n}$ fractile of the standard normal distribution which can be obtained using a scientific calculator.

Suppose, for example, that it were again decided to fix a minimum limit such that a producer conforming to a standard of quality represented by $\mu = 100$ dg, $\sigma = 3{,}5$ dg will tend to have one sample in 20 rejected. Then the appropriate factors, k, are shown in Table 17, together with the resulting critical limits and the number of rejections among the same series of samples, i.e. those of Figure 5 in 4.3.2. The results of choosing a higher standard of quality, $\mu = 103$ dg, are shown in the right-hand side of the same table.

Table 17 — Fabric specimens — Testing rule based on smallest mass in sample

Size of sample	**Factor k**	Case $\mu = 100$ dg		Case $\mu = 103$ dg	
		Limit in dg	**Number of samples in Figure 5 rejected**	**Limit** in dg	**Number of samples in Figure 5 accepted**
4	2,234	92,18	2 out of 32	95,18	23 out of 32
8	2,490	91,28	1 out of 16	94,28	11 out of 16
16	2,726	90,46	0 out of 8	93,46	6 out of 8

It is instructive to compare Table 16 and Table 17. As in the case of basing the test on the sample mean, the left-hand side of Table 17 shows that, even when basing the test on the *smallest* mass in the sample, a producer whose material lies on the borderline will receive broadly similar treatment whatever the sample size.

The most noticeable difference between the tables is how much better the user safeguards himself against receiving material of inferior quality by using the sample *mean* mass rather than the *smallest* of the masses in the sample.

It would not be justifiable to draw general conclusions from a single practical example. This is particularly true with this example as the test specimens were cut from the same roll, so the variation from specimen to specimen would not have been entirely random. However, the general conclusions that this example suggests are in accordance with what would have been predicted by statistical theory.

The focus of this example has been on protecting the user against receiving material of low mass. Uniformity of mass may, however, be an important characteristic, that is to say it may also be desirable to protect against the acceptance of fabric with a large variation in mass from specimen to specimen. For this purpose, it would be possible to specify some upper limit either to the standard deviation, s, or to the range (i.e. the difference between the heaviest and lightest specimens) in a sample. The question of control of variation is discussed later in connection with control charts (see Clause 10).

8.4 Tests and confidence intervals for means and standard deviations

8.4.1 Confidence intervals for means and standard deviations

A sample provides an estimate of the mean and the standard deviation of a variable x in the population from which the sample is drawn, that is to say $\bar{x}$ and s provide estimates of μ and σ. It has been indicated that if the sample does not contain many items, then these estimates may not be very accurate; the inaccuracy is measured by the standard deviations, expressions for which have been given in 8.1. For some practical purposes, a rather more precise method of expressing the uncertainty of estimation may be desirable. This can be provided by statistical theory, but only on certain assumptions, which are summarized below, and which must not be overlooked in using Table 9. The problem and its solution may be put in the following form.

Given a sample of size n having, for a certain measured variable, a mean $\bar{x}$ and a standard deviation s, to determine

a) the limits μ_1 and μ_2 between which the population mean μ is likely to lie, and

b) the limits σ_1 and σ_2 between which the population standard deviation σ is likely to lie.

The expression “is likely” needs to be defined in terms of probability. For example, the limits may be chosen in such a way that, using limits derived in the same way on repeated occasions, we would be wrong only one time in 50 (i.e. 2 % of the time) in the long run. Such limits may be calculated as follows:

for the population mean, $\mu_1 = \bar{x} - as, \quad \mu_2 = \bar{x} + as$ (24)

for the population standard deviation, $\sigma_1 = b_1 s, \quad \sigma_2 = b_2 s$ (25)

where the factors a, b_1 and b_2 are given in Table 9 for four levels of probability and for sample sizes n from 5 to 30.

For larger values of n, the following approximations to a, b_1 and b_2 are reasonably accurate:

$$a = \frac{u}{\sqrt{n-3}}, \; b_1 = \frac{1}{1+\frac{u}{\sqrt{2n}}} \text{ and } b_2 = \frac{1}{1-\frac{u}{\sqrt{2(n-2)}}}$$

where the values of u are related to the chance of error for two-sided intervals for a standard normal distribution as shown in Table 15.

For example, the value of u for a 10 % chance of error with a two-sided confidence interval is 1,644 9. For a sample of size 30, the values of a, b_1 and b_2 for a 10 % chance of error are calculated from these approximations as 0,317, 0,825 and 1,282, which are reasonably close to the correct values 0,311, 0,825 and 1,280.

These limits on the population mean and population standard deviation are called *confidence limits*, because they are associated with a stated measure of confidence. If, for instance, we have a sample of 10 items, and assert that in the sampled population the mean lies in the range:

$\bar{x} - 0{,}580s$ to $\bar{x} + 0{,}580s$

then it can be seen from Table 9 that we should expect such predictions to be correct about 90 % of the time in the long run, and in error 10 % of the time. About half of the 10 % of erroneous assertions would be because $\bar{x} - 0{,}580s$ exceeds the population mean μ and the other half because $\bar{x} + 0{,}580s$ was less than μ. On the other hand, if we take the wider range:

$\bar{x} - 0{,}893s$ to $\bar{x} + 0{,}893s$

we shall be 98 % confident that we are correct, knowing that there is only a 1 % chance that $\bar{x} - 0{,}893s$ will exceed μ and a 1 % chance that $\bar{x} + 0{,}893s$ will be less than μ. The interpretation will be similar for the case of σ.

One-sided confidence intervals may be appropriate if either an upper or a lower limit may be fixed by definition and observed values are close to such a limit. This can be, for example, the case for the standard deviation, which has to be positive. In other situations, one may be interested in only the high or small values of a parameter, e.g. possible high values of the mean value of a normal distribution.

Two-sided confidence intervals are appropriate when all possible values of a parameter are of interest. Table 9 was designed for two-sided intervals, but can be used for one-sided intervals simply by doubling the chance of error. For example, if an upper confidence limit on σ was required at 99 % confidence when the sample size is 15, the chance of error of 1 % is doubled to 2 % to locate the appropriate value of b_2, which is 1,734. Thus, we would have 99 % confidence that σ is less than 1,734*s*.

The validity of such confidence limits depends upon certain assumptions, *viz.*:

1) *that the variation is under statistical control*. For instance, whilst the sample might be any one of those in case 1 or case 2 of Figure 35 in 6.1.1, we clearly could not expect to derive meaningful limits from any one of the samples in cases 3, 5 or 6;

2) *that the form of variation among the items is represented approximately by the normal curve* (see 5.3.8);

3) *that the sample has been drawn at random from a much larger population*. This condition is generally satisfied if the sample size is no more than 5 % of the population size. If, for example, a sample of 10 items were to be drawn at random from a lot containing only 20 items, then it would be possible to estimate the μ and σ of this lot from the $\bar{x}$ and s of the sample within considerably narrower limits.

The first assumption is of particular importance. It cannot be emphasized too strongly that if the variation is not under statistical control, then it is foolhardy to attempt to predict the characteristics of the population from a sample chosen at random.

In practice, statistical control will rarely be perfect, so it is advisable not to pay too much attention to the precise risks associated with the limits. It is better to regard the constants in Table 9 as part of a useful working tool, whose value will be tested by experience. For the same reason, although the constants are given to three decimal places of accuracy, some common sense is necessary in determining how many figures are worth retaining in the calculated confidence limits.

The following illustration is based on the canned tomatoes data given in Table 11. The unit of measurement throughout is the mass in grams.

The first group of three shifts provide a total of 12 observations, with $\bar{x} = 404{,}16$ and $s = 1{,}681$. If we assume that the process is in statistical control, we may use these values to define limits within which we would feel confident that the mean and standard deviation of production lies. Choosing a 98 % confidence level (i.e. a 2 % chance of error), we find from Table 9 that $a = 0{,}785$, $b_1 = 0{,}667$ and $b_2 = 1{,}899$, giving the following values:

i) for the mean of production, $404{,}16 \pm 0{,}785 \times 1{,}681$, i.e. 402,8 to 405,5;

ii) for the standard deviation of production, $0{,}667 \times 1{,}681$ to $1{,}899 \times 1{,}681$, i.e. 1,12 to 3,19.

The range of uncertainty is clearly very large. Adding further observations to the group can narrow this range. Suppose that we base the confidence limits on the first six shifts, doubling the number of observations to 24. Calculating from the original data in Table 11, it is found that $\bar{x} = 404{,}22$ and $s = 1{,}598$. The constants for 98 % confidence limits are found from Table 9 to be $a = 0{,}511$, $b_1 = 0{,}743$ and $b_2 = 1{,}502$, giving limits as follows:

I) for the mean, $404{,}22 \pm 0{,}511 \times 1{,}598$, i.e. 403,4 to 405,0;

II) for the standard deviation, $0{,}743 \times 1{,}598$ to $1{,}502 \times 1{,}598$, i.e. 1,19 to 2,40.

The extra information has narrowed the limits, as expected. In both cases, the limits include the values $\mu = 403{,}8$ and $\sigma = 1{,}91$ calculated from the first 160 test records.

8.4.2 Tests for means and standard deviations

8.4.2.1 Terminology

In constructing statistical tests, it is important to be precise about the hypotheses under consideration. Some terminology is helpful here. Generally speaking, the hypothesis of “no difference” or equality of population values is called the *null hypothesis*, and is usually denoted by H_0. The hypothesis against which this hypothesis is to be compared is called the *alternative hypothesis*, and is usually denoted by H_1. The test is performed on the value of a *test statistic* that is calculated from the sample data. The region of variation of the test statistic that leads to rejection of the null hypothesis is called the *critical region*. The probability that the test statistic falls in the critical region when the null hypothesis is true (thereby leading to the erroneous decision to reject the null hypothesis in favour of the alternative hypothesis) is called the *size* or *significance level* of the test, usually denoted by α. Finally, the probability that the test statistic falls in the critical region when the alternative hypothesis is true (thereby leading to the correct decision to reject the null hypothesis in favour of the alternative hypothesis) is called the *power* of the test. The power is usually denoted by $1-\beta$. A good test will have high power and a low value for the significance level.

A rejection of the null hypothesis when the null hypothesis is true is called a Type I error, or an error of the first kind. A rejection of the alternative hypothesis when the alternative hypothesis is true is called a Type II error, or error of the second kind. It follows that the probabilities of Type I and Type II errors are α and β respectively.

8.4.2.2 Test of a population mean against a given value

A simple example will illustrate these concepts. Suppose that the standard deviation of a normal population, σ, is known but that the mean, μ, is unknown. We wish to test the null hypothesis:

H_0 : the mean of the normal population is μ_0;

against the alternative hypothesis:

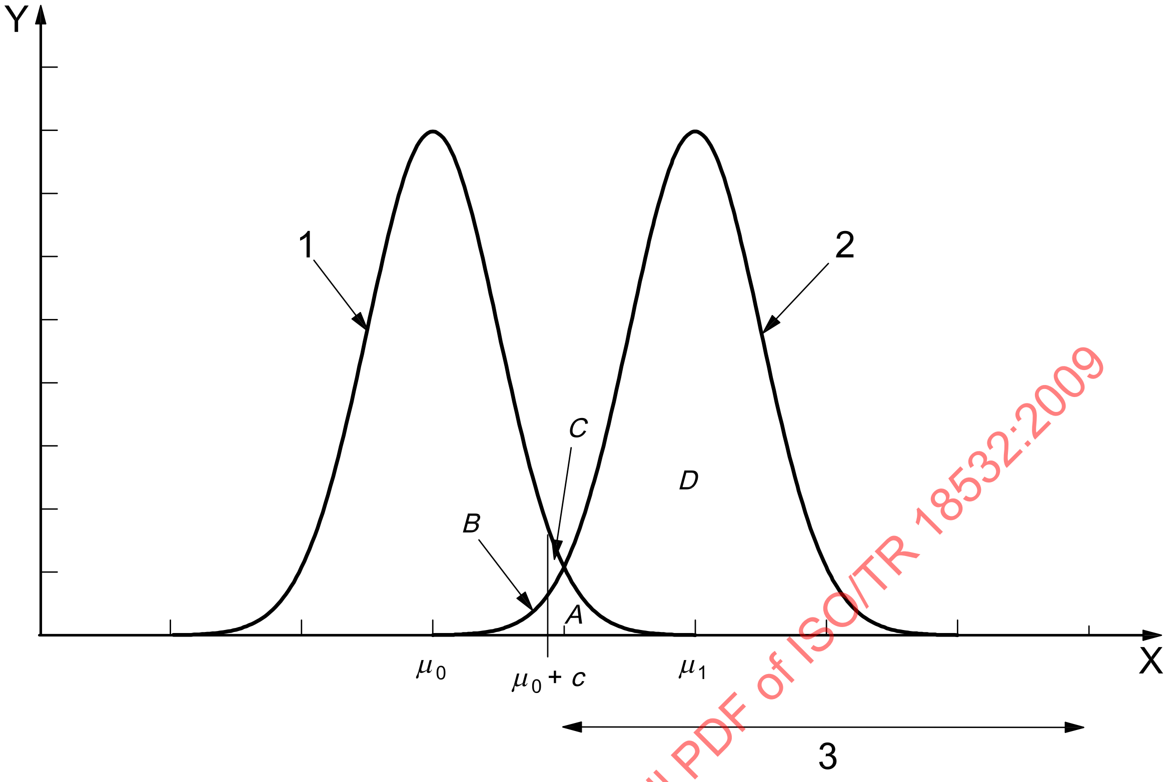
H_1 : the mean of the normal population is greater than μ_0;

The significance level of the test is to be 5 %. A random sample of size n is drawn from the population, and its mean $\overline{x}$ calculated.

It is intuitively obvious for this example that the critical region should lie entirely to the right of $\mu_0 + c$, where c is some constant that is greater than zero. For a 5 % significance level, we require $\mu_0 + c$ to be the upper confidence limit on μ with a 95 % confidence level. The appropriate standard normal fractile for a one-sided confidence interval at confidence level 95 % is found from Table 15 to be 1,644 9. As $\overline{x}$ is normally distributed with a mean μ_0 and a standard deviation $\sigma / \sqrt{n}$ under hypothesis H_0, it follows that $c = 1{,}644\ 9 \times \sigma / \sqrt{n}$.

Figure 37 shows the critical region. The area $A + C$ represents the significance level, in this case 5 % or 0,05. The area $A + D$ $(= 1 - B)$ represents the power of the test when $\mu = \mu_1$. To calculate the power of the test, we first calculate the standardized difference between $\mu_0 + c$ and μ_1, i.e. $z = \left[(\mu_0 + c) - \mu_1\right] / (\sigma / \sqrt{n})$. The power of the test is then the area to the right of z under the standard normal curve, which may be found from Table 7, for example.

A more common situation would be where the population mean and standard deviation are both unknown. In this case, the multiplier $1{,}644\ 9 / \sqrt{n}$ in the above (one-sided) example would be replaced by the value of a given in Table 9 corresponding to the sample size and a $2 \times 5\ \% = 10\ \%$ chance of error. Thus, for sample size 9, the multiplier $1{,}644\ 9 / \sqrt{n} = 0{,}548\ 3$ would be replaced by 0,620. The increase in the value of the multiplier reflects the increased uncertainty due to not knowing the value of σ.



Key

1 distribution of $\bar{x}$ when H_0 is true

2 distribution of $\bar{x}$ when H_1 is true

3 critical region

X $\bar{x}$

Y probability density of $\bar{x}$

Figure 37 — Illustration of one-sided test

8.4.2.3 Test of the difference between two population means; degrees of freedom

A few remarks are in order at this point about the concept of *degrees of freedom*. Degrees of freedom are parameters of a number of important statistical distributions, and therefore form a natural quantity by which to tabulate them. For the straightforward case of a standard deviation in a single sample, it makes very little difference whether the tabulation is in terms of sample size n or degrees of freedom ν, for in that case $\nu = n - 1$. But tabulating the distribution of the appropriate statistic in terms of degrees of freedom can facilitate the use of the tables for other cases, for example:

a) for a single sample where the number, say k, of independent constraints is greater than 1; the table could be used in such a case with $\nu = n - k$;

b) for k samples of sizes n_1, n_2, ..., n_k, with different means but equal standard deviations that are to be combined for the purposes of estimating their common standard deviation; the table could be used in such a case with $\nu = (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) = n - k$ where $n = n_1 + n_2 + \cdots + n_k$.

The appropriate statistic to use in such problems when σ is unknown is the t-statistic, a tabulation of which is provided in Table A.6.

Consider case b) with $k = 2$, the comparison of the means of two populations when neither the population means nor the population standard deviations are known. Suppose the hypotheses are as follows:

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2$$

The sample data are a random sample of size n_1 from the first population and an independent random sample of size n_2 from the second population. The sample means are $\bar{x}_1$ and $\bar{x}_2$, and the sample variances (i.e. squares of the sample standard deviations) are s_1^2 and s_2^2, given by:

$$s_1^2 = \frac{\sum(x_1 - \bar{x}_1)^2}{n_1 - 1} \text{ and } s_2^2 = \frac{\sum(x_2 - \bar{x}_2)^2}{n_2 - 1}$$

Consider the statistic $d = \bar{x}_1 - \bar{x}_2$. Assuming that the sample sizes are small by comparison with their respective population sizes, we know from statistical theory that the population mean of d is as follows:

$$\mu_d = \mu_1 - \mu_2$$

and that the population standard deviation of d is as follows:

$$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where μ_1 and μ_2 are the population means and σ_1 and σ_2 are their standard deviations. The hypotheses can be restated as $H_0 : \mu_d = 0$ and $H_1 : \mu_d \neq 0$. The test is two-tailed, as the critical region of the test (where the truth of H_0 would be in doubt) will clearly consist of the large positive and negative values of d.

We shall consider only the case where the two population dispersions are the same, i.e. $\sigma_1 = \sigma_2 = \sigma$. The formula for σ_d then simplifies slightly to the following:

$$\sigma_d = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

An estimate of σ is s obtained by adding the numerators of the expressions given above for s_1^2 and s_2^2, dividing by the sum of the denominators, and then taking the square root, i.e.:

$$s = \sqrt{\frac{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)}}$$

with degrees of freedom $\nu = (n_1 - 1) + (n_2 - 1)$. Substituting for σ in the previous formula, the following estimate is obtained:

$$s_d = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

of σ_d. The upper fractile $t_{\nu,1-\alpha/2}$ corresponding to a two-tailed test with the required significance level α is read from the t-table (Table A.6). The null hypothesis is rejected in favour of the alternative hypothesis if the confidence interval for μ_d, with limits $d \pm t_{\nu,1-\alpha/2} s_d$ does not include the value zero.

8.4.2.4 Power of the test

As with all statistical tests, the power of the test as defined in 8.4.2.1 should be assessed in advance of collection of the data. Too often, in practice, the power is wholly disregarded in planning a trial and the results turn out to be inconclusive. In this example, the power would have to be assessed rather than determined, as the value of σ is unknown. If the assessed power turned out to be lower than required for detecting a difference of a given magnitude between the population means, consideration should be given to increasing one or both sample sizes. If this is not possible, it may be decided not to waste resources in carrying out the trial or experiment, as the outcome is so unlikely to tell us anything we did not know (or thought we knew) already.

There is always the possibility that the power will be found to be higher than required in which case the size of the proposed trial could be reduced.

In many cases, and this example is no exception, the calculation of the power involves relatively advanced statistics, which may explain why considerations of power are often avoided. However, the speed of modern desk-top computers now enables the power of any proposed test to be assessed by means of simulation using simple statistical concepts.

The following formula combines all the quantities that are important for the power of the two sample t-test:

$$(t_{1-\alpha/2}+t_{1-\beta})^2=\frac{(\mu_1-\mu_2)^2}{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{26}$$

The quantities are the level α of the test, the power $1-\beta$, the unknown variance of the populations, the sample sizes, and, finally, the differences between the means of the populations. If the sample sizes are identical, Equation (26) can be converted into a useful formula that can give the common size of each sample n that is needed to obtain a specified power:

$$n=2(t_{1-\alpha/2}+t_{1-\beta})^2\frac{\sigma^2}{(\mu_1-\mu_2)^2} \tag{27}$$

The formula is used as follows. First, the smallest difference in means $\mu_1-\mu_2$ that it is important to verify is chosen. This quantity is called the *minimum relevant difference* in some disciplines. Then the level α of the test and the desired power $1-\beta$, and finally an informed guess of the unknown variance is made. Guessing the variance is often the difficult part and may require some experimentation. There is a slight complication in applying Equation (27) because the degrees of freedom of the t fractiles on the right-hand side depend on the sample size that we want to determine. But it only means that one may have to make a couple of iterations in the formula, as the following example shows.

Suppose the minimum relevant difference is $\mu_1-\mu_2=25$ and the population standard deviation is $\sigma=20$. Suppose further that the level of the test is $\alpha=0{,}05$ and the power $1-\beta$ is 0,9. We then start by pretending that we have an infinite number of degrees of freedom and find the fractiles in the bottom row of Table A.6. We find $t_{1-\alpha/2}=t_{0{,}975}=1{,}960$ and $t_{1-\beta}=t_{0{,}9}=1{,}2816$. Using Equation (27) gives

$$n=2(t_{1-\alpha/2}+t_{1-\beta})^2\frac{\sigma^2}{(\mu_1-\mu_2)^2}=2(1{,}96+1{,}281\,6)^2\frac{20^2}{(25)^2}=13{,}45 \tag{28}$$

Applying the formula again with degrees of freedom 26 corresponding to a total sample size of 28 gives

$$n=2(t_{1-\alpha/2}+t_{1-\beta})^2\frac{\sigma^2}{(\mu_1-\mu_2)^2}=2(2{,}055\,5+1{,}315)^2\frac{20^2}{(25)^2}=14{,}54 \tag{29}$$

Using the formula again, with a total sample size of 30 and 28 degrees of freedom, changes the sample size very little (n = 14,46), so the conclusion is that a total sample size of 30, with 15 observations in each sample, will give a power of 0,90.

8.4.2.5 Comparison of two means in the case of paired observations

Increasing the sample sizes is not the only way to increase the power. Another is to improve the precision of the comparisons by eliminating or reducing the effects due to differences between the samples of raw material on which measurements or treatments are carried out. For example, suppose we wish to compare the effect of two fertilizers, A and B, on a certain crop. One approach would be to apply fertilizer A to one random sample of n test plots and fertilizer B to a second random sample of n test plots, the $2n$ plots all coming from the same fairly homogeneous field, and then to compare the two yields. But plots will inevitably differ with regard to drainage, levels of nutrients, etc. and by pure bad luck we could select two samples such that most of the plots in one sample were inferior to most of the plots in the other. This could compromise the conclusions from the trial. Even if both samples were similar, the plot to plot variation may be enough to reduce the power of the test to an unacceptable level, with commercially significant differences between fertilizers having too high a chance of not being detected.

A simple way of reducing the effect of plot to plot variation from such a test would be to select adjacent *pairs* of plots, applying fertilizer A to one of each pair chosen at random, and fertilizer B to the other.

Suppose the yield from the ith pair of plots is x_i for fertilizer A and y_i for fertilizer B. Then the difference $d_i = x_i - y_i$ would be affected hardly at all by plot to plot differences, assuming that adjacent plots are nearly identical. This is called the method of paired comparisons (see ISO 3301 [11]). Provided x and y are independent and have approximately normal distributions, the differences d will be approximately normally distributed about the population mean difference μ_d with a population standard deviation σ_d. The sample mean $\bar{d}$ and sample standard deviation s_d provide estimates of these parameters.

The precise nature of the test will depend on the null and alternative hypotheses and the significance level of the test. If the two fertilizers are new and untried, we may simply wish to determine if one is superior. This could be done using a two-tailed test of $H_0 : \mu_d = 0$ against $H_1 : \mu_d \neq 0$; alternatively, and equivalently, a two-sided confidence interval for d could be calculated, to see whether or not it included the value zero. On the other hand, fertilizer B may be the standard against which a more expensive new fertilizer A is being tested; in this case, the new fertilizer may need to improve yield by more than an amount c to justify its extra cost. This could be done by means of a one-tailed test of $H_0 : \mu_d = c$ against $H_1 : \mu_d > c$. Or fertilizer B may be a null treatment, i.e. no treatment at all, in which case we would use a one-tailed test of $H_0 : \mu_d = 0$ against $H_1 : \mu_d > 0$ to determine if fertilizer A is effective. In all of these cases, Table 9 could be used.

The method of paired comparisons can be even more effective when the same item of raw material can be used for both treatments, e.g. in comparing the results of two test methods or two measuring instruments or two laboratories on the same product. The test or measurement process would clearly have to be non-destructive, e.g. measuring the strand widths of samples of tobacco for the purpose of classifying a consignment as pipe or cigarette tobacco, to determine the rate of duty payable. The method of paired comparisons is particularly suitable for comparing the responses before and after a certain type of treatment.

Any information from previous trials of the likely size of σ_d should be utilized to determine the sample size n that will provide sufficient power.

8.4.2.6 Comparisons of standard deviations

We have already seen in 8.4.1 how to set confidence limits for a population standard deviation σ. Testing whether σ is equal to a given value σ_0 can be effected by calculating a confidence interval (σ_1, σ_2) and seeing if σ_0 lies inside the interval. A problem we have not yet addressed is testing for differences between two population standard deviations σ_x and σ_y. The problem is tackled by determining if s_x / s_y differs significantly from unity.

Suppose we have a sample of size n_1 from the first population and one of size n_2 from the second population. The respective sample standard deviations are s_x and s_y. Then a 100$(1-\alpha)$ % two-sided confidence interval on σ_x / σ_y is:

$$\frac{s_x / s_y}{\sqrt{F_{n_1-1,\, n_2-1,\, 1-\alpha/2}}} \text{ to } \frac{s_x / s_y}{\sqrt{F_{n_1-1,\, n_2-1,\, \alpha/2}}}$$

where $F_{n_1-1,\, n_2-1,\, 1-\alpha/2}$ and $F_{n_1-1,\, n_2-1,\, \alpha/2}$ are the upper and lower $(\alpha/2)$-fractiles of the F-distribution, with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

Tables of the F-distribution are three-way and therefore too extensive to reproduce here, so to illustrate the method we will simply quote the appropriate fractiles from published tables. Suppose a sample of size 10 from the first population yields $s_x = 10{,}5$ and a sample of size 16 from the second population yields $s_y = 6{,}8$. Is this evidence sufficient to conclude with 95 % confidence that there is a difference between σ_x and σ_y ?

The significance level of the test is 5 %, or 0,05. From tables of the F-distribution, it is found that $F_{9,\, 15,\, 0{,}975} = 3{,}12$ and $F_{9,\, 15,\, 0{,}025} = 0{,}265$. The ratio $s_x / s_y = 10{,}5/6{,}8 = 1{,}544$. A 95 % confidence interval on σ_x / σ_y is therefore $1{,}544/\sqrt{3{,}12}$ to $1{,}544/\sqrt{0{,}265}$, i.e. 0,874 to 3,00.

Since this range encloses the value 1, we cannot conclude that there is a difference between σ_x and σ_y.

One-sided tests can be treated in a similar way by calculating either of the one-sided confidence intervals:

$$\left(0, \frac{s_x / s_y}{\sqrt{F_{n_1-1,\, n_2-1,\, \alpha}}}\right) \text{ or } \left(\frac{s_x / s_y}{\sqrt{F_{n_1-1,\, n_2-1,\, 1\text{-}\alpha}}}, 0\right)$$

and determining whether or not the value 1 lies in the interval.

ISO 2854 [3] and ISO 3494 [12] deal with estimation and tests for means and variances with power functions for tests.

8.4.3 Equivalence of methods of testing hypotheses

It should be noted that there are three strategies for testing a null hypothesis, all of which are equivalent. All reject H_0 in favour of H_1

- if the value of the test statistic falls in the critical region,
- if the p-value for the test statistic (i.e. the probability that the test statistic takes the observed value or a more extreme value given that H_0 is true) is equal to or smaller than the chosen significance level α, and
- if the appropriate type of confidence interval (one-sided or two-sided depending on whether the alternative hypothesis is one-sided or two-sided) with confidence coefficient $(1 - \alpha)$ does not include the value of the parameter specified under H_0.

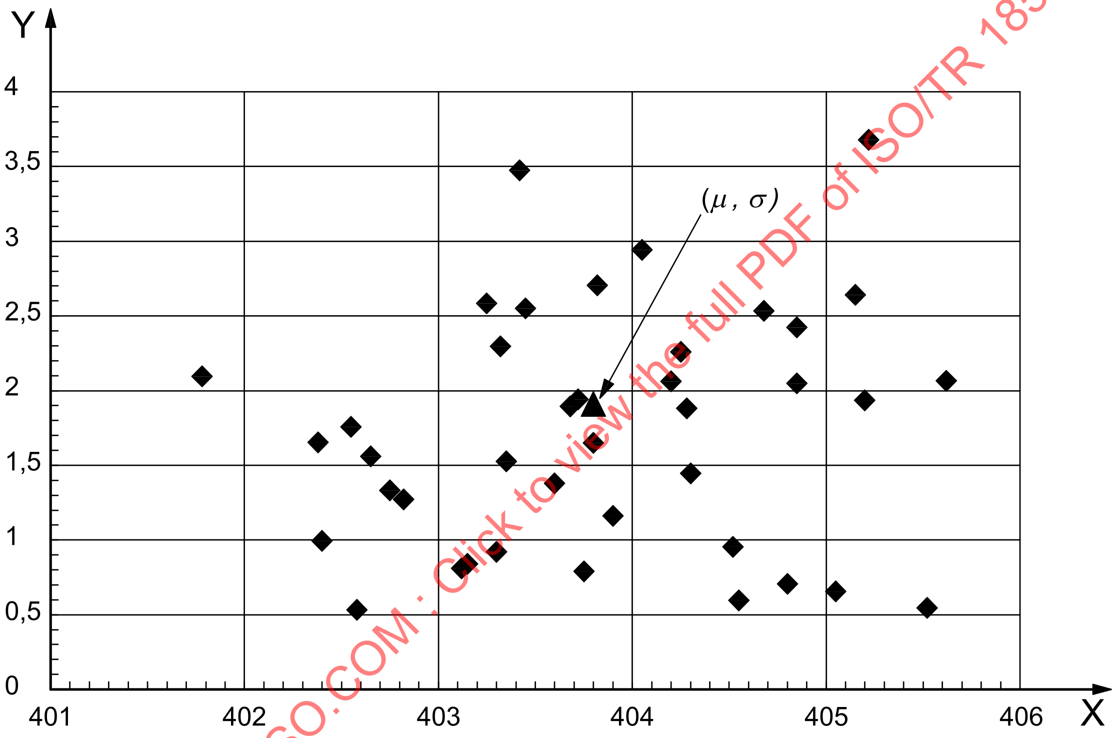
The appropriate strategy to choose from among these depends on the actual situation.

8.5 Simultaneous variation in the sample mean and in the sample standard deviation

Until now, we have considered the variation in sample means and standard deviations separately, but for problems of the type to be discussed later, the two need to be treated together. The nature of their relationship can be illustrated most clearly by plotting a point $(\overline{x}, s)$ to represent each sample on a diagram having $\overline{x}$ and s as co-ordinate axes. Such a diagram for the canned tomatoes data is shown in Figure 38, where each of the 40 dots represents a set of four test results. In the centre of the field, shown by a triangle, lies the point (μ, σ) representing the whole aggregate of items sampled; the values used are for all 160 original observations, namely $\mu = 403{,}8$ g and $\sigma = 1{,}91$ g.

Suppose now that an increasingly large number of shifts were sampled, the can contents weighed, the mean and standard deviation for sets of 4 results calculated and the points $(\bar{x}, s)$ plotted on the diagram. An increasing swarm of points would surround the central spot (μ, σ). Assuming production to be under statistical control, theory would allow the prediction, at least approximately, of what may be called the density of this swarm of points at different distances and in different directions from (μ, σ). In other words, we could express the chance that a sample point $(\bar{x}, s)$ would fall in any prescribed region of the diagram.

An illustration of where most of the sample points $(\bar{x}, s)$ would be expected to fall is given by the standardized $(\bar{x}, s)$ control charts of Kanagawa, Arizono and Ohta [103]. These charts, which are based on information theory, are useful for determining the type of departure from control by considering $\bar{x}$ and s simultaneously. A standardized $(\bar{x}, s)$ control chart for sample size 4, with limits for which there is only a 27 in 10 000 risk per observation of a false out-of-control signal, is shown in Figure 39 for the canned tomatoes data; the 40 sample points $[(\bar{x}-\mu_0)/\sigma_0, s/\sigma_0]$ have been plotted on the chart for target values of $\mu_0 = 404{,}0$ g and $\sigma_0 = 1{,}90$ g. (The reason for the strange choice of probability, 27 in 10 000, or 2 in 741, will become evident in 10.5 and 10.6.)

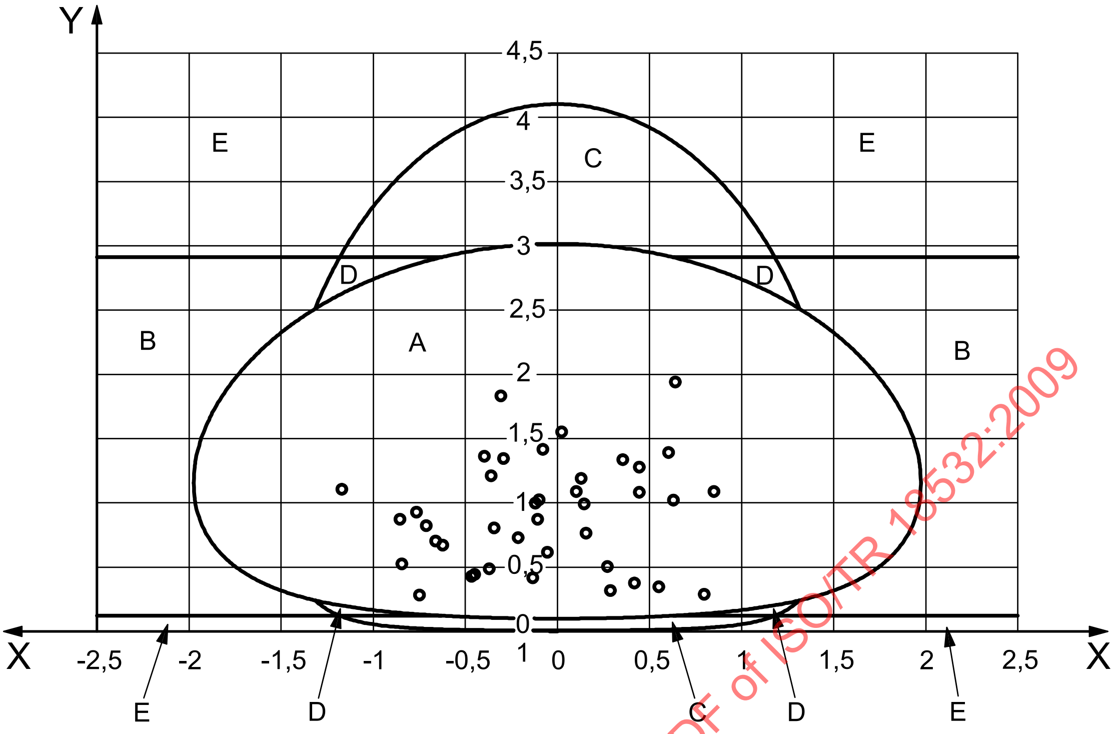


Key

X sample mean, $\bar{x}$

Y standard deviation, s

Figure 38 — Scatter chart for sample means and standard deviations in canned tomatoes data



Key

X $(\bar{x} - \mu)/\sigma_0$

Y s/σ_0

Figure 39 — Standardized control chart for mean and standard deviation

The regions on the chart are identified as follows.

A. The process is in control.

B. The process is out of control because of a change in the process mean.

C. The process is out of control because of a change in the process standard deviation.

D. The process is out of control because of a slight change to both the process mean and the process standard deviation.

E. The process is out of control in both the process mean and the process standard deviation.

All of the 40 plotted points lie in region A, indicating that the tomato canning process is in control with respect to net mass.

The chart is standardized so that the same chart can be used regardless of the values of μ_0 and σ_0. The chart would only need to be changed if the sample size or false signal rate were changed. Increases in either of these quantities shrink the boundary lines and curves closer to the point with co-ordinates $(\bar{x} - \mu_0)/\sigma_0 = 0$, $s/\sigma_0 = \sqrt{n/(n-1)}$.

What is particularly interesting about this type of chart is the shape of the region A in which most of the standardized sample points $[(\bar{x} - \mu_0)/\sigma_0,\ s/\sigma_0]$ are expected to lie when a process is under control. Note that the closer s is to the target value σ_0, the more latitude is allowed in $\bar{x}$; similarly, the closer $\bar{x}$ is to the target value μ_0, the more latitude is allowed in s. In other words, there is, in a sense, a natural trade-off between estimated departures of μ from μ_0 and estimated departures of σ from σ_0. Traditional control charts, by contrast, treat $\bar{x}$ and s separately (see Clause 10). Superficially, the latter approach may seem logical from the point of view that $\bar{x}$ and s are known from statistical theory to be independent in samples from a normal distribution. But it allows no trade-off between the estimated departures from the target values μ_0 and σ_0 and is equivalent to having a rectangular in-control region on an $(\bar{x}, s)$ chart.

The *joint* consideration of $\bar{x}$ and s will be a recurring theme later in this clause and also in discussing methods of determining conformity to specification in Clause 9.

8.6 Tests and confidence intervals for proportions

8.6.1 Attributes

For many quality characteristics, it is either impossible or impracticable to obtain a measure of the characteristic on a continuous scale. For example, consider office cleaning services. A random sample of rooms could be inspected after cleaning to check that waste bins had been emptied, filing cabinets and desks had been dusted, and carpets had been vacuum-cleaned. For each of these three characteristics an experienced inspector would have little difficulty determining if the operation had been carried out to a satisfactory standard, and deciding that a room had been satisfactorily cleaned if it passed on all three checks. He would have rather more difficulty in grading the extent to which these tasks had been done on a meaningful continuous scale from, say, zero to one, and combining the grades in a coherent way to come to a decision on whether the cleanliness of the room was satisfactory.

Characteristics such as these, the realizations of which can most naturally be considered to fall into one of two states (pass/fail, go/no-go, ignites/fails to ignite), are called “attributes”.

For critical characteristics, e.g. those that may affect the safety of personnel, every effort should have been made to ensure that the proportion of nonconforming items in the population is as near zero as possible. 100 % inspection would be used where practicable, and the critical items removed, in which case the proportion of critical items remaining in the population would be known to be zero. This assumes, of course, that the inspection is 100 % effective.

For non-critical characteristics, there may be a need to estimate the proportion of nonconforming items in the population, to calculate confidence limits on the proportion in the population, to test the proportion against a given value, or to compare two or more proportions.

8.6.2 Estimating a proportion

To continue the office cleaning example, suppose that a contractor is responsible for cleaning N rooms, i.e. the size of the population is N. On a particular day, suppose that R of the rooms would fail inspection, i.e. they have not been cleaned to a satisfactory standard. The proportion of the population that would fail inspection is therefore:

$$P = R / N$$

R is unknown, so n rooms are chosen at random and inspected, with r failing inspection. The question is how best to estimate P.

A close analogy between the treatment of attributes and variables is possible here. Suppose the state of a room is characterized by a variable X taking the value 0 if a room is satisfactorily cleaned and 1 otherwise.

The population of values of X then consists of R ones and $(N-R)$ zeros, while the sample consists of r ones and $(n-r)$ zeros. Denote the sample values of X by x. Then the sum of the sample values of X is equal to r, i.e.:

$$\sum x = r$$

The sample mean, $\bar{x}$, is therefore given by the following:

$$\bar{x} = \sum x / n = r / n = p \qquad (30)$$

The sum of X in the population is equal to R, i.e.:

$$\sum X = R$$

The population mean, $\bar{X}$, is therefore given by the following:

$$\bar{X} = \sum X / N = R / N = P \tag{31}$$

It was stated in 8.1.1 that the sample mean is an unbiased estimator of the population mean. Here, we use $\bar{x}$ from Equation (30) as an unbiased estimator of $\bar{X}$ from Equation (31), which translates into using the sample proportion p as an unbiased estimator of the population proportion P. The lack of bias holds good even when the sample size is a large proportion of the population size. For our example, if a sample of 50 rooms reveals that 2 were inadequately cleaned, it would be estimated that 2 out of 50, i.e. 4 % of the rooms in the population of rooms under consideration were unsatisfactory.

8.6.3 Confidence intervals for a proportion

Given that a random sample of size n contains r nonconforming items, it may be required to provide an interval, say P_1 to P_2, within which we may have a given confidence that the true proportion, P, of nonconforming items from the production process lies. Suppose that the confidence is denoted by $100(1-\alpha)$ %, with the chance of error of 100α % being equally divided between the case $P < P_1$ and the case $P > P_2$. These limits can be interpreted as follows.

i) If P were as low as P_1, there would be a probability of only $\alpha/2$ of finding r or more unsatisfactory items in a sample of size n.

ii) If P were as high as P_2, there would be a probability of only $\alpha/2$ of finding r or fewer unsatisfactory items in a sample of size n.

The probabilities in i) and ii) are calculated from a distribution called the *binomial* distribution. For small sample sizes, the values of P_1 and P_2 satisfying i) and ii) may be found in published tables. For larger sample sizes, approximate values may be read from published charts.

It is instructive to see how far the analogy between the treatment of attributes and variables can be extended to provide approximate confidence limits. To emulate the procedure in 8.4, we require an estimate of the standard deviation of p, the estimated proportion. As the values of x are all zero or one and the square of zero is zero and the square of 1 is 1, we have the following situation:

$$\sum x^2 = \sum x = r$$

The sample standard deviation of x (see 5.2) is therefore given by the following expression:

$$s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{r - r^2/n}{n-1}} = \sqrt{\frac{r(1-r/n)}{n-1}}$$

so an estimator of the standard deviation of $\bar{x}$ (or p) is as follows:

$$s/\sqrt{n} = \sqrt{\frac{\frac{r}{n}\left(1-\frac{r}{n}\right)}{n-1}} = \sqrt{\frac{p(1-p)}{n-1}}$$

This is the point from which the analogy becomes rather stretched. Confidence limits on P can be obtained by assuming the distribution of p is approximately normal. A two-sided confidence interval for P would then be of the form (P_1, P_2) where:

$$P_1 = p - u\sqrt{\frac{p(1-p)}{n-1}} \text{ and } P_2 = p + u\sqrt{\frac{p(1-p)}{n-1}}$$

where u is the upper $(\alpha/2)$-fractile of the standard normal distribution. If a one-sided confidence interval were required, then only P_1 or P_2 would be required; the chance of error would then only apply at one end of the interval, so the appropriate value of u would be the upper α-fractile of the standard normal distribution (see Table 15). Unfortunately, the normal approximation to the distribution of p is poor unless either P is close to one half or the sample size is quite large. In cases where this is not true, the use of this approximation should be confined to cases where only rough approximations to P_1 and P_2 will suffice.

Much effort has been devoted in the past to obtaining accurate approximations for P_1 and P_2, generally involving methods of improving the closeness of the normal approximation. See, for example, Molenaar [113] and Blyth [76].

8.6.4 Comparison of a proportion with a given value

Another common problem is how to determine whether a sample proportion differs from a given population value by more than can be attributable to chance. For example, would three substandard items in a sample of size 30 be sufficient to provide 95 % confidence that the percentage of substandard items in the population under consideration exceeded 3 %?

There are two ways of answering this question. The first is to determine from the sample results the lower one-sided 95 % confidence limit on the percentage in the population, answering "yes" if 3 % was below this value and "no" otherwise. The second, a kind of inversion of the first, is to determine the probability of finding three or more substandard items in a sample of 30 when the percentage in the population is 3 %, answering "yes" if this probability is below 0,05 and "no" otherwise. The methods are essentially equivalent, but the latter has the advantage that it provides an actual measure of the confidence, rather than the result that it exceeded or did not exceed 95 %.

To answer the question in this specific case, the first method produces a lower confidence limit at 95 % confidence of 2,7 % substandard items in the population. (This can be found from published tables, e.g. ISO 11453 [41].) As 3 % lies within the confidence interval, we would conclude that the sample result is compatible with a population percentage of 3 %. Alternatively, the probability of a random sample of 30 items containing three or more substandard items when the percentage in the population is 3 % can be shown to be 0,06. As this is greater than 0,05, the sample result does not provide sufficient evidence to conclude with 95 % confidence that the percentage of substandard items in the population exceeds 3 %. In fact, we would only have confidence $100(1-0{,}06) = 94$ % that such a conclusion was correct.

It may seem somewhat surprising that a sample result of 3 in 30, i.e. 10 %, is insufficient to provide very high confidence that a population percentage exceeds 3 %. This goes to show how important it is to take into account the sample size when assessing a sample result.

8.6.5 Comparison of two proportions

Another related group of questions that can be answered by the use of appropriate statistical methods concerns whether the difference between two sample proportions is more than can be attributable to chance. Suppose that a random sample of size n_1 is taken from one population and a random sample of size n_2 from another population. (Usually n_1 and n_2 will be chosen to be equal.) Suppose further that the numbers of items with a given characteristic in the samples are determined to be r_1 and r_2. The two sample proportions are therefore $p_1 = r_1/n_1$ and $p_2 = r_2/n_2$. If P_1 and P_2 denote the unknown population proportions, the various questions that may be asked on the basis of the sample evidence are:

a) what confidence may we have that P_1 is different from P_2?

b) what confidence may we have that P_1 exceeds P_2? or

c) what confidence may we have that P_1 is less than P_2?

If p_1 is less than p_2 in case b), or p_1 is greater than p_2 in case c), then we could answer “not very much” without having to carry out any statistical calculations at all. The same would be true if p_1 and p_2 are approximately equal. In all other cases the answer may be determined by the use of tables of a distribution called the *hypergeometric* distribution. Special tables have been developed for directly determining the significance of any differences between p_1 and p_2 when n_1 and n_2 are small. For larger values of n_1 and n_2, a number of approximate methods have been devised.

8.6.6 Sample size determination

For tests on proportions, as with tests on means and variances, it is important to keep in mind the probability of detecting a difference of a size that would be considered important in practice, i.e. the power of the test. There may be technical or economic reasons why the sample or samples have to be limited in size, in which case it is useful to know to what extent this limits the power. If not, joint consideration of the required power and significance level of a test will enable an appropriate sample size (or sizes) to be determined, either from published tables or from approximate formulae. The formulae can look somewhat daunting at first, but most are straightforward to use, albeit requiring a little care.

For example, suppose we wish to test the hypothesis that two population proportions, P_1 and P_2, are equal against the hypothesis that P_1 is greater than P_2, assuming that the sample size is to be the same from both populations. The common sample size is required that will provide a confidence $100(1-\alpha)$ % of accepting the equality hypothesis when it is true, and provide a power $100(1-\beta)$ % of concluding that there is a difference when P_1 and P_2 take certain different values (with P_1 greater than P_2). Walters [135] has shown that the approximate sample size can be found as the solution in n to Equation (32):

$$n = \frac{1}{2}\left(\frac{u_{1-\alpha} + u_{1-\beta}}{\sin^{-1}\sqrt{P_1 - 1/(2n)} - \sin^{-1}\sqrt{P_2 + 1/(2n)}}\right)^2 \qquad (32)$$

where $u_{1-\alpha}$ and $u_{1-\beta}$ are respectively the upper α and β fractiles of the standard normal distribution.

Consider the case of a significance level of 5 % and a power of 90 % to detect a difference if $P_1 = 0{,}8$ and $P_2 = 0{,}6$. Setting $\alpha = 0{,}05$ and $\beta = 0{,}10$ and inserting the values of u from the left-hand side of Table 15, then Equation (32) becomes:

$$\begin{aligned} n &= \frac{1}{2}\left(\frac{1{,}644\ 9 + 1{,}281\ 6}{\sin^{-1}\sqrt{0{,}8 - 1/(2n)} - \sin^{-1}\sqrt{0{,}6 + 1/(2n)}}\right)^2 \\ &= \frac{4{,}282\ 2}{\left(\sin^{-1}\sqrt{0{,}8 - 1/(2n)} - \sin^{-1}\sqrt{0{,}6 + 1/(2n)}\right)^2} \end{aligned} \qquad (33)$$

Equation (33) can be solved iteratively by first guessing a value of n, then evaluating the right-hand side to give a new value of n, and repeating until the values of n converge. Suppose that we start with an initial guess of $n = 50$. It can be verified that successive iterations of Equation (33) give $n = 109$, 96, 98, 98. Further iterations are pointless, as they will evidently all produce $n = 98$. Thus, 98 is the appropriate size of random sample from each population.

Not only does this approximate method provide a solution in a matter of only a few iterations, but it is also very accurate. Equation (32) can also be used for two-sided alternative hypotheses by replacing α by $\alpha / 2$.

ISO 11453 [41] provides methods of estimation, testing, setting of confidence limits and sample size determination for problems relating to proportions.

8.7 Prediction intervals

8.7.1 One-sided prediction interval for the next m observations

We may sometimes wish to determine the value of an upper limit, T_U, based on the results of a random sample of size n from a normal population, in such a way that we would have a given level of confidence that none of the next m random observations from the same normal population will exceed T_U. In general, this upper limit is given by the formula:

$$T_U = \overline{x} + qs$$

where $\overline{x}$ and s are respectively the sample mean and the sample standard deviation and q is a factor that depends on the sample size n, on the number m of future observations and on the level of confidence required.

Table 18 shows the values of this factor for a range of values of n and m for a confidence level of 95 %.

Note the way in which the factor inflates as n decreases (due to having less information on which to base the prediction) or as m increases (due to being more ambitious in what the interval is to include).

Table 18 — Factors, q, for calculating one-sided prediction intervals — Confidence level 95 %

Sample size n	Number of future observations, m							
	5	10	20	50	100	200	500	1 000
5	3,787 9	4,417 8	5,028 8	5,793 8	6,337 5	6,852 5	7,493 4	7,951 5
10	2,886 8	3,284 1	3,669 9	4,159 3	4,512 5	4,851 0	5,276 9	5,583 8
20	2,574 4	2,890 7	3,194 0	3,577 7	3,855 7	4,123 7	4,463 2	4,709 7
50	2,415 3	2,689 8	2,948 8	3,272 0	3,504 4	3,727 8	4,011 2	4,217 5
100	2,366 1	2,627 7	2,872 7	3,175 9	3,392 5	3,599 8	3,861 6	4,051 7
200	2,342 2	2,597 6	2,835 7	3,129 0	3,337 6	3,536 5	3,786 9	3,968 0

From the symmetry of the normal distribution, it will be evident that a value of q that provides a given confidence that none of the next m observations *exceed* the upper limit T_U provides the same confidence that none of the next m observations are *less than* a lower limit, T_L, given by:

$$T_L = \overline{x} - qs$$

To illustrate the use of one-sided prediction intervals, suppose that a retailer has complained to its supplier that several size 12 ladies' jumpers of a particular style had bust sizes above the nominal maximum of 92 ½ cm. The supplier has 1 100 jumpers remaining out of a batch of this size and style, all of which were made under the same conditions, and decides to check the bust sizes of a random sample of 100 of them.

None of the 100 measurements was found to exceed 92½ cm. Past supplier data suggests that the bust sizes tend to be approximately normally distributed, and a normal plot of the 100 measurements gives no grounds to doubt the assumption of normality. The sample mean and standard deviation turn out to be $\overline{x} = 90{,}1$ cm and $s = 0{,}4$ cm respectively. The factor for a one-sided prediction interval with $n = 100$ and $m = 1\ 000$ is seen from Table 18 to be 4,051 7. The supplier can therefore be roughly 95 % confident that none of the remaining 1 000 garments have bust measurements in excess of $90{,}1 + 4{,}051\ 7 \times 0{,}4 = 91{,}7$ cm.

As 91,7 cm is well below the nominal maximum of 92 ½ cm, the supplier continues to supply the retail trade with jumpers from this batch.

8.7.2 Two-sided prediction interval for the next m observations

Alternatively, it may be required to determine both a lower limit T_L and an upper limit T_U from our initial sample of size n, such that we have a given confidence that none of the next m observations will lie outside the interval (T_L, T_U). These limits are given by the following:

$$T_L = \overline{x} - rs \text{ and } T_U = \overline{x} + rs$$

where r, like q, is a factor depending on n, m and the required level of confidence. Table 19 shows the values of r for two-sided prediction intervals with confidence 95 % for some values of n and m.

Table 19 — Factors, r, for calculating two-sided prediction intervals — Confidence level 95 %

Sample size	Number of future observations, m							
n	5	10	20	50	100	200	500	1 000
5	4,577 3	5,228 6	5,851 7	6,624 0	7,169 8	7,685 4	8,326 1	8,783 7
10	3,321 0	3,717 3	4,102 5	4,590 5	4,942 4	5,279 3	5,702 9	6,008 2
20	2,902 1	3,207 6	3,502 9	3,878 4	4,151 4	4,415 1	4,749 8	4,992 9
50	2,693 4	2,952 3	3,198 9	3,509 0	3,733 5	3,950 2	4,226 0	4,427 3
100	2,629 8	2,874 3	3,105 5	3,394 1	3,601 6	3,801 2	4,054 3	4,238 6
200	2,599 0	2,836 6	3,060 3	3,338 3	3,537 2	3,727 9	3,968 9	4,144 0

As an example of two-sided prediction intervals, suppose that it is required to verify on a sample basis that a batch of 250 pairs of size L men's trousers have waistbands all in the range 86 cm to 92 cm. A random sample of 50 pairs yields $\overline{x} = 88{,}8$ cm and $s = 0{,}78$ cm, with none of the individual measurements outside the specified range. The appropriate factor from Table 19 with $n = 50$ and $m = 200$ is 3,950 2. Assuming a normal distribution of waistband measurements, a two-sided prediction interval is found to be 88,8 ± 3,950 2 × 0,78, i.e. 85,7 cm to 91,9 cm. Since the lower limit of this prediction interval violates the lower specification limit of 86 cm, the supplier decides it is in his best interests to check the other 200 pairs individually before shipping them.

For both one-sided and two-sided prediction intervals, it is also possible to provide factors that assure with a given confidence that no more than 1, or no more than 2, etc. of the next m observations will fall outside the limits. However, the case of zero out of m is generally of the most interest. It is also possible to provide factors for the case where the process standard deviation s is known, or at least presumed to be so, which will tend to lead to smaller prediction intervals. See ISO 16269-8 [49] or Hahn [93], [94], Hahn and Nelson [96] and Hahn and Meeker [95] for further details.

8.7.3 One and two-sided prediction intervals for the mean of the next m observations

It may on the other hand be required to provide a prediction interval for the *mean* of the next m observations. Prediction intervals for the mean can be determined more readily from standard tables, being based upon the t-distribution. (Percentage points of the t-distribution are provided in Table A.6.) A one-sided upper limit for the mean at confidence $100(1-\alpha)$ % is given by:

$$T_U = \overline{x} + t_{n-1,1-\alpha} \times s\sqrt{\frac{1}{n} + \frac{1}{m}}$$

where $t_{n-1,1-\alpha}$ is the upper α-fractile of the t-distribution with $n-1$ degrees of freedom.

A corresponding one-sided lower limit for the mean of the next m observations is given by:

$$T_L = \bar{x} - t_{n-1,1-\alpha} \times s\sqrt{\frac{1}{n} + \frac{1}{m}}$$

Note that $t_{n-1,1-\alpha}$ only depends on two quantities, namely the sample size and the required confidence level, as a result of which tables of t are rather more compact than the three-way tables needed for q and r.

Two-sided prediction intervals on the mean are given by $(T_L,\ T_U)$ where:

$$T_L = \bar{x} - t_{n-1,1-\alpha/2} \times s\sqrt{\frac{1}{n} + \frac{1}{m}}$$

and

$$T_U = \bar{x} + t_{n-1,1-\alpha/2} \times s\sqrt{\frac{1}{n} + \frac{1}{m}}$$

It should always be borne in mind that departures from normality could cause considerable errors in the prediction intervals, particularly when these intervals extend far outside the range of the sample values.

8.8 Statistical tolerance intervals

8.8.1 Statistical tolerance intervals for normal populations

We have seen in 8.7 that a prediction interval is an interval, derived from a sample, within which a specified finite *number* of future observations may be asserted to lie with a given confidence. A statistical tolerance interval is also derived from a sample, but is an interval within which a specified *proportion* of the population values may be asserted to lie with a given confidence.

The name of these intervals is unfortunate, as it can be misconstrued to mean the interval between the tolerance limits specified by the user. In fact, the limits of a statistical tolerance interval, like those of a prediction interval, will vary from sample to sample. As a statistical tolerance interval is asserted to include, or *cover*, a proportion of the population, an alternative name that is sometimes used is statistical *coverage* interval.

For populations that are normally distributed, the intervals are constructed in much the same way as prediction intervals, but with different values for the factors by which the standard deviation is multiplied.

Published tables address four cases:

a) one-sided limits when the process standard deviation is known, of the form $\bar{x} - b_1\sigma$ or $\bar{x} + b_1\sigma$;

b) two-sided intervals when the process standard deviation is known, of the form $(\bar{x} - b_2\sigma,\ \bar{x} + b_2\sigma)$;

c) one-sided limits when the process standard deviation is unknown, of the form $\bar{x} - b_3 s$ or $\bar{x} + b_3 s$;

d) two-sided intervals when the process standard deviation is unknown, of the form $(\bar{x} - b_4 s,\ \bar{x} + b_4 s)$

where the constants b_1, b_2, b_3 and b_4 depend on the sample size, the coverage and the required level of confidence.

A simple example will illustrate this type of interval. A customer who has received a batch of 12 000 bobbins of cotton yarn decides to check on its breaking load distribution. He takes a random sample of 24 bobbins, and cuts from each a test piece of length 50 cm at about 5 m distance from the free end. The central part of each test piece is tested for breaking load. The unit of measurement is the centinewton. The sample mean and standard deviation turn out to be $\bar{x} = 249{,}8$ and $s = 31{,}4$. From previous experience, it is known that the distribution of breaking loads closely approximates a normal distribution. From tables, it is found that the constant for a one-sided statistical tolerance interval for sample size 24 for coverage 95 % and confidence level 95 % is $b_3 = 2{,}310$. The lower statistical tolerance limit is therefore $249{,}8 - 2{,}310 \times 31{,}4 = 177{,}3$. The customer can therefore be 95 % confident that at least 95 % of the breaking loads are in excess of 177,3 centinewtons.

Suppose that the customer felt confident enough from previous batches from the same supplier to assume that it was only the mean that varied from batch to batch, and that the cotton yarn coming from the production process had a breaking load with a constant standard deviation. Then the appropriate constant would be $b_1 = 1{,}981$. Note that this is considerably smaller than the corresponding value 2,310 of b_3 for the case of unknown process variability. This is because the extra information leads to a smaller safety margin being required. Suppose that σ is known to be 33,2. This produces a statistical tolerance limit of $249{,}8 - 1{,}981 \times 33{,}2 = 184{,}0$. The customer could now be 95 % confident that at least 95 % of the breaking loads are in excess of 184,0 centinewtons.

If there is any doubt about the constancy of σ, it should be assumed to be unknown and case c) or d) used as appropriate.

8.8.2 Statistical tolerance intervals for populations of an unknown distributional type

Even if the form of the distribution of values of the characteristic in the population is in doubt, it is still possible to construct one- and two-sided statistical tolerance intervals. Instead of being based on statistics such as $\bar{x}$ and s, they are based on what are known as the *order statistics*, that is to say individual sample values after they have been sorted and numbered in ascending order. Any single or pair of order statistics can be used to provide a statistical tolerance interval but, of course, the largest and/or the smallest provide the greatest coverage. The penalty in not knowing the distributional form is that the statistical tolerance intervals will be rather wider than they would otherwise have been, or require larger sample sizes. To give some idea of the numbers involved, a sample of size 93 is required in order to have 95 % confidence that the interval formed by the largest and smallest observations covers 95 % of the population values. This rises to a sample of size 473 when the coverage increases to 99 %, and to a sample of size 4 742 for coverage of 99,9 %.

8.8.3 Tables for statistical tolerance intervals

Tables of factors for statistical tolerance limits for the normal distribution may be found in ISO 16269-6 [47], Odeh and Owen [117] and Hahn and Meeker [95]. ISO 16269-6 also provides tables of minimum sample sizes required for a selection of coverages and confidence levels, for both the one- and two-sided cases, when the population distribution is of unknown form.

8.9 Estimation and confidence intervals for the Weibull distribution

8.9.1 The Weibull distribution

8.9.1.1 Parameter estimation

The best estimate of the parameters of the two-parameter Weibull distribution is obtained using the method of maximum likelihood. There is no closed form expression for these estimates which require an iterative procedure. It is recommended to use a statistical package.

Many numerical procedures have been proposed for the estimation of the parameters of the two-parameter Weibull distribution. Some involve iterative solution, others involve the use of special tables. A rough graphical approach is as follows.

a) Plot the data on Weibull probability paper and, if appropriate, fit a straight line to the points by eye.

b) Read off from this line the value of t at which the cumulative probability is 1,0 %. Denote this by $t_{0{,}010}$.

c) Read off the value of t at which the cumulative probability is 63,2 %. Denote this by $t_{0,632}$.

d) Estimate α as $t_{0,632}$.

e) Estimate β as $4,6 / \ln(t_{0,632} / t_{0,010})$.

8.9.1.2 Confidence intervals

Detailed discussion of point estimation and confidence interval determination for quantities related to the Weibull distribution is beyond the scope of this Technical Report. We content ourselves with merely listing the most important ones. A variety of methods and computer packages for the calculation of point estimates and confidence intervals exist for:

a) the parameters α and β;

b) the mean time to failure;

c) fractiles of the time to failure;

d) the reliability at time t.

IEC 61649 [1] and EN 12603 [64] provide procedures for the calculation of point estimates and confidence intervals for the Weibull distribution. IEC 61649 also contains a valuable annex explaining the reasons for the particular choice of methods. Two well-known books on the subject are Mann *et al* [111] and Lawless [106]. See 5.3.9 for further discussion of the Weibull distribution.

8.10 Distribution-free methods: estimation and confidence intervals for a median

So far, we have considered inference from a sample about characteristics of a population when the population distribution is known to belong to a particular family of distributions, e.g. the normal or Weibull families. However, if the form of the population distribution is unknown, statistical methods can still be brought to bear in drawing inferences about a population distribution. Such methods, because they do not depend on the form of population distribution, are called *distribution-free* methods. The advantage of distribution-free methods is that they have greater integrity when there is any doubt at all about the form of the population distribution. The disadvantage is that confidence intervals for probabilities and fractiles are wider than would be the case using methods specially tailored to the specific family of distributions.

An example of the use of a distribution-free method is the determination of confidence limits for the population median (i.e. the value of the characteristic under consideration that divides the total frequency into two halves) when the distributional form is unknown. The median of a population may be of more interest than the mean when the distribution is highly skewed, which can cause the mean to be unduly affected by a small number of extreme values as is the case, for example, for income distributions. To obtain a distribution-free confidence interval, the values of the characteristic in a random sample of size n are first ranked in ascending order of magnitude to give the order statistics $x_{[1]}, x_{[2]}, ..., x_{[n]}$. A symmetrically positioned pair of order statistics $(x_{[k]}, x_{[n+1-k]})$ is then used as the pair of confidence limits.

The smaller the value of k, the larger the confidence that the population median will be included in the interval. For example, consider the case with $k = 1$, providing confidence limits $x_{[1]}$ and $x_{[n]}$. These limits will only fail to include the population median if all n sample values lie above the median or all lie below. As the chance of each original observation lying below the population median is one half and the chance of lying above it is one half, the chance of the population median not being included in the interval is $(\tfrac{1}{2})^n + (\tfrac{1}{2})^n = (\tfrac{1}{2})^{n-1}$.

Our confidence on any one occasion that the population median is included between $x_{[1]}$ and $x_{[n]}$ is therefore $1-(\frac{1}{2})^{n-1}$, which as n takes the values 2, 3, 4, 5, ..., etc. gives confidence levels (in percentage terms) of 50 %, 75 %, 87,5 %, 93,75 %, ..., etc. One-sided distribution-free confidence intervals can be constructed, of the form $(a, x_{[n]})$ or $(x_{[1]}, b)$ where a and b are the smallest and largest possible values of the characteristic in the population. These confidence levels represent the *largest* confidence levels for the given sample sizes; in other words, the confidence that distribution-free confidence intervals can provide is limited by the sample size.

Note that these confidence levels are entirely independent of the distributional form of the population. The only assumption made in the above argument is that the probability of a sample value lying on either side of the population median is one half, which requires the distribution to be continuous at that point. Note also that the effect of increasing k is to decrease the width of the confidence interval at the cost of decreasing the confidence level. Published tables provide, for moderate sample sizes and popular confidence levels, the largest value of k that will provide at least the required confidence. For large sample sizes, a number of approximations have been developed. ISO 16269-7 [48] gives tables for k for sample sizes up to 100 and, for use with larger samples, approximations for confidence level $1-\alpha$ of the following form:

$$k = \tfrac{1}{2}\left[n + 1 - u(1 + 0{,}4/n)\sqrt{n-c}\right]$$

where

k is to be rounded down to the next whole number;

u is the standard normal deviate corresponding to an upper tail area of α for a one-sided confidence interval and $\alpha/2$ for a two-sided interval;

c is a constant depending on u.

The values of u and c are provided for eight different confidence levels and for one- and two-sided intervals. It is claimed in ISO 16269-7 that in all these cases the formula yields the correct value of k for sample sizes up to at least 280 000, enough for most practical purposes! For illustration, for a two-sided confidence interval with confidence level 99 %, the values of u and c are given as 2,575 829 30 and 1,74. Thus, for a sample size of 200, the appropriate value of k is:

$$\tfrac{1}{2}\left[200 + 1 - 2{,}575\ 829\ 30 \times (1 + 0{,}4/200)\sqrt{200 - 1{,}74}\right] = \tfrac{1}{2}(201 - 36{,}341\ 5) = 82{,}33$$

rounded down, i.e. $k = 82$. It can therefore be asserted in general with at least 99 % confidence that the confidence interval $(x_{[82]}, x_{[119]})$ from a sample of size 200 includes the population median.

The eight decimal places for u are only necessary when obtaining k with high accuracy for very large sample sizes, and may be reduced to two or three decimal places if an approximate value for k is adequate.

Similar methods can be used to determine confidence intervals on other fractiles of the population.

9 Acceptance sampling

9.1 Methodology

In Clause 8, a variety of statistical tests and intervals have been described. In the example in 8.3, it was shown how easy it can be to select an inferior criterion for assessing the quality of a lot, even in the simplest case where there is a single-sided requirement on a single quality characteristic. The difficulties in selecting a sound criterion are compounded if there are multiple quality characteristics, perhaps some with single and the others with double specification limits, particularly if not all of these characteristics are independent. In the face of the multiplicity of potential applications and the many techniques from which to choose, a general approach to the problem of assessing quality is desirable.

The supplier will naturally concentrate his attention on keeping the mean of each quality characteristic as close as possible to a target value, and the standard deviation as small as possible (see Clause 10). The customer, on the other hand, will be principally concerned with the quality level of submitted product, i.e. the fraction of nonconforming items or the number of nonconformities per 100 items. Nonconformity is defined as departure of a characteristic from specification, and the probability of such a departure will generally depend on the mean, μ, and standard deviation, σ, of the characteristic in the population. For example, it can be seen from Table 15 that if a lower specification limit, L, has been set then, provided $\mu - 2{,}053\,7\sigma \geqslant L$, no more than about 2 % of product will be outside specification if the distribution of the characteristic is normal. If L was equal to 100, say, then a combination such as $\sigma = 1$, $\mu = 102{,}6$ would provide a similar quality level to the customer as the combination $\sigma = 1{,}2$, $\mu = 103{,}1$.

This suggests the following general approach to the assessment of product acceptability: use the sample information to estimate the proportion of product that is outside specification, and accept the batch only if the estimate is below a given maximum value. A judicious choice of this maximum value will provide a given level of assurance that not more than a given proportion of product is outside specification. It turns out that this approach does indeed lead to efficient use of the sample information and to intuitively sensible sampling procedures.

Sometimes the testing of the estimated quality level against a maximum value is done implicitly. For example, for sampling by attributes, the unbiased estimate of the fraction nonconforming is:

$$\hat{p} = \frac{r}{n}$$

where r is the number of nonconforming items in a sample of size n from the lot.

Suppose that the maximum value of $\hat{p}$ for which lot acceptance takes place is denoted by p*. Then a lot is only accepted if $\hat{p} \leqslant p$*, i.e. $r/n \leqslant p$*, i.e. $r \leqslant np$*, i.e. $r \leqslant c$ where c is the largest whole number less than or equal to np*. In practice, the acceptance criterion in this situation is always expressed as $r \leqslant c$ (or $r \leqslant$ Ac, where Ac is the acceptance number, see 9.4.1), so it is not immediately obvious that it conforms to the suggested general approach.

Because these sampling methods are used to determine whether or not a population (lot, batch, consignment) of product should be accepted, they are referred to as *acceptance sampling* methods. The methods described in Clause 9 are primarily for application to a continuing series of lots from the same supplier, although the case of isolated lots or short series is also considered.

9.2 Rationale

In the 1930s and 1940s, the theoretical foundations of acceptance sampling were originally laid. The outbreak of World War II in 1939 created an almost immediate need to put these acceptance sampling methods into practice due to the following circumstances. Many of the skilled workforces of the countries involved in the war were recruited into military service, leaving less skilled people to fill the void in factories for the production of materials needed for the war effort. The demand for these materials was both large and urgent, and the situation inevitably led to degradation in product quality. To address this problem, acceptance sampling standards were created and implemented with the dual objective of protecting the customers (i.e. the troops) from receiving materials of inferior quality while ensuring that they had sufficient material for their missions.

In the years following the war, these military standards found their way into general use in industry and were further developed with the philosophy of motivating suppliers, through lot rejection, tightened inspection, and discontinuation of sampling inspection, to improve their production processes if inferior quality was being provided. In addition, industry began to direct its attention to more general quality management improvement strategies to prevent rather than detect and correct the occurrence of nonconformance. The emphasis moved from inspection of final product towards improving the manufacturing processes and controls and producing more robust product designs.

When the supplier's quality control system can provide assurance that a process is in statistical control with sufficiently low variability in the quality characteristic or characteristics, acceptance sampling of the final output from the process is redundant, merely confirming what was already known. Indeed, modern quality management standards focus on achieving this goal through ongoing activities of continuing process improvement and have achieved great success in many areas. In such an environment, a company that relies on acceptance sampling to assure the customer that he is getting what he wants is nowadays rightly seen as operating in an inferior state, being akin to an admission of failure to get the production processes into shape.

So under what scenarios should acceptance sampling be used? There are several. The first is when the process may be immature, with unexpected teething troubles arising from time to time that would not necessarily be picked up in production under the existing process controls. Furthermore, for quality characteristics of a qualitative type, the application of statistical-process-control (SPC) procedures like Shewhart control charts may be either unfeasible or unprofitable. Another is when some of the processes involved may be state-of-the-art, perhaps using materials whose properties are not yet fully understood. (This is sometimes the case in defence industries, due to the constant push at the limits of technology and changes to the specification in order to produce devices that are superior to those of the perceived adversary.) Another reason is that it may be necessary to guard against human fallibility and unpredictability, for example where the production item is a complex and delicate assembly of components and where a warranty system is inappropriate. Some products may be affected by handling and shipping activities, making acceptance sampling at the point of receipt necessary. Other products such as in-service utility meters must be tested periodically to determine whether quality has degraded to an unacceptable state and acceptance sampling is an efficient means of doing this. There are also many products of a non-industrial nature where the room for process improvement is limited such as harvested quantities of fish or produce that must be inspected for acceptance or grading on a sampling basis.

9.3 Some terminology of acceptance sampling

9.3.1 Acceptance quality limit (AQL)

The acronym AQL was originally an abbreviation of “acceptable quality level”, and most international standards on acceptance sampling are being indexed by the AQL. In 1998, in order to emphasize that the AQL should not be interpreted as a desirable quality level, the meaning of the acronym was revised by ISO/TC 69/SC 5, the ISO subcommittee responsible for developing and maintaining International Standards on acceptance sampling. It now stands for “acceptance quality limit” defined as the “quality level that is the worst tolerable process average when a continuing series of lots is submitted for acceptance sampling”.

There are two main objections to the continued usage of AQL as an abbreviation of acceptable quality level. One was the name itself; the idea that any quality level other than perfection should be considered acceptable or satisfactory in the modern era became outmoded, as one of the basic tenets for surviving in a global economy is the need to strive for continuous quality improvement. The other was the definition of an AQL as “when a continuing series of lots is considered, a quality level which for the purposes of acceptance sampling is the limit of a satisfactory process average”. This definition had been deliberately worded to indicate that the acceptability was to be construed as only for the purpose of identifying a suitable sampling plan, not in any absolute sense. But for the most part the words “for the purposes of acceptance sampling” have been either ignored or misunderstood by commentators.

9.3.2 Limiting quality (LQ)

When acceptance sampling is applied to a single lot, or to a short series of lots, the concept of an AQL is inappropriate as there is no longer a continuing series. The principal index to classical sampling plans for isolated lots is the limiting quality (LQ), which is the quality level for which the probability of acceptance is a specified low value, usually 10 %. [This is the same thing as the lot tolerance percent defective (LTPD), a term that is now rarely used.] The LQ is chosen by the customer to be an unsatisfactory lot quality level at which lots would be expected to have a high probability of failing the acceptance criterion. For a continuing series of lots, the corresponding unsatisfactory *process* quality level is called the limiting quality level (LQL).

More generally, for both lots and processes, this quality level is referred to as the consumer's risk quality (CRQ).

9.3.3 Classical versus economic methods

Classical acceptance sampling methods are typically selected with little attempt to balance the costs of sampling and inspection against the savings due to more reliably accepting good products and rejecting bad.

The reasons for this are fourfold. Firstly, experience has indicated that the classical methods provide sampling procedures that are not that far from the economic optimum over a wide range of scenarios. Secondly, in order to be able to determine the optimal level of inspection for an economic plan, the cost of non-acceptance of good lots and the cost of acceptance of bad lots need to be known. The latter cost is typically particularly difficult to ascertain in most cases, partly because it depends on the extent to which the nonconforming items are out of specification. Thirdly, a presumption has to be made about the distribution of incoming quality. Fourthly, even when an assessment can be made of all these quantities, the economic sampling procedure generally depends on finding the minimum of a complicated formula, requiring too high a level of mathematical sophistication for the typical user.

For the above reasons, we shall only consider classical types of acceptance sampling scheme below. Readers interested in further details of the economic approach are referred to Wetherill and Chiu [138] and von Collani [131]. Some flexibility to lower or raise the amount of inspection on economic or other grounds is provided by the choice of inspection level, which is described next.

9.3.4 Inspection levels

The two best-known acceptance sampling systems (ISO 2859-1 [4] for sampling by attributes and ISO 3951 [13] for sampling by variables) provide three general inspection levels (I, II and III), and four special inspection levels (S-1 to S-4). If the inspection level is not specified, it is assumed that general inspection level II is to be used. If better discrimination between good and bad quality is required, perhaps because the supplier has a history of erratic quality, inspection level III may be chosen. Conversely, if a lower level of discrimination is adequate, inspection level I may suffice. Sometimes even the lower sample sizes required by inspection level I are uneconomic when the inspection is expensive or involves destructive testing, or unnecessarily costly in view of the excellent quality history of similar products, the reputation of the supplier or the low importance of the characteristics under consideration. In such cases, one of the levels S-1 to S-4 may be selected, as long as it is understood that the discriminatory ability (i.e. the power) of the sampling scheme tends to diminish as one moves from S-4 to S-1.

The inspection level in combination with the lot size determines a sample size code letter, which is then used in conjunction with the AQL to look up the parameters of the sampling plans of an acceptance sampling scheme.

9.3.5 Inspection severity and switching rules

At the start of sampling inspection, when it is believed that the quality level of a process is satisfactory, so-called *normal inspection* is used in ISO 2859-1 [4] and in ISO 3951 [13]. If the results from a predetermined number of lots under normal inspection indicate that the quality level of the process is less than satisfactory, then the severity of the inspection is increased to *tightened inspection.* A tightened inspection plan will usually have the same sample size as the corresponding normal inspection plan, but with a stricter acceptability criterion.

If the results from a predetermined number of lots under normal inspection indicate that the quality level of the process is very good, then the severity of the inspection may be decreased to *reduced inspection.* Thus, each normal inspection plan has a corresponding tightened and reduced inspection plan. Each group of three such plans is called a *sampling scheme*. The International Standards ISO 2859-1 [4] and ISO 3951-1 [13] are collections of sampling schemes of a particular type, and are called *sampling systems*. The rules for moving between the plans that make up a scheme are called the *switching rules*. Whereas the normal inspection and tightened inspection plans are mandatory parts of a ISO 2859-1 or ISO 3951 scheme, the reduced inspection plan is discretionary.

Thus, a sampling scheme from these standards consists of at least two plans. (It is sometimes overlooked that the plans were not designed to be used in isolation.)

A switch is also made from tightened inspection to discontinuation of inspection if the sample quality levels fail to improve sufficiently quickly. The supplier then needs to act to resolve any problems with the process before acceptance sampling may be resumed. Tightened inspection is then applied with the switching rules reset in the same way as if there had just been a switch from normal inspection.

9.3.6 Use of “non-accepted” versus “rejected”

New users of acceptance sampling standards may be bemused to find the word “non-accepted” used where the word “rejected” may seem more appropriate. There is a good reason for distinguishing these terms. The term “rejected” implies that the user is not prepared to accept the lot under any circumstances. However, when a lot is non-accepted, this only means that it has failed the acceptance criterion of the sampling inspection plan. It does not preclude the customer and the supplier from coming to some arrangement to accept the lot on concession, e.g. at a reduced price, or for a different use such as training, or after some remedial action.

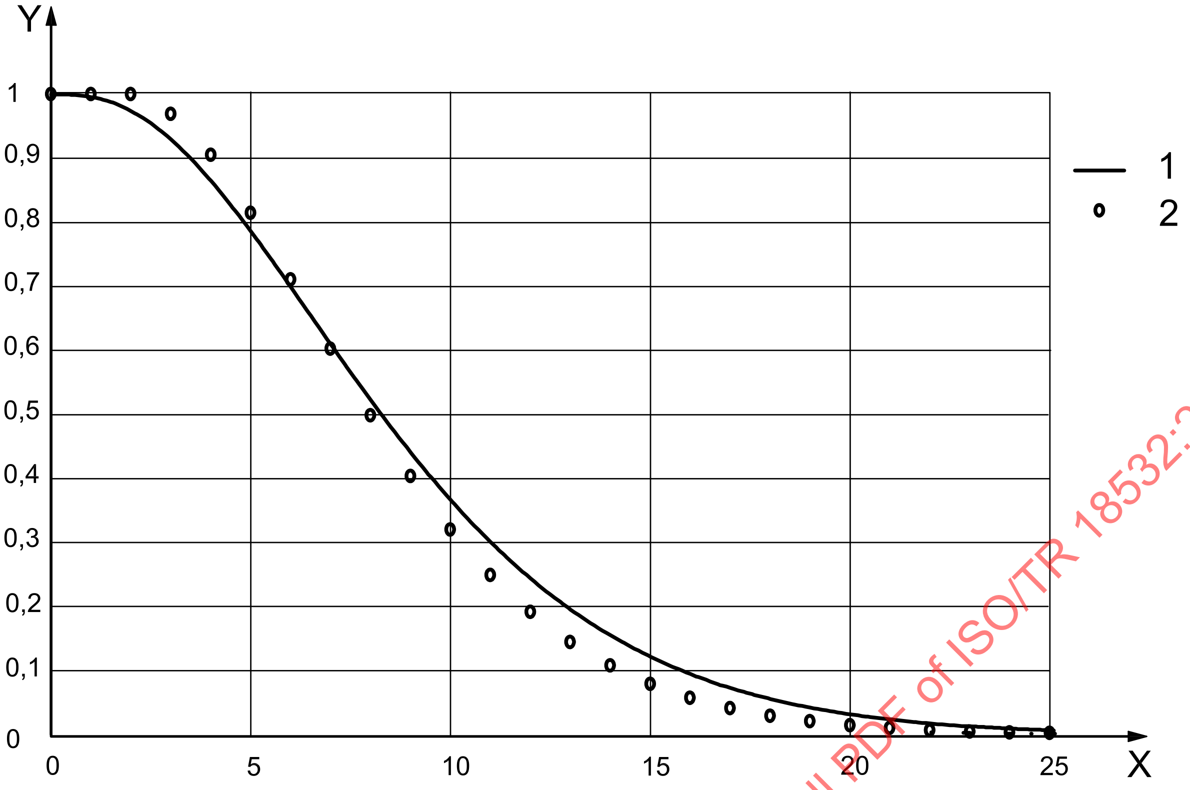
9.4 Acceptance sampling by attributes

9.4.1 General

The concept of an attribute has already been discussed in 8.6.1. For the moment, suppose that items have a single quality characteristic that is an attribute; multi-attribute cases are discussed in 9.6.3 and 9.6.4. For the purposes of discussion, suppose also that quality is measured in terms of percent nonconforming rather than nonconformities per 100 items. (The methods for nonconformities per 100 items for a single attribute are very similar.)

It is important to distinguish the case of an isolated lot from that of a continuing series of lots. For an isolated lot, the primary purpose of acceptance sampling will be to provide assurance that the *lot* percentage nonconforming is no worse than the limiting quality (LQ). Tabulated plans for isolated lots or short series of lots are therefore indexed by LQ and lot size. For a continuing series of lots, AQL-indexed plans, which are also indexed by lot size, are designed to provide protection against lots being accepted when the *process* quality level is worse than the AQL.

The performance of a sampling plan can be assessed in both cases by considering its *operating characteristic* (OC) curve. This is a graph of probability of acceptance against quality level. Note that if the lot is of size N, there are only a finite number of possible values of the lot percent nonconforming, namely 0, $1/N$, $2/N$,…, $(N-1)/N$, 1. Strictly speaking, the operating characteristic curve for isolated lots is therefore not really a curve, for it will only exist at these values, i.e. it will appear as a series of dots. This type of OC curve is said to be of Type A. By comparison, the process quality level could be any value in the range from 0 % to 100 %, so the operating characteristic appears as a curve, called Type B. Figure 40 shows both types of OC curve for plans with a sample of size $n = 32$ drawn from a lot of size $N = 100$, where the lot is accepted when there are no more than Ac = 2 nonconforming items in the sample. Ac is called the *acceptance number* of the plan.



Key

1 type B

2 type A

X percent nonconforming

Y probability of acceptance

Figure 40 — Type A and B OC curves for $n = 32$, Ac = 2, $N = 100$

9.4.2 Single sampling

The plans in 9.4.1 are both called *single sampling* plans, as the decision whether to accept or not to accept the lot depends on the results of a single sample. ISO 2859-1 provides master tables of single sampling plans indexed by AQL for a continuing series of lots. (See 9.4.10 for LQ-indexed plans.) To find the appropriate plan, first the lot size and inspection level are used to determine the appropriate *sample size code letter* from Table 1 of ISO 2859-1:1999. Then this code letter together with the AQL are used to determine the sample size and acceptance number from Table 2-A for normal inspection, Table 2-B for tightened inspection or Table 2-C for reduced inspection.

These master tables have a very simple structure. The sample sizes are restricted to the set of 17 *preferred* sample sizes 2, 3, 5, 8, 13, 20, 32, 50, 80, 125, 200, 315, 500, 800, 1 250, 2 000, 3 150. These roughly form a geometric series with common ratio $10^{1/5} = 1{,}585$. The AQLs run from 0,01 % to 1 000 % also roughly as a geometric series with the same common ratio. (Above the AQL of 10 %, the plans are for use only for nonconformities per 100 items.) The result of this is that the acceptance numbers, which are also restricted to a series of preferred values, are the same along diagonals of the tables.

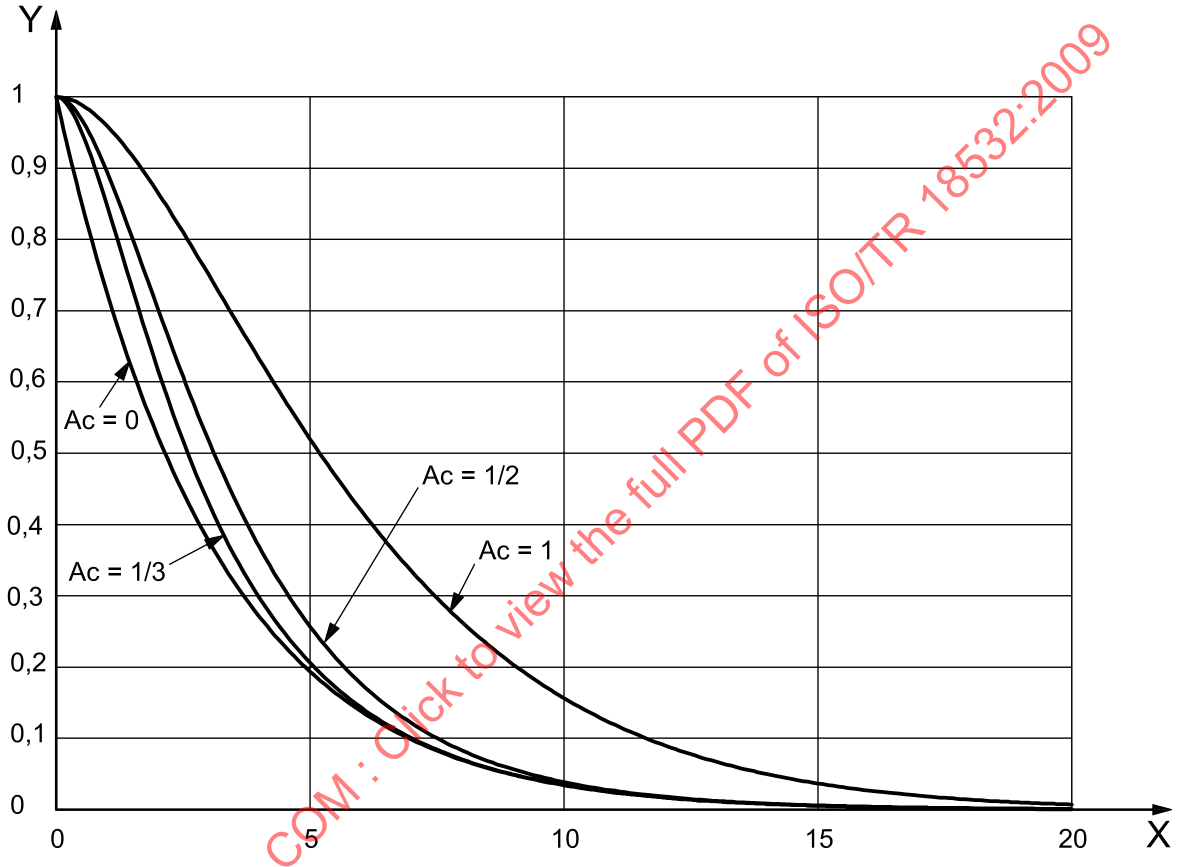
ISO 2859-1 underwent substantial revision between the first and second editions. One of the changes is the introduction of optional *fractional acceptance number* plans between the diagonals for acceptance numbers zero and one, where in the previous edition there were arrows pointing upwards or downwards. They operate as follows. An acceptance number of $1/k$ when the sample size remains constant from lot to lot means that the present lot can be accepted if:

a) the sample from the present lot contains no nonconforming items; or

b) the sample from the present lot contains one nonconforming item, and the samples from the immediately preceding $(k-1)$ lots contain no nonconforming items between them.

The fractional acceptance number plans are given in Tables 11-A, 11-B and 11-C of ISO 2859-1:1999 for normal, tightened and reduced inspection. The fractions used are $\frac{1}{3}$ and $\frac{1}{2}$ for normal and for tightened inspection; for reduced inspection, where there are three diagonals between the acceptance numbers zero and one, the fractions used are $\frac{1}{5}$, $\frac{1}{3}$ and $\frac{1}{2}$.

The reason for introducing fractional acceptance number plans is that there is such a big difference between the OC curves for acceptance numbers zero and one, and often the desired OC curve is somewhere in between. Figure 41 illustrates how rapidly OC curves change shape in this range by showing the Type B OC curves for the plans with sample size 32 and acceptance numbers 0, $\frac{1}{3}$, $\frac{1}{2}$ and 1.



Key

X percent nonconforming

Y probability of acceptance

Figure 41 — Type B OC curves for Ac = 0, 1/3,1/2 and 1

If the lot sizes change sufficiently to cause the sample sizes to vary from lot to lot, then the determination of whether the acceptance number for the present lot should be zero or one becomes rather more complicated.

It is based on a cumulative *acceptance score.* This score is reset to zero whenever there is a switch to a different severity of inspection or whenever a nonconforming item is found. For the current lot, it increases by 2, 3, 5 or 7 whenever the tabulated acceptance number is $\frac{1}{5}$, $\frac{1}{3}$, $\frac{1}{2}$ or at least 1 respectively, but remains unchanged if the tabulated acceptance number is 0. If the tabulated acceptance number is a whole number, it is used irrespective of the acceptance score, but if the tabulated acceptance number is a fraction, the acceptance number is taken to be zero if the score is 8 or below, and one otherwise. An added complication when the sample size changes is that the switching rules also become more complicated, requiring the maintenance of a *switching score*.

9.4.3 Double sampling

Double sampling plans provide one means by which the average amount of sampling may be reduced. Note that it is only the *average* that is reduced; for some lots the amount of sampling will be greater than for single sampling. A double sampling plan by attributes works as follows. In general it has five parameters, which may be denoted n_1, n_2, c_1, c_2 and c_3. A random sample of size n_1 is taken from the lot and the number of nonconforming items d_1 is counted. There are three possible outcomes at this stage.

a) If $d_1 \leqslant c_1$, then the lot can be accepted without further sampling.

b) If $d_1 \geqslant c_2$, then the lot can be non-accepted without further sampling.

c) If $c_1 < d_1 < c_2$, then no immediate decision can be taken on lot acceptability.

In case c), another random sample, this time of size n_2, is selected and the number of nonconforming items, d_2, in the sample is counted. The total number of nonconforming items found in the two samples is $d_3 = d_1 + d_2$. If $d_3 \leqslant c_3$, then the lot is accepted, otherwise it is non-accepted.

In summary, if the evidence from the first sample is very good or very bad, then an immediate decision can be taken. When the evidence is inconclusive, then a further sample is necessary to resolve the matter.

The integers c_1 and c_3 are the acceptance numbers of the plan. The integers c_2 and $c_3 + 1$ are the rejection numbers. Note that $c_2 - c_1$ has to be at least equal to 2, otherwise a decision on lot disposition will always be reached from the results of the first sample.

The procedure for nonconformities is the same as this except that “nonconforming items” is replaced throughout by “nonconformities”.

In standards on sampling by attributes, the five parameters of each double sampling plan are chosen so that the OC curve of the double sampling plan roughly matches the OC curve of the corresponding single sampling plan. For simplicity and ease of operation, this matching is generally constrained so that the sample sizes n_1 and n_2 are equal to one another and so that the acceptance and rejection numbers are identical along diagonals of the master tables. Denoting the corresponding single sample size by n_0, it turns out that the double sample sizes are typically given by $n_1 = n_2 \cong 0{,}63 n_0$. It follows that average savings in inspection effort of up to about 37 % of the single sample size may be achieved by using double sampling instead of single sampling, depending on the submitted quality.

The disadvantages of double sampling are their increased administrative and logistical requirements, which often lead to double sampling being impracticable. For example, suppose the acceptance test is to determine whether a device can survive 1 000 h at 200°C. It may be possible to test the devices simultaneously, so the test of a single sample would take 1 000 h. However, if double sampling were used and the result from the first sample was inconclusive, then a second sample would be necessary. Testing of the second sample may not even be able to start at once if the test facility needs to be booked in advance. Coupling this with the time that the second sample requires to be tested, a decision on lot disposition will be substantially delayed. Meanwhile the lot will need to be stored somewhere, awaiting shipment.

Double sampling plans for sampling by attributes may be found for normal, tightened and reduced inspection in Tables 3-A, 3-B and 3-C of ISO 2859-1:1999 respectively, and in equivalent standards.

9.4.4 Multiple sampling

Multiple sampling takes the idea of double sampling a stage further. A k-stage multiple sampling plan has sample size n_i and acceptance and rejection numbers Ac_i, Re_i at the ith stage, for $i = 1, 2, ..., k$. ISO 2859-1:1999 has five-stage multiple plans with the sample sizes the same at each stage, and each equal to about one quarter of the corresponding single sample size. These five-stage plans represent an improvement over the seven-stage plans of the previous edition of ISO 2859-1, in terms of both practicality and match with the OC curves of the corresponding single sampling plans. Again, the sets of acceptance and rejection numbers are kept the same along diagonals of the master tables, which are given as Tables 4-A, 4-B and 4-C in ISO 2859-1:1999.

As may be expected, multiple sampling plans provide a further reduction in average inspection requirements compared to double sampling plans. They are worthwhile provided the gains are not outweighed by logistical and administrative difficulties. At perfect quality, there may be as much as a 75 % saving in inspection costs when compared with single sampling plans with acceptance number greater than 5. For multiple sampling plans matching single sampling plans with acceptance numbers of 5 or lower, the maximum saving will be nearer to 50 % as a decision to accept will not be possible after the first multiple sample.

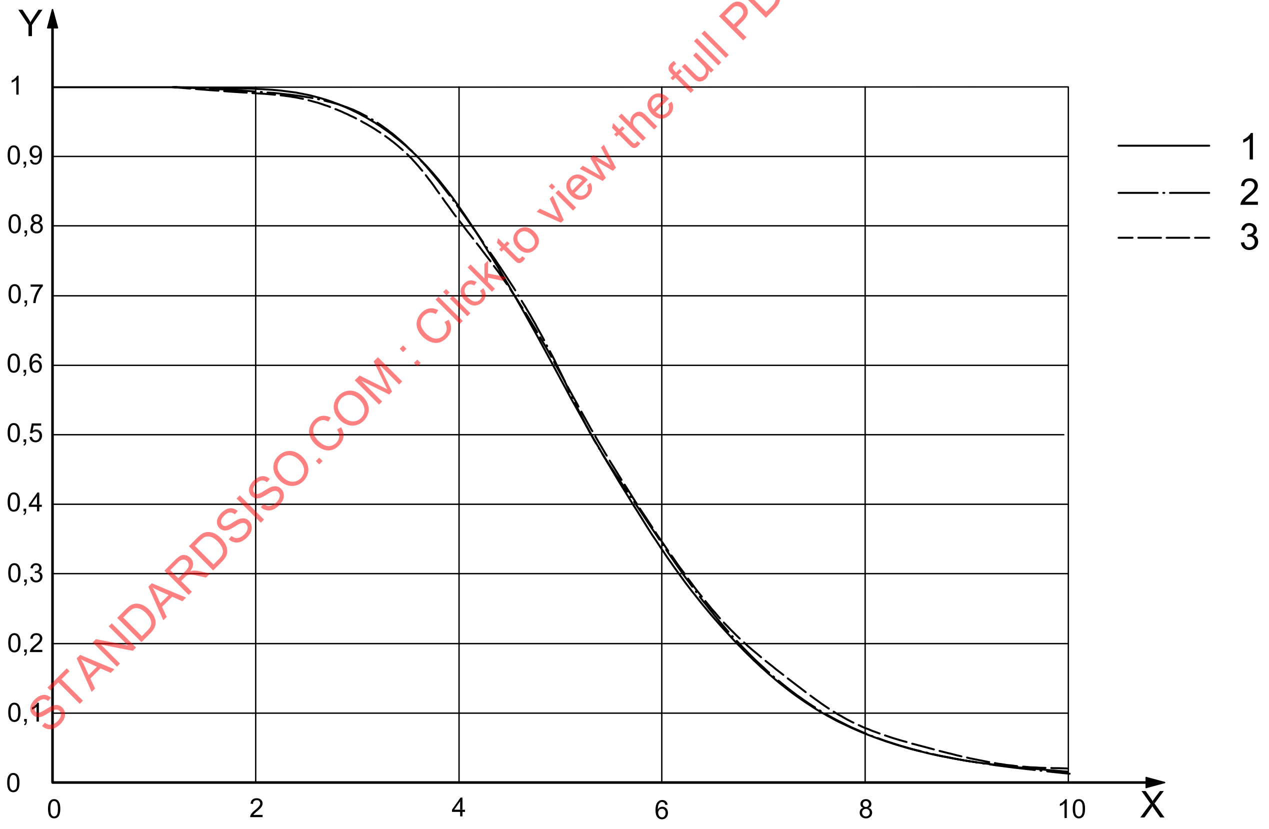
It is instructive to compare the properties of single, double and multiple sampling plans at different quality levels. Consider, for example, sample size code letter L in combination with an AQL of 2,5 %. The plans to be compared in Figure 42 are:

— Single sampling: $n = 200$; Ac = 10, Re = 11;

— Double sampling: $n_1 = n_2 = 125$; $\text{Ac}_2 = 5$, $\text{Re}_2 = 9$; $\text{Ac}_2 = 12$, $\text{Re}_2 = 13$;

— Multiple sampling: $n_1 = n_2 = n_3 = n_4 = n_5 = 50$; $\text{Ac}_1 = 0$, $\text{Re}_1 = 5$; $\text{Ac}_2 = 3$, $\text{Re}_2 = 8$; $\text{Ac}_3 = 6$, $\text{Re}_3 = 10$; $\text{Ac}_4 = 9$, $\text{Re}_4 = 12$; $\text{Ac}_5 = 12$, $\text{Re}_5 = 13$.

Figure 42 shows the OC curves of the three plans. It can be seen that there is a very good match between all three.



Key

1 single sampling

2 double sampling

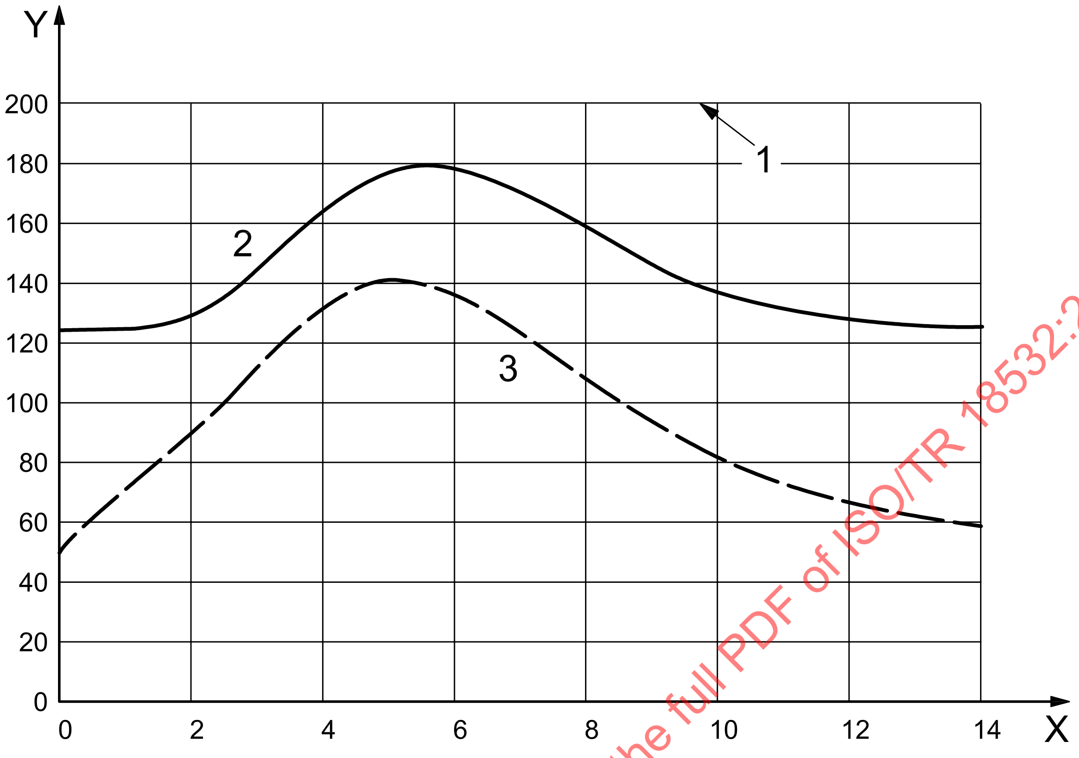
3 multiple sampling

X percent nonconforming

Y probability of acceptance

Figure 42 — OC curves for single, double and multiple sampling size code letter L and AQL 2,5 %

Figure 43 shows the average number of items that will be inspected at different quality levels, the so-called average sample size (ASSI) curves, for the three types of plan.



Key

1 single sampling

2 double sampling

3 multiple sampling

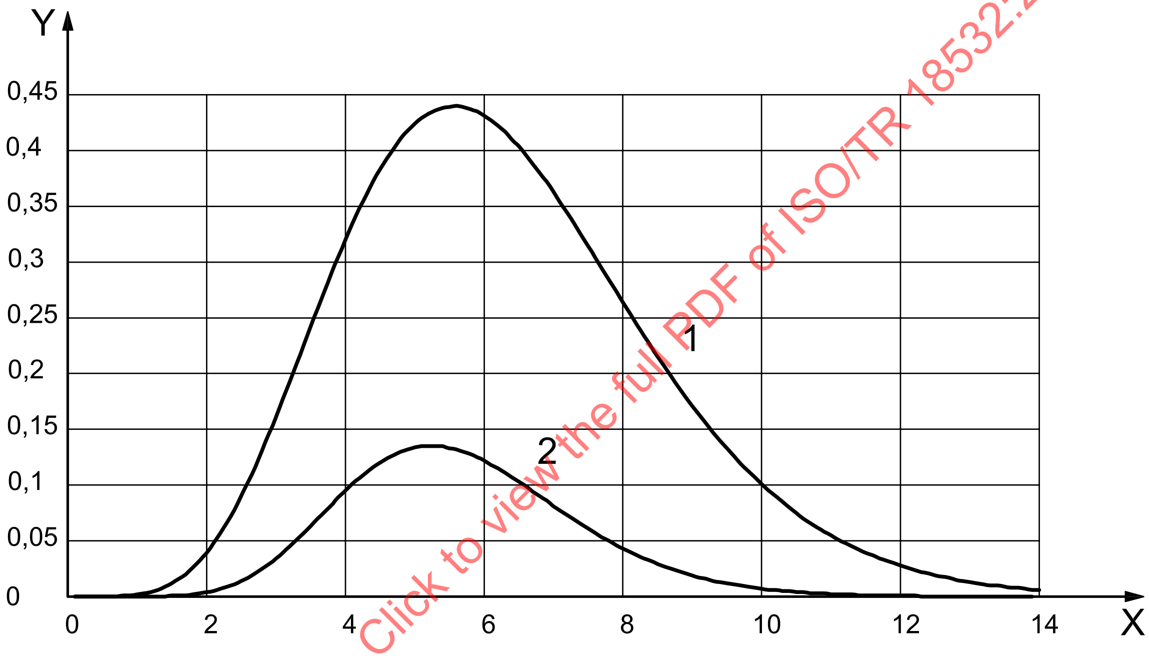
X percent nonconforming

Y average sample size

Figure 43 — Average sample size (ASSI) curves for single, double and multiple sampling plans for sample size code letter L and AQL 2,5 %

Sometimes, particularly with destructive inspection, the main disincentive to using double and multiple sampling plans is the possibility that more items will need to be inspected than would be the case with single sampling. Figure 44 shows the probability that the corresponding single sample size is exceeded for the double and multiple sampling plans. In the case of double sampling plans in general, this is the probability of needing a second sample to come to a decision. In the case of this multiple sampling plan, it is the probability of needing all five samples to come to a decision. (Note that for multiple sampling plans where the single sample size is not divisible by four, needing the fourth sample from the ISO 2859-1 multiple sampling plans may lead to the single sample size being exceeded, but not significantly.)

Figure 44 shows clearly that another advantage that multiple sampling has over double sampling is a very substantial reduction in the chance of needing to inspect significantly more items than under single sampling.



Key

1 double sampling

2 multiple sampling

X percent nonconforming

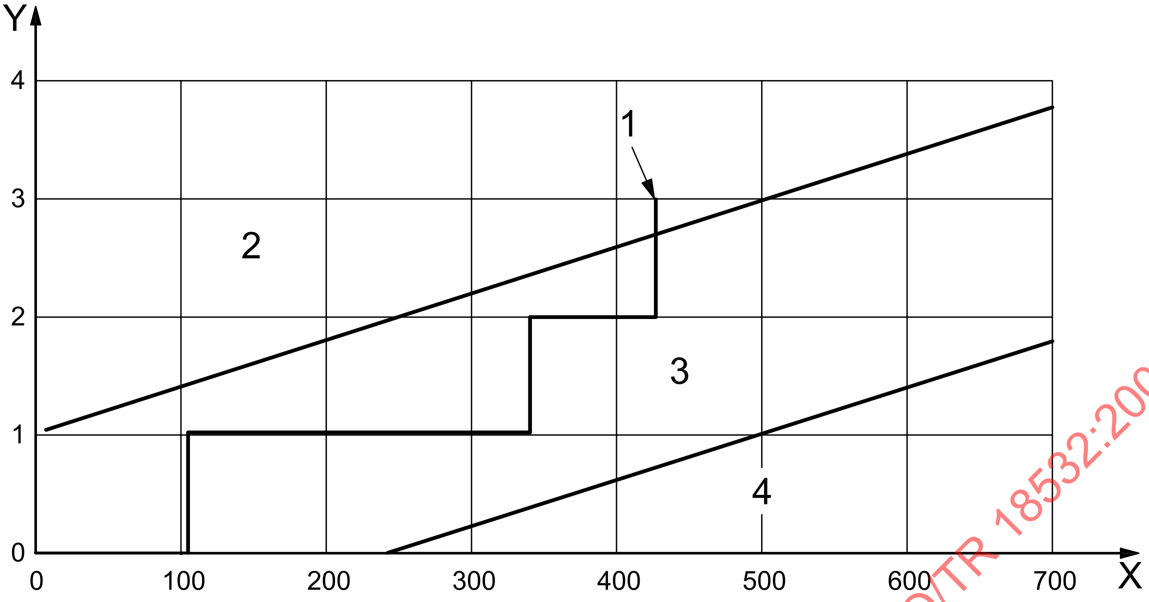
Y probability of exceeding the single sample size

Figure 44 — Curves for the double and multiple sampling plans for sample size code letter L and AQL 2,5 % showing the probability of needing to inspect significantly more sample items than under single sampling

9.4.5 Sequential sampling

The ultimate multi-stage procedure is to inspect items one at a time, making a decision after each inspection either to accept the lot, not to accept the lot or to continue sampling. This is called sequential sampling.

Wald [134] devised an approximate method of determining the acceptance and rejection numbers at each cumulative sample size that provide specified values of the overall supplier's and customer's risks. It turns out that, on a graph of cumulative number of nonconforming items against cumulative sample size, the boundaries of the “continue sampling” region are parallel straight lines, as shown in Figure 45.



Key

1 reject lot

2 reject zone

3 continue sampling

4 accept zone

X cumulative sample size

Y cumulative number of nonconforming items

Figure 45 — Example of sequential sampling by attributes for percent nonconforming

The diagram for nonconformities looks similar, except that jumps of more than one nonconformity on the vertical axis are possible for an increase of one in the cumulative sample size.

The first edition of ISO 8422 [30] was based upon the Wald approximation. Baillie [68],[71] demonstrated that, although Wald's method works well when the supplier's and customer's risks are no more than about 1 % or 2 %, the method can be very inaccurate when these risks are 5 % and 10 % respectively, as they were designed to be in ISO 8422. In fact, the resulting plans often have a supplier's risk much less than 5 %, while the customer's risk sometimes exceeds 10 %, the net effect being to require more inspection than necessary on average, i.e. a higher average sample size (ASSI).

Another complication is that the plans in ISO 8422 are curtailed at 1½ times the corresponding single sampling size, further distorting the risks from the design values.

For example, the Wald approximation to the plan for nonconforming items for a nominal 5 % supplier's risk at a quality level of 0,1 % nonconforming and a nominal 10 % customer's risk at 1,0 % nonconforming turns out to have a 2,92 % customer's risk and an 11,13 % supplier's risk. However, by suitable choice of the plan parameters (to three decimal places of accuracy), it is possible to achieve risks of 4,99 % and 10,00 % respectively and consequentially lower values of the ASSI.

By taking the approximate values of the parameters from the Wald approximation and iteratively adjusting the gradient and position of the parallel lines, plans in the second edition of ISO 8422 (i.e. ISO 8422:2006 [30]) have been obtained that have supplier's and customer's risks much closer to the nominal values.

9.4.6 Continuous sampling

When items are produced in a continuous stream, there may be no natural way of dividing production into lots for the purposes of acceptance sampling. It is for such cases that continuous sampling plans were devised. The first of these, and the best known, is the Dodge [84] CSP-1 plan. This works as follows. First, a sampling frequency, f, and a clearance number, i, are specified. Then 100 % inspection begins. Once i successive conforming items have been inspected, 100 % inspection ceases and sampling inspection begins, with items being selected for inspection with probability f. As soon as a nonconforming item is found, inspection reverts to the 100 % level.

Many variations on this theme have been developed subsequently, with extra sampling frequencies and different rules for returning to 100 % inspection. The US Military Standard MIL-STD 1235A [139] contained five types of continuous sampling plan, indexed by the average outgoing quality limit (AOQL), i.e. the worst possible average outgoing quality over all values of incoming quality p. These plans were designed to have AOQLs that matched the AOQLs of the standard single sampling attributes schemes such as those in ISO 2859-1. All suffer from the same disadvantages: phases of 100 % inspection, which may be impracticable or uneconomic, and rapidly changing requirements for inspection personnel.

Beattie [73] proposed a different type of continuous sampling plan based on cumulative sums (cusums, see 10.7.3) on the number of nonconforming items. A cumulative sum is begun at a specified sampling frequency of one item in f. At the conclusion of accumulating and inspecting each sample of given size, n, the cumulative sum is increased by an amount $(d-k)$, where d is the number of nonconforming items in the sample and k is a specified target reference value. Until the cusum reaches or exceeds a specified upper limit, h, product is accepted. When h is reached or exceeded, a new cumulative sum is started at a specified value, $h+h'$; this could be at a different sampling frequency, but inspection requirements can be kept constant by keeping to the same frequency. Product is non-accepted until the second cusum reaches as low as, or goes below, h. Then a new cusum is started at zero, and the process starts all over again.

The problem with designing a Beattie-type system of acceptance sampling plans is determining how to index them, i.e. determining which performance requirements should be mapped into values of n, f, h and h'. The probability of acceptance at a given quality level does not mean quite the same as for lot-by-lot inspection, as it is the probability of acceptance of an item, rather than a lot. Read and Beattie [120] introduced this probability of acceptance as the Type C OC curve. They defined it, for any quality level p, as the ratio of the average run length (ARL) for the cusum in the acceptance zone to the sum of the ARLs in the acceptance and rejection zones, i.e.:

$$P_a(p) = \frac{\text{ARL}_A(p)}{\text{ARL}_A(p) + \text{ARL}_R(p)}$$

Points on this OC curve could be specified as a way of identifying performance requirements. A related requirement, assuming rectification of nonconforming items found in samples in the acceptance zone and all product in the rejection zone, would be a specified AOQL.

Wadsworth and Wasserman [132], based on the work of Wasserman [136], devised design guidelines for Beattie-type cusum procedures for normally distributed variables and for variables with the Poisson distribution. They proposed these as the basis of a national or international standard.

9.4.7 Skip-lot sampling

It has already been mentioned that inspection severity can be optionally switched to reduced inspection when the quality of successive lots remains consistently at a high level. For some types of product, the savings in switching to reduced inspection may not be very great. For example, the lot may still be delayed while inspection takes place. Moreover, the inspectors may still need to travel to the place of manufacture; or alternatively the sample, however much smaller than under normal inspection, may still need to be transported to a test facility. In short, many of the fixed costs may still have to be borne.

In response to these concerns, Dodge [85] developed the first skip-lot plan. In essence, it was a CSP-1 plan applied to lots instead of items; it was intended for use on homogeneous bulk materials from a reliable source where a single test result from each lot would determine its acceptability. Dodge and Perry [86] extended the idea as an overlay to a reference plan for lots consisting of discrete items. It operates as follows. The reference plan is used on successive lots until i successive lots have been accepted, at which point skip-lot sampling is introduced with a fraction f of lots, which are chosen at random, undergoing inspection still in accordance with the reference plan. As soon as a lot is non-accepted, application of the reference plan to every lot is resumed, and the process is repeated. By this means, some of the fixed costs may be avoided.

Liebesman and Saperstein [109] and Liebesman [110] developed a more sophisticated three-state procedure, which has been implemented as ISO 2859-3 [6]. To quote from Liebesman [110]:

"Three states are defined as part of the skip-lot standard: (1) state 1, lot-by-lot sampling, (2) state 2, skip-lot sampling, and (3) skip-lot interrupt. Qualification takes place during state 1 and requires acceptance of 10 lots in a row, the last two satisfying an individual lot criterion and the cumulative number of defects in the 10 lots satisfying a limit number criterion. When the program for a product is in state 2, interrupt occurs when a lot fails to satisfy the individual lot criterion. The program then transfers to state 3. During state 3, the product either re-qualifies for skip-lot having 4 lots in a row accepted with the last two satisfying the individual criterion; or the product becomes disqualified if a lot is rejected or the product is in state 3 for 10 lots."

Within the skip-lot state, there are four levels for the sampling frequency f, namely $1/2$, $1/3$, $1/4$ and $1/5$. The procedure is designed to encourage the supplier to maintain a quality level at half the AQL or better.

9.4.8 Audit sampling

There is a need for sampling procedures suited to formal, systematic inspections such as reviews or audits, to relieve the user from the problem of determining the appropriate sample size from formulae. ISO 2859-4 [7] has been developed in response to this need. The plans are designed to provide a risk of less than 5 % of wrongly contradicting a correctly declared quality level. In order to keep the sample sizes to reasonable levels, a relatively large risk is allowed of failing to contradict an erroneously declared quality level. Three levels of discrimination are provided, and the plans are recognizable as a subset of the familiar single sampling plans of ISO 2859-1 [4]. ISO 2859-4 is couched in terms of percent nonconforming items, but by a simple change of wording throughout, it can also be applied to nonconformities per 100 items.

Note that, strictly speaking, audit sampling is hypothesis testing rather than acceptance sampling despite the fact that the plans have been drawn from ISO 2859-1, which is an acceptance sampling standard.

9.4.9 Sampling for parts per million

For very good quality levels, typically measured in nonconforming items per million items, there are two difficulties with sampling inspection by attributes. One is the very large sample sizes that would be required to have supplier's and customer's risks as small as is generally the case with acceptance sampling plans. The second is that, quite often, no nonconforming items will be found; the result of this is that the unbiased estimator of the process fraction nonconforming, formed by dividing the number of nonconforming items in the sample by the sample size, will often take the unrealistic value zero.

ISO 14560 [46] presents plans for this situation for lot-by-lot sampling. The first problem is overcome by allowing larger supplier's and customer's risks. Instead of the usual values for these risks of around 5 % and 10 % respectively, they are increased to as much as 10 % and 20 %. The second problem is overcome by using an estimator that will overestimate p, the process fraction nonconforming, about half the time and underestimate it about half the time. The approximate formula for this estimator given in ISO 14560 is:

$$p = \frac{d + 0{,}7}{n + 0{,}4} \times 10^6 \text{ items per million items}$$

where d is the number of nonconforming items found in a sample of size n.

The plans are indexed by limiting quality level. Smaller sample sizes are required at better quality levels, thereby providing an incentive to the supplier to improve quality.

9.4.10 Isolated lots

When only one lot is being supplied, or a short series of lots, the protection afforded to the customer by the switching rules no longer applies. Attention then focuses on ensuring that any individual lot of a quality worse than a specified value has a low probability of acceptance. ISO 2859-2 [5] is a sampling system indexed by the limiting quality. The probability of acceptance at the LQ is usually no greater than 10 %, but in some cases, it is as high as 13 %. Two procedures are given. Procedure A is intended for use when the supplier and the customer both wish to regard the lot in isolation. Procedure B is for use when the supplier considers the lot to be one of a continuing series while the customer regards the lot to be received in isolation. In such a case, both the consumer (for the purpose of the acceptance sampling of one particular lot) and the producer (for the purpose of the acceptance sampling of the process) may use the same sampling plan.

The first edition of ISO 2859-2 was designed for inspection for percent nonconforming. The second edition is planned to also include inspection procedures for nonconformities per 100 items.

9.4.11 Accept-zero plans

Two of the quality world's buzz-phrases that originated in the late 20th century are “get it right first time” and the “zero defects philosophy”. In conjunction with the need to strive for ever-higher levels of quality, these words have often been taken to imply that if one or more nonconforming items are found in a sample, then the lot should not be accepted. In other words, the implication for sampling by attributes was interpreted to be that only plans with acceptance number zero should be admissible. Such plans are commonly referred to as *accept-zero* plans.

A system of accept-zero sampling plans should consist of rules for switching from one sample size to another in response to quality history, if necessary supported by master tables of sample sizes. Ideally, the switching rules would guarantee some property of the outgoing product. Squeglia [126] roughly matches accept-zero plans to the normal inspection single sampling plans of ISO 2859-1 [4] at the LQ, but provides no switching rules whatsoever. ISO 21247 [51] provides accept-zero plans for normal, tightened and reduced inspection with similar switching rules to those of ISO 2859-1. The ISO 21247 plans provide seven “verification” levels (rather than AQLs) and five sample size code letters. Unfortunately, the guidance as to choice of verification level is merely that higher numbered levels require more inspection and should be applied to more important characteristics; it is not clear precisely what each verification level is designed to achieve.

Klaassen [105] derived a remarkably simple formula for determining the accept-zero sample sizes from successive lots that would be needed to guarantee an average outgoing quality limit (AOQL). Defining K, the “credit”, as the total number of items accepted since the last lot non-acceptance, he showed that the sample size required to guarantee an AOQL of a is given by the smallest integer, n, such that:

$$n \geqslant \frac{N}{(K+N)a+1}$$

where N is the next lot size. This assumes that lots that are non-accepted when $K = 0$ are 100 % inspected and that all conforming items found in such lots are accepted. The advantages of this method are threefold:

a) no tables are required for its implementation;

b) a single quantity, K, is sufficient to summarize the quality history; and

c) the AOQL guarantee is valid regardless of the sizes of successive lots or the sequence of lot qualities, rendering the method virtually abuse-proof.

To illustrate this method, suppose for a certain item that it is required that the average quality reaching the market place does not exceed 1,5 % nonconforming in the long term, i.e. $a = 0{,}015$. Suppose that a supplier always submits lots of the same size, N. K is initialized to zero. Suppose first that $N = 200$. If no nonconforming items are found in each sample, the sample sizes found by using the Klaassen formula are:

50, 29, 20, 16, 13, 11, 10, 8, …

Even if the lot sizes are huge, the sample sizes for this AOQL for successive lots never increase above:

67, 34, 23, 17, 14, 12, 10, 9, …

However, whenever a nonconforming item is found in a sample, the lot has to be screened and conforming items accepted, K has to be reset to zero and the sample size for the next lot needs to immediately return to the one at the beginning of the sequence.

Note that the method does not guarantee that the outgoing quality will not exceed the AOQL over any particular sequence of lots. The guarantee applies to the *long-term average*, or *expected*, outgoing quality over the whole sequence. Over a short series of lots there will be an appreciable probability that the AOQL will be exceeded, but this probability will tend to zero as the length of the series increases.

The method is embodied in ISO 18414 [50].

9.5 Acceptance sampling by variables — Single quality characteristic

9.5.1 General

For quality characteristics that are variables distributed according to a known family of probability distributions, it is possible to utilize this extra information to produce sampling plans that are more efficient. Most procedures for acceptance sampling by variables are for data from normal distributions, and discussion in this clause will be confined to the normal distribution case. Where possible, the procedures will be explained with reference to ISO 3951-1 [13].

In ISO 3951-1, as in ISO 2859-1, the choice of inspection level and lot size determines a sample size code letter which, in conjunction with an AQL and an inspection severity, determines the sampling plan.

The procedures are classified as Form k or Form p^*, depending on whether the process fraction nonconforming is estimated implicitly or explicitly. In all cases, it is assumed that there is a continuing series of lots and that the process fraction nonconforming is the subject of assessment. Consequently, the methods have much in common with statistical tolerance intervals (see 8.8.1). Type A OC curves are not relevant to acceptance sampling by variables, as the presumption that the sampled population is normal cannot be true if the population is a finite lot.

The control of double specification limits on a variable is treated in one of three ways:

a) *combined control* is when a single AQL is applied to the sum of the process fractions nonconforming below the lower specification limit or above the upper specification limit;

b) *separate control* is when one AQL is applied to the lower specification limit and another AQL is applied at the upper specification limit;

c) *complex control* is when the one AQL is applied at either the lower or upper specification limit and a larger AQL is applied to the sum of the process fractions nonconforming beyond both of the specification limits.

The 1989 edition of ISO 3951 only contained Form k procedures for a single variable. ISO 3951 is being developed as a multi-part standard. ISO 3951-1 [13] is concerned with Form k procedures for a single variable and a single AQL. ISO 3951-2 [14] provides a more comprehensive coverage, using Form p^*.

The symbols L and U are used to denote the lower and upper specification limits.

Finally, a distinction is made between the "s" method and the "σ" method. The expression "s" *method* indicates that neither the process mean nor the process standard deviation is known. The expression "σ" *method* indicates that the process mean is unknown but the process standard deviation may be presumed to be known; in practice, this will mean that σ has been estimated from a large data set, and therefore is known within a small margin of error.

9.5.2 Single sampling plans by variables for known process standard deviation — The "σ" method

For a single, normally distributed quality characteristic with known σ, the acceptability of a lot may be determined as follows. Suppose that the Form k acceptance sampling plan has been determined, say by reference to ISO 3951-1. The plan will consist of a sample size, n, and an *acceptability constant*, k. A random sample of size n is drawn from the lot, and the sample mean, $\bar{x}$, is calculated.

The *quality statistic, Q,* is calculated for a lower specification limit:

$$Q_L = \frac{\bar{x} - L}{\sigma}$$

and/or for an upper specification limit:

$$Q_U = \frac{U - \bar{x}}{\sigma}$$

For a single specification limit, the lot is acceptable only if the quality statistic exceeds k. For double specification limits, and before inspection begins, it first needs to be checked that σ is not so big that the AQL requirements are impossible to meet under tightened inspection. This is done by comparing σ with $(U - L)$ times the tabulated value of the standardized maximum (allowable) process standard deviation (MPSD). The reason for this is that it is pointless to begin sampling inspection if the process variation is too large for the switching rules to function. For separate control of double specification limits, there will be two acceptability constants, say k_L and k_U, corresponding to the AQLs at each limit; in this case, the lot is acceptable only if both quality statistics exceed their respective acceptability constants. For combined control, both quality statistics have to exceed k. For complex control involving a separate AQL requirement on, say, the lower specification limit, the acceptance criterion would be similar to that for separate control except that k_U would correspond to the combined part of the requirement. Similar remarks apply to complex control involving a separate AQL requirement on the upper specification limit.

For the σ known case, the calculation of the quality statistics for each sample may be avoided. For example, the acceptability criterion $Q_L \geqslant k_L$ may be converted into $\bar{x} \geqslant L + \sigma k_L = \bar{x}_L$, which can be calculated in advance. Similarly, $Q_U \geqslant k_U$ may be converted into $\bar{x} \leqslant \bar{x}_U$.

For Form p^*, the acceptability constants p^* are maximum acceptable values of the estimated process quality level. The quality statistics are calculated in the same way as for Form k. Denoting a quality statistic in general by Q, the process fraction beyond a single specification limit is estimated by the area under the standard normal curve above the value $Q\sqrt{n/(n-1)}$. Denoting these estimates at the lower and upper specification limits by $\hat{p}_L$ and $\hat{p}_U$, the lot acceptance criteria become:

a) for a single lower specification limit, $\hat{p}_L \leqslant p^*$;

b) for a single upper specification limit, $\hat{p}_U \leqslant p^*$;

c) for combined control of double specification limits, $\hat{p}_L + \hat{p}_U \leqslant p^*$;

d) for separate control of double specification limits, $\hat{p}_L \leqslant p_L^*$ and $\hat{p}_U \leqslant p_U^*$;

e) for complex control of double specification limits, either $\hat{p}_L \leqslant p_L^*$ and $\hat{p}_L + \hat{p}_U \leqslant p^*$ or $\hat{p}_U \leqslant p_U^*$ and $\hat{p}_L + \hat{p}_U \leqslant p^*$.

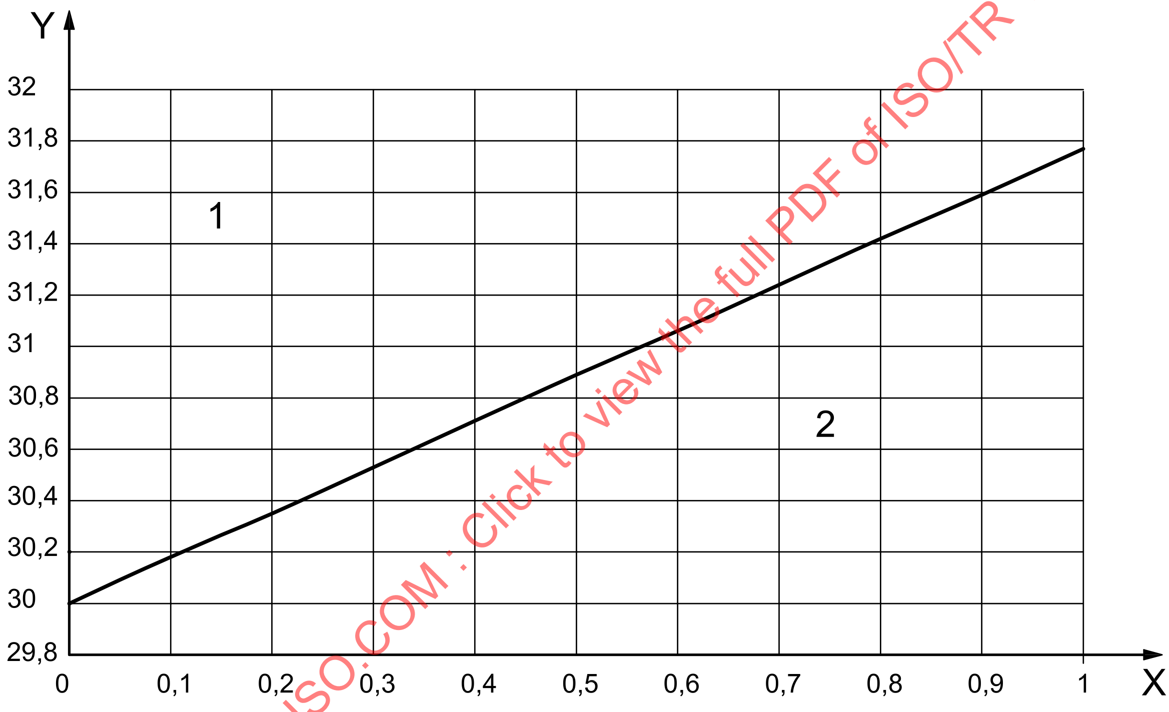
9.5.3 Single sampling plans by variables for unknown process standard deviation — The "*s*" method

When the process standard deviation cannot be presumed to be known, it is estimated by the sample standard deviation, s. The quality statistics become the following:

$$Q_L = \frac{\bar{x} - L}{s} \text{ and } Q_U = \frac{U - \bar{x}}{s}$$

The Form k acceptability constants become larger than for the "σ" method and the Form p^* acceptability constants become smaller, to allow for the increased uncertainty in the estimation of the process quality.

Consider first Form k. For a single specification limit, the acceptance criterion is similar to those for the "σ" method, i.e. the lot is acceptable if $Q \geqslant k$. Figure 46 shows an acceptance chart for a lower specification limit on a graph of $\bar{x}$ against s, for sample size code letter G on normal inspection (giving sample size 18) with an AQL of 1 % (giving $k = 1{,}77$). The accept zone is bounded by the line $\bar{x} = L + ks$, where the lower specification limit L has been taken to be 30 units.



Key

1 accept zone

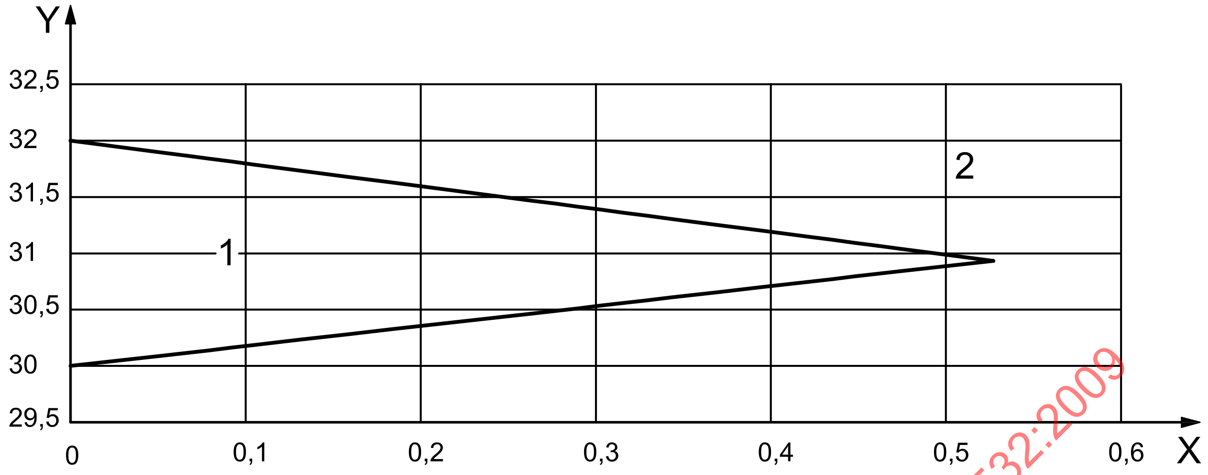
2 reject zone

X standard deviation s

Y sample mean $\bar{x}$

Figure 46 — Acceptance chart for a lower specification limit

For double specification limits, no check on σ can be carried out before inspection begins because σ is unknown. Nevertheless, an initial test may still be carried out on the process potential by comparing s with a maximum (allowable) sample standard deviation (MSSD). The MSSD is found by multiplying $(U - L)$ by a tabulated standardized value. For separate control of double specification limits, the lot is acceptable only if $Q_L \geqslant k_L$ and $Q_U \geqslant k_U$. An acceptance chart for separate control is shown in Figure 47; it can be seen in this case that the accept zone is bounded by two straight lines.



Key

1 accept zone

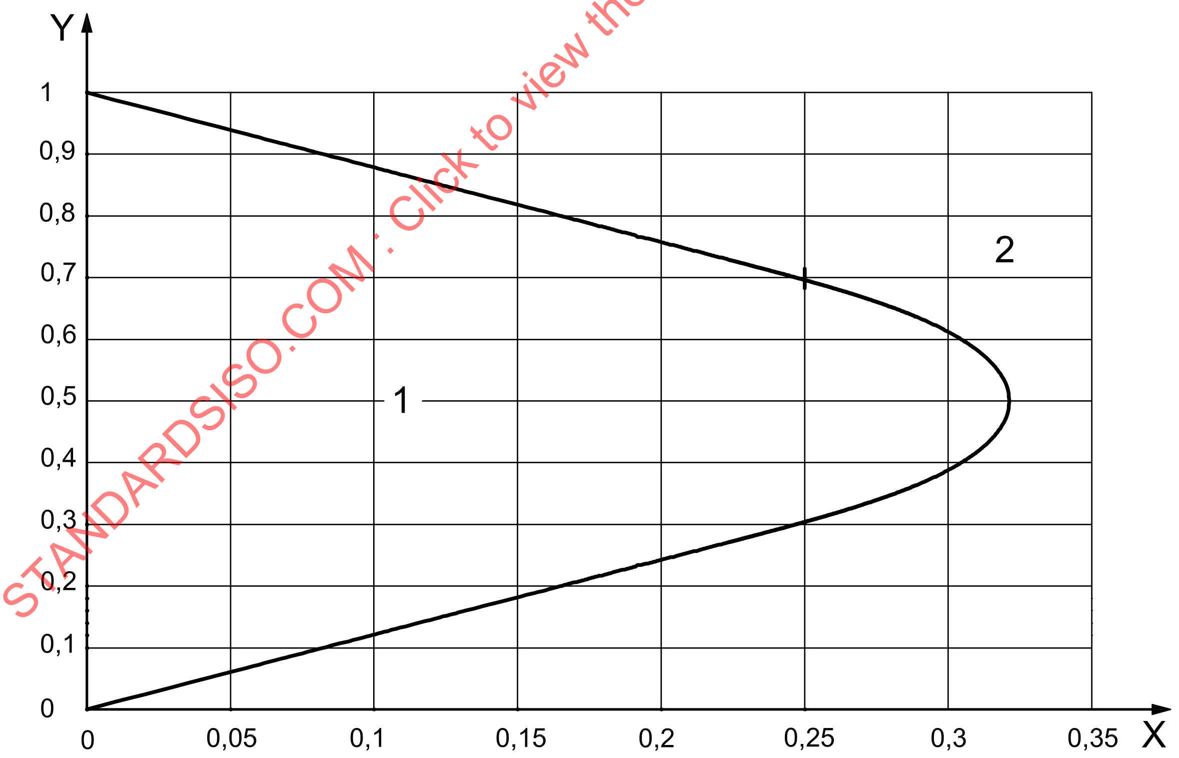
2 reject zone

X s

Y $\bar{x}$

Figure 47 — Acceptance charts for double specification limits with separate control

For combined control, the acceptability of the lot is determined by plotting the point with standardized co-ordinates $[s/(U-L), (\bar{x}-L)/(U-L)]$ on a standardized chart for the given sample size and AQL. The lot is accepted if the point lies within the acceptance region. Figure 48 shows such a chart for sample size 18 with an AQL of 4 %.



Key

1 accept zone

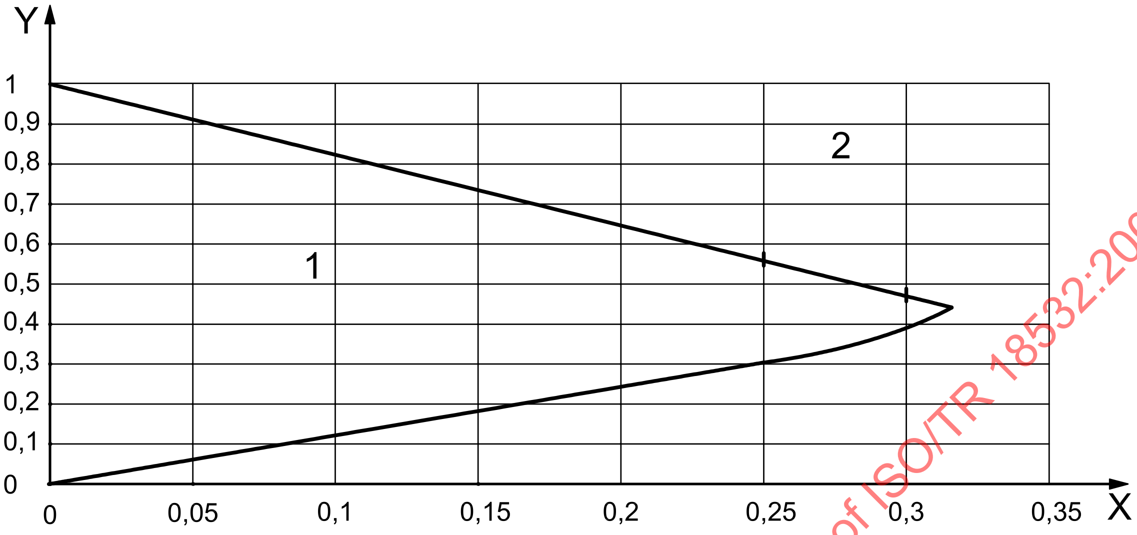
2 reject zone

X $s/(U-L)$

Y $(\bar{x}-L)/(U-L)$

Figure 48 — Standardized acceptance chart for sample size 18 for double specification limits with combined control at an AQL of 4 % under normal inspection

For complex control, lot acceptability is determined in the same way as for combined control except that part of the acceptance region is eliminated in accordance with the requirement on the single specification limit. Figure 49 shows the acceptance region for sample size code letter G on normal inspection for a 1 % AQL at the upper specification limit and a 4 % AQL overall.



Key

1 accept zone

2 reject zone

X $s/(U-L)$

Y $(\bar{x}-L)/(U-L)$

Figure 49 — Standardized acceptance chart for sample size 18 for double specification limits with combined control at an AQL of 1 % for the upper limit and an AQL of 4 % overall under normal inspection

For Form p^*, the procedures are identical to a) through e) of 9.5.2 except in one respect. For the "s" method, the process fraction nonconforming beyond a single specification limit is estimated by the area to the left of the value $\frac{1}{2}-\frac{1}{2}Q\sqrt{n}/(n-1)$ under a symmetrical *beta* distribution that has both parameters equal to $\frac{1}{2}(n-2)$. In order to avoid the need to use tables of the beta distribution to implement Form p^* plans, Baillie [66] has developed a normal approximation for use when $n > 4$, which works as follows.

a) Set $x=\frac{1}{2}-\frac{1}{2}Q\sqrt{n}/(n-1)$. If $x \leqslant 0$, then $\hat{p}=0$, or if $x \geqslant 1$, then $\hat{p}=1$; in both cases, no further steps are necessary. Otherwise, continue to step b).

b) Set $y=d_n \ln\left[x/(1-x)\right]$ where $d_n=\frac{1}{2}\sqrt{(n-3)\left\{1+\dfrac{1}{3\left[(n-3)^2+1\right]}\right\}}$

c) Set $w=y^2-3$

d) If $w>0$, set $t=\dfrac{y}{1+w/\left[12(n-1)\right]}$, otherwise $t=\dfrac{y}{1+w/\left[12(n-2)\right]}$

e) Then $\hat{p}$ is approximated by the area to the left of t under the standard normal curve [usually denoted $\Phi(t)$].

NOTE "ln" means natural logarithm (i.e. logarithm to base e).

This approximation is quite accurate, guaranteeing a maximum absolute error of not more than 0,000 4 for sample size 5, 0,000 2 for sample size 6 and 0,000 1 for sample sizes of 8 or more. Values of d_n for selected values of n are given in Table 20.

To illustrate, suppose that there is a single, lower specification limit $L = 32$ and a sample of size 18 has a mean $\bar{x}$ = 34,1 and a standard deviation $s = 0{,}93$. Then $Q = (\bar{x} - L)/s$ = (34,1 – 32)/0,93 = 2,258. The value of $x = \frac{1}{2} - \frac{1}{2}Q\sqrt{n}/(n-1)$ is found to be 0,218. From tables or a computer program, it may be found that the area to the left of 0,218 under a symmetric beta distribution with both parameters equal to $\frac{1}{2}(n-2) = 8$ is 0,007 4.

Table 20 — Values of d_n for estimating the process fraction nonconforming

Sample size	d_n	Sample size	d_n	Sample size	d_n
3	0,318 310	15	1,734 040	70	4,092 828
4	0,551 329	18	1,937 919	75	4,242 777
5	0,731 350	20	2,062 737	95	4,795 926
6	0,880 496	25	2,346 014	100	4,924 516
7	1,009 784	30	2,598 669	125	5,522 742
8	1,125 182	35	2,828 887	150	6,062 225
9	1,230 248	40	3,041 751	160	6,265 024
10	1,327 276	45	3,240 676	200	7,017 865
13	1,583 745	50	3,428 086	250	7,858 138

If neither the appropriate tables nor software are available, the normal approximation is found as follows, starting at step b). d_{18} is found from Table 20 to be 1,937 919. Then y is calculated as

$$y = 1{,}937\ 919\ 4 \ln(\frac{0{,}218}{0{,}782}) = -2{,}475\ 4$$

Then $w = y^2 - 3 = 3{,}128$. As $w > 0$, we set

$$t = y/\left\{1 + w/\left[12(n-1)\right]\right\} = 2{,}475\ 4/(1 + 3{,}128/204) = -2{,}438$$

From normal tables, it is found that the area under the normal curve to the left of $-2{,}438$ is 0,007 4. Thus, $\hat{p} = 0{,}007\ 4$, which is in complete agreement with the exact method.

9.5.4 Double sampling plans by variables

A double sampling plan by variables can be formulated in either Form k or Form p*. Consider for illustration the case of σ unknown for a single lower specification limit L. Form k will be described here. A double sampling plan by variables has five parameters, namely the two sample sizes n_1 and n_2 and the three acceptability constants k_1, k_2, and k_3. A sample of size n_1 is drawn, and the quality statistic $Q_1 = (\bar{x}_1 - L)/s_1$ is calculated. If $Q_1 \geqslant k_1$, the lot is accepted. If $Q_1 \leqslant k_2$, the lot is non-accepted. If $k_2 < Q_1 < k_1$, another sample, this time of size n_2, is drawn and its mean $\bar{x}_2$ and standard deviation s_2 are calculated. The combined mean is calculated as:

$$\bar{x}_\mathrm{c} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

and the combined standard deviation as:

$$s_c = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

The combined quality statistic is calculated as:

$$Q_c = (\bar{x}_c - L)/s_c$$

If $Q_c \geqslant k_c$, the lot is accepted; otherwise, it is non-accepted.

The procedure for the "σ" method is similar except that $Q_1 = (\bar{x}_1 - L)/\sigma$ and $Q_c = (\bar{x}_c - L)/\sigma$.

Hamaker [99] investigated the matching of σ-known double sampling plans by variables to σ-known single sampling plans. He observed that there were three requirements that are partly contradictory:

a) a reasonably close match between the OC curves;

b) a worthwhile reduction in the average sample number;

c) a low frequency of second sample (FSS);

and developed rules that provided a sensible balance between them. Baillie [69] extended Hamaker's results for $n_1 = n_2$ to the case of unknown σ. ISO 3951-3 [15] provides double sampling plans by variables.

9.5.5 Sequential sampling plans by variables for known process standard deviation

ISO 8423 [31] provides curtailed sequential sampling plans for inspection by variables when the process standard deviation is known, for single and for double specification limits. The acceptance chart for a single specification limit is similar in appearance to the attributes sequential chart (see Figure 45), the difference being that the cumulative sum of the *leeway* is plotted on the vertical scale. (The leeway is defined as $U - x$ for an upper specification limit and $x - L$ for a lower specification limit, where x is the measured value of the variable.) Thus, the increments on the vertical scale are not constrained to be integers, and can even be negative if x lies outside specification.

It was observed in 9.4.5 that the Wald approximation has been found to be poor for sampling by attributes when the supplier's and customer's risks are of the order of 5 % and 10 %. The same has been found to be true for sampling by variables so, like ISO 8422, ISO 8423 was revised to provide a better match with the corresponding single sampling plans.

9.5.6 Accept-zero plans by variables

Accept-zero plans provide for lot acceptance if there are no nonconforming items in the sample. Denoting the smallest and largest observations in a sample by $x_{[1]}$ and $x_{[n]}$, this requires $x_{[1]} \geqslant L$ for a lower specification limit, $x_{[n]} \leqslant U$ for an upper specification limit, and both inequalities are to be satisfied in the case of double specification limits.

Klaassen's [105] credit-based method of guaranteeing an AOQL for accept-zero plans for sampling by attributes was described in 9.4.11. Effectively, it provides a switching rule between sample sizes in response to perceived quality history. Baillie and Klaassen [72] have generalized this result to the case of guaranteeing an AOQL for any acceptance sampling plan that includes an accept-zero requirement, and applied the general method to the following three cases, with c any positive constant:

a) $x_{[1]} \geqslant L$ (for sampling by attributes);

b) $x_{[1]} \geqslant L + c\sigma$ (for sampling by variables with known σ); and

c) $x_{[1]} \geqslant L + cs$ (for sampling by variables with unknown σ).

Again, as with sampling by attributes, the AOQL guarantee requires lots that are rejected when the credit is zero to be 100 % inspected, with acceptance of all conforming items found in such lots. As expected, sampling by variables when the value of σ is presumed known requires smaller sample sizes than sampling by variables when the value of σ is unknown, which in turn requires smaller sample sizes than sampling by attributes.

9.6 Multiple quality characteristics

9.6.1 Classification of quality characteristics

Most products have more than one quality characteristic, all of which need to conform to specification if an item is to be classed as conforming. Some of these characteristics may be of greater importance, and may therefore need to be controlled more tightly. This is achieved by classifying the quality characteristics into class A for those of the highest level of importance, class B for the next level of importance, etc., and applying a low AQL to class A, a larger AQL to class B, etc. Sampling inspection schemes are then applied to the classes independently; for example, it would be possible for classes A and C to be on normal inspection while class B is on tightened inspection. The acceptance criteria for all classes have to be satisfied for a lot to be classified as acceptable. The following discussion of multiple quality characteristics is on the treatment of a single class of quality characteristics, where by definition all the quality characteristics in the class are of approximately equal importance to the integrity of the product.

9.6.2 Unifying theme

As before, the unifying theme of the discussion for nonconforming items will be the comparison of $\hat{p}$, the estimated process fraction nonconforming from the sample, with p^*, a specified maximum value. As a rough rule, we can set:

$$p^* = (a + \tfrac{1}{2})/n \tag{34}$$

where the reference single sampling plan by attributes has sample size n and acceptance number $\text{Ac} = a$.

9.6.3 Inspection by attributes for nonconforming items

9.6.3.1 Independent attributes

Consider first the simplest case where there are k quality characteristics, all of which are attributes. First suppose that the attributes are independent, i.e. the probability of any one of the attributes in the class being out of specification is constant, regardless of the state of any of the other attributes in the class. Suppose also that in a sample of size n it is found that there are r_1 items that are nonconforming on attribute 1, r_2 items that are nonconforming on attribute 2,…, r_k items that are nonconforming on attribute k. The estimate of the probability of *conformance* on the ith attribute is estimated by $(1 - r_i / n)$. As the attributes are independent, the estimated overall probability of an item conforming to all the specifications is the product of such estimated probabilities, *viz.* $(1 - r_1 / n)(1 - r_2 / n)\cdots(1 - r_k / n)$. Subtracting this from 1, it is seen that the overall probability of an item *not* conforming to at least one of the specifications is estimated by:

$$\hat{p} = 1-(1-r_1/n)(1-r_2/n)\cdots(1-r_k/n)$$

The acceptance criterion $\hat{p} \leqslant p^*$ therefore becomes:

$$1-(1-r_1/n)(1-r_2/n)\cdots(1-r_k/n) \leqslant p^*$$

Provided all the fractions r_i/n are small, it can be shown by expanding the product term that this inequality is approximately the same as:

$$(r_1+r_2+\cdots+r_k) \leqslant np^* \text{ i.e. } r \leqslant c$$

where r is the total number of items that are out of specification with respect to each attribute, summed over attributes, and c is the largest whole number less than or equal to np^*.

9.6.3.2 Dependent attributes

Now suppose that the attributes are dependent. The estimate of the process fraction nonconforming is $\hat{p} = d/n$, where d is the number of nonconforming items in the sample. The acceptance criterion $\hat{p} \leqslant p^*$ then becomes $d \leqslant np^*$ i.e. $d \leqslant c$.

9.6.3.3 Example

To illustrate the difference between the independent and dependent cases, suppose that a single sampling plan under normal inspection is to be used, with a sample size code letter F and an AQL of 4 %. From Table 2-A of ISO 2859-1:1999, it is found that the sampling plan is $n = 20$, Ac $= 2$. From Equation (34):

$$p^* = (2+\tfrac{1}{2})/20 = 0{,}125$$

Suppose that an item has two quality characteristics that are both attributes. A sample of size 20 yields one item that is nonconforming on both attributes and one item that is nonconforming on one attribute. Assuming independence between the attributes, the estimate of the process fraction nonconforming would be:

$$\hat{p} = 1-(1-2/20)(1-1/20) = 0{,}145$$

On the other hand, assuming dependence, there are only two nonconforming items in the sample of size 20, so the estimate of the process fraction nonconforming would be:

$$\hat{p} = 2/20 = 0{,}100$$

As $p^* = 0{,}125$, we see that the lot would be non-accepted if the attributes were considered to be independent, but accepted if they were considered to be dependent. On reflection, this is not a surprising result. On the evidence from the example, when the two attributes are dependent there seems to be a tendency for both attributes to be out of specification at the same time. Treating the nonconformities as independent in such a situation leads to some double counting.

9.6.4 Inspection by attributes for nonconformities

Suppose that the sample contains a total of r_1 nonconformities on attribute 1, r_2 nonconformities on attribute 2,…, r_k nonconformities on attribute k. The rate of nonconformity on the ith attribute is estimated by r_i/n. If the attributes are independent, these estimated rates are added to give an estimated overall rate of nonconformity per item of r/n, where $r = r_1+r_2+\cdots+r_k$ is the total number of nonconformities in the sample.

On the other hand, if the attributes are dependent, the estimated rate is simply the total number of nonconformities divided by the sample size, i.e. r/n again. It follows that the multi-attribute acceptance criterion is $r \leqslant$ Ac; this is regardless of the number of attributes and whether or not they are independent.

9.6.5 Independent variables

The principle for k independent variables is the same as for k independent attributes inspected for nonconforming items. The process fraction nonconforming is estimated as:

$$\hat{p} = 1 - (1 - \hat{p}_1)(1 - \hat{p}_2)\cdots(1 - \hat{p}_k)$$

where $\hat{p}_i$ is calculated as explained in 9.5.2 for σ known and in 9.5.3 for σ unknown.

Form p* plans by variables are presented in ISO 3951-2 [14].

9.6.6 Dependent variables

For dependent variables, it is theoretically possible to carry out acceptance sampling, but impracticable without the use of suitable software, as the formula for $\hat{p}$ is a multidimensional integral over a complicated region. For further information, the reader is referred to Baillie [70]. If the correlation between the variables is not strong, the variables may be treated as independent without much danger of reaching the wrong decision on lot acceptability if the decision is not marginal. If the correlation between the variables is strong, then the variables can be converted to dependent attributes, and treated as described in 9.6.3.2, although this is an inefficient use of the data.

9.6.7 Attributes and variables

Baillie [67] presented master tables of “attriables” plans and procedures for use when the quality characteristics in a class consist of at least one attribute and at least one variable. The plans are only suitable when it is practicable to have a larger sample size for the attributes than for the variables. The implementation of the plans is necessarily complicated, and would need to be supported by suitable software, particularly if there are two or more dependent variables.

10 Statistical process control (SPC)

10.1 Process focus

The question to what extent it is possible to obtain from a sample a reliable estimate of the quality characteristics of product lots has been discussed from various points of view in the preceding clauses. It has been shown that if the variation among individual units is considerable, it may not be an economic proposition for the customer to sample and test sufficient items to provide the desired degree of assurance regarding the consignment.

Furthermore, what if the correct technical decision, on the basis of an “after the event” sample, is to “reject” the consignment? All too often, the correct business decision has to be “accept” because of logistics, time and other constraints. It is therefore inevitable that attention should be focused on ways of securing and demonstrating conformity to specification that involve the requirement that statistical methods be deployed at the place and time of the process activity giving rise to the product or service. This is recognized in the following circumstances.

a) **Generic quality management system requirements such as ISO 9001:2008**:

This International Standard recognizes that any activity that receives inputs and converts them to outputs can be considered as a process. For organizations to function effectively, they have to identify and manage numerous linked processes. Often the output from one process will directly form the input of the next process. ISO 9001:2008 is based on the “process approach” to management, which involves the systematic identification and management of the processes employed within an organization, and the interactions between such processes. The model shown in Figure 50 covers the requirements of ISO 9001:2008, but does not show the process at a detailed level.

NOTE The term product in the ISO 9000 family has four generic categories: hardware, software, services and processed materials.

More specifically, ISO 9001:2008, 8.1, states the following.

ISO 9001:2008, Quality management systems — Requirements

8 Measurement, analysis and improvement

8.1 General

The organization shall plan and implement the monitoring, measurement, analysis and improvement processes needed

a) to demonstrate conformity to product requirements,

b) to ensure conformity of the quality management system, and

c) to continually improve the effectiveness of the quality management system.

This shall include determination of applicable methods, including statistical techniques, and the extent of their use.

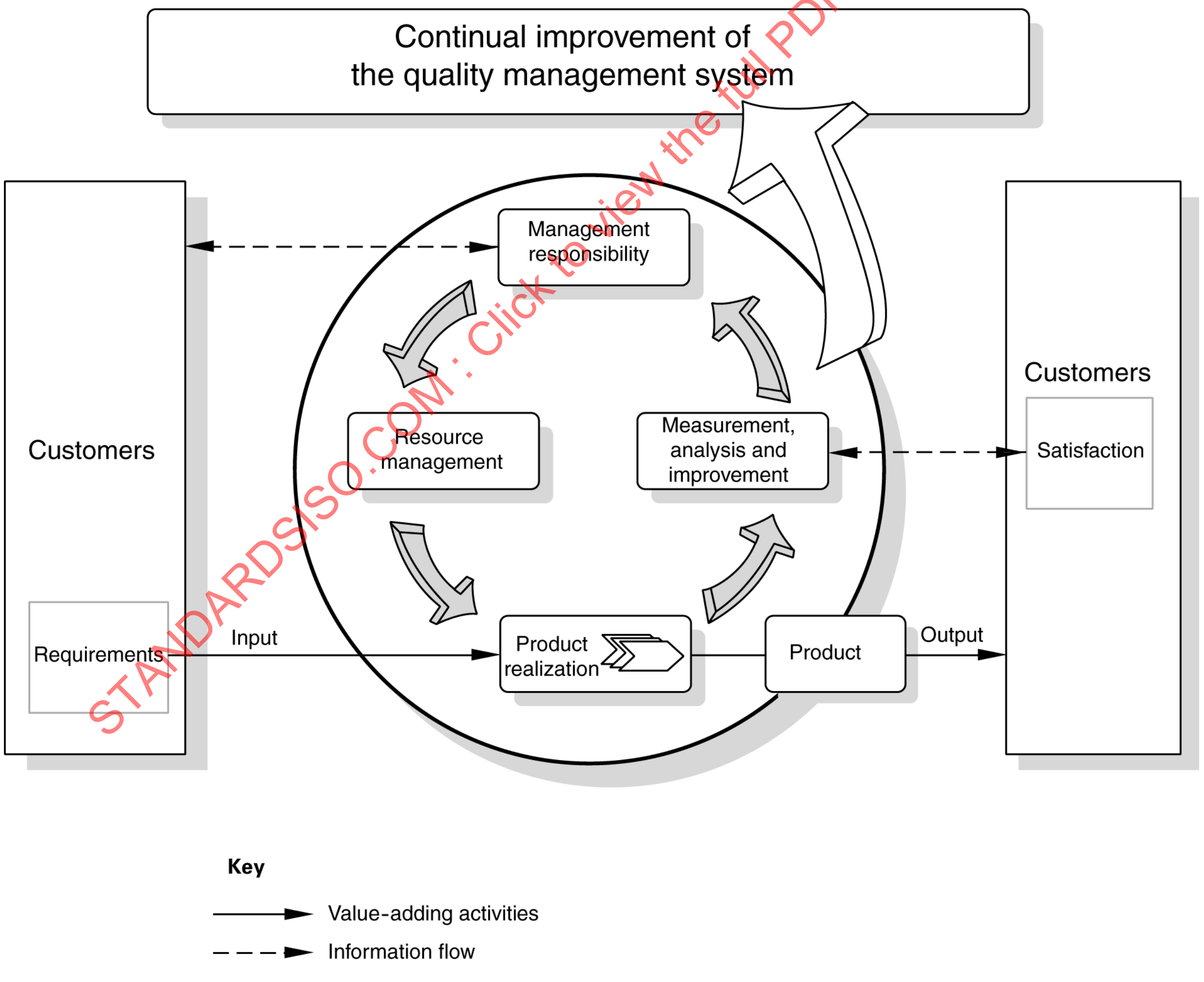


Figure 50 — ISO 9001:2008 — Model of a process-based quality management system

For example, ISO 9001:2008, 4.1, states the following.

> **ISO 9001:2008, Quality management systems — Requirements**
>
> ## 4 Quality management system
>
> ### 4.1 General requirements
>
> The organization shall establish, document, implement and maintain a quality management system and continually improve its effectiveness in accordance with the requirements of this International Standard.
>
> The organization shall
>
> a) determine the processes needed for the quality management system and their application throughout the organization (see 1.2),
>
> b) determine the sequence and interaction of these processes,
>
> c) determine criteria and methods needed to ensure that both the operation and control of these processes are effective,
>
> d) ensure the availability of resources and information necessary to support the operation and monitoring of these processes,
>
> e) monitor, measure where applicable, and analyse these processes, and
>
> f) implement actions necessary to achieve planned results and continual improvement of these processes.
>
> These processes shall be managed by the organization in accordance with the requirements of this International Standard.
>
> Where an organization chooses to outsource any process that affects product conformity to requirements, the organization shall ensure control over such processes. The type and extent of control to be applied to these outsourced processes shall be defined within the quality management system.

Explanatory notes are provided in ISO 9001:2008.

b) **Numerous sectors, such as medical devices, aerospace and automotive have prescriptive quality system requirements**.

Taking the automotive sector as an example, three major USA-based suppliers have jointly produced quality systems requirements, together with supporting documentation that includes manuals on: *statistical process control* (SPC) [79] and *measurement system analysis* (MSA) [89] that provide a unified formal approach to both SPC and MSA in the automotive industry.

Regardless of whether or not the application of statistical process control is explicitly stated in system or product requirements, the organization dedicated to “never ending improvement” or aiming for “world class” will recognize its key role in improving business performance. This is illustrated by an example from the aircraft supply industry.

EXAMPLE Steel tube dimensions:

A steel tube supplier to the aircraft industry buys steel strip from the steel-maker by the kilogram and converts strip into tube to sell by the metre. This organization recognized that, by managing variation better, more metres could be produced per kilogram of strip. It aimed for preferred minimal values for outside diameter and wall thickness commensurate with the need to maintain dimensional specification requirements. This aim decreased as they identified and progressively reduced variation using statistical process control methodology. Controlling the new minimum size and its variation using statistical process control, it was then able to produce a lighter, more consistent product with a saving of some $ 400 000 per year.

This example demonstrates the ability of statistical process control to both increase profits for the user organization and value to its customers.

These considerations, amongst others, have given rise to a growth in the development and widespread application of statistical process control methods.

10.2 Essence of SPC

The primary operational tool of SPC is the control chart. The first question to be answered is: what is to be its basic purpose? The reason for this is that there are two fundamentally different approaches to control charting. One approach aims at directly controlling to specification using control limits based upon, and set *inwards* from, specified tolerance limits. Such an approach is described in ISO 7966:1993 [28]. The other approach uses control limits based entirely on process performance. These control limits are set *outwards* from the mean value of the characteristic plotted to an extent based on the inherent variability of the process with no regard to specified tolerance limits. Such an approach is described in ISO 8258:1991[29].

Many organizations using tolerance-based control charts have had to abandon their use in favour of performance-based statistical process control, for contractual reasons, to meet customer requirements expressed in current quality systems standards. Others have chosen to adopt performance-based control charts, for a number of reasons, such as the following.

a) The recognition that first class quality for a characteristic is achieved only by realizing a preferred value and that there is a progressive deterioration in quality as one moves away from this value towards a specification limit, even though one may, technically, still be “in tolerance”. Meeting tolerance then becomes a “minimum” standard that may just be tolerated. It is not a standard of excellence. In a competitive climate, there can be considerable advantage in aiming for a preferred value with minimum variation.

b) The acknowledgement that a tolerance-based control chart does not provide information on the sources of variation in the process essential for control and improvement purposes. The focus on the classification of the process output purely in terms of specification is in direct contrast to the focus of the performance-based control chart, which is on the discrimination between common and special cause variation in the process.

c) An appreciation that there are two kinds of people and two types of variation:

   1) *technical and managerial people* who are responsible for the process and for the presence of *inherent/common cause* variation and its reduction;

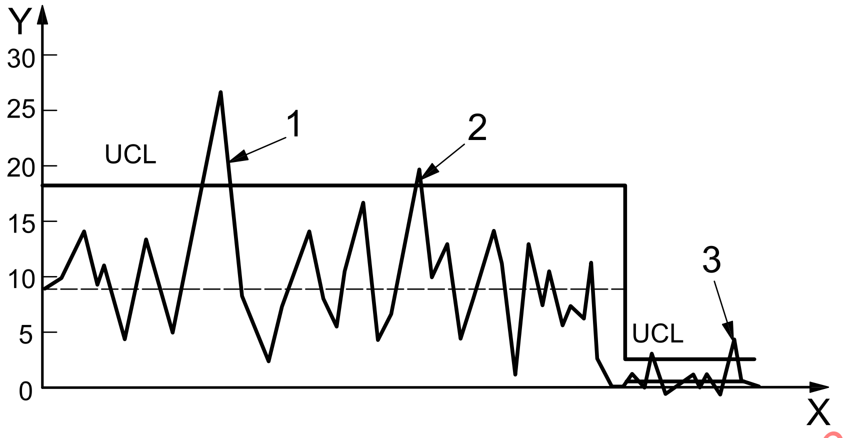
   2) *operational people* who work in the process who can best observe and report on *special cause* variation through the use of performance-based control charts.

Point b) recognizes that the primary operational role of a control chart is to discriminate between special and common cause variation. Common cause variation is generally outside the remit of people who work in the process. Suppose that *operational people* have established, using a performance-based control chart, that a process is in statistical control, namely, that no special causes of variation are present. Then, and only then, *technical and managerial people* can use the control chart data to compare the magnitude of the residual common cause variation present with specified limits, using capability analyses as described in Clause 11. Standardized quality capability indices for the characteristic may then be generated and any necessary improvement actions initiated.

Such process capability analysis brings out another very important aspect of the overall role of SPC. A primary role of a control chart, in an operational sense, by its very name, is to control; namely to inhibit change. The removal of special cause variation to bring a process back into control does not actually improve the process, it only returns it to its original state. This, however should not blind one to the fact that often the objective of SPC, in an overall sense, is to improve process performance by inducing change.

Such betterment, through common cause reduction does not have to await special cause removal. A significant improvement in process performance is evidenced in a control chart by an “out of control” situation, as is a significant deterioration. Hence, the control chart has a built-in statistical test of significance.

These features are demonstrated in Figure 51 for an underwear making-up process.



Key

1 crotch stitch (broken needle)
2 shop reorganization
3 oil leak

X production sequence of men's briefs
Y fault rate percent

Figure 51 — Control chart for nonconforming underwear

Figure 51 shows:

- undesirable "out-of-control" situations (points above the UCLs ± upper control limits) due to special causes: a broken needle, a sewing shop reorganization and oil leak;
- a desirable "out-of-control" situation (more than 9 consecutive points below the original centre line) due to a management-led major training and personal development initiative that gave rise to a reduction in common cause variation from a nearly 10 % fault rate to less than 1 %.

This example brings out why it is important to differentiate between special and common cause variation. The sporadic special cause variation is due to specific assignable activities attributable to a machinist and direct support personnel. The overall performance of nearly 10 % fault rate, however, is a result of common causes endemic in the system, which is a management responsibility. Without this perspective, using a control chart, it is usual, in such a situation, not to consider the impact of common causes on the performance of individuals.

10.3 Statistical process control or statistical product control?

At this stage it is necessary, too, to distinguish between statistical process control and statistical product control. Much so-called statistical process control is, strictly speaking, after the event statistical product control. Figure 52 illustrates this point.

In such a process, SPC is quite frequently applied to the product characteristics such as, image density and curl. Superficially, from the standpoint of the customer, this may appear quite acceptable. However, it is clearly not nearly as effective as control of the process parameters and process inputs that affect these product characteristics. After the event, detection of unsatisfactory product may give rise to delays in shipment and increased production cost which, in turn, results in a decreased profit margin for the supplier and/or increased price to the customer.

Why then is true process control not practiced more often? The primary reason is that for a large proportion of processes the technical relationship between process parameters/process inputs and product characteristics is not known. This is why the prior application of statistical experimental designs (commonly termed DoE: design of experiments) often leads to a more purposeful and effective application of SPC. This aspect is dealt with in Clause 12.

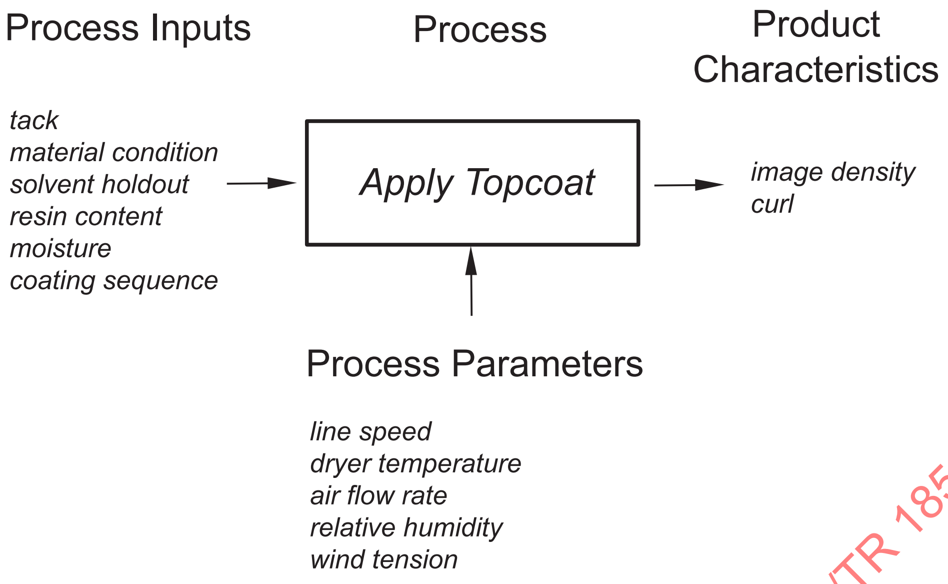


Figure 52 — Outline of process of applying a topcoat to a photographic film

This shows that it is necessary to realize that:

a) every process generates information that can be used to control and improve its performance;

b) there is a need to develop informed, perceptive observers using appropriate statistical methodology;

c) there are two sources of information and two primary statistical tools for dealing with them:

1) natural variation: use SPC, a listening tool;

2) induced variation: use DoE, a conversational tool.

10.4 Over-control, under-control and control of processes

10.4.1 General

A process monitoring system may give rise to the following situations:

a) over-control: action is taken when it should not be;

b) under-control: action is not taken when it should be;

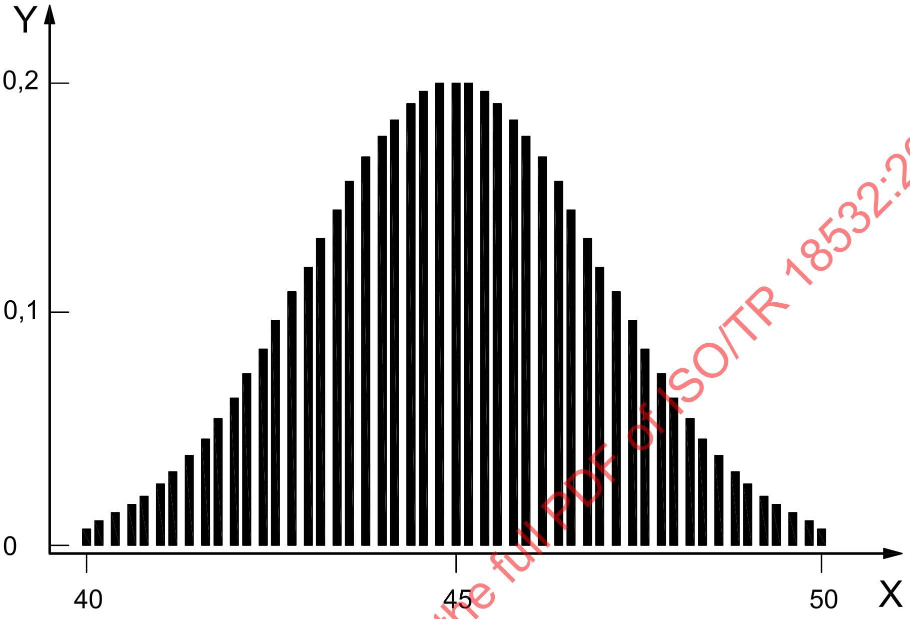
c) control: action is taken when it should be and not taken when it should not be. A process is said to be under a state of (statistical) control when no special causes of variation are present. Variation can then be attributed purely to “common causes”. Control is not a natural state but it is an achievement, arrived at by elimination, one by one, by determined effort, of special cause variation. To achieve that, it is essential to use SPC charts that set out to provide a signal when a special cause of variation is present, and to avoid giving false signals when a special cause is not present.

Sometimes “assignable cause” is taken to be synonymous with “special cause”. However, a distinction should be recognized. In practice, not all special causes are assignable. A state of control does not imply that the common cause variation is large or small, within or outside of specification, but rather that it is predictable using statistical techniques.

10.4.2 Scenario 1: Operator reacts to each individual sample giving rise to process over-control

Suppose a particular preform extrusion process has a stable variation about the target mass value of 45 as shown in Figure 53.

The operator takes one measurement at intervals and decides, from each particular observation, whether or not to adjust.



Key

X value

Y odds

NOTE Pattern of process variation around stable mean of 45 shows chance of obtaining a particular value with a single sample.

Figure 53 — Probability of setter/operator observing a single mass value when mean = 45

The setter/operator takes one mass measurement at 20-minute intervals and compares the result with the preferred, target or reference value of 45. Mass is controlled by adjustment of the speed feed. Adjustment is in steps of 1 so an appropriate adjustment is made if the result differs from 45 by 1 or more. Table 21 shows what may be expected in a process whose *actual* level is initially at the preferred level and which is also stable throughout with respect to variation about the various actual process levels experienced. A typical result from this monitoring plan is shown in Table 21.

Table 21 — Control plan (take one measurement at intervals and adjust or do not adjust)

Time	Measurement value	Subsequent adjustment made	Actual process level
08:00	46	–1	45
08:20	42	+3	44
08:40	46	–1	47
09:00	48	–3	46
09:20	44	+1	43
09:40	43	+2	44
10:00	47	–2	46
10:20	44	+1	44
10:40	47	–2	45
11:00 etc.	44	+1	43
NOTE Measured values were obtained by taking values at random from a process with constant variation about actual process levels. This simulates a real-life situation.			

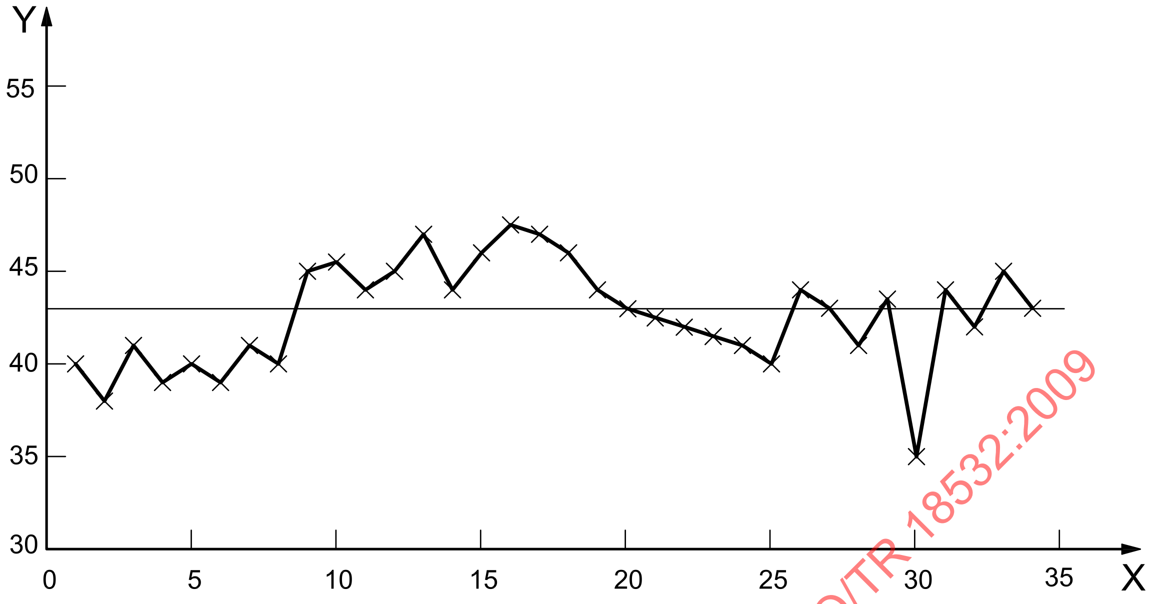
At 08:00, the setter/operator sees 46 and increases feed speed to decrease mass, thus over-controlling and bringing the actual mean mass down to 44. At 08:20, the setter/operator measures 42 and decreases feed speed to increase mass by 3 units. The mass then overshoots to a mean of 47; again over-control. And so on.

The consequence of this monitoring plan is to increase overall variation from 10 units of mass (45 ± 5, see Figure 54) to 17 units of mass [(43 – 5) to (47 + 5)] in Table 21.

This is an example of *process over-control*. Here, the penalty of over-adjustment is some 40 % increase in variation over the short time period considered. The general conclusion is that continual adjustment of a stable process will increase variation.

10.4.3 Scenario 2: Operator monitors a process using a run chart giving rise to haphazard control

Suppose a process is being monitored using a run chart as shown in Figure 54. A *run chart* is a graph that displays observed data in a time sequence. Whether or not reaction is made to changes in the results monitored will depend solely on the operators. Control is thus not likely to be effective. Under-control and/or over-control could thus be expected to arise. Control here is likely to be inconsistent and capricious as no guidance is given on how to interpret the variability.



Key

X run number

Y length

Figure 54 — Example of process run chart with variation, but with no guidance on how to interpret and deal with variation

10.4.4 Scenario 3: Monitoring using SPC chart with a potential for effective control

Here, *under-control* is the result if improper use is made of the control chart such as:

- "out-of-control" signals are not reacted to, as they arise, and a completed SPC chart is analysed purely on a retrospective basis;
- the data used for plotting do not represent process reality; for example, data is selected to make the process "look good".

Figure 55 shows an example of the use of an SPC chart with the data of Figure 54. Four "out-of-control" situations are flagged on the chart. If these are reacted to positively at the time they are signalled, then the process will be effectively controlled.

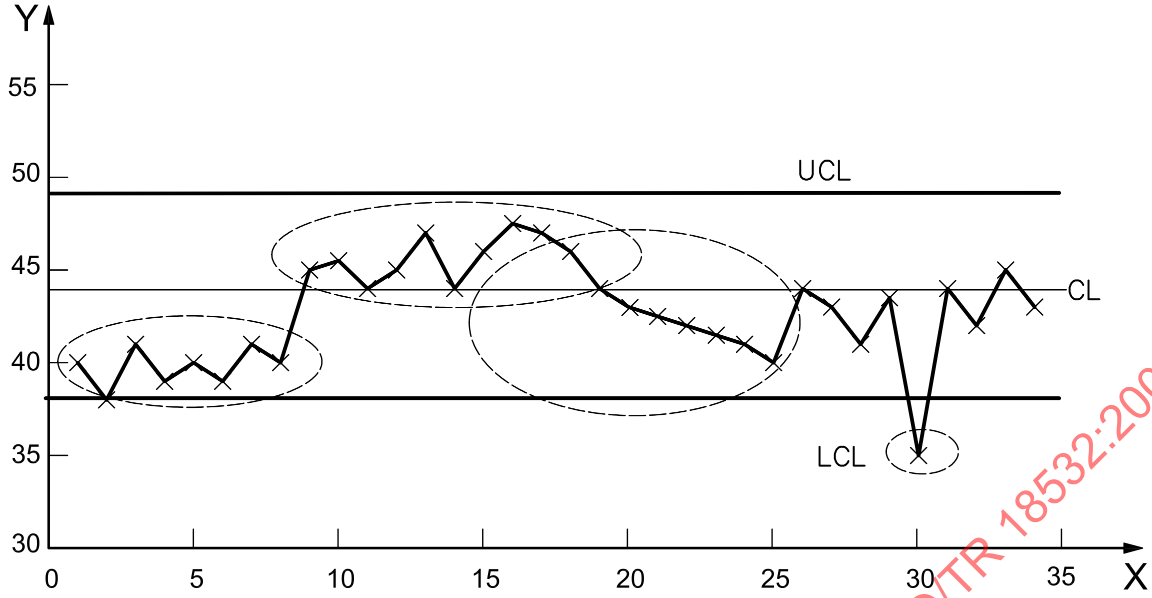
Typical criteria for "out-of-control" include the following:

1) any point outside of the control limits (upper and lower: UCL and LCL);

2) any run of 9 consecutive points above or below the centre-line (CL);

3) any run of 7 consecutive points up or down;

4) any obvious non-random patterns (based on technical and operational knowledge of the process).



Key

X sample number

Y length

Figure 55 — Example of process control chart with criteria for "out-of-control" signals

10.5 Key statistical steps in establishing a standard performance-based control chart

10.5.1 General

Having identified the process parameter or product characteristic to be observed, it is first necessary to decide on a monitoring strategy, (such as how to constitute a sample or subgroup, how many observations to take, and how frequently) followed by the setting up and interpreting of the control chart (such as how to set control limits and establish out-of-control criteria). These are now discussed.

10.5.2 Monitoring strategy

10.5.2.1 Subgroup constitution

In constituting a subgroup (see 4.2.4 and 4.2.6), a number of factors need consideration.

The concept of subgrouping is that the variation *within* a subgroup is made up only of common causes, with all special causes of variation occurring *between* subgroups. As the primary role of a control chart is to distinguish between common and special cause variation, the choice of rational subgroup has a considerable bearing on the usefulness of a control chart for a given purpose. For instance, if a subgroup is made up of, say, the diameter of three consecutive parts on a high-precision honing operation, the common cause variation within the subgroup may be miniscule. However, if the subgroup is made up of three parts, each of which is selected from consecutive wheel dressings, the common cause variation will be much greater. There will be far less homogeneity in the subgroup. This will have considerable impact on control limits. Hence the constitution of a subgroup will depend on the primary purpose of the control chart and a thorough knowledge of the process.

Frequently, the term rational subgroup is used. This highlights the need for further care in subgroup constitution. Consider a multi-headed machine that is to be sampled at the rate of 1 per 15 min to make up a subgroup of four. It would not be rational to take one measurement on head 1 at 08:00, one from head 2 at 08:15, one from head 1 at 08:30 and one from head 3 at 08:45 as it would be difficult to separate out within-head, between-heads and between-times variation.

Summarizing, the basic mean ($\bar{X}$) and range (R) control chart can be looked upon as a two-factor nested experimental design that separates out within-subgroup (common cause) variation from between-subgroup (special cause) variation. This is shown diagrammatically in Figure 56.

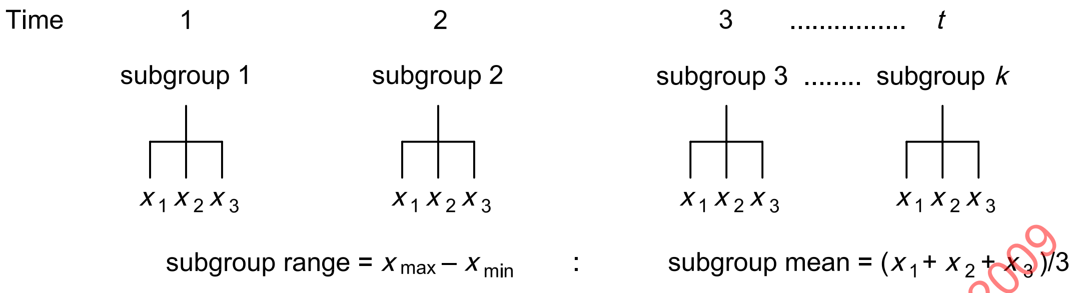

Time
1
2
3 .... t
subgroup 1
subgroup 2
subgroup 3 .... subgroup k
$x_1\,x_2\,x_3$
$x_1\,x_2\,x_3$
$x_1\,x_2\,x_3$
$x_1\,x_2\,x_3$
subgroup range = $x_{\max} - x_{\min}$: subgroup mean = $(x_1 + x_2 + x_3)/3$

Figure 56 — A two-factor nested design is the basis of an $\bar{X}\,R$ chart (illustrated with a subgroup size of 3)

The mean and range for each subgroup are calculated. These are then plotted in time sequence. The R chart evaluates the variation *within* a subgroup. The $\bar{X}$ chart assesses the variation *between* subgroups.

It is often said that the measurements in a subgroup should be independent of one another. However, in practice, this is frequently not achieved in a real-life process. A measurement in a subgroup is often influenced by another to some degree. Hence data for control charts often exhibit serial correlation (autocorrelation). What impact does this have? A consensus view is that:

a) for most situations “significant” autocorrelation will have minimal impact upon control chart limits;

b) whilst severe autocorrelation may contaminate the control limits, the control chart may usually be safely interpreted at face value.

This indicates that one need not be overly concerned about the effects of autocorrelation on control chart interpretation in most situations. Hence, recourse to complex techniques, such as the use of variograms and correlograms, to distinguish between random, cyclic, trend and correlated variations, as expounded in ISO 11648 is usually not required.

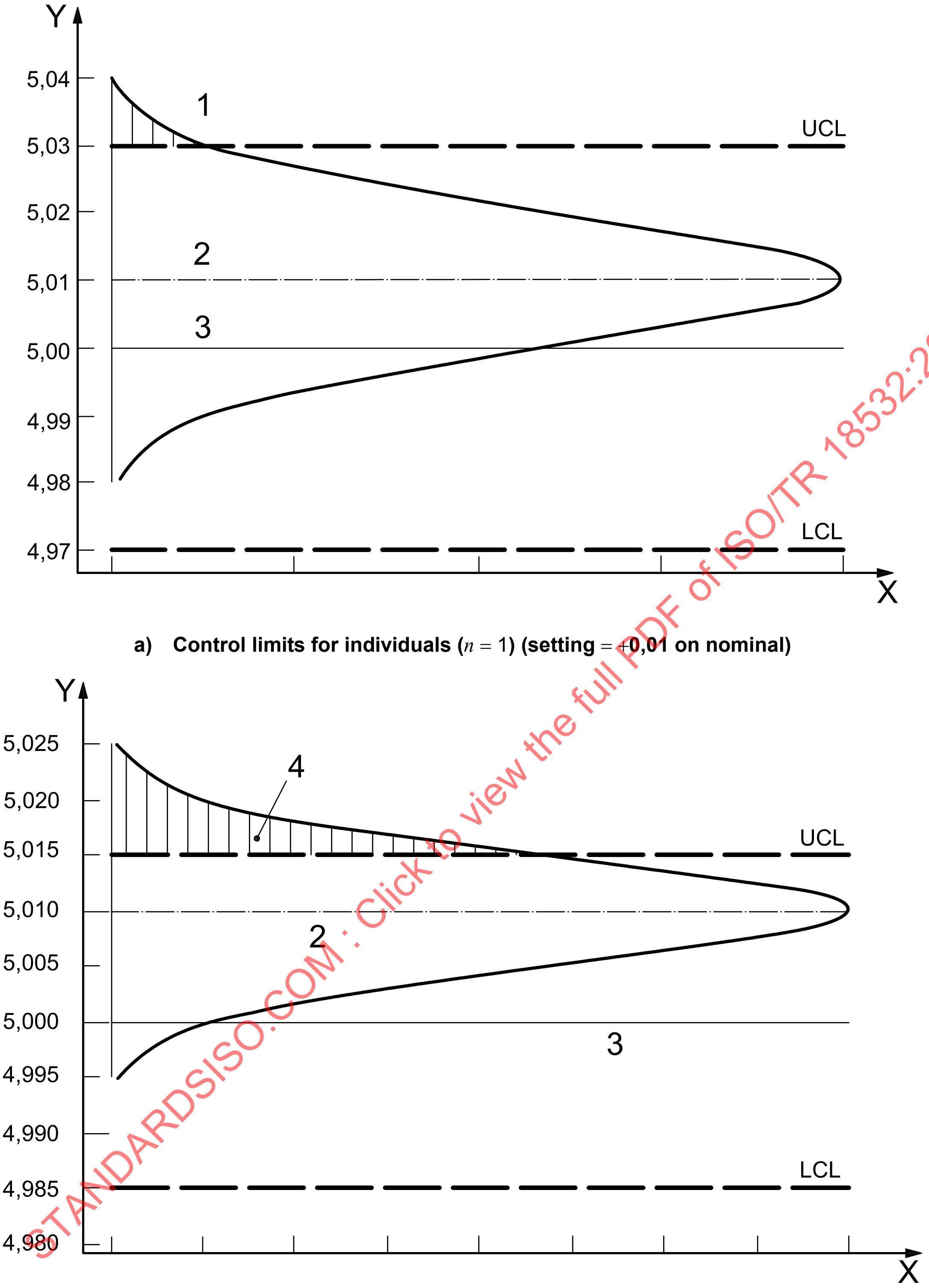
10.5.2.2 Subgroup size

Sometimes the sample or *subgroup size* may be dictated by circumstances. If the measurement or test is destructive or expensive, or on a process parameter, such as curing time or flow rate, the subgroup size may be necessarily small, for example, $n = 1$ or $n = 2$. However, larger subgroups have certain technical advantages:

a) even if the individual measurements are not normally distributed, the distribution of the mean of the subgroups tends to normality as the subgroup size increases (central limit theorem); a sample size of 5 is generally adequate to achieve this;

b) the larger the subgroup size, the greater the ability of the control chart to detect changes in the mean.

This is depicted graphically in Figure 57.



a) **Control limits for individuals (*n* = 1) (setting = +0,01 on nominal)**

Key

1 2 % above UCL

2 set at 5,01

3 nominal value

4 16 % above UCL

X probability density

Y measurement

b) **Control limits for averages (*n* = 4) (setting = +0,01 on nominal)**

Figure 57 — Effect of subgroup size on ability to detect changes in process mean (process nominal = 5,00, process standard deviation = 0,01)

Figure 57a) shows that with a subgroup size of $n = 1$, a shift in mean of 0,01 will only be expected to be detected some 2 % of the time if the control limits are the only criteria for control. By contrast, Figure 57b) shows that if the subgroup size is increased to $n = 4$, the same shift in mean is expected to be detected almost 16 % of the time.

This is because the variation of means is less than the variation of individuals according to the following relationship:

$$\text{Standard deviation of mean} = \frac{\text{Standard deviation of individual}}{\sqrt{\text{Subgroup size}}}$$

10.5.2.3 Frequency of sampling

The frequency of sampling is a compromise between sampling cost and value of the timely detection of process changes. A useful guide is to consider taking about six subgroups between anticipated changes in a process.

10.5.3 Construction of a standard control chart

10.5.3.1 Common features

The generic control chart for both measured data and attributes (count and classified data) shares similar features. Typically it consists essentially of five lines and a series of plotted points:

a) a vertical scale of values of a chosen statistic, X, say (e.g. mean, range, standard deviation, number of nonconformities) of the subject characteristic;

b) a horizontal scale depicting subgroup sequence numbers;

c) a centre-line (CL) , where CL = mean of $X = \bar{X}$;

d) an upper control limit: $\text{UCL}_\text{S} = \bar{X} + 3s_\text{S}$ (where s_S = standard deviation of the statistic plotted);

e) a lower control limit: $\text{LCL}_\text{S} = \bar{X} - 3s_\text{S}$;

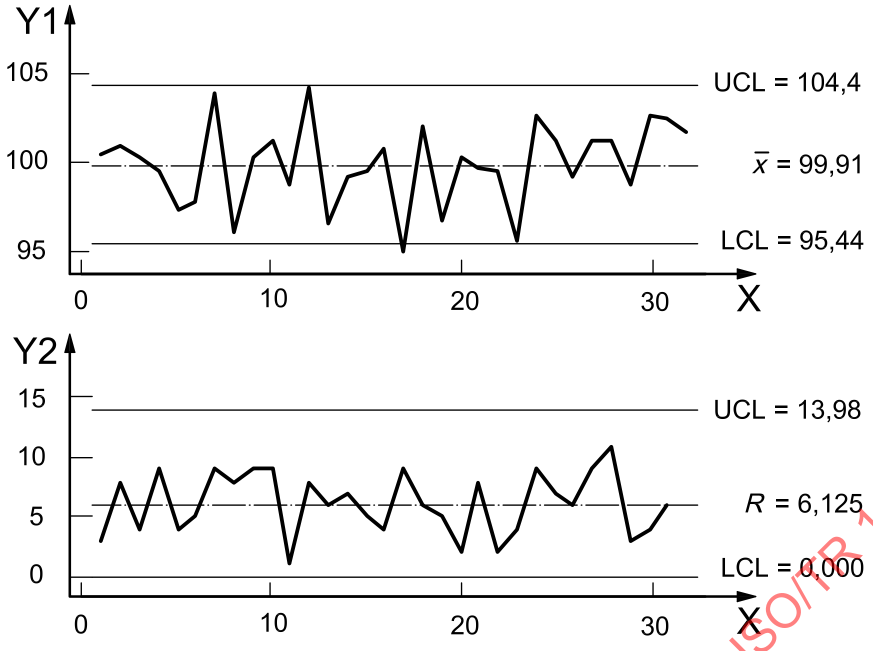
f) plotted points representing the calculated values of the statistic, X, of rational subgroups sequentially formed from measurements of the chosen characteristic.

Standard formulae and tabled constants are available for the calculation of standard limits. These are given in Annex A.

10.5.3.2 Example of typical mean and range control chart for measured data

Unlike attribute charts that are formed from a single statistic (see Figure 51), standard control charts for measured data are made up of two statistics; the mean or median, to monitor changes in the level of a characteristic between subgroups; and the standard deviation or range, to monitor variability within a subgroup.

An example of a mean and range ($\bar{X}$ and R) control chart for measured data is shown in Figure 58 for the masses of specimens of fabric given in Table 2.



Key

X sample number
Y1 sample mean, $\bar{x}$
Y2 range, R

Figure 58 — Mean and range chart for masses of standard specimens of fabric

The calculations required for such a chart are as follows:

— first plotting point of Xbar chart is given by: $\bar{x}_1 = \frac{(101+99+100+102)}{4} = 100,5$

— first plotting point of R chart is given by: $R_1 = 102 - 99 = 3$

— average range $=$ Rbar $=$ $\bar{R} = \frac{(3+8+4+\cdots)}{32} = 6,112$

— average of Xbars $= \bar{\bar{x}} = \frac{(100,5+101+100,25+\cdots)}{32} = 99,92$

— UCLrange $= D_4 \times \bar{R} = 2,282 \times 6,112 = 13,95$ (where $D_4 =$ constant for $n = 4$, see Table A.1)

— LCLrange $= D_3 \times \bar{R} = 0$

— UCLmean $= \bar{\bar{x}} + A_2 \times \bar{R} = 99,92 + 0,729 \times 6,112 = 104,4$ (where $A_2 =$ constant for $n = 4$, see Table A.1)

— LCLmean $= \bar{\bar{x}} - A_2 \times \bar{R} = 99,92 - 0,729 \times 6,112 = 95,5$

10.5.3.3 Rationale for control limits

Traditionally, there are two distinct approaches to the setting of control limits for performance-based control charts. One approach was developed by Walter Shewhart, who chose control limits formed by adding to and subtracting from the “expected” value, 3 (2 for warning limits) times the standard deviation. This was based on his experience that this was an “acceptable economic value”. At about the same time, control limits were formed by adding to, and subtracting from, the “expected” value, 3,09 (1,96 for warning limits) times the standard deviation. The reason for this difference is brought out in Table 22. One approach focused on rounded probabilities and the other on rounded multiples of standard deviations.

Table 22 shows that the difference is trivial. The International Organization for Standardization (ISO) has adopted the Shewhart system as the world standard (ISO 8258:1991).

From a rational viewpoint, the use of ± 3 standard deviations for action control limits can be argued, for a normal distribution, as striking a reasonable balance between:

- looking for trouble when it does not exist; and

- not looking for trouble when it does exist.

A normal distribution can usually be expected for a means chart based on subgroups of 5 or more, even if the distribution of individual values is non-normal. However, for smaller subgroup sizes in a means chart, and also for charts of individuals, ranges, standard deviations and attributes, the distribution of the plotted statistic can be decidedly non-normal. In such cases, limits based on probabilities for the representative distribution (e.g. skew) are sometimes used.

Table 22 — The two traditional systems for calculating control limits

Control limits (action)		Warning limits (sometimes used)	
Formula [a]	**Probability of being above the upper control limit** [b,c]	**Formula** [a]	**Probability of being above the upper warning limit** [b,c]
mean ± 3 standard errors	0,135 %	mean ± 2 standard errors	2,28 %
mean ± 3,09 standard errors	0,1 % (1/1 000)	mean ± 1,96 standard errors	2,5 % (1/40)

[a] The mean and standard error used for control limits are derived from prior process knowledge or a trial run of sufficient duration for the major sources of variation to manifest themselves. As the control chart is a model, or exemplar, of common cause variation, data arising from special cause variation should not be used in the calculation of control limits. Once calculated, there is no logic in routinely recalculating centre-lines and control limits as is sometimes common practice. Recalculation is only required when there has been a significant change in nominal value or common cause variation.

[b] Probabilities were obtained from the standard normal distribution (see Table 7).

[c] The probability of being below the lower control limit is equal to the probability of being above the upper control limit. Similarly, the probability of being below the lower warning limit is equal to the probability of being above the upper warning limit.

10.6 Interpretation of standard Shewhart-type control charts

The lines in a control chart reflect common cause variation of the statistic. If the plotted points do not adhere to that model, the presence of special causes is indicated. To test for an out-of-control situation, certain guidelines are provided. Typical such guidelines for a normal distribution of the statistic plotted are shown in Table 23.

Table 23 — Probabilities associated with different decision criteria

Rule	Description	Probability
1	Point outside of upper or lower control limit (action)	0,002 70
2	Nine consecutive points on the same side of the centre-line	2/512 = 0,003 91
3	Six consecutive points increasing or decreasing (including the first and last)	2/720 = 0,002 78
4	Any obvious non-random variation [a]	

[a] Based on process technical or operational knowledge rather than probability considerations.

NOTE By way of illustration, take rule 1. If a value exceeds the upper control limit, say, either the process is:

— in control; in which case one has witnessed an extremely unusual phenomena, an event that has a very remote chance of occurrence, namely, 1 chance in 741, 0,001 35 being approximately 1/741; or it is

— out of control because of the presence of a special cause that needs investigation with a view to elimination.

Using rules 1, 2 and 3 for the example shown in Figure 58, an out-of-control situation is indicated for the mean at subgroup 17. The cause should be sought and eliminated as soon as such an indication is seen on the control chart. The control limits should be recalculated, after discarding the data from this subgroup, to form the basis for ongoing control.

10.7 Selection of an appropriate control chart for a particular use

10.7.1 Overview

There are many classes and types of Shewhart-type control charts available. In addition, the cumulative sum (CUSUM) chart is becoming more widely recognized as a very useful diagnostic and control tool.

Standard Shewhart-type SPC control charts normally require a different chart for each process parameter or product characteristic. This was recognized by Bothe [77] who has developed universal charts to apply to situations where continuity of charting is required in short run situations and across parameters and characteristics having different nominal or target and/or average range values.

The cumulative sum chart is, in many ways, superior to conventional Shewhart methods. It is appropriate for examining all forms of numerical data relative to a reference value, on a retrospective or current basis. It has three main uses: control, diagnosis and prediction.

10.7.2 Shewhart-type control charts

Principal kinds of Shewhart-type measured data (i.e. variables) charts are shown in Table 24 and attribute data charts in Table 25.

Table 24 — Principal Shewhart-type measured data control charts

	Subgroup size (n)					
	$n = 1$		$1 < n \leqslant 10$		$n > 10$	
Chart	X chart	MR chart	$\bar{X}$ chart	R chart	$\bar{X}$ chart	s chart
Plot point	X	Moving R	$\bar{X}$	R	$\bar{X}$	s
Centre-line	$\bar{X}$	$\bar{R}$	$\bar{\bar{X}}$	$\bar{R}$	$\bar{\bar{X}}$	$\bar{s}$
UCL	$\bar{X} + E_2\bar{R}$	$D_4\bar{R}$	$\bar{\bar{X}} + A_2\bar{R}$	$D_4\bar{R}$	$\bar{\bar{X}} + A_3\bar{s}$	$B_4\bar{s}$
LCL	$\bar{X} - E_2\bar{R}$	$D_3\bar{R}$	$\bar{\bar{X}} - A_2\bar{R}$	$D_3\bar{R}$	$\bar{\bar{X}} - A_3\bar{s}$	$B_3\bar{s}$

NOTE 1 The standard individual and moving range (X and MR) chart would be suitable in those situations where it is only practicable, or desirable, to take a single measurement at a time. Examples are process parameters such as temperature or pressure and where destructive testing is involved. Moving ranges are constructed from progressive sets of individuals, for example, of size two or greater. The constants E_2, D_3 and D_4 are based on the size of the set that constitutes the range. An alternative is to use a moving average and moving range chart. Prior to calculating the control limits, the resulting distribution should be checked for normality. The lesser sensitivity of the individuals chart, compared with the average chart, for detecting shifts in the level of the process should be noted (see Figure 57). An example of a standard individuals chart is shown for fraction silicon in % in Figure 64.

NOTE 2 The standard average and range chart is recommended for its simplicity, where manual charting is concerned, for subgroup sizes up to about 10. However, it should be borne in mind that the range is based only on the two extreme values in a subgroup and its efficiency falls off, in comparison with the standard deviation, as the subgroup size increases. An example of a standard average and range chart is shown for fabric mass specimens in Figure 58.

NOTE 3 The standard average and standard deviation chart can be used instead of the average and range for all subgroup sizes greater than 1.

The A, B and D constants depend on the subgroup size. They are tabulated in Annex A. In the case of the moving range chart, the equivalent subgroup size is the number of individuals making up each successively plotted range.

10.7.3 Cumulative sum (cusum) charts

A cusum is essentially a running summation of deviations from some preselected reference value. The mean of any group of consecutive values is represented visually by the current slope. Its principal features are:

a) its greater sensitivity than the Shewhart-type chart in detecting small changes in the mean;

b) any changes in the mean, and the extent of the change, are indicated visually by a change in slope of the graph:

— horizontal graph, i.e. zero slope: process is on target or reference value;

— downwards slope: process mean is less than the reference or target value; the greater the slope the bigger the difference;

— upwards slope: process mean is more than the reference or target value; the greater the slope the bigger the difference;

c) it can be used retrospectively for investigative purposes, on a running basis for process control, and for prediction of process performance in the immediate future.

Table 25 — Principal Shewhart-type attribute data control charts

Chart	Events: nonconformities		Nonconforming units	
	Constant sample size: “c” chart	Variable sample size: “u” chart	Constant sample size: “np” chart	Constant sample size: “p” chart
Plot point	C	U	$n{\cdot}p$	p
Centre-line	$\bar{c}$	$\bar{u}$	$n \cdot \bar{p}$	$\bar{p}$
UCL	$\bar{c} + 3\sqrt{\bar{c}}$	$\bar{u} + 3\sqrt{\dfrac{\bar{u}}{n}}$	$n \cdot \bar{p} + 3\sqrt{n \cdot \bar{p}(1 - \bar{p})}$	$\bar{p} + 3\sqrt{\dfrac{\bar{p}(1 - \bar{p})}{n}}$
LCL	$\bar{c} - 3\sqrt{\bar{c}}$	$\bar{u} - 3\sqrt{\dfrac{\bar{u}}{n}}$	$n \cdot \bar{p} - 3\sqrt{n \cdot \bar{p}(1 - \bar{p})}$	$\bar{p} - 3\sqrt{\dfrac{\bar{p}(1 - \bar{p})}{n}}$

NOTE 1 There are four types of standard attribute charts:

- c: number of incidences, events or nonconformities in a sample that is of constant size;
- u: number of incidences, events or nonconformities per unit in a sample that is of variable size;
- np: number of nonconforming units in a sample that is of constant size;
- p: proportion of nonconforming units in a sample that is of variable size.

The choice of which to use depends on whether the sample size (n) is constant or variable and whether incidences/nonconformities or nonconforming units are involved. It is advisable, from the point of view of simplicity, to keep sample sizes constant if possible. Incidences/nonconformities charts frequently provide more technical information than nonconforming units ones; however, certain logistics information may be lost. For example, if one has 14 nonconformities in a sample of 50 units, it would not be known, from the chart, how many units were affected. On the other hand, if 8 units were involved, some with multiple nonconformities, diagnostic information would be lost on some nonconformities.

NOTE 2 The measured data chart is preferred whenever possible. An example would be a diameter, say, which could be checked with either a go/no-go gauge or measured with a micrometer. Another illustration is where subjective judgements made on a particular characteristic are converted into a rating scale of, say, 1 to 10. This permits the use of a measured data chart rather than an attribute one. An example is a scale of 1 to 5 for degree of fabric pilling.

NOTE 3 This preference is for two principal reasons: one, the measured data chart provides more information, and two, in the quality field the attribute chart often requires nonconformities or nonconforming units to occur before plotting can take place.

NOTE 4 Having made the decision of which type of attribute chart to use, a second choice is either to use a single characteristic chart (see Figure 51) or a multiple characteristic chart. The multiple characteristic chart facilitates prioritizing the sources of variation and diagnosis with a view to improving process capability.

NOTE 5 The capability of a standard attribute chart is given by the overall average (centre-line) value; nearly 10 % initially in Figure 51, and ultimately improves to less than 1 %.

NOTE 6 A much larger sample size is needed with attributes charting compared with variables control charting to determine significant change in the process.

NOTE 7 If the LCL calculates to be 0 or negative, there is no lower control limit.

11 Process capability

11.1 Overview

In Clause 10, the performance-based control chart was discussed purely in terms of process control. No regard was given to the acceptability, or otherwise, of the characteristic in respect to an imposed standard of performance. A further important technical role of the performance-based control chart is the provision of the basis for the assessment of process capability against the requirements of a specification and for the formulation of standardized benchmarks of performance. The four states of any process are shown in Table 26.

Table 26 — The four possible states of any process

		Control (stability)	
		Not ok	**Ok**
Capability (performance)	**Not ok**	eliminate special causes reduce common causes	reduce common causes
	Ok	eliminate special causes	ideal situation: monitor at low level

The performance-based control chart provides answers to the following three significant business questions.

- Question 1: *Is the process in control?* This is directed, primarily, at operational people working in the process. If not in control, there is a need to seek out and eliminate detrimental special cause variation.

- Question 2: *What is the process capability in relation to the specified requirement or customer expectation?* This is directed, primarily, at technical people responsible for the process. If this is not at an acceptable level, there is a need to make fundamental changes to the process to reduce common cause variation, to use a more capable process or to relax the specification.

- Question 3: *Is there evidence of improvement?* This will be signalled as follows:

  - in a measured data chart, by an “out-of-control” movement in the mean towards the preferred value and/or an “out-of-control” reduction in the within-subgroup variation indicated by the range or standard deviation chart;

  - in an attribute chart by an “out-of-control” change in the mean towards the preferred value: seven consecutive points below the centre-line (rule 2) in the case of the plot of nonconformities in Figure 51.

This is directed, primarily, at management who, in a best practice organization, are responsible for the continual improvement of processes.

11.2 Process performance versus process capability

ISO 3534-2 distinguishes between process performance and process capability as follows:

a) process performance and its related P_p (Performance$_\text{process}$) indices relate to the statistical estimate of the outcome of a characteristic from a process that may *not* have been demonstrated to be in a state of statistical control in relation to that characteristic; whereas

b) process capability and its related C_p (Capability$_\text{process}$) indices have an identical definition, with the exception that here the process has been demonstrated to be *in* a state of statistical control.

Arising from these international definitions, process performance measures are preliminary indicators confined to early development activities in developing the potential of new processes or more mature processes that are not in a state of statistical control. They are thus unrelated to the quality of product, process or service offered to the ultimate customer. Concentration here is therefore focused on process capability measures that are based on prior demonstration of process stability. Having stated that the calculations associated with process capability are identical to those of process performance, the significant difference is the stability of the data used in the calculations and the reliability of subsequent predictions.

11.3 Process capability for measured (i.e. variables) data

11.3.1 General

Process capability is calculated for a particular process parameter or product characteristic *only* after process stability has been confirmed *and* the distribution pattern of individual values has been determined.

The confirmation of process stability is first established from a control chart. Only then may reliable predictions be made and the distribution pattern be determined by recourse to tally charts, histograms, probability papers or computer-based distribution techniques.

11.3.2 Estimation of process capability (normally distributed data)

For normally distributed data from a stable process, an estimate of the capability of a particular characteristic is given by the formula:

$$\overline{\overline{X}} \pm (z \cdot s)$$

where

$\overline{\overline{X}}$ is the overall mean;

z is the chosen constant, often equal to 4;

s is the estimated standard deviation of individuals.

EXAMPLE If data taken from a stable process exhibits a normal pattern of variation and $\overline{\overline{X}} = 10{,}01$, $z = 4$ and $s = 0{,}01$, then the estimated capability is quoted as:

10,01 ± 0,04

EXAMPLE within which (from Table 7) nearly 99,994 % of values are predicted to lie (as 2 × 32 parts per million are expected outside of this range of values).

If, on the other hand, the capability standard is less stringent and z is taken to be 3, then the estimated capability is quoted as:

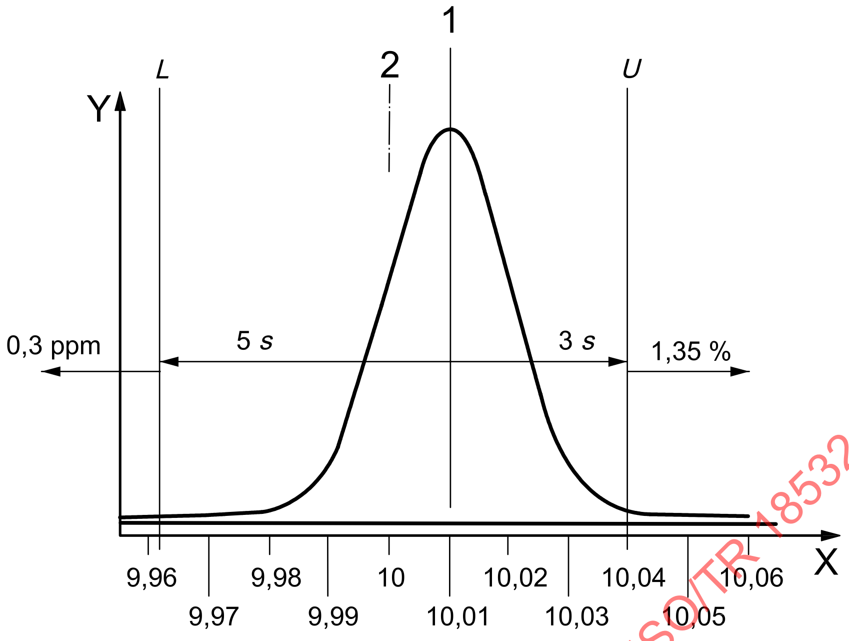
10,01 ± 0,03

within which 99,73 % of values are expected to lie.

Capability can then be referenced to any imposed specification limits and the proportion expected outside of these limits estimated using Table 7. For instance, if the specification is 10,00 ± 0,04:

— enter Table 7 at $z = \frac{(U - \overline{\overline{X}})}{s} = (10{,}04 - 10{,}01)/0{,}01 = 3$, to give 0,135 % above the upper specification limit;

— enter Table 7 at $z = \frac{(\overline{\overline{X}} - L)}{s} = (10{,}01 - 9{,}96)/0{,}01 = 5$, to give 0,3 parts per million below the lower specification limit. The pictorial expression of this is shown in Figure 59.



Key

1 mean

2 CL

X data

Y function value of probability density

NOTE 1 Mean = 10,01; s = 0,01; Specification = 10 ± 0,04

NOTE 2 CL = centre-line = mid-distance between upper specification limit (U) and lower specification limit (L)

Figure 59 — Graphical comparison of process capability with specified tolerance

It can be seen that the choice of z does not affect the proportion of values predicted to lie outside of the specified tolerance. However, if $z = 3$, rather than $z = 4$, is mandated as a minimum standard, the implication is that, provided the capability expressed by 10,01 ± 0,03 lies within the specified tolerance band, the process capability is acceptable. For $z = 4$, this becomes 10,01 ± 0,04. In other words, with $z = 3$, up to 0,135 % is tolerated outside each specification limit as opposed to 32 parts-per-million with $z = 4$. Thence in the example:

- if $z = 3$ is the minimum reference standard, the process is deemed *capable*;
- if $z = 4$ is the minimum reference standard, the process is deemed *incapable*. It can be made capable by either reducing the standard deviation from 0,01 to 0,007 5, by adjusting the mean from 10,01 to 10,00 or by changing the specification.

11.3.3 Estimation of process capability (non-normally distributed data)

11.3.3.1 General

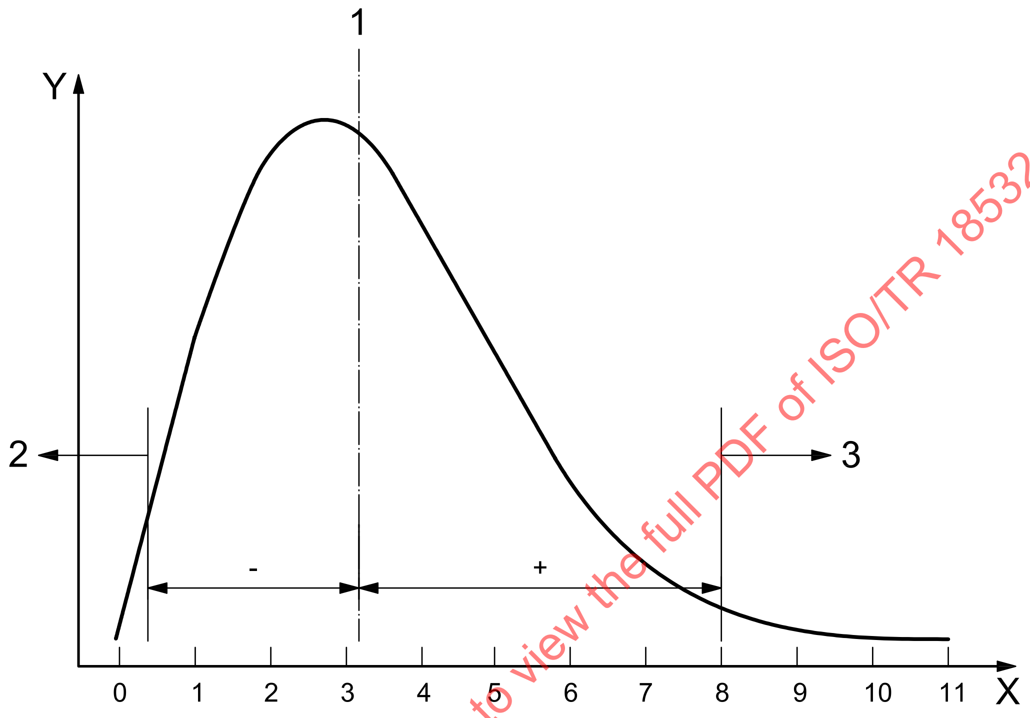
From the central limit theorem, it is known that averages of subgroups tend to normality as the subgroup size increases. However, many processes quite naturally produce patterns of variation for individuals that are non-normal. For example, dimensions with a natural zero such as eccentricity, parallelism and taper are likely to be skewed. So are such things as times to pay, arrival times and length of time to resolve a query. Similarly, there may be patterns that have natural upper bounds and thus produce a negatively skewed distribution. It is essential that any statistical statement of capability be based upon the pattern of variation exhibited by the process.

Statistical expressions for capability of non-normal distributions are best expressed in probability terms rather than in terms of standard deviations.

A typical expression of capability for a skew distribution, equivalent to the ± 3 standard deviations for the normal distribution, would be:

$$\text{mean} \begin{array}{l} + \text{ range of values between the mean and upper 0,135 distribution percentile} \\ - \text{ range of values between the mean and lower 0,135 distribution percentile} \end{array}$$

This is indicated graphically in Figure 60.



Key

1 mean

2 0,135 % below the lower specification limit

3 0,135 % above upper specification limit

X data

Y function value of probability density

Figure 60 — Illustration of the estimation of capability with a skew distribution (equivalent to a range of ± 3σ in a normal distribution)

A simple graphical procedure is the use of an appropriate probability paper. See 5.3.7 and 5.3.9.4. Alternatively, good SPC computer programs use distribution-fitting routines.

11.3.3.2 Example of assessing process capability of a skew distribution using probability paper

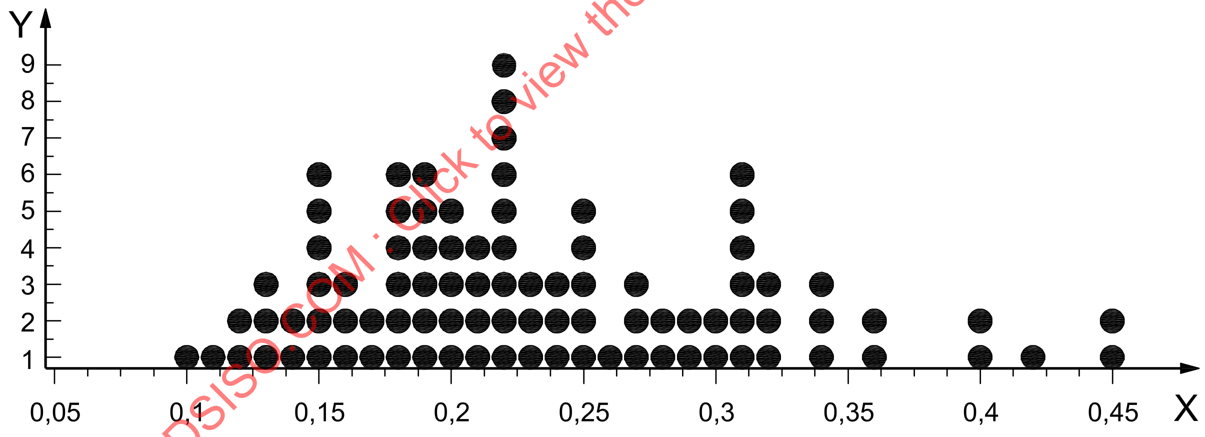
An example will best illustrate the procedure. The data of Table 27, relating to the measurement of silicon through the tap hole of a blast furnace, are to be used.

Table 27 — Values of fraction silicon in % taken in sequence from a blast furnace

0,13	0,15	0,19	0,22	0,20	0,20	0,18	0,26	0,40
0,10	0,22	0,29	0,18	0,13	0,16	0,28	0,34	0,20
0,19	0,21	0,28	0,25	0,15	0,22	0,23	0,32	0,20
0,22	0,24	0,21	0,19	0,12	0,31	0,24	0,30	0,42
0,45	0,19	0,29	0,22	0,21	0,18	0,18	0,31	0,31
0,25	0,22	0,15	0,17	0,22	0,16	0,22	0,31	0,36
0,13	0,15	0,32	0,15	0,23	0,14	0,31	0,27	0,27
0,14	0,17	0,20	0,18	0,34	0,16	0,25	0,12	0,36
0,25	0,22	0,30	0,15	0,32	0,19	0,31	0,24	0,27
0,23	0,25	0,19	0,11	0,18	0,34	0,45	0,40	0,21
NOTE Sequence of readings: first read downwards in column 1, then downwards in column 2, etc.								

The tap hole of a blast furnace is opened at 3-hour intervals and the fraction silicon in % is measured and recorded. Ninety of these values, taken in sequence, are shown in the table. (Data reference: ISO 11648-1 [43].)

Before carrying out any detailed calculations, it is always advisable to do some plotting of the data. A simple dot plot is shown in Figure 61 and there appears to be a longer positive tail or positive skewness. A normal probability plot follows in Figure 62, which casts severe doubt that the distribution is normal.



Key

X fraction silicon, in %

Y number of observations

Figure 61 — Dot plot for percent of silicon data showing overall pattern of variation

A simple transformation is to make a log transform of the data. The probability plot of the log transform in Figure 63 shows essentially a straight line of the data.

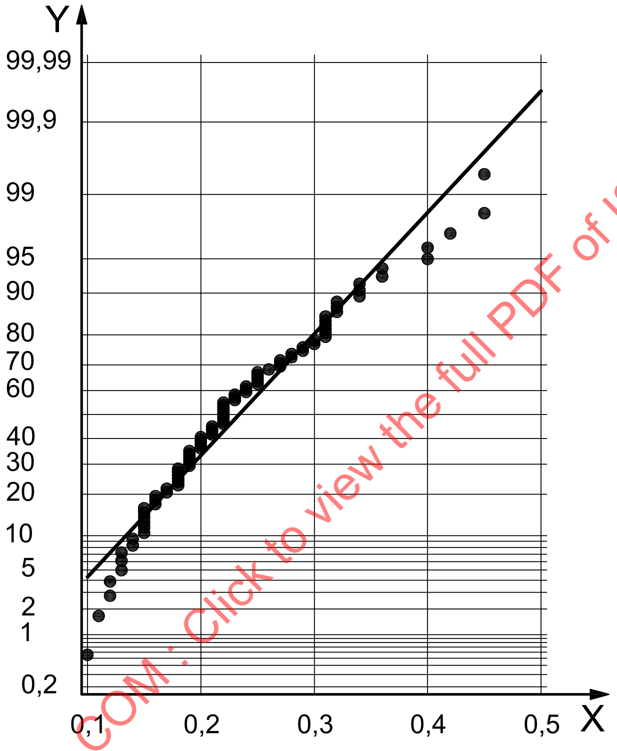
Although other transformations may be applicable, the log transform is easily understood. Thus, further analysis of the data will be on the log transform. An individual's control chart in Figure 64 is shown next.

A detailed analysis will not be made of the control chart other than to state that there probably are more trends seen than would be expected if the process were in perfect control. An EWMA (exponentially weighted moving average) chart and a CUSUM chart confirm this.

Further examination of the original data and of the log transformed data: The mean of the original data is 0,234 and its median is 0,220. If the log transform is an appropriate transform, then the mean of the log transformed data, when converted back to original units, should be essentially the same as the median of the original data. The mean of the log data is −1,5083 and $e^{-1,5083} = 0,221$! Very good.

Working with the log-transformed data, the approximate C_p limits can be calculated and then converted back to the original units. The results are: 0,08 and 0,60. However, since these are based on only 90 data points, they should be used with great caution.

When the data are non-normal, it is always a challenge to find an appropriate transformation to ease many of the necessary calculations. The challenge of finding that transformation is beyond the scope of this Technical Report, but there are excellent textbooks available to help. Likewise, some of the modern statistical software packages can be of great aid.

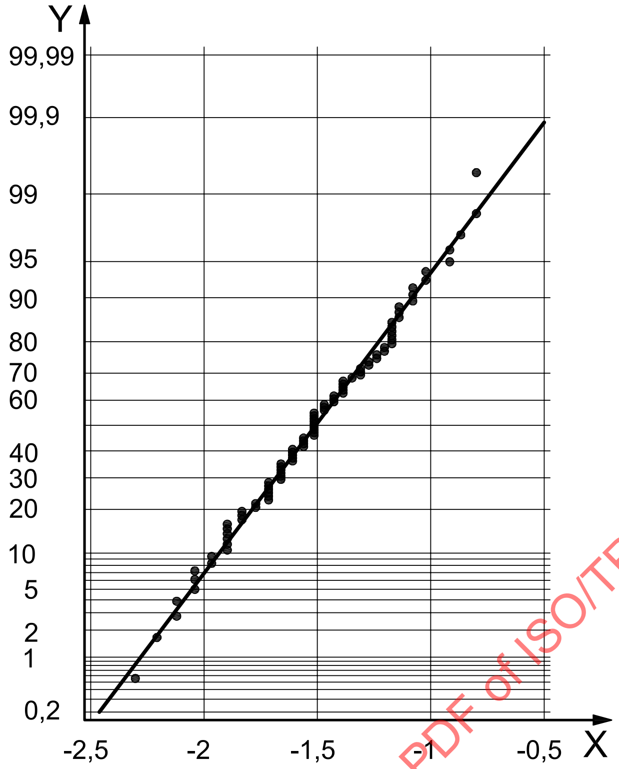


Key

X fraction silicon, in percent

Y percent

Figure 62 — Probability plot for percent of silicon data showing overall pattern of variation

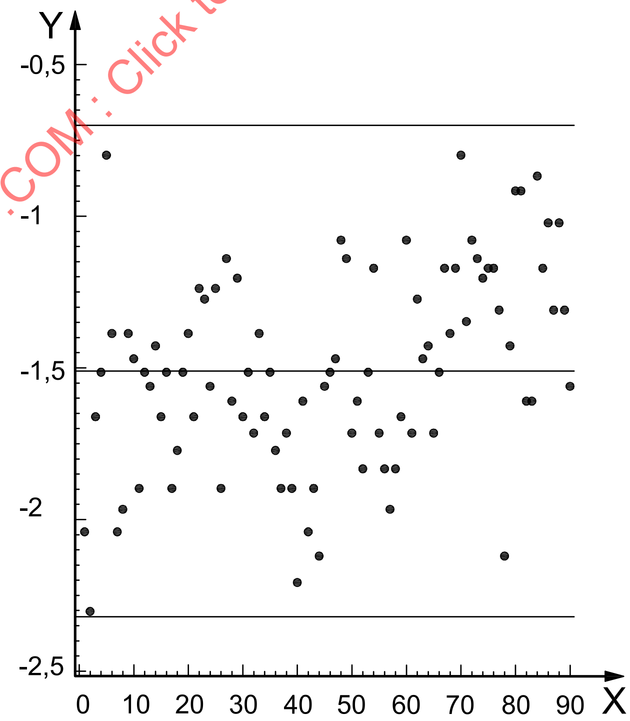


Key

X logarithm of fraction silicon, in percent

Y percent

Figure 63 — Probability plot for the logarithm of percent of silicon data showing overall pattern of variation



Key

X observation number

Y logarithm of fraction silicon, in percent

Figure 64 — Individuals control chart of ln percent of silicon with limits

11.4 Process capability indices

11.4.1 General

Process capability indices provide simple standardized metrics, in worldwide use, which assess the capability of measured characteristics in relation to specified requirements. The application of these indices is growing rapidly with an increasing number of customers requiring documented proof of first time quality through:

- the achievement and demonstration of appropriate control chart stability, together with;
- the realization and confirmation of minimum value capability indices for significant process parameters and product characteristics.

Of equal consequence is the use of SPC and capability indices to provide suppliers themselves with the means to use “first time quality health profiling” as a business tool within their organization and those of their subsuppliers.

Originally, capability indices were intended for use with normally distributed data. Unfortunately, there are those who still calculate and declare indices based on normality even when the distribution is patently non-normal. This has arisen, to a large degree, by the equations for the indices often appearing to be generic when indeed they are specific to the normal distribution.

Discussion of specific capability indices is confined to standardized ones that are in general use.

11.4.2 The C_p index

The C_p process capability index relates a standardized process spread to the specified tolerance interval. It does not take the location (e.g. mean) of the distribution into account. Generically, for a process in control, it is given by:

$$C_p = \frac{\text{Permissible range of values}}{\text{Actual standardized range of values}} = \frac{\text{Specified tolerance}}{(99{,}865 \text{ percentile} - 0{,}135 \text{ percentile})}$$

NOTE The C_p index is referenced worldwide, quite arbitrarily, against the probability equivalent to 6 standard deviations for the normal distribution. Figure 60 indicates the significance of the 99,865 and 0,135 percentiles.

For a normal distribution, this expression reduces to:

$$C_p = \frac{\text{Specified tolerance}}{\text{6 standard deviations}} = \frac{U - L}{\text{6 standard deviations}}$$

As C_p does not take the location of the distribution into account, it provides a value for the relative capability of a centred process. For a non-centred process, it represents the potential capability of the process parameter or product characteristic. Hence C_p should always be used in conjunction with other indices that do take location into account.

The minimum acceptable value of C_p will depend on the appropriate customer contractual requirement or benchmark set internally by the supplier for a given application. In some business sectors $C_p \geqslant 1{,}33$. For a centred process having a normal distribution, a C_p of 1,33 can be expected to give rise to 32 parts per million above and 32 parts per million below specification limits. Substitution in the equation:

$$C_p = \frac{\text{Specified tolerance}}{\text{6 standard deviations}} = \frac{U - L}{\text{6 standard deviations}}$$

indicates that:

— a C_{p} of 1,33 equates to: $U - L = 8$ standard deviations. If centred and normal, from Table 7, this can be expected to give rise to 32 parts per million above and 32 parts per million below specification limits; namely, nearly 99,994 % conforming to specification;

— a C_{p} of 1,00 equates to: $U - L = 6$ standard deviations. If centred and normal, this will give rise to 0,135 % above and 0,135 % below specification limits; namely 99,73 % conforming to specification.

Table 28 — Factors for calculation of the 95 % lower confidence limit, $C_{\text{p min}}$, of C_{p} for various numbers of subgroups and subgroup sizes

Subgroup size	Number of subgroups				
	1	5	10	20	30
3	0,255	0,631	0,735	0,811	0,845
4	0,369	0,697	0,783	0,845	0,873
5	0,443	0,735	0,811	0,865	0,890
6	0,494	0,760	0,829	0,879	0,901
7	0,533	0,779	0,843	0,888	0,908
8	0,562	0,793	0,853	0,895	0,914
9	0,586	0,804	0,861	0,901	0,919
10	0,605	0,813	0,867	0,906	0,923

C_{p} is an estimate. It is thus subject to sampling variation. Strictly speaking, confidence intervals should be computed to provide a range of $C_{\text{p}}s$ that include the true C_{p} with a given probability. A centred process is then deemed capable if the minimum acceptable $C_{\text{p}} \geqslant$ lower confidence limit. In practice, it is the exception rather than the rule to use such confidence limits. Table 28 (from Li, Owen and Borrego) [107] provides values with which to factor the estimated C_{p} to obtain the 95 % lower confidence limit, $C_{\text{p min}}$, of C_{p} for a range of subgroup sizes in terms of number of subgroups.

Multiply the tabulated value by the C_{p} estimate to obtain the 95 % lower confidence limit, $C_{\text{p min}}$, of C_{p}.

EXAMPLE C_{p} has been calculated as 1,60, based on the average subgroup range of 20 subgroups of 5, for a stable process having a normal distribution.

From Table 28, $C_{\text{p min}}$ (at the 95 % confidence level) $= 0{,}865 \times 1{,}60 = 1{,}38$

11.4.3 The C_{pk} family of indices

There are three indices from the C_{pk} family in general use. These are:

$$C_{\text{pk}U} = \frac{U - \text{mean}}{\text{Range between mean and upper 0,135 distribution percentile}}$$

$$C_{\text{pk}L} = \frac{\text{Mean} - L}{\text{Range between mean and lower 0,135 distribution percentile}}$$

$$C_{\text{pk}} = \text{Minimum of } C_{\text{pk}U} \text{ and } C_{\text{pk}L}$$

For a normal distribution, $C_{\mathrm{pk}U}$ and $C_{\mathrm{pk}L}$ reduce to:

$$C_{\mathrm{pk}U} = \frac{U - \text{mean}}{3 \text{ standard deviations}}$$

$$C_{\mathrm{pk}L} = \frac{\text{Mean} - L}{3 \text{ standard deviations}}$$

The C_{pk} family of indices relates both the process variability and the location (setting) of the process to specification limits.

$C_{\mathrm{pk}U}$ is an index that relates process variability and location to the upper specification limit, whereas $C_{\mathrm{pk}L}$ relates process variability and location to the lower specification limit. This is shown graphically in Figure 65.

For a single-sided specification limit, only one of these indices can be calculated. Knowing $C_{\mathrm{pk}U}$ and/or $C_{\mathrm{pk}L}$, the proportion lying outside a specification limit can be determined using Table 28.

C_{pk}, the lowest of $C_{\mathrm{pk}U}$ and/or $C_{\mathrm{pk}L}$, is sometimes quoted alone as a minimum standard in contractual requirements. Typical such values are:

$C_{\mathrm{pk}} \geqslant 1{,}33$

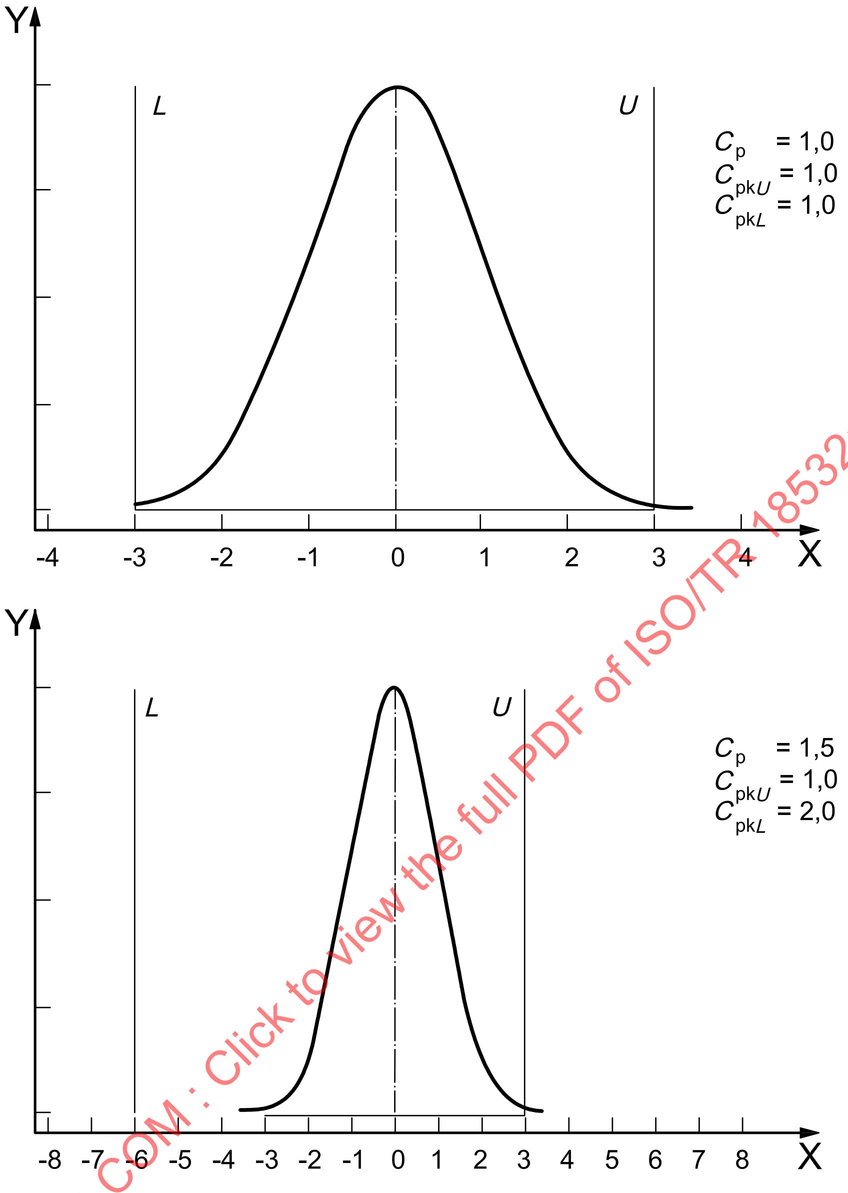
However, it should be borne in mind that C_{pk}, on its own, gives no indication of the direction in which the process is biased, if at all; the location of the distribution; or the extent of the variation. It thus degrades the information conveyed. This is particularly relevant to a supplier if the penalty of transgressing one limit is different from transgressing the other. Such a situation could arise, for example, with a characteristic such as a length; too short could give rise to scrap and too long to less expensive reworking. Neither is the preferred value always on nominal: for example, if minimum is best then one would aim for the minimum acceptable $C_{\mathrm{pk}L}$ whilst, at the same time, maximizing C_{p} and $C_{\mathrm{pk}U}$.

In practice, the minimum standard for $C_{\mathrm{pk}U}$ and $C_{\mathrm{pk}L}$ is often taken to be 1,33. However, this will depend on contractual requirements or self-imposed benchmarks currently in place in a given sector or organization.

These indices are becoming more widely used:

- by customers for supplier process certification/accreditation in certain industrial sectors (for example, automotive and aerospace sector technical requirements);

- by suppliers to provide quality health profiles for their organizations. A typical example of such a profile is shown in Table 29.



Key

X standard deviations

Y probability density

Figure 65 — Relationship between C_p and C_{pkU} and C_{pkL} for two sets of process variability and locations of specification limits

To avoid confusion, or worse, it is recommended that any statement of capability using these indices should contain at least five items of information, viz. C_p, C_{pkU}, C_{pkL} distribution shape and an indication of the preferred value, namely maximum, minimum or nominal is best. Table 29 provides such information.

As C_{pk} is an estimate, from a statistical viewpoint, it should be qualified by confidence bounds between which the true value is expected to lie. In such cases, frequently the lower confidence limit only is quoted. Table 30, from Chou, Owen and Borrego [80], provides such limits for a range of sample sizes and values of C_{pk}.

Table 29 — Steel works quality health profile for selected process characteristics

Characteristic	Aim	In control	Distribution	Capability		
				$C_{\text{pk}L}$	C_{p}	$C_{\text{pk}U}$
Silicon	nominal	yes	skew	1,3	1,0	0,9
Aluminium	nominal	yes	normal	1,4	1,5	1,6
Teeming temperature	nominal	yes	normal	1,3	1,3	1,3
Teeming time	nominal	yes	normal	1,6	1,7	1,8
Injuries per workforce per week	minimum	yes	attribute	0,73 %		
Cobbles	minimum	yes	attribute	0,14 %		
Billet rhomboidity	minimum	yes	normal	2,4	—	—
Time to charge	minimum	no	bi-modal	disparity between steelmen [a]		

[a] Subject of investigation.

Table 30 — Approximate 95 % lower confidence limits for C_{pk}

Estimated C_{pk}	Sample size, n						
	20	30	40	50	100	200	300
1,0	0,66	0,72	0,76	0,79	0,85	0,89	0,91
1,2	0,81	0,88	0,93	0,95	1,03	1,08	1,10
1,4	0,95	1,04	1,09	1,12	1,20	1,26	1,29
1,6	1,10	1,20	1,25	1,29	1,38	1,45	1,47
1,8	1,25	1,35	1,41	1,46	1,56	1,63	1,66
2,0	1,39	1,51	1,58	1,62	1,74	1,81	1,85

The confidence limits are given for a normal distribution in terms of estimated C_{pk} and sample size.

11.4.4 The C_{pm} index

11.4.4.1 Current specification practice versus optimal design values

The point value of a measured characteristic expressed in a design specification is intended to reflect preferred value. This focus has been diffused by two practices:

- the setting of acceptable tolerances around the preferred value to reflect the presence of some variation in the realization process, for example, 20,0 mm ± 0,1 mm;
- the quoting of the range of permissible values, for example, 20 Nm to 80 Nm, without any reference to a preferred value. This leaves it open as to whether nominal, minimum or maximum is best.

These practices can give rise to two types of response:

- no emphasis or regard is placed on achieving preferred value: the "goal post mentality" prevails; namely, anything within the specified tolerance represents acceptable, or even premium quality;

— aiming at a value that is most cost effective from the supplier's point of view, often to the detriment of the customer. If this is coupled with a drive to minimize variation in the process to maximize the gain to the supplier, then this can be even more detrimental to the customer. An example would be to offset the aim of the process towards the specification limit that provides the greatest saving in material. As an illustration, a garment manufacturer who buys wool by the kilogram could knit more jumpers or cardigans per kilogram if the wool is on the thinner side (higher "count" wool) of the specified tolerance. Whilst this would be cost advantageous to the manufacturer, it would be to the detriment of the retailer and of the ultimate wearer.

11.4.4.2 Expression for C_{pm} index

An index, C_{pm}, has been devised to provide a single quantitative measure of diminished utility that can arise in terms of process offset from preferred value and the extent of process variability.

$$C_{\text{pm}} = \frac{U - L}{6\sqrt{s^2 + (\overline{X} - T)^2}}$$

where

s is the sample standard deviation;

$\overline{X}$ is the process mean;

T is the target value.

When $\overline{X} = T$, $C_{\text{pm}} = C_{\text{p}}$. As the mean drifts from the target and/or the standard deviation increases, the C_{pm} value declines.

Thus C_{pm} is a measure of both process spread and level in relation to the target value. The use of C_{pm} refocuses on the targeting of optimal values rather than a degraded minimum requirement of conformance to specified tolerances. The C_{pm} index is based on some fundamental loss concepts as illustrated in Figure 66.

11.4.4.3 Basis of C_{pm} index

Quality is frequently defined as "conformance to specification". Traditionally, such specifications for measured characteristics embrace an allowable tolerance band. This widely practiced approach is based on the *"goal post*" mentality, as indicated in Figure 66.

In terms of loss function it assumes, in the model of Figure 66 a), that there is no loss for values of a characteristic anywhere within a specified tolerance band, but there is an incremental loss for those beyond the specification limits.

The implications of this mind set are the following.

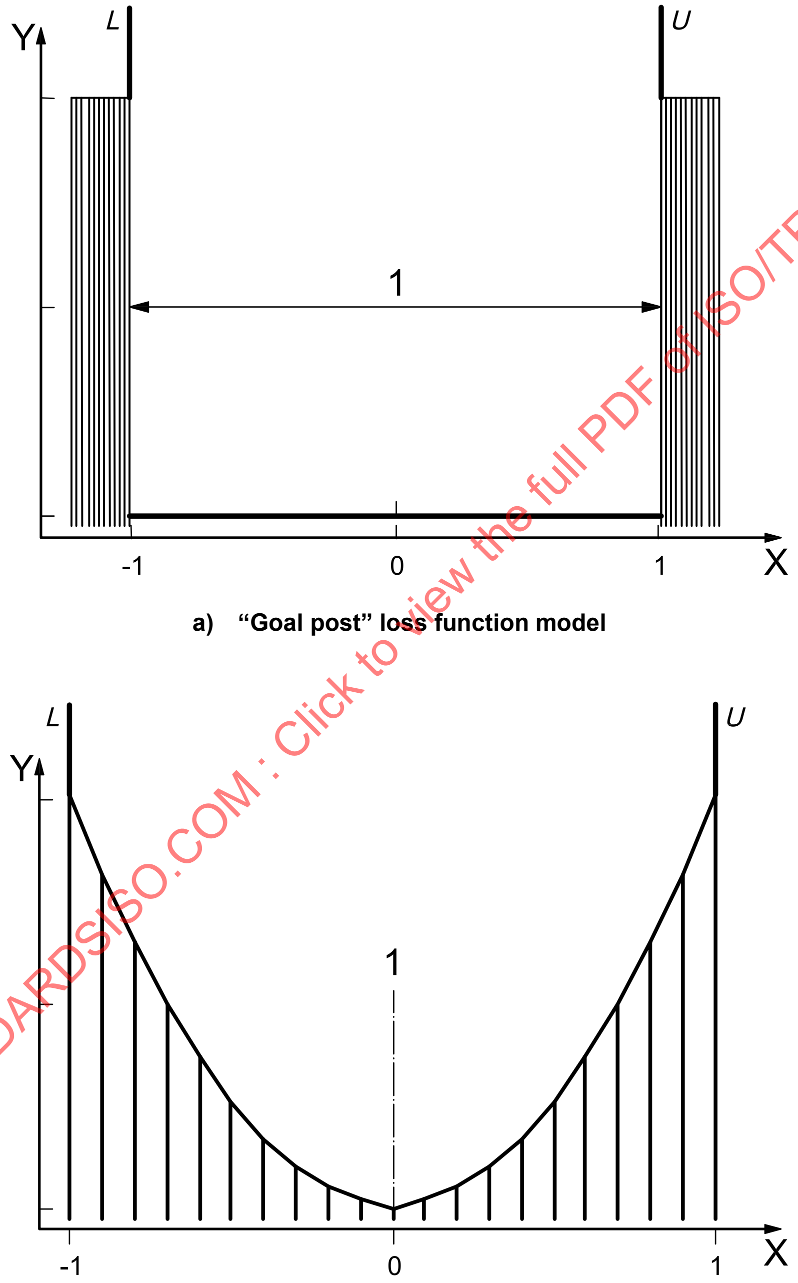
a) All characteristic values within the specified tolerance range are equally acceptable so there is no or little incentive to aim at an optimal design value; namely, it produces a mindset that inhibits quality improvement. However, it does enable clear-cut decisions to be made about conformity.

b) Exploitation of a) is acceptable to the detriment of the customer. One example of deliberately *offsetting* the process to achieve gain to the supplier at the expense of the customer has already been given in 11.4.4.1 for the wool garment manufacturer. Another illustration of this exploitation of relatively low process variation relative to specified tolerance is the practice of permitting the process mean to *drift* across the specified range. This can arise fortuitously due to lack of statistical control of a process or deliberately in situations, for instance, involving physical tool wear or progressive diminishing in the strength of a solution. In such cases, this will result in marginally acceptable characteristics at the start and end of each cycle of tool replacement or topping up. Such practices frequently give rise to a decrease in utility to the customer.

This goal post mentality model is contrasted with the Taguchi [127] economic loss model shown in Figure 66 b). To obviate the need for extensive calculations for each and every design characteristic, Taguchi advocates the use of a generic quadratic loss function. The resulting parabola has its minimum point at zero at the optimal design value and rises on either side in proportion to the square of the distance from the preferred or target value. This quadratic loss function can be conveniently split into two elements:

— the process variance, σ^2;

— the square of the offset of the process mean from the target, $(\mu - T)^2$.



a) "Goal post" loss function model

Key

X value

Y loss

b) Taguchi generic loss function model

Figure 66 — Comparison of conformance to toleranced specification with optimal value approach

Thence, the loss is equal to $k\sigma^2 + k(\mu - T)^2$, where k is the loss parameter. This gives rise to the function for C_{pm}.

Complications arise in the use of C_{pm} if the optimal value is not the mid-point between specification limits, if the distribution is non-normal or if the process is not under statistical control.

Although the generic quadratic loss function may be difficult to quantify in specific instances, one should not lose sight of the very important message it conveys. That is that:

Quality as perceived by a customer is not a go/no-go situation. There is an optimum or target value. As a characteristic varies from this point, the perception of quality progressively deteriorates until at some point, possibly a specification limit, the condition becomes untenable.

A simple example of this is ambient temperature. Although in an industrial situation there may be statutory maximum and minimum values laid down, any deviation from the perceived ideal value may cause a degree of discomfort depending on the extent of that deviation.

11.5 Process capability for attribute data

The capability of an attribute process is obtained simply from the centre-line of the attribute control chart of a stable process. It is typically expressed as:

a) average nonconformities or faults per unit for c and u charts;

b) average proportion of units nonconforming for p and np charts.

Removing special causes of variation from an attribute process, through elimination of *sporadic* causes of variation, restores the status quo. It does not improve it.

A process in statistical control reflects *systemic* causes of variation, the extent of which is indicated by the status quo or on-going level of performance, namely the capability of the process. Reduction of systemic causes demands fundamental changes in approach to that adopted for the removal of sporadic causes. An example of this is shown in Figure 51. Changing the capability to a more favourable value improves attribute process performance. This is seen to be achieved when the level of performance moves significantly (shown by an “out-of-control” control chart) towards the preferred or targeted value of capability. This preferred value is frequently zero where nonconformities or nonconforming items are concerned. However, the targeted value should be realistic and reflect the specific diagnostic and improvement programme established to achieve it.

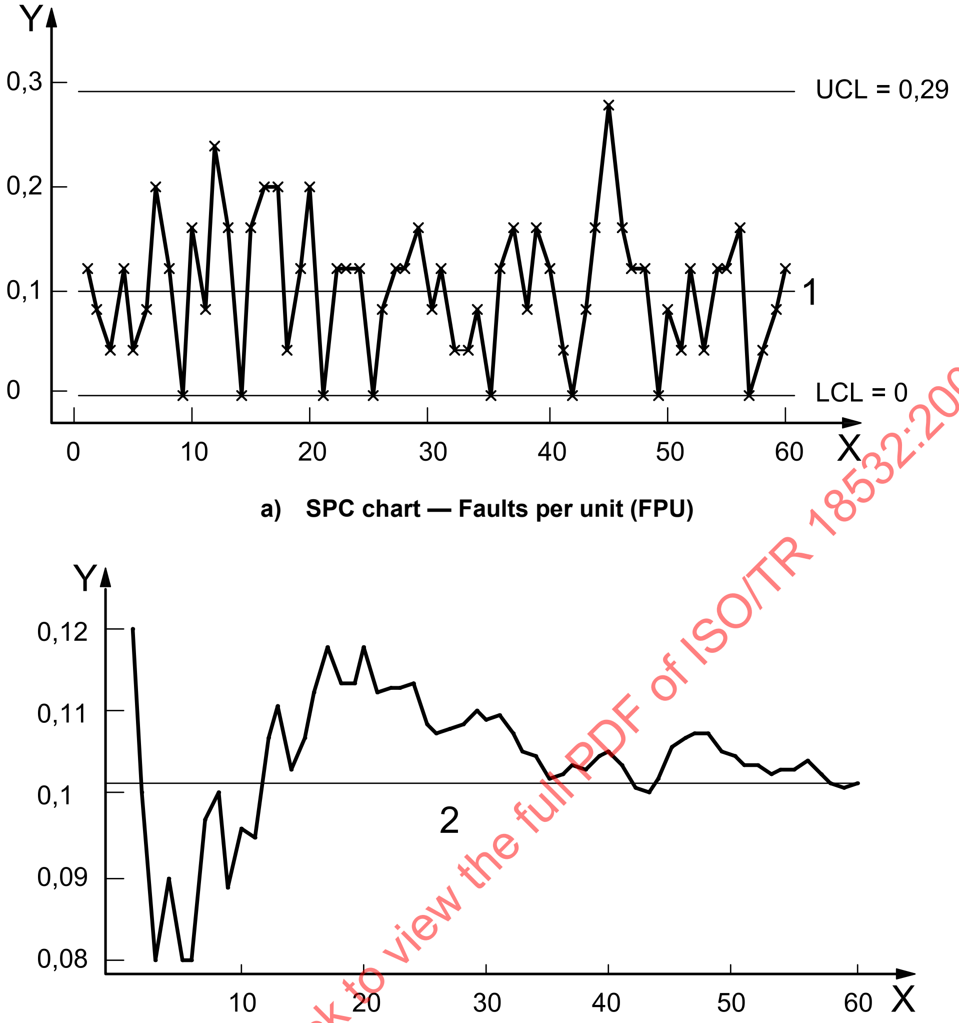
EXAMPLE Printed circuit boards of a particular type are assembled in batches of 25 units. These are 100 % inspected in assembly sequence and the number of nonconformities/faults per batch recorded. The results of 60 such batches are shown in Table 31. Determine the capability of the printed circuit board assembly process.

Table 31 — Faults per batch on printed circuit boards (60 batches of 25)

Batch no.	Faults	Batch no.	Faults	Batch no.	Faults
1	3	21	0	41	1
2	2	22	3	42	0
3	1	23	3	43	2
4	3	24	3	44	4
5	1	25	0	45	7
6	2	26	2	46	4
7	5	27	3	47	4
8	3	28	3	48	3
9	0	29	4	49	0
10	4	30	2	50	2
11	2	31	3	51	1
12	6	32	1	52	3
13	4	33	1	53	1
14	0	34	2	54	3
15	4	35	0	55	3
16	5	36	3	56	4
17	5	37	4	57	0
18	1	38	2	58	1
19	3	39	4	59	2
20	5	40	3	60	3

The attributes control chart for this data shown in Figure 67 is seen to be in statistical control. Hence, capability may be calculated from the control chart mean. This is given as 0,102 faults per unit.

A plot is also shown in Figure 67 of cumulative faults per unit. In practice, it is recommended that this figure be plotted as well as the control chart to determine whether or not the capability value has stabilized. It is seen that it starts to stabilize at about the 35th batch in this particular case. Any prediction of capability prior to this could be unreliable.



a) **SPC chart — Faults per unit (FPU)**

b) **Cumulative FPU**

Key

1 overall average FPU, $\bar{p} = 0,102$

2 PCB capability = 0,102

X batch number

Y FPU

NOTE 1 The SPC chart is a plot of number of faults per batch divided by batch size, namely faults per unit or faults per printed circuit board. Hence the first point plotted is 3/25 = 0,12, the second point plotted is 2/25 = 0,08 and so on. Denoting the overall average FPU by $\bar{p}$, the upper control limit is given by:

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{25}} = 0,102 + 3\sqrt{\frac{0,102 \times 0,898}{25}} = 0,283\,6$$

NOTE 2 The faults per unit (FPU) chart is a plot of the cumulative faults per unit. Hence the first point plotted is 3/25 = 0,12, the second one is (3 + 2) / (25 + 25) = 0,10, the third one is (3 + 2 + 1)/(25 + 25 + 25) = 6/75 = 0,08, and so on.

Figure 67 — Printed circuit board faults SPC chart and cumulative faults per unit (FPU) chart

In summary:

1) the control chart confirms that the process is in a state of statistical control and provides a value of the overall mean;

2) the cumulative chart indicates when enough data has been collected to provide a stable estimate of the process capability.

12 Statistical experimentation and standards

12.1 Basic concepts

12.1.1 What is involved in experimentation?

An experiment involves changing things that are believed to have an effect on the performance of the process, product or service. By changing from one set of conditions to another, to a predetermined pattern, the actual effect can be estimated. In an experiment:

a) the things that are changed are called *factors*;

b) the conditions to which the factors are changed/set are known as *levels*;

NOTE The term "level" is normally associated with a quantitative characteristic such as temperature, in which case differing experimental levels could be 200 °C and 220 °C, say. In experimentation, it also serves as the term describing the setting of a qualitative characteristic, for example, the absence or presence of a catalyst, compound A or compound B and matt or gloss ink base.

c) the value of the performance characteristic outcome is called the *response*;

d) the change in the response as a result of a change in factor level is termed an *effect*.

12.1.2 Why experiment?

Experimentation has many practical uses. It enables one to determine how standards of performance, dependability, acceptability and affordability of products and services, processes, materials and mixtures are influenced by:

a) features of products and services (e.g. tolerances, nominal values);

b) parameters of processes (e.g. temperature, pressure);

c) properties of materials (e.g. hardness, machinability);

d) formulations of mixtures (e.g. of alloys, fuels, concrete, cloth).

Whilst experimentation plays a major role in problem solving, there is a need to progressively shift the emphasis to its integration in the mainstream activities of design and development. Genichi Taguchi [128] has proposed a two-step approach, which uses experimentation to "tune in" a basic prototype design, which he terms "parameter" design and "tolerance" design.

Parameter design is concerned with the identification and exploitation of three types of design factor:

— control factors: those that affect the variability of the response;

— signal factors: those that affect only the level of the response;

— null factors: those that do not materially affect either the variability or level of response.

Firstly, control factors are identified and adjusted to achieve design “robustness”. A robust design is one that is insensitive to so-called noise factors that are impossible, inconvenient or impractical to manage.

Examples of noise factors are: environmental, ambient temperature, humidity, vibration, supply tension and dust; deterioration, wear, drift and fatigue; and imperfections in manufacture, delivery or use, deviations from nominal.

Secondly, signal factors are adjusted to bring the response on target.

Thirdly, the null factors are adjusted to the most economic level.

The overall effect in identifying and setting nominal values of design factors in this way is to achieve optimal performance over a wide range of conditions with economy.

Tolerance design is concerned with specifying the most liberal tolerances and controls to meet a given performance. This is achieved by experimentation that seeks to take advantage of any non-linear relationship between factors and responses.

12.1.3 Where does statistics come in?

Today's statistical experimental designs emanate from R.A. Fisher's [90] work in England in the 1920s. Prior to this, it was deemed scientifically sound to conduct a multi-factor experiment by varying the level of one factor at a time, keeping the levels of all other factors constant. Fisher introduced the concept of a *factorial experimental design* in which all factors are varied simultaneously. The principal motivators for using such statistically designed experiments include:

a) increase in information for a given number of experimental runs, including the separation of main effects, interactions and experimental “noise”;

b) potential for cost and time savings through the reduction in the number of experimental runs required for a given effectiveness and the ability to plan and execute tests more efficiently;

c) ability to predict optimal combinations of factor levels even when they do not form part of the actual experimental plan;

d) ability to adopt a sequential rather than a one-shot approach;

e) relative ease of analysis and interpretation of the results.

12.1.4 What types of standard experimental designs are there and how does one make a choice of which to use?

12.1.4.1 Full factorial experiments

Full factorial experiments in the form of orthogonal (balanced) arrays are well suited for determining the extent to which the effect on the response of a change in level of a factor differs at different levels of other factors.

However, when the number of factors and/or their levels become large, the size of a full factorial experiment can become prohibitively large. For example, to test all combinations of 6 factors each at 4 levels would require a minimum of $4^6 = 4\ 096$ experimental runs. Additional runs would still be required to investigate variation in the response at each combination and to estimate experimental noise. In such an event, *fractional factorial designs* often provide an economic solution that is technically adequate, particularly in situations where higher order interactions or non-linearity can be safely ignored.

12.1.4.2 Fractional factorial experiments

12.1.4.2.1 General

Fractional factorial designs stem from the work of Tippett, Finney and Rao in the 1930s and 1940s. More recently, they have been popularized by Taguchi [128]. A number of orthogonal arrays are available together with simple, mainly pictorial, instructions for selection, application and analysis. The versatility of the most popular basic two-level orthogonal array is shown in Table 32. It is seen that if technical considerations indicate that some interactions are not likely to be important, then considerable economy in experimental effort is possible. At least a three-level design is required to investigate non-linearity.

The L8 design of Table 32 is a standard orthogonal (balanced) array with seven columns and eight rows. Factors A, B, C, etc. can be assigned to the columns. Factor levels are indicated by a 1 or a 2. In some texts, minus and plus signs are used instead. Each row indicates a combination of factor levels to run in the experiment. The design is such that four independent estimates can be made of the effect of each factor on the response, at each level, under different operating conditions of other factors. These four estimates can then be averaged for each factor level. This is illustrated in the design validation and development example given later within this subclause.

In using these factorial designs, a number of features need to be considered.

a) The statistical desirability of *randomizing* the run sequence to protect against bias due to factors not included in the experiment. For example, without randomization, take the situation if the first two runs of L8 were performed on Saturday morning, the next two on Saturday afternoon with the further four runs done similarly on Sunday. It would not be possible to separate out the day-to-day effect present in column 1 from the factor A effect. Statisticians would say the effects are confounded. Neither would it be possible to separate out the morning to afternoon effect in column 2 from the factor B effect. On the other hand, operational interests would prefer to retain the order given in Table 32 if some factor levels are more difficult to change than others. The most difficult factor to change would be put into column 1, which has the minimum number of changes and the easiest factor to change in column 4, which has the maximum number of changes. Hence the actual run order will often be based on a trade-off between statistical and operational considerations.

b) *Replication/repetition* of the experiment for each specified combination of factor levels. This is desirable for two principal reasons: one, to estimate the value of any noise or error and; two, to provide a measure of the variability of the response at each combination. The latter is required if the aim of the experiment is to optimize the response with minimum variation.

c) *Sequential experimentation*, as opposed to one-shot experiments. This is possible with the L8 design. This flexibility facilitates the building of knowledge as the experiment progresses in order to meet the experimental objectives with the minimum of effort and cost. For instance, if the L8 does not yield the information required, say, with four or seven factors, it may be extended into an L16, which has 15 factor columns and 16 runs.

12.1.4.2.2 Design validation and development example

This example shows an application of experimental design 2 of Table 32. It has two roles; one, as a design validation tool to determine the suitability of a sintered part for a particular application and two, as a development tool in the sense of searching for preferred operating conditions. Four design factors were investigated each at two levels as indicated in Table 33.

The experimental layout chosen uses columns 1, 2, 4 and 7 of a standard L8 array. Strength of fit, in kN, at minimum interference conditions was recorded for each part subjected to each experimental combination.

Three parts were used for each run in order to separate out means from variation in order to permit a search for design factors that would enhance mean strength (signal factors) and those that would reduce variation (control factors). Variation is expressed in terms of standard deviation. The results are shown in Table 34.

Table 32 — Alternative useful designs with the “Taguchi” (Lattice) L8 two-level array

L8 lattice	Column for factors						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
Run no. 4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2
Design 1:							
full factorial 3-factor							
design with all	A	B		C			
interactions isolated			AB		AC	BC	ABC
Design 2:							
4-factor design with	A	B		C			D
main effects clear of							
all 2-factor			AB		AC	BC	
interactions			CD		BD	AD	
Design 3:							
7-factor design with	A	B	C	D	E	F	G
each factor		AC	AB	AE	AD	AG	AF
confounded[a] with	BC			BF	BG	BD	BE
2-factor				CG	CF	CE	CD
interactions (only 2	DE	DF	DG				
factors shown)		EG	EF				
	FG						

[a] A factor is said to be confounded with another factor, or factors, when their separate effects cannot be isolated.

Table 33 — Sintered part design factors and their levels

Design factor	Level 1	Level 2
A. Surface finish	Fine turned	Microlled
B. Lubrication	Yes — number 2 oil	No
C. Speed	Low	High
D. Density	6,5	6,8