Designation: F3172 – 15 (Reapproved 2021)

Standard Guide for Design Verification Device Size and Sample Size Selection for Endovascular Devices[1]

This standard is issued under the fixed designation F3172; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This guide provides guidance for selecting an appropriate device size(s) and determining an appropriate sample size(s) (that is, number of samples) for design verification testing of endovascular devices. A methodology is presented to determine which device size(s) should be selected for testing to verify the device design adequately for each design input requirement (that is, test characteristic). Additionally, different statistical approaches are presented and discussed to help guide the developer to determine and justify sample size(s) for the design input requirement being verified. Alternate methodologies for determining device size selection and sample size selection may be acceptable for design verification.

1.2 This guide applies to physical design verification testing. This guide addresses in-vitro testing; in-vivo/animal studies are outside the scope of this guide. This guide does not directly address design validation; however, the methodologies presented may be applicable to in-vitro design validation testing. Guidance for sampling related to computational simulation (for example, sensitivity analysis and tolerance analysis) is not provided. Guidance for using models, such as design of experiments (DOE), for design verification testing is not provided. This guide does not address sampling across multiple manufacturing lots as this is typically done as process validation. Special considerations are to be given to certain tests such as fatigue (see Practice E739) and shelf-life testing (see Section 8).

1.3 Regulatory guidance may exist for endovascular devices that should be considered for design verification device size and sample size selection.

1.4 *Units*—The values stated in SI units are to be regarded as the standard. No other units of measurement are included in this standard.

1.5 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.6 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E739 Practice for Statistical Analysis of Linear or Linearized Stress-Life (*S-N*) and Strain-Life (ε-*N*) Fatigue Data

F2914 Guide for Identification of Shelf-life Test Attributes for Endovascular Devices

2.2 *ISO Standards:*[3]

ISO 14971:2012 Medical devices—Application of risk management to medical devices

3. Terminology

3.1 *Definitions:*

3.1.1 *attribute data, n*—data that identify the presence or absence of a characteristic (for example, good/bad or pass/fail).

3.1.2 *design input requirements, n*—physical and performance requirements of a device that are used as a basis for device design (typically defined as test characteristics such as balloon burst pressure, shaft tensile strength, and so forth).

3.1.3 *design output, n*—features of the device (that is, dimensions, materials, and so forth) that define the design and make it capable of achieving design input requirements.

3.1.4 *design subgroup, n*—set defined by the device sizes within the device matrix in which the essential design outputs do not vary for a specified design input requirement (that is, device sizes that share the same design for a specified design input requirement).

[1] This guide is under the jurisdiction of ASTM Committee F04 on Medical and Surgical Materials and Devices and is the direct responsibility of Subcommittee F04.30 on Cardiovascular Standards.

Current edition approved Aug. 1, 2021. Published August 2021. Originally approved in 2015. Last previous edition approved in 2015 as F3172 – 15. DOI: 10.1520/F3172-15R21.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

[3] Available from American National Standards Institute (ANSI), 25 W. 43rd St., 4th Floor, New York, NY 10036, http://www.ansi.org.

3.1.5 *design validation, n*—establishing by objective evidence that the device conforms to defined user needs and intended use(s).

3.1.6 *design verification, n*—confirmation by examination and provision of objective evidence that the device design (design output) fulfills the specified requirements (design input).

3.1.7 *device matrix, n*—entire range of available models/sizes for the device family.

3.1.8 *device size, n*—individual model/size (for example, 6 mm diameter by 25 mm length balloon on 135 cm length catheter or a 6Fr 100 cm length guide catheter).

3.1.9 *endovascular device, n*—device used to treat vascular conditions from within the vessel.

3.1.10 *essential design output, EDO, n*—design feature(s) or characteristic(s) of the device that affects its ability to achieve the design input requirements (that is, design output(s) that has a relevant effect on the test results).

3.1.11 *process validation, n*—establishment by objective evidence that a process consistently produces a result or device achieving its predetermined requirements.

3.1.12 *safety factor, n*—ratio of the device performance to the specification requirement (for example, how much stronger the device is than it needs to be to meet its specification requirement).

3.1.13 *sample size, n*—quantity of individual specimens of a device tested.

3.1.14 *variables data, n*—data that measure the numerical magnitude of a characteristic (how good/how bad).

4. Significance and Use

4.1 The purpose of this guide is to provide guidance for selecting appropriate device size(s) and determining appropriate sample size(s) for design verification of endovascular devices. The device size(s) and sample size(s) for each design input requirement should be determined before testing. The device size(s) selected for verification testing should establish that the entire device matrix is able to achieve the design input requirements. If testing is not performed on all device sizes, justification should be provided.

4.2 The sample size justification and statistical procedures used to analyze the data should be based on sound scientific principles and should be suitable for reaching a justifiable conclusion. Insufficient sample size may lead to erroneous conclusions more often than desired.

4.3 Guidance regarding methodologies for determining device size selection and appropriate sample size is provided in Sections 5 and 6.

5. Selection of Device Size(s)

5.1 Design input requirements are the physical and performance requirements of a device that are used as a basis for device design. Once the device design is defined, testing is typically performed to verify that the design input requirements are met. The appropriate device size(s) for verification testing should be determined for each design input requirement. Testing the same device size(s) is typically not appropriate to verify all design input requirements. Differences in the device design throughout the device matrix will drive which device size(s) is selected for verification of each design input requirement.

5.1.1 As explained in subsequent sections, when determining device size(s) for testing, the following should be considered for each design input requirement:

5.1.1.1 Essential design outputs,

5.1.1.2 Design subgroups, and

5.1.1.3 Other considerations.

5.2 *Define Essential Design Outputs (EDOs)*—The design outputs of the device are the features of the device (that is, dimensions, materials, and so forth) that define the design and make it capable of achieving design input requirements. Not all design outputs are essential for each design input requirement. Therefore, for each design input requirement, the essential design outputs (EDOs) should be identified. In Table 1, example EDOs for design input requirements of a balloon catheter device are provided.

5.3 *Define Design Subgroups:*

5.3.1 The design subgroups should be defined for each design input requirement based on the EDOs identified.

5.3.2 For a specific design input requirement, the design subgroups can be defined as one of the following:

5.3.2.1 The entire device matrix if the EDOs for the design input requirement are constant throughout the entire device matrix,

5.3.2.2 Subsets of the device matrix if the EDOs for the design input requirement vary in groups or stages throughout the device matrix, or

5.3.2.3 Each individual device size of the device matrix if EDOs for the design input requirement are different for each individual device size.

5.3.3 Fig. 1 represents the device matrix (entire range of available device sizes) for a 135 cm length balloon catheter device that has balloon diameters ranging from 3 to 7 mm and balloon lengths ranging from 10 to 50 mm. Balloon catheters are available in any combination of balloon diameter and length resulting in 25 unique device sizes in the device matrix.

5.3.4 Figs. 2-4 illustrate how the device matrix in Fig. 1 is defined by different design subgroups for different design input

TABLE 1 Example EDOs for Design Input Requirements for a Balloon Catheter Device

Design Input Requirement	EDOs
Manifold connection/Luer lockability	Luer thread dimensions Manifold material
Catheter shaft tensile strength for a single lumen catheter	Shaft material Shaft cross-sectional area (diameter and wall thickness) Shaft bond design
Balloon compliance (diameter versus pressure)	Balloon diameter Balloon material Balloon wall thickness
Balloon deflation time	Balloon volume Shaft deflation lumen design

Device Matrix		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm	x	x	x	x	x
	4 mm	x	x	x	x	x
	5 mm	x	x	x	x	x
	6 mm	x	x	x	x	x
	7 mm	x	x	x	x	x

FIG. 1 Device Matrix for a Balloon Catheter Device (25 unique device sizes)

Design Subgroup: EDOs Constant		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm	Manifold/Luer Lock "A"				
	4 mm					
	5 mm					
	6 mm					
	7 mm					

FIG. 2 Design Subgroup for Manifold Connection/Luer Lockability Testing (EDOs remain constant throughout the device matrix)

requirements. Fig. 2 represents a design subgroup that is defined by the entire device matrix because all device sizes share the same design for the specified design input requirement (that is, the EDOs remain constant for all device sizes). The design input requirement is manifold connection/luer lockability testing, and the EDOs (luer thread dimensions and manifold material) are the same for all sizes in the device matrix.

5.3.5 Figs. 3 and 4 represent design subgroups that are subsets of the device matrix because the EDOs for the design input requirement vary throughout the device matrix. Fig. 3 represents design subgroups for shaft tensile strength for a device that contains two different shaft designs in the device matrix, but the other EDOs that were identified (shaft material and shaft bond design) are the same for the entire device matrix. Therefore, there is a design subgroup that is defined by the device sizes that have shaft design "A" and a design subgroup that is defined by the device sizes that have shaft design "B." Fig. 4 represents design subgroups for balloon compliance in which each balloon diameter defines a unique design subgroup.

5.4 *Design Input Requirements and Other Considerations*—In addition to design subgroup definition, design input, device labeling, or regulatory requirements may make it necessary to test additional sizes.

5.5 *Device Size Selection Approach:*

5.5.1 *Approach*—Once the design subgroups are defined for a given design input requirement, the device size(s) to be tested for design verification testing can be appropriately selected by using one of the following approaches:

5.5.1.1 Test each design subgroup,

5.5.1.2 Test the worst-case design subgroup, or

5.5.1.3 Test a subset of the design subgroups.

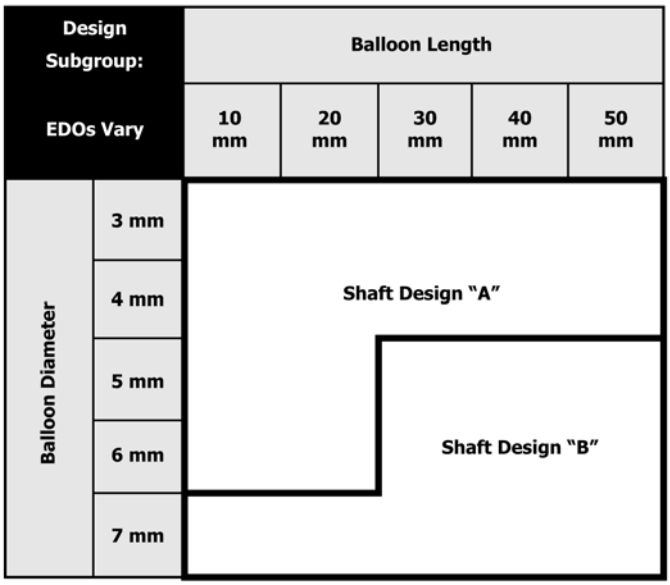


FIG. 3 Design Subgroups for Shaft Tensile (EDOs vary throughout the device matrix but are constant within each design subgroup)

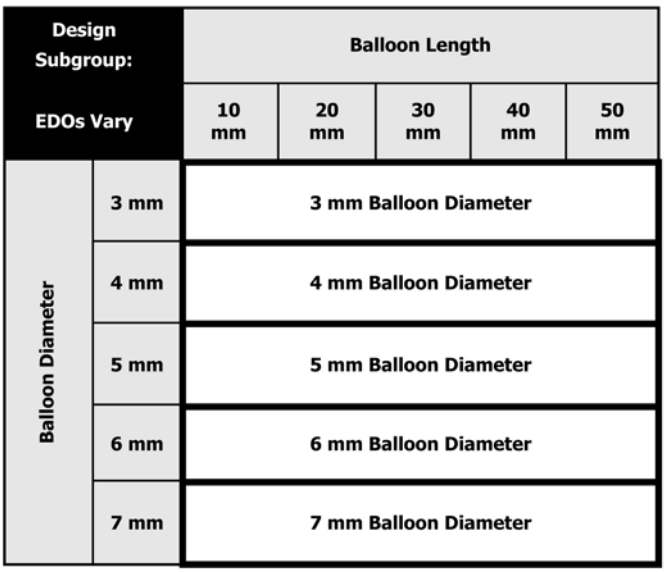


FIG. 4 Design Subgroups for Balloon Compliance (EDOs vary throughout the device matrix but are constant within each design subgroup)

5.5.2 *Test Each Design Subgroup:*

5.5.2.1 Depending on the design subgroup definition, testing each design subgroup may translate into testing one device size or multiple device sizes to verify the entire device matrix.

5.5.2.2 When the design subgroup is defined by the entire device matrix and the requirement is the same throughout the device matrix, any device size may be selected for verification testing to represent the entire device matrix. This approach is appropriate since all device sizes share the same design for the specified design input requirement (that is, the EDOs are the same for all device sizes). Fig. 5 illustrates the design subgroup and example device size selection for verification testing for manifold connection/luer lockability. Since any device size represents the entire device matrix, factors such as device sizes used for other testing to minimize total test units or device size with the highest sales volume may be considered.

5.5.2.3 When the design subgroups are defined by subsets of the device matrix, a device size should be selected from within each design subgroup to verify the design adequately since EDOs vary throughout the device matrix. Fig. 6 illustrates the design subgroups and example device sizes selected for verification testing for shaft tensile strength. Note that the shaft tensile strength requirement is the same for all device sizes and the other EDOs identified (shaft material and shaft bond design) are the same for all device sizes.

EDOs: Manifold Material and Luer Lock		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm	Manifold/Luer Lock "A"				
	4 mm					
	5 mm					
	6 mm					
	7 mm					

Verification Device Sizes		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm					
	4 mm		X			
	5 mm					
	6 mm					
	7 mm					

FIG. 5 Example Design Subgroup and Verification Device Size Selection for Manifold/Luer Lockability Testing

EDO: Shaft Design		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm	Shaft Design "A"				
	4 mm					
	5 mm					
	6 mm					
	7 mm	Shaft Design "B"				

Verification Device Sizes		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm					
	4 mm					X
	5 mm					
	6 mm				X	
	7 mm					

FIG. 6 Example Design Subgroups and Verification Device Size Selection for Shaft Tensile Strength

		Stent Fatigue Safety Factor	Worst-Case Size for Verification Test
Stent Diameter	3 mm	2.0	
	4 mm	2.1	
	5 mm	1.9	X
	6 mm	2.2	

FIG. 7 Worst-Case Size May Be Selected Based on a Safety Factor Calculation

5.5.2.4 An alternate approach to selecting one device size to represent each design subgroup would be to pool multiple sizes within a design subgroup for testing. Refer to Section 7 for more information on data pooling.

5.5.3 *Test the Worst-Case Design Subgroup:*

5.5.3.1 For certain design input requirements, testing only the worst-case design subgroup adequately verifies the entire device matrix. The worst-case design subgroup is determined by considering how the EDOs impact performance to the design input requirements. If the design input requirement limit varies throughout the device matrix (for example, different rated burst pressure (RBP) requirements for different diameter balloon catheters), a worst case could be tested for each specification limit or one worst-case subgroup could be tested by performing a worst-case analysis that accounts for the differences in the specification limits, such as a safety factor

EDO: Shaft Deflation Lumen		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm	Shaft Design "A"				
	4 mm					
	5 mm					
	6 mm					
	7 mm	Shaft Design "B"				

EDO: Balloon Volume (Length and Diameter)		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm	Volume "A"	Volume "B"	Volume "C"	Volume "D"	Volume "E"
	4 mm	Volume "F"	Volume "G"	Volume "H"	Volume "I"	Volume "J"
	5 mm	Volume "K"	Volume "L"	Volume "M"	Volume "N"	Volume "O"
	6 mm	Volume "P"	Volume "Q"	Volume "R"	Volume "S"	Volume "T"
	7 mm	Volume "U"	Volume "V"	Volume "W"	Volume "X"	Volume "Y"

FIG. 8 Example Design Subgroups to Consider for Balloon Deflation Time

Verification Device Sizes		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm					
	4 mm					X
	5 mm					
	6 mm					
	7 mm					X

FIG. 9 Xs Represent Worst Case (Largest Balloon Volume) Within Each Shaft Design Subgroup (these are the device sizes selected to verify that the entire device matrix can achieve the design input requirement)

calculation. Additionally, if the design input requirement has both an upper and a lower specification limit, there may be a worst case for the upper specification and a different worst case for the lower specification.

5.5.3.2 Testing the worst-case design subgroup is a commonly used verification method when EDOs vary throughout the device matrix and their impact to the design input performance is well understood/defined (for example, increasing diameter has a negative impact on achieving the design input requirement and decreasing the diameter has a positive impact on achieving the design input requirement).

5.5.3.3 The worst-case design subgroup may be determined by one of the following methods:

(1) Historical data (similar predicate device or development characterization of current device) or

(2) Engineering judgment, analysis, computational simulation, or safety factor calculation.

NOTE 1—While the engineering or computational analysis, or both, may be applied to determine the worst-case size selection, additional considerations that could impact which device size to test may exist. For example, manufacturing process variations between device sizes could result in an actual worst-case device size that is different than the theoretical worst case. Additionally, the assembly of a multi-component device could result in failures that would not be predicted by an engineering analysis applied to only one component of the device. Use of historical knowledge of failures can be used to justify whether these factors should be considered in the device size selection. The following

		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm	X				X
	4 mm					
	5 mm					
	6 mm					
	7 mm	X				X

FIG. 10 A 2-by-2 Factorial May Be Selected for Evaluation When One Device Size Does Not Represent the Entire Device Matrix or a Worst-Case Device Size is Not Known

are a couple examples of types of analysis to determine worst case:

(a) Hoop stress calculation—The highest balloon hoop stress may represent the worst-case situation for balloon burst testing when it is known that the finished device always fails in the balloon.

(i) By using a thin-walled pressure vessel assumption, the hoop stress of a cylindrical balloon could be calculated by:[4]

$$Hoop\ Stress = \frac{P*D}{(2 * T)} \quad (1)$$

where:

P = pressure (rated burst pressure (RBP) design input requirement),
D = diameter (EDO), and
T = wall thickness (EDO).

(ii) By using the rated burst specification requirement for P in the hoop stress formula, the worst-case size (that is, the size with the highest hoop stress at rated burst pressure) can be calculated.

(b) Fatigue safety factor calculation—Appropriately validated finite element analysis may be used on each implant diameter or other relevant property (for example, design platform, length) to determine the fatigue safety factor as well as the critical stress and strain values and locations. The predicted stresses and strains are compared to the fatigue life line to determine the fatigue safety factor. The implant with the lowest fatigue safety factor may be tested as the worst case in design verification (see Fig. 7 for a stent example).

5.5.3.4 Balloon deflation time is an example of a design input requirement for which a worst-case design subgroup approach may be acceptable to verify the entire device matrix. The EDOs defined for deflation time are balloon volume and shaft deflation lumen design. For the example in Fig. 8, the balloon volume and the shaft deflation lumen design both vary throughout the device matrix; therefore, there are multiple design subgroups that should be considered when selecting the device size(s) for testing. Fig. 8 illustrates the design subgroups to consider for balloon deflation time testing (two different shaft design subgroups and 25 different balloon volume design subgroups).

5.5.3.5 Since the relationship between deflation time and balloon volume for a constant shaft design is well understood (that is, the larger the balloon volume, the longer the deflation time), a worst-case approach can be used to verify each shaft design subgroup. Fig. 9 illustrates that the worst-case balloon volume device is selected within each shaft design to verify that the entire device matrix has acceptable deflation times.

5.5.3.6 Note that if the design input requirement for deflation time is not the same for all balloon sizes, then additional sizes may need to be tested to verify the worst case for each specification requirement.

5.5.3.7 Other examples of tests that may rely on the worst-case device size rationale for selecting the device sizes for testing are the following: accelerated durability, particulate generation, corrosion, and magnetic resonance imaging (MRI) compatibility. The rationale and device size for each test is different because each test evaluates a different aspect of device performance.

5.5.4 *Test a Subset of the Design Subgroups*—For certain design input requirements, a subset of the design subgroups may be required for verification testing. This approach may be used when EDOs vary throughout the device matrix and a worst-case device size is not known. For example, a two-by-two factorial (Fig. 10) of the largest and smallest diameters and lengths may be an approach to device size selection to capture the performance at the corners of the design space.

6. Statistical Approaches for Sample Size Determination

6.1 Once the device size(s) has been selected for verification testing per the methodology presented in Section 5, the sample size needs to be defined. The sample size justification and

[4] Hibbele, R. C., *Mechanics of Materials*, Third Edition, 1997.

statistical procedures used to analyze the data are to be based on sound scientific principles and suitable for reaching a justifiable conclusion. An insufficient sample size may lead to erroneous conclusions more often than desired.

6.2 This section provides an overview of determining statistically based sample sizes for the following design verification test methodologies:

6.2.1 Attribute testing to a predefined specification (pass/fail), and

6.2.2 Sampling by variables for proportion nonconforming.

6.3 If a predefined specification is not initially available, a comparison to a predicate or similar device may be performed to justify acceptability. It is recommended that before design verification, the predicate or similar device should be characterized and a specification limit should be defined based on that characterization. Once the specification limit is defined, either the attribute or variable testing approaches can be used to verify the design. In Table 2, some of the considerations/limitations of the approaches discussed in this guide are summarized.

6.3.1 *Risk and Sample Size*—Selection of a sample size is a three-step process:

6.3.1.1 First, determine the risk level based on the perceived risk to the patient as a result of the failure of the specific design input requirement. Risk level may be defined as the combination of the severity outcome of the device failure and the likelihood of that failure happening (occurrence). For additional detail regarding how to determine the risk level, refer to ISO 14971.

6.3.1.2 Next, determine the confidence and reliability (proportion of the population) levels based on the risk level. The risk level should point to a specific confidence and reliability combination that should be demonstrated by the output. Typically, the higher the risk level associated with failing the design input requirement, the higher the confidence and reliability combination required. This confidence and reliability combination defines a statistical tolerance limit. For example, one can state that at least 99 % of the population needs to be below the output's upper specification limit and then calculate the 99th percentile based on a sample of 30 values. However, this is only a point estimate since there is still uncertainty in the true location of the 99th percentile because of the random draw of the samples. Therefore, an upper confidence limit on the estimate of the 99th percentile is needed, which will take into account the sample size and provide a margin of error on the percentile estimate in the direction of the specification limit(s). Placing a confidence limit on this percentile estimate creates a statistical tolerance limit.

6.3.1.3 Finally, select a sampling plan that meets the tolerance limit requirement. There are usually many different sampling plans that can satisfy the same tolerance limit requirement. Assuming the true population characteristic meets the percent tolerance limit, the selection of the sampling plan is a trade-off between efficiency (lower sample size) and likelihood of passing (higher sample size). Therefore, the best practice is to select a sampling plan that efficiently provides evidence that the product meets tolerance requirements and is unlikely to give false conclusions.

TABLE 2 Considerations/Limitations of the Approaches in This Guide

Approach	Advantages	Limitations
Attribute testing (to a predefined specification)	Data distribution model does not impact results (that is, non-normal data can be assessed as attribute).	Sample sizes tend to be larger than when using variable assessments. If continuous data are not obtained, the safety factor is unknown unless the test is run at multiple levels of severity.
Variable testing (to a predefined specification limit)	For a given confidence/reliability, sample sizes tend to be smaller than when using attribute assessments. Variable assessments allow for a predictive model of the entire population performance.	Sample size or development time may increase if the data contain unexpected outliers which do not permit an acceptable distribution fit.

6.4 *Attribute Testing:*

6.4.1 *Description*—When data are assessed as attribute, each unit returns a result of pass or fail. The data collected may be binary data (for example, successful or not successful) or variable data individually assessed against the criteria (for example, is the measured value greater than the specified limit). This section provides guidance on how to determine a sample size for attribute data to establish a desired confidence and reliability level.

6.4.2 *Sample Size Determination:*

6.4.2.1 For attribute data assessment, sample sizes are chosen to demonstrate passing a predetermined specification at a desired confidence and reliability level. Acceptance criteria for attribute sampling plans include a sample size (n) and a number of allowable failures (a). In a single-stage sampling plan when $a = 0$, the number of samples needed for testing may be calculated using:[5]

$$n = \frac{\ln(1 - \text{confidence})}{\ln(\text{reliability})} \tag{2}$$

6.4.2.2 Additional variations of attribute sampling plans include single-stage plans in which the acceptance number *(a)* > 0 and multi-stage plans in which the potential outcomes of a given stage include passing the plan, failing the plan, or taking additional samples per the prescribed plan so that the total probability of acceptance meets the quoted confidence level. In both of these variations, the test can result in a pass even when failures are observed, which seems counterintuitive. However, the statistics behind the sampling plan demonstrate that the plan will be more likely to pass populations that achieve the performance requirements and more likely to fail populations that do not achieve the performance requirements. Reference Appendix X1 for a more in-depth explanation of sampling plans where $a > 0$. Table 3 provides an example of an attribute sample size calculation.

[5] Rothbart, H. A., *Mechanical Design and Systems Handbook*, McGraw-Hill, New York, 1985.

TABLE 3 Example of an Attribute Sample Size Calculation

Design Input Requirement	Data Output	Confidence[A]	Reliability[A]	Sample Size Acceptance	Sample Size Calculation[B]
Device shall withstand 20 cycles to RBP	Pass/fail (20 inflations met)	95 %	90 %	a = 0 plan: n = 29, a = 0 a >0 plans, for example, n = 45, a = 1	n = ln (1-0.95) / ln (0.90) Use software package to calculate

[A] Confidence and reliability recommended from *Guidance for Industry and FDA Staff: Class II Special Controls Guidance Document for Certain Percutaneous Transluminal Coronary Angioplasty (PTCA) Catheters*, Sept. 8, 2010.
[B] Sample size equation referenced from Rothbart, Harold A., *Mechanical Design and Systems Handbook*, McGraw-Hill, New York, 1985.

6.5 *Variable Testing:*

6.5.1 *Description*—Variables data consist of measuring the numerical magnitude of a characteristic. Common methods of assessing variable data include statistical tolerance limits and process performance indices (Ppk and Pp), which can both be defined by the combination of a confidence level and reliability level. This section provides guidance on how to assess variables data against a predetermined specification requirement.

6.5.2 *Sample Size Determination*—Since variables data contain more information per data point than attribute data, variables sampling plans typically require fewer samples than attribute sampling plans to achieve the same level of confidence and reliability. In fact, there are virtually no restrictions on the number of samples needed for variable data. However, normality testing typically requires a minimum sample size to provide enough power for the test, and the probability of passing a given confidence and reliability combination does increase with sample size, as described in 6.3.1.

6.5.3 *Normality Testing:*

6.5.3.1 The normal distribution is a common statistical distribution for modeling variable data and is known for its diverse applications. Tolerance limits tables are based on the normal distribution, so a good practice is to test whether the sample data adequately fit a normal distribution or whether the normal distribution would at least be a conservative model to use. Many normality tests are available in commercial statistics packages. If the data pass the normality test, the tolerance limit will provide a good, buffered estimate of the quoted percentile of the population. If the data fail to fit a normal distribution, options include:

(1) Test the fit to a different statistical distribution,

(2) Assume a normal distribution if the sample skewness is away from the specification limit,

(3) Apply a normalizing transformation to the data, or

(4) Analyze the sample as attribute data (often the choice of last resort since more samples may be needed).

6.5.3.2 Test data populations within the realm of engineering processes follow a number of different distributions besides the normal distribution. Since various distributions carry differing amounts of mass in the distribution tails, this can drastically impact the accuracy of the tolerance limit estimate. Therefore, characterization of the sample data's distribution may be necessary with the understanding that the normal distribution is not always the correct model. Consider the examples in Table 4.

6.5.3.3 If test sample sizes are not large enough to support distribution analysis, the following options may be used to provide support:

(1) Historical data review,

(2) Literature research, and

(3) Engineering rationale (for example, physics based).

6.5.4 *Statistical Tolerance Limits:*

6.5.4.1 When assessing variables data using tolerance limits and assuming normality, the following equations are used to determine whether data pass the acceptance criteria. The sampling plan passes if:

(1) One-sided:

$$\mathrm{UTL} = \mathrm{Avg} + k_1 * s \leq \mathrm{USL} \tag{3}$$

or

$$\mathrm{LTL} = \mathrm{Avg} - k_1 * s \geq \mathrm{LSL} \tag{4}$$

(2) Two-sided:

$$\mathrm{UTL} = \mathrm{Avg} + k_2 * s \leq \mathrm{USL} \text{ and } \mathrm{LTL} = \mathrm{Avg} - k_2 * s \geq \mathrm{LSL} \tag{5}$$

where:

Avg = sample average,
s = sample standard deviation,
LSL and USL = lower specification limit and upper specification limit, respectively,
LTL and UTL = lower tolerance limit and upper tolerance limit, respectively,
k_1 = one-sided k-factor = f (confidence, reliability, sample size) given one tail, and
k_2 = two-sided k-factor = f (confidence, reliability, sample size) given two tails.

6.5.4.2 Table 5 provides some commonly used k-factors for when the population standard deviation is unknown, that is, the standard deviation is estimated from the collected data. There is a different k-factor table for situations in which the population standard deviation is known, but prior knowledge of the standard deviation is extremely rare at the design verification stage of the product life cycle. Note that the k-factor decreases as the sample size increases as a result of reduced uncertainty of the quoted percentile's value. A lower confidence or reliability level also lowers the required k-factor. Table 6 provides an example of a variable sample size calculation for a tolerance limit.

6.5.4.3 If the sample does not fit a normal distribution, the user may choose to normalize the data using traditional transformation tools such as the Box-Cox or Johnson transformations. The tolerance limits are then calculated using the normalized data and compared against the specification limits that have been transformed using the same normalization function. The Box-Cox method uses a power series transformation to normalize the data, and the Johnson transformation

TABLE 4 Skewed Populations

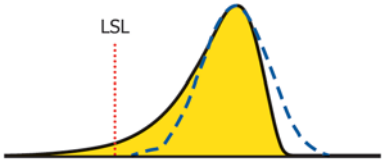
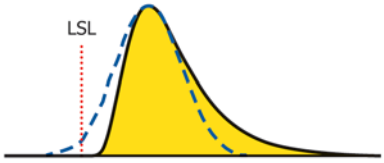
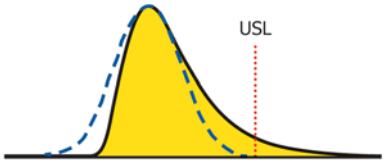
Population Description Relative to the Normal Distribution	Spec Limit	True Condition	Outcome when Assuming Normality
Long heavy tail to the left 	Lower	Should fail	Pass—the normal distribution underestimates lower tail mass
Long heavy tail to the right 	Lower	Should pass	Fail—the normal distribution overestimates lower tail mass
Long heavy tail to the left 	Upper	Should pass	Fail—the normal distribution overestimates upper tail mass
Long heavy tail to the right 	Upper	Should fail	Pass—the normal distribution underestimates upper tail mass

TABLE 5 Commonly Used *k*-Factors

		One-Sided *k*-Factors[A]			
Confidence	90 %	95 %	95 %	95 %	95%
Reliability	90 %	90 %	95 %	99 %	99.9%
n = 10	2.066	2.355	2.911	3.981	5.203
n = 20	1.765	1.926	2.396	3.295	4.318
n = 50	1.559	1.646	2.065	2.862	3.766
n = 120	1.452	1.503	1.899	2.649	3.495
		Two-Sided *k*-Factors[A]			
Confidence	90 %	95 %	95 %	95 %	95%
Reliability	90 %	90 %	95 %	99 %	99.9%
n = 10	2.546	2.856	3.393	4.437	5.640
n = 20	2.158	2.319	2.760	3.621	4.616
n = 50	1.918	1.999	2.382	3.129	3.995
n = 120	1.805	1.851	2.206	2.899	3.703

[A] The *k*-factor values for 90, 95, and 99 % reliability levels from Gerald J. Hahn and William Q. Meeker, *Statistical Intervals: A Guide for Practitioners*, John Wiley and Sons, 1991. The *k*-factor values for 99.9 % reliability level from Robert E. Odeh and D. B. Owen, *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*, ProQuest Co., Ann Arbor, MI, 2002.

uses a more complex formula to normalize the data. Refer to Sleeper (2006)[6] for instructions and examples using each of these transformations.

[6] Sleeper, A., *Design for Six Sigma Statistics: 59 Tools for Diagnosing and Solving Problems in DFSS Initiatives*, McGraw-Hill, New York, 2006.

6.5.5 *Process Capability Indices:*

6.5.5.1 Process capability indices describe the relationship between process (or product) variation and the specification limits. The metrics Cp and Cpk are process capability indices that account for only within-lot variation, whereas the metrics Ppk and Pp account for the total variation, consisting of both

TABLE 6 Example: Variable Sample Size Calculation for a Tolerance Limit (that is, *k*-Factor)

Design Input Requirement	Data Output	Lower Spec Limit (LSL)	Confidence/ Reliability[A]	Test Parameters[B]	Acceptance Calculation[C]
Device shall withstand inflation to RBP	Burst pressure (each test unit is inflated until it bursts)	RBP	95/99.9 %	k_1 = 4.318 n = 20	LTL = Avg – 4.318 * $s \geq$ RBP

[A] Confidence and reliability requirements from *Guidance for Industry and FDA Staff: Class II Special Controls Guidance Document for Certain Percutaneous Transluminal Coronary Angioplasty (PTCA) Catheters*, Sept. 8, 2010.
[B] The value for the *k*-factor is from Robert E. Odeh and D. B. Owen, *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*, ProQuest Co., Ann Arbor, MI, 2002.
[C] LTL shall be ≥RBP for the data set to pass the acceptance criteria. Avg is the data set average and *s* is the data set standard deviation.

TABLE 7 Example Variable Sample Size Calculation for Ppk

Design Input Requirement	Data Output	Confidence/ Reliability[A]	Test Parameters[B]	Ppk Calculation	Ppk Requirement
Device shall withstand inflation to RBP	Burst pressure (each test unit is inflated until it bursts)	95/99.9 %	n = 15 k = 4.607	Ppk = 4.607 / 3	Ppk ≥ 1.54

[A] Confidence and reliability requirements from *Guidance for Industry and FDA Staff: Class II Special Controls Guidance Document for Certain Percutaneous Transluminal Coronary Angioplasty (PTCA) Catheters*, Sept. 8, 2010.
[B] The value for the *k*-factor is from Robert E. Odeh and D. B. Owen, *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*, Marcel Dekker, Inc., 1980.

within-lot and between-lot variation. The metric Ppk describes how close the sample average is to the specification limit, relative to the total variation.

6.5.5.2 For a normal distribution with one specification limit:

$$\mathrm{Ppk}_{\mathrm{estimated}} = \frac{\mathrm{USL} - \mathrm{Avg}}{3*s} \text{ or } \frac{\mathrm{Avg} - \mathrm{LSL}}{3*s} \tag{6}$$

where:

Avg = sample average,
s = sample standard deviation, and
LSL and USL = lower specification limit and upper specification limit, respectively.

6.5.5.3 The metric Pp is calculated when the specification limit is two-sided. Pp describes the ratio of the specification range over the total variation. For a two-sided specification assuming a normal distribution, the equations for Ppk and Pp are:

$$\mathrm{Ppk}_{\mathrm{estimated}} = \mathrm{Min}\left[\frac{\mathrm{USL} - \mathrm{Avg}}{3*s}, \frac{\mathrm{Avg} - \mathrm{LSL}}{3*s}\right] \tag{7}$$

$$\mathrm{Pp}_{\mathrm{estimated}} = \frac{\mathrm{USL} - \mathrm{LSL}}{6*s} \tag{8}$$

6.5.5.4 In a variables acceptance sampling plan for a one-sided specification, Eq 9 serves as the acceptance criteria.[7] The sampling plan passes if:

$$\mathrm{Ppk}_{\mathrm{estimated}} \geq \mathrm{Ppk}_{\mathrm{required}} \tag{9}$$

where:

$$\mathrm{Ppk}_{\mathrm{required}} = \frac{k_1}{3} \tag{10}$$

where:

k_1 = one-sided *k*-factor = f (confidence, reliability, sample size) given one tail.

6.5.5.5 This *k*-factor for k_1 is the same parameter used for a one-sided statistical tolerance limit in the preceding section. Table 7 provides an example of a one-sided variables sampling plan calculation for Ppk assuming a normal distribution.

6.5.5.6 In a variables acceptance sampling plan for a two-sided specification, there are two acceptance criteria to meet. The sampling plan passes if:

$$\mathrm{Ppk}_{\mathrm{estimated}} \geq \mathrm{Ppk}_{\mathrm{required}} \tag{11}$$

$$\mathrm{Pp}_{\mathrm{estimated}} \geq \mathrm{Pp}_{\mathrm{required}} \tag{12}$$

where:

$$\mathrm{Ppk}_{\mathrm{required}} = \frac{k_2}{3} \tag{13}$$

$$\mathrm{Pp}_{\mathrm{required}} = \frac{1}{6*\mathrm{MSD}} \tag{14}$$

where:

k_2 = two-sided *k*-factor = f (confidence, reliability, sample size) given two tails, and
MSD = maximum standard deviation = f (confidence, reliability, sample size).[8]

6.5.5.7 Unfortunately, there is not a closed-form solution for MSD, although it may be calculated using appropriate software packages. Note that this k_2 value is slightly smaller than the *k*-factor used earlier in the two-sided statistical tolerance limit. This is because the tolerance limit approach demonstrates with the quoted confidence that a quoted proportion of the product is within the calculated tolerance interval, whereas the process capability approach demonstrates with the quoted confidence that a quoted proportion of the product is within specification. In other words, the tolerance limit approach rigidly places half of the quoted proportion on each side of the sample mean and

[7] Taylor, W. A., *Distribution Analyzer Software User's Guide (Version 1.2)*, 2006, http://www.variation.com/files/da/da12man.pdf and Negrin, I., Parmet, Y., and Schechtman, E., "Developing a Sampling Plan Based on Cpk—Unknown Variance," *Qual. Reliab. Engng. Int.*, doi: 10.1002/qre.1094, Vol 27, 2011, pp. 3–14.

[8] Schilling, E. G. and Neubauer, D. V., *Acceptance Sampling in Quality Control*, Second Edition, Chapman and Hall/CRC, Boca Raton, FL, 2009.

compares this endpoint against the specification, whereas the process capability approach is more flexible since it is able to compare the total quoted proportion against the specification limits. As a result, sampling plans using the process capability approach are slightly more likely to pass than sampling plans using tolerance limits, especially if a small percentage is out in one tail and there is a negligible percentage in the other tail.

6.5.5.8 Estimates for Ppk and Pp of non-normal distributions may be calculated using a variety of statistical software packages. Table 8 lists Ppk and Pp values for commonly used two-sided tolerance limits.

6.5.5.9 Eq 6-8 are only appropriate for normally distributed populations, since the mean and standard deviation are not useful predictors of tail probabilities for non-normal distributions. When a population follows a non-normal distribution, two methods are widely accepted to estimate Pp and Ppk:

(1) First, the user may use a modified normal capability metric, such as the ISO method or the Bothe percentage method.[9] The ISO method is based on specific percentiles in the distribution tails, and the Bothe method employs the Z-score of each specification limit. Refer to Sleeper (2007) for a more detailed discussion of the ISO and Bothe methods.

(2) Second, the user may choose to transform the data to follow a normal distribution using the Box-Cox or Johnson transformations, and then calculate Pp and Ppk using the transformed data and specification limits.[6] Refer to Sleeper (2006) for instructions and examples using each of these transformations.

[9] Sleeper, A., *Six Sigma Distribution Modeling*, McGraw-Hill, New York, 2007.

7. Data Pooling

7.1 Data pooling on multiple device sizes and/or models can be justified when EDOs for a given design input requirement are the same for the device sizes and/or models being pooled. Data pooling allows the total sample size across multiple device sizes and/or models to meet the required sample size as opposed to each individual device size and/or model meeting the required sample size. Fig. 11 illustrates the differences in sample size between pooling four device sizes and testing four device sizes individually.

8. Shelf Life

8.1 The methodologies presented in Sections 5 and 6 may be applicable to shelf-life testing. If there is sufficient knowledge of the device performance at time zero and shelf-life conditions, a reduced number of device sizes and reduced sample size may be appropriate to verify the device shelf life adequately. Additionally, all design input requirements may not need to be tested at shelf life (see Guide F2914).

9. Keywords

9.1 attribute; normality; sample size; sampling plan; statistics; variables; verification

TABLE 8 Ppk, Pp Values for Commonly Used Two-Sided Tolerance Limits[A]

Confidence	90 %	90%	90%	95 %	95 %	95 %	95%
Reliability	90 %	95%	99%	90 %	95 %	99 %	99.9%
$n = 10$	0.70, 0.74	0.86, 0.87	1.18, 1.18	0.79, 0.81	0.97, 0.97	1.33, 1.33	1.73, 1.73
$n = 15$	0.64, 0.71	0.79, 0.84	1.07, 1.08	0.70, 0.76	0.86, 0.90	1.17, 1.17	1.53, 1.55
$n = 20$	0.61, 0.69	0.75, 0.81	1.02, 1.05	0.66, 0.73	0.81, 0.87	1.11, 1.13	1.43, 1.45
$n = 25$	0.58, 0.67	0.72, 0.79	0.99, 1.04	0.63, 0.71	0.78, 0.84	1.06, 1.10	1.38, 1.39
$n = 30$	0.57, 0.66	0.71, 0.79	0.97, 1.02	0.61, 0.70	0.75, 0.82	1.03, 1.07	1.34, 1.36
$n = 35$	0.55, 0.65	0.69, 0.77	0.96, 1.01	0.59, 0.68	0.74, 0.81	1.01, 1.06	1.32, 1.35
$n = 40$	0.55, 0.65	0.68, 0.76	0.94, 1.00	0.58, 0.67	0.72, 0.80	0.99, 1.04	1.30, 1.33
$n = 50$	0.53, 0.63	0.67, 0.76	0.92, 0.98	0.56, 0.66	0.70, 0.78	0.96, 1.02	1.26, 1.30
$n = 60$	0.52, 0.62	0.66, 0.75	0.91, 0.97	0.55, 0.65	0.69, 0.77	0.95, 1.01	1.24, 1.28
$n = 80$	0.51, 0.62	0.64, 0.73	0.89, 0.96	0.53, 0.63	0.66, 0.75	0.92, 0.99	1.21, 1.26
$n = 100$	0.50, 0.61	0.63, 0.72	0.88, 0.95	0.52, 0.63	0.65, 0.74	0.90, 0.97	1.19, 1.24
$n = 150$	0.48, 0.59	0.61, 0.71	0.85, 0.93	0.50, 0.61	0.63, 0.72	0.88, 0.95	1.16, 1.21
$n = \infty$	0.43, 0.55	0.55, 0.65	0.78, 0.86	0.43, 0.55	0.55, 0.65	0.78, 0.86	1.03, 1.10

[A] Pp and Ppk values generated using Sampling Plan Analyzer software, v2.0.

Device Sizes Pooled for Testing (n = 20)		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm	5				5
	4 mm					
	5 mm					
	6 mm					
	7 mm	5				5

Device Sizes Tested Individually (n = 20)		Balloon Length				
		10 mm	20 mm	30 mm	40 mm	50 mm
Balloon Diameter	3 mm	20				20
	4 mm					
	5 mm					
	6 mm					
	7 mm	20				20

FIG. 11 Four Device Sizes Are Pooled Together to Meet the Sampling Plan Requirements of n = 20 versus Testing Four Device Size Individually to Meet the Sampling Plan Requirements of n = 20

APPENDIX

(Nonmandatory Information)

X1. SAMPLING PLANS WITH $a > 0$

X1.1 A common misconception surrounding attribute sampling plans is that plans with $a > 0$ offer less protection than $a = 0$ plans. In this appendix, it will be explained why $a > 0$ plans actually offer more protection than $a = 0$ plans. For purposes of simplicity, in this appendix, $a = 0$ and $a = 1$ options will be specifically compared for an attribute sampling plan with 90 % confidence and 90 % reliability. Table X1.1 shows the $a = 0$ plan appears to be the better choice since it offers the possibility of zero defectives in the population.

X1.2 The two plans can also be compared using an operating characteristic (OC) curve, as shown in Fig. X1.1. OC curves present the probability of passing a sampling plan as a function of the population defective rate. Intuitively, when the defective rate is very low, the probability of passing a sampling plan will be near 100 %, but as the defective rate increases, the probability of passing will dwindle until the plan is very unlikely to pass. A perfect sampling plan is depicted by the step function in Fig. X1.1 in which the sampling plan has a 100 % probability of passing when the defective rate is below the maximum acceptable rate and a 0 % chance of passing above the maximum acceptable rate. The perfect sampling plan has perfect information, knowing the exact defective rate of the population, so it is able to make perfect decisions.

X1.3 Unfortunately, given the random nature of sampling, sometimes populations with defective rates below the threshold will fail the sampling plan, and populations with defective rates higher than the threshold will pass. Fig. X1.1 demonstrates the $a = 1$ plan to be closer to the perfect sampling plan than the $a = 0$ plan. The $a = 1$ plan has a higher probability of passing below the maximum allowable defective rate (fewer false alarms) and a lower probability of passing above the maximum allowable rate (fewer escapes). The $a = 1$ OC curve performs better than the $a = 0$ curve because it has more information about the population due to its higher sample size.

TABLE X1.1 Lower Confidence Bound Example

Sampling Plan	n = 22, a = 0	n = 38, a = 1
90 % Lower confidence bound	0 %	0.28 %

X1.4 Another disadvantage of an $a = 0$ sampling plan is that it can lead to a false sense of security. Passing an $a = 0$ plan can lead the user to believe there are zero defectives in the population regardless of how low the sample size is even when zero defectives might be highly unlikely. An $a = 1$ plan acknowledges the potential existence of defective units and plans for their potential appearance in the sample. That being said, note that a sampling plan with $a > 0$ may not be acceptable for a design output possessing a "zero tolerance" status for defectives, even though the $a > 0$ plan provides more information about the population. In this scenario, the full knowledge that a defective unit has been observed may deem the design unacceptable for release.

X1.5 In summary, here is what can be stated about $a = 0$ plans:

X1.5.1 Fewest samples needed,

X1.5.2 Lowest chance of passing when the true defective rate is lower than the maximum allowable defective rate,

X1.5.3 Highest chance of passing is when the true defective rate is higher than the maximum allowable defective rate, and

X1.5.4 Can lead the user to believe there are zero defectives in the population regardless of how low the sample size is.

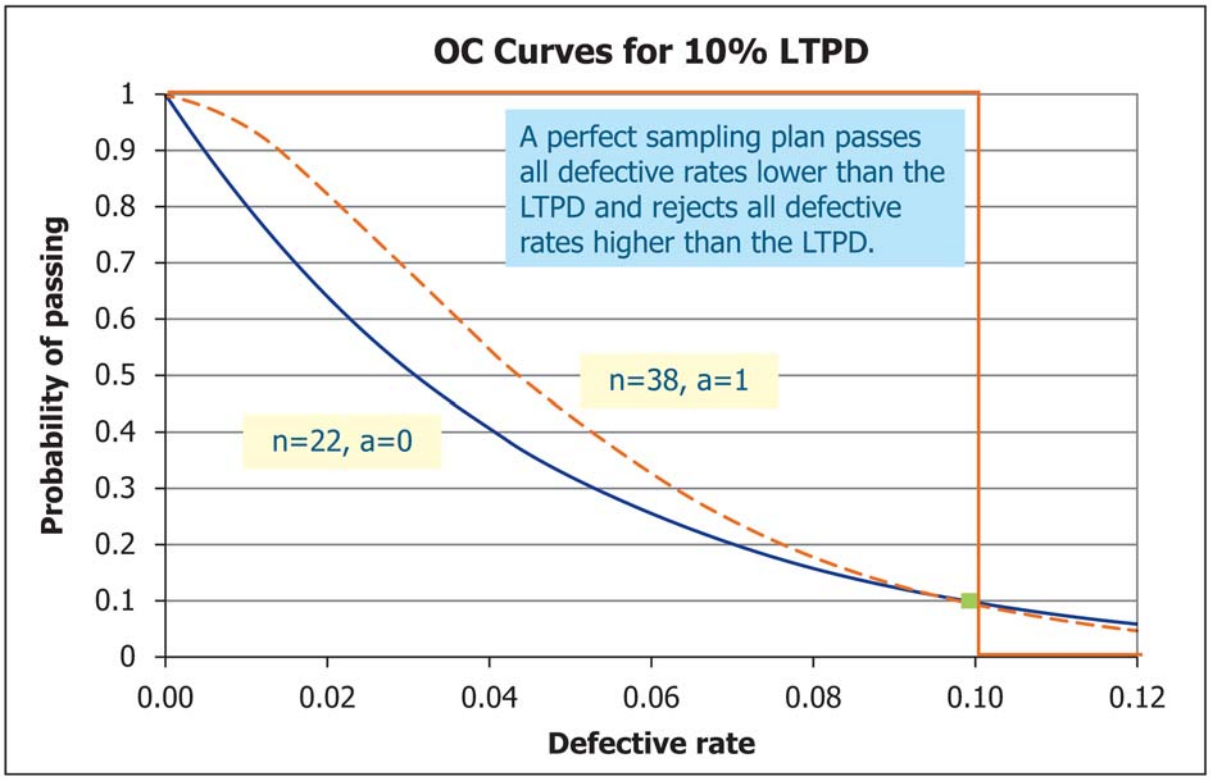


FIG. X1.1 Operating Characteristic Curve

X1.6 Here is what can be said about $a = 1$ plans:

X1.6.1 More samples needed than with an $a = 0$ plan,

X1.6.2 Better chance of passing than an $a = 0$ plan when the true defective rate is lower than the maximum allowable defective rate,

X1.6.3 Lower chance of passing than an $a = 0$ plan when the true defective rate is higher than the maximum allowable defective rate, and

X1.6.4 If the plan passes with an observed defective unit in the sample, the population is known to contain at least one defect.