INSTITUTE FOR DEFENSE ANALYSES

Sample Size Determination Methods Using Acceptance Sampling by Variables

Kerry Walzl, Project Leader
Lindsey A. Davis
Thomas H. Johnson
Heather M. Wojton

August 2019

Approved for public release.
Distribution is unlimited.

IDA Document NS D-10666

Log: H 2019-000359

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task BD-9-1833(10), "Personal Protective Equipment," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule, Director, and consisted of Rebecca Medlin from the Operational Evaluation Division.

For more information:
Kerry N. Walzl, Project Leader
kwalzl@ida.org • (703) 845-6927

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-10666

Sample Size Determination Methods Using Acceptance Sampling by Variables

Kerry N. Walzl, Project Leader
Lindsey A. Davis
Thomas H. Johnson
Heather M. Wojton

Executive Summary

Acceptance Sampling by Variables (ASbV) is a statistical testing technique used in Personal Protective Equipment programs to determine the quality of the equipment in First Article and Lot Acceptance Tests. This article intends to remedy the lack of existing references that discuss the similarities between ASbV and certain techniques used in different sub-disciplines within statistics. Understanding ASbV from a statistical perspective allows us to provide DOT&E with customized test plans, beyond what is available in MIL-STD-414. We plan to submit this article to a statistics journal. This paper does not include any real test data, and does not mention any specific program by name.

Sample Size Determination Methods for Acceptance Sampling by Variables

Thomas H. Johnson, Lindsey Davis
Kerry Walzl, and Heather Wojton

August 28, 2019

Abstract

The sample size for an Acceptance Sampling by Variables experiment is often determined using a standard such as MIL-STD-414 or one of its many derivatives. These standards specify sample sizes for a series of experiments to be used on a stream of incoming lots. They are not intended to be used to plan a stand-alone experiment, which requires an alternative approach. In this paper we focus on three alternatives: the operating characteristic (OC) curve approach, the Faulkenberry and Weeks approach, and a power analysis for a hypothesis test on a quantile. These three methods originate from different scientific sub-disciplines, but their concepts are quite similar and are easily confounded. To illustrate this, we review their development, highlight their similarities and differences, and present them in a consistent notation that demonstrates their mathematical equivalence.

Keywords: Acceptance Sampling by Variables, Design of Experiments, Sample Size Determination, Reliability, Operating Characteristic Curve, Tolerance Intervals, Power Analysis

1 Introduction

Determining the sample size for an Acceptance Sampling by Variables (ASbV) test can be challenging. ANSI/ASQ Z1.9-2003, a modern derivative of MIL-STD-414, is over 100 pages long and includes a "flow chart for use" that contains over 50 arrows (see Figure 1 in ANSI/ASQ Z1.9-2003). Neubauer and Luko (2013, 182) state, “MIL-STD-414 is complex”, while Horsnell opines, “It is my view that there are probably more frustrated designers of acceptance sampling schemes than in any other branch of applied statistics” (Gascoigne and Hill 1976, 312).

Wetherhill criticizes BS 6002, another MIL-STD-414 derivative, for its lack of guidance involving the development of individual plans. He states, “[it] suffers from the characteristic failing of Defence Sampling Schemes, of a lack of guidance as to when the plans should be used, and of what to do if they are not appropriate” (Gascoigne and Hill 1976, 308). One such example is when the stream of lots is too short to provide an effective use of the switching rules. In this case, a sampling plan that is independent of the overall scheme may be desired, and Schilling (2017, 216) recommends the use of an OC curve to develop this plan. BS 6002 provides the same recommendation as Schilling, and even though Wetherhill states “...BS 6002 is a great improvement on MIL-STD-414” he complains that “the brief reference to looking at the OC-curve [in BS 6002] is totally unsatisfactory [for developing an individual plan]” (Gascoigne and Hill 1976, 308).

If one is accustomed to using a standard to develop a variables sampling plan, then it can be challenging to develop an individual plan using a traditional OC curve approach. Textbooks that cover ASbV (e.g., Duncan [1959] or Schilling [2017]) include an abundance of content on standards and their associated switching rules. As a result, a bit of searching is required to find the theory that underpins the OC curve to develop an individual, stand-alone plan.

Another challenge involved with designing a stand-alone ASbV plan comes from the overlapping of concepts between separate scientific sub-disciplines. It is not uncommon to find nearly identical concepts nested within scientific sub-disciplines that use different terminology to address the same problems (e.g., the overlap of statistical learning, machine learning, and data mining, as well as the overlap of reliability analysis and survival analysis). This paper aims to disentangle commonly confounded approaches for developing stand-alone ASbV plans.

The first example of such confounding involves the methodology developed both in ASbV literature and in statistical tolerance interval literature. Statistical tolerance intervals share a close history with ASbV. They both grew in popularity around the same time that look-up tables for the non-central t distribution were becoming available (e.g., Owen [1963]). An approach for sizing tolerance intervals, called the Faulkenberry and Weeks approach (abbreviated FW; [1968]), was developed more than a decade after ASbV and is some times used instead of the OC curve approach to size an individual sampling plan. For example, Young (2016) uses the FW approach to design a sampling plan and uses historical data to inform the setting of the parameters involved in the FW computation.

Mixing of sub-disciplines also occurs between the OC curve approach in ASbV literature and a power analysis for a hypothesis test in statistics literature. In particular, associated with the use of OC curves are the symbols α and β, which are respectively referred to as the producer and consumer risk (Montgomery 2009, 642). This implies the setup of a hypothesis test, yet this is absent in ASbV literature (for example, hypothesis tests are not included in Shilling's textbook [2017], or Montgomery's textbook [2009]). This raises the question: can the ASbV sample size determination problem be set up as a traditional power analysis? This concept was recently investigated for the Acceptance Sampling by Attributes problem (Samohyl 2017), but we have yet to see a similar exposition for ASbV.

Acceptance sampling has received less attention in literature in the last few decades. Jenson et al. (2018) showed that articles on acceptance sampling commonly appeared in the Journal of Quality Technology in the 70's and 80's, but in the 90's the field shifted towards capability analysis, Stewart control charts, and classical Design of Experiments. Despite this trend, acceptance sampling remains a statistically defensible approach, and is still commonly used in defense tests.

In this paper we clarify the equivalence between three sample size determination methodologies that can be used to plan a stand-alone ASbV test: the OC curve approach, the FW approach, and a power analysis for a hypothesis test on a quantile. We review the development of these methods, highlight their similarities and differences, and present them in a consistent notation that demonstrates their mathematical equivalence.

Assumptions throughout this paper follow the classic setup of the Acceptance Sampling by Variables problem. Let X denote the response variable, where $X \sim \mathcal{N}(\mu, \sigma^2)$ and both μ and σ^2 are unknown. Let $X_1, X_2, ...X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$. The sample mean $\bar{x}$ and variance s^2 are

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 \quad . \tag{1}$$

Let Z_p denote the p quantile of a standard normal distribution, where the p quantile of $\mathcal{N}(\mu, \sigma^2)$ is given by

$$Q_p = \mu + Z_p\sigma \quad . \tag{2}$$

Acceptance criteria for variables sampling plans can be based on an upper, lower, or double specification limit, but for the sake of simplicity, in this paper we focus solely on the upper specification limit, U.

2 Operating Characteristic Curve

Jennet and Welch (1939), Johnson and Welch (1940), the Statistical Research Group (abbreviated as SRG; see Eisenhardt et al. [1947]), Bowker and Goode (1952), and Liebermann and Resnikoff (1955) are often credited as the architects of ASbV. The following review primarily comes from *Techniques of Statistical Analysis* prepared by Eisenhardt et al. (1947) that reflects the work of the SRG.

The SRG states, "An ASbV procedure may be analyzed into three phases" (1947, 13). The first phase is "the *plan of action*, that is, the set of rules on the basis of which to accept or reject the lot." The plan of action that we focus on in this paper for assessing an upper specification limit, U, is called the "Standard Deviation Method - Variability Unknown" by MIL-STD-414, and is also sometimes referred to as the "k-method" (Schilling and Neubauer 2017). The Standard Deviation Method acceptance criteria is

$$\frac{U - \bar{x}}{s} \geq k \quad , \tag{3}$$

where k is referred to as an "acceptability constant (e.g. MIL-STD-414)" that must be determined for a given plan.

The SRG notes that other measures of central tendency could be used instead of $\bar{x}$ and s in Equation 3, such as the median or range. Measures such as these have found their way into MIL-STD-414 and subsequent derivative standards because they were thought to have been easier to implement (i.e., the range is easier to calculate than the standard deviation), but the SRG notes that "the gains in computational simplicity that may be afforded by other measures are likely to be unimportant in the situation for which variables inspection is appropriate." For an additional discussion on the controversy involving the inclusion of "other" measures, see Acheson and Duncan (1975, 40).

The SRG defines the next two phases as follows. The second phase is "the *amount of inspection* required by the plan, that is, the number of items that must be inspected from each lot," and the third and final phase is "the *operating characteristics* of the plan, that is, the proportion of submitted lots of various qualities that will be accepted and rejected if the plan is used."

The amount of inspection and operating characteristics can be investigated using an OC curve. SRG states, "it is important to know what proportion of submitted lots will be accepted for each possible quality...as a plan is clearly unsuitable if it passes too many of the lots of unsatisfactory quality or rejects too many of the lots of acceptable quality that are submitted to it." For a given sample size, an OC curve displays the probability of lot acceptance, denoted as P_A, versus the assumed, or anticipated, lot proportion defective, denoted as $\tilde{p}$, where $\tilde{p}$ is related to p (see Equation 2) by $\tilde{p} = 1 - p$.

OC curves were challenging to produce in the middle of the twentieth century because computing the probability of acceptance involved the evaluation of the non-central t distribution. In the paper that is credited with some of the first theory on ASbV (Jennett and Welch 1939), the authors state, "[the distribution of k] is a particular example of what is termed the non-central t distribution. Tables of this distribution, in a form suitable for the present problem, do not exist, but are in process of calculation." This task was accomplished in a follow-up paper (Johnson and Welch 1940).

The derivation of the distribution of k in Equation 3 leads to the equation for the OC curve. Jennet and Welch (1939) note, "The true distribution of [k] is not difficult to find," while Johnson and Welch (1940) and Liebermann and

Resnikoff (1955) seemed to have shared the same sentiment, as they showed no intermediate steps in re-expressing Equation 3 as

$$\left(\frac{\sqrt{n}(U-\mu)}{\sigma}-\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\right)\frac{\sigma}{s}\geq\sqrt{n}k \quad . \tag{4}$$

For more details on obtaining Equation 4, one can consult Resnikoff and Lieberman (1957) or Schilling and Neubauer (2017, 238).

In Equation 4, by definition of the non-central t distribution, the quantity on the left hand side of the equation is distributed as a non-central t random variable, with $n-1$ degrees of freedom and non-centrality parameter equal to

$$\sqrt{n}\frac{U-\mu}{\sigma} \quad . \tag{5}$$

Thus, the probability of acceptance is

$$P_A = Pr\left(\frac{U-\bar{x}}{s}\geq k\right) = 1-t\left(\sqrt{n}k, n-1, \sqrt{n}\frac{U-\mu}{\sigma}\right) \quad , \tag{6}$$

where $t(\cdot,\cdot,\cdot)$ is the noncentral t cumulative distribution function with quantile $\sqrt{n}k$, $n-1$ degrees of freedom, and noncentrality parameter $\sqrt{n}\frac{U-\mu}{\sigma}$. Equation 6 is not yet the equation of the OC curve, since it requires one more assumption.

SRG stresses that the OC curve cannot predict the future, because the true population parameters are unknown. They state, "...the OC curve does *not* show the probability that an accepted lot will be of quality p" (1947, 16). The OC curve simply shows, for a given sample size, the probability of accepting a lot of an assumed (or anticipated) quality. This means that the OC curve assumes that the fraction of the normal population exceeding U is equal to $\tilde{p}$, or equivalently

$$\frac{U-\mu}{\sigma} = Z_{1-\tilde{p}} \quad . \tag{7}$$

Finally, substituting this assumption into Equation 6 gives the equation of the OC curve as

$$P_A = 1-t\left(\sqrt{n}k, n-1, \sqrt{n}Z_{1-\tilde{p}}\right) \quad . \tag{8}$$

The equation of the OC curve can be uniquely defined in a number of different ways. One way is to specify two points that the OC curve passes through.

Let these two points be $(\tilde{p} = 1 - p_1, P_A = 1 - \alpha)$ and $(\tilde{p} = 1 - p_2, P_A = \beta)$. Substituting these points, respectively, into Equation 8 yields

$$1 - \alpha = 1 - t\left(\sqrt{n}k, n-1, \sqrt{n}Z_{p_1}\right) \quad , \tag{9}$$

$$\beta = 1 - t\left(\sqrt{n}k, n-1, \sqrt{n}Z_{p_2}\right) \quad . \tag{10}$$

which can be respectively expressed as

$$\sqrt{n}k = t'\left(\alpha, n-1, \sqrt{n}Z_{p_1}\right) \quad , \tag{11}$$

$$\sqrt{n}k = t'\left(1-\beta, n-1, \sqrt{n}Z_{p_2}\right) \quad . \tag{12}$$

Here, $t'(\cdot,\cdot,\cdot)$ denotes the non-central t quantile function, where the first, second, and third arguments to this function are the cumulative density, degrees of freedom, and non-centrality parameter.

Theoretically, the sample size that causes the OC curve to pass through the two points can be found as the solution for n that satisfies

$$t'\left(\alpha, n-1, \sqrt{n}Z_{p_1}\right) = t'\left(1-\beta, n-1, \sqrt{n}Z_{p_2}\right) \quad , \tag{13}$$

but in nearly all cases the discrete nature of n prevents the OC curve from passing through those exact two points. Instead, one can show that the OC curve passes through the two points $(\tilde{p} = 1-p_1, P_A = 1-\alpha^*)$ and $(\tilde{p} = 1-p_2, P_A = \beta)$, where α^* is the actual value of α that the OC curve passes through. That is, the OC curve passes through three of the four coordinates of the intended two points, but the fourth coordinate, α, has a source of error that is introduced due to the discrete nature of n. Additionally, to *control the risk* associated with the parameter α, we constrain the solution for n such that $\alpha* < \alpha$. That is, the sample size solution is the minimum value of n that satisfies

$$t'\left(\alpha, n-1, \sqrt{n}Z_{p_1}\right) \geq t'\left(1-\beta, n-1, \sqrt{n}Z_{p_2}\right) \quad . \tag{14}$$

After numerically solving for n, we solve for k by substituting n and the coordinates of one of the points that the OC curve passes through into Equation

8. Then, using n and k, the OC curve displays the probability of acceptance, P_A, versus the anticipated fraction defective, $\tilde{p}$.

Before we move on we should note that ASbV plans typically assign names to the parameters of the OC curve. According to Montgomery (2009), α is the "producer's risk", $1 - p_1$ is the Acceptable Quality Limit (AQL), β is the "consumer's risk", and $1 - p_2$ is the Rejectable Quality Limit (RQL). That is,

$$P_A\left(\tilde{p} = 1 - p_1 = AQL\right) = 1 - \alpha \quad , \tag{15}$$

and

$$P_A\left(\tilde{p} = 1 - p_2 = RQL\right) = \beta \quad . \tag{16}$$

In terms of these variables, the OC curve takes the usual form (e.g., Juran and Godfrey [1999, 46.46]) as shown in Figure 1.

3 OC Curve Notional Example

The Army buys body armor from a supplier. The Army has established an upper specification on a bullet's penetration depth into the body armor that is equal to 5 mm. If 1 percent or more of the bullets fired demonstrate a penetration depth above this limit, the Army wishes to accept the lot with probability 0.95 ($p_1 = 1 - AQL = 0.99, 1 - \alpha = .95$), whereas if 6 percent or more of the bullets fired demonstrate a penetration depth above this limit, the Army would like to reject the lot with probability .90 ($p_2 = 1 - RQL = .94, \beta = .10$). Determine the sample size for this variables sampling plan.

Begin by finding the minimum value of n that satisfies Equation 14. A brute-force way of doing this is to simply plot the quantity on the left and right hand side of Equation 14 as a function of n, as shown in Figure 1, and visually locate the minimum value of n that satisfies 14 . It is clear from this figure that the solution is $n = 42$.

Then, calculate k using Equation 10 with $\beta = 0.1$ and $p_2 = .94$ to obtain $k = 1.898$. Substituting this value of k, $n = 42$, and $p_1 = .99$ into Equation 9, we obtain $\alpha^* = 0.047$, which means we are properly *controlling the risk*, since

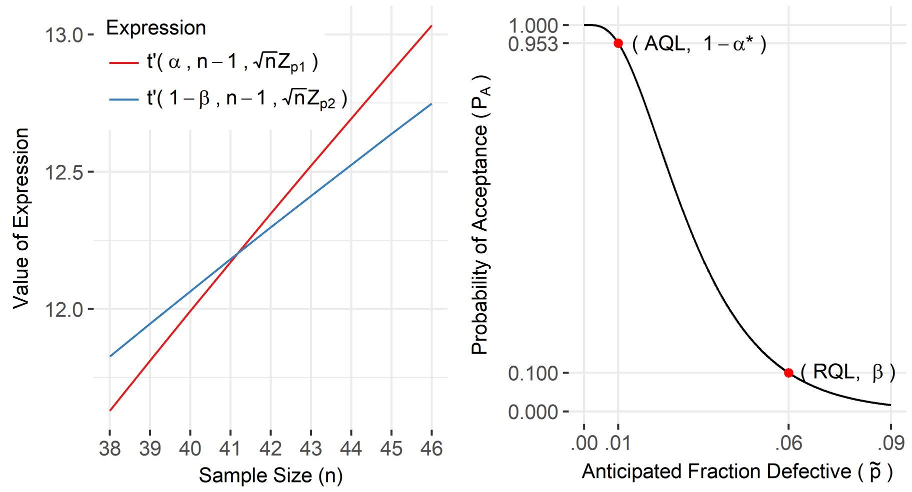

Figure 1: Left side of figure demonstrates the numerical solution of the body armor example. Right side of figure shows the OC curve.

$\alpha^* < \alpha$. The OC curve can be plotted using Equation 8 as shown on the right side of Figure 1.

4 Faulkenberry and Weeks Approach

The Faulkenberry and Weeks approach, a sample size determination methodology quite similar to the OC curve approach, comes from the field of statistical tolerance regions. The following review is taken from Krishnamoorthy (2009). A tolerance interval is constructed using the random sample $X_1, X_2, ..., X_n$, and is required to contain a proportion p or more of the sampled population, with confidence level $1 - \Gamma$. Formally, a $(p, 1 - \Gamma)$ upper tolerance interval has the form $\bar{x} + ks$, where k is to be determined such that it satisfies the condition

$$Pr\left(\bar{x} + ks \geq \mu + Z_p\sigma\right) = 1 - \Gamma \quad . \tag{17}$$

Krishnamoothy shows the derivation for the solution of k using a "classical approach" and a "generalized variable approach" that both result in

$$k = \frac{1}{\sqrt{n}} t'\left(1 - \Gamma, n - 1, \sqrt{n} Z_p\right) \quad . \tag{18}$$

Faulkenberry and Weeks (1968) developed a procedure for determining the sample size that produces a tolerance interval that meets a "goodness criterion." We summarize their approach as follows.

Consider two different tolerance intervals that are constructed from the same random sample, $X_1, X_2, ...X_n$. For the first tolerance interval, let $p = p_1$ and $\Gamma = \alpha$, and for the second tolerance interval, let $p = p_2$ and $\Gamma = 1 - \beta$. That is, consider a (p_1, α) tolerance interval, and a $(p_2, 1 - \beta)$ tolerance interval, where $p_1 > p_2$, and α and β are typically values between $.01 - .20$, and p_1 and p_2 are typically values between $.80 - .99$.

Faulkenberry and Daly (1970) define the "goodness criterion" in a compact form, stating, "The criterion used for determining sample size is as follows: For a tolerance limit such that $Pr\,(\text{coverage} \geq P) = \gamma$, choose $P' > P$ and δ (small) and require $Pr\,(\text{coverage} \geq P') \leq \delta$." For the sake of comparison between sample size determination methods in this paper, we use a different set of variables than FW. That is, let $\alpha = \delta$, $\beta = 1 - \gamma$, $p_1 = P'$, and $p_2 = P$.

Thus, for the $(p_2, 1 - \beta)$ tolerance interval that can be expressed as

$$t\left(\sqrt{n}k, n-1, \sqrt{n}Z_{p_2}\right) = 1 - \beta \quad , \tag{19}$$

according to the "goodness criterion", the sample size solution is the minimum value of n that satisfies

$$t\left(\sqrt{n}k, n-1, \sqrt{n}Z_{p_1}\right) \leq \alpha \quad . \tag{20}$$

Or equivalently, the FW sample size solution is the minimum value of n that satisfies

$$t'\left(\alpha, n-1, \sqrt{n}Z_{p_1}\right) \geq t'\left(1-\beta, n-1, \sqrt{n}Z_{p_2}\right) \quad . \tag{21}$$

Equation 21 is identical to the sample size solution using the OC curve approach (same as Equation 14).

4.1 FW Example

The following brief example is taken from Faulkenberry and Daly, (1970, 818). Suppose we are interested in $p_1 = .95$, $\alpha = .10$, $p_2 = .90$, and $\beta = .10$. Using

Equation 21, and solving for n in a similar manner as the OC curve example, we obtain $n = 104$, which is used to obtain $k = 1.466$. Substituting these values and $p_1 = .95$ into Equation 9, we obtain $\alpha = .099$. Thus, the goodness inequality, $Pr(\text{coverage} \geq p_1) \leq \alpha$, is satisfied.

5 Hypothesis Test on Quantile Approach

The last approach in this paper is a power analysis for the hypothesis that tests whether a population quantile is different from a constant. In developing the power function that we use to determine sample size, we follow a procedure similar to that used in the classical examples, such as the test involving $H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$ (e.g. Mathews [2010, 31]). Lenth (2001) outlines this general procedure as follows.

The procedure starts with the definition of the null and alternative hypotheses, and the definition of the test statistic (or as Lenth states, "the underlying probability model for the data"). This is followed by the definition of the effect size (what Casella and Berger [2002, 382] call "defining the rejection region"), the solution for the power function, and finally the sample size that provides a desired level of power.

We implement this general procedure for the hypothesis that tests whether a population quantile is different from a constant. To begin, define the null and alternative hypothesis as

$$H_0 : Q_{p_1} = U \implies \text{Accept Lot} \quad , \tag{22}$$

$$H_a : Q_{p_1} > U \implies \text{Reject Lot} \quad . \tag{23}$$

Here, Q_{p_1} denotes the p_1 quantile, and U, the upper specification limit, is treated as a constant. The null hypothesis implies the lot is accepted, while the alternative implies the lot is rejected.

The test statistic is conveniently obtained from the equation for the confidence interval on Q_{p_1}. Chakraborti and Li (2007) present a derivation of this equation (originally shown by Lawless [2002]) based on the biased estimator $\hat{Q}_{p_1} = \bar{x} + Z_{p_1}\hat{\sigma}$, where $\hat{\sigma}$ is the biased MLE of σ. It is straightforward to perform a similar derivation, based on a slightly different estimator that is also biased, $\hat{Q}_{p_1} = \bar{x} + Z_{p_1} s$, to obtain an equation for a confidence interval on Q_{p_1} that satisfies

$$Pr\left(\bar{x} + t'\left(1-\alpha, n-1, \sqrt{n}Z_{p_1}\right)\frac{s}{\sqrt{n}} \geq Q_{p_1}\right) = 1-\alpha \quad . \tag{24}$$

This equation also satisfies the definition of a $(p_1, 1-\alpha)$ upper one-sided tolerance limit, as it is well known that a one-sided upper tolerance limit is equivalent to a one-sided upper confidence interval on a quantile (e.g. Krishnamoorthy [2009, 27]). We can rearrange Equation 24 to obtain

$$Pr\left(\frac{Q_{p_1} - \bar{x}}{s/\sqrt{n}} \leq t'\left(1-\alpha, n-1, \sqrt{n}Z_{p_1}\right)\right) = 1-\alpha \quad . \tag{25}$$

which implies that the pivotal quantity,

$$\frac{Q_{p_1} - \bar{x}}{s/\sqrt{n}} \quad , \tag{26}$$

is distributed as a non-central t random variable with $n-1$ degrees of freedom and non-centrality parameter equal to $\sqrt{n}Z_{p_1}$.

Under H_0 we assume $Q_{p_1} = U$ and apply this assumption to Equation 26 to obtain the test statistic

$$T = \frac{U - \bar{x}}{s/\sqrt{n}} \quad . \tag{27}$$

In practice, once data collection is complete, we reject H_0 if $T < T_{crit}$, where the critical value is

$$T_{crit} = t'\left(\alpha, n-1, \sqrt{n}Z_{p_1}\right) \quad . \tag{28}$$

Prior to conducting the experiment, and for the purpose of determining sample size, we need to assume a value for U assuming that we reject H_0. That is, we need to define the effect size. Thus, when H_a is true, let $U = Q_{p_2}$. The difference between Q_{p_1} and Q_{p_2} can be interpreted as the effect size.

If in Equation 27 we substitute U with Q_{p_2}, then the test statistic assuming H_a is true is distributed as a non-central t random variable with $n-1$ degrees of freedom and non-centrality parameter equal to $\sqrt{n}Z_{p_2}$.

Power is the probability that the test statistic (assuming H_a is true) is less than the critical value, and represents the probability of correctly rejecting the lot. The power function is

$$1 - \beta = t\left(T_{crit}, n-1, \sqrt{n}Z_{p_2}\right) \quad . \tag{29}$$

Additionally, if we *control the risk* associated with α as we did in the previous sections, then we can constrain the sample size solution such that the integer solution for n yields an actual value, α^*, that is less than the intended value, α. This implies that the integer solution for n satisfies

$$1 - \beta \leq t\left(T_{crit}, n-1, \sqrt{n}Z_{p_2}\right) \quad . \tag{30}$$

Applying the inverse t quantile function with $n-1$ degrees of freedom and noncentrality parameter $\sqrt{n}Z_{p_2}$ to both sides of Equation 30, results in

$$t'\left(1-\beta, n-1, \sqrt{n}Z_{p_2}\right) \leq T_{crit} \quad , \tag{31}$$

which is identical to the sample size equations (14 and 21) from the previous sections.

6 Power Analysis Example

The Air Force buys guided missiles from a manufacturer. The Air Force has established an upper specification on the missile's radial miss distance that is equal to 5 feet. If 4 percent or more of the missiles fired demonstrate a radial miss distance above this limit, the Air Force wishes to accept the lot with probability 0.95 ($p_1 = 1 - AQL = 0.96, 1 - \alpha = .95$), whereas if 12 percent or more of the missiles fired demonstrate a radial miss distance above this limit, the Air Force would like to reject the lot with probability .90 ($p_2 = 1 - RQL = .88, \beta = .10$). Determine the sample size for this variables sampling plan.

As in the previous sections, the solution for n can be found numerically, which is $n = 53$. This results in $\alpha^* = .0499$, which satisfies $\alpha < \alpha^*$. It is instructive to plot the density of the test statistic assuming H_0 is true, and assuming H_a is true, as shown in Figure 2.

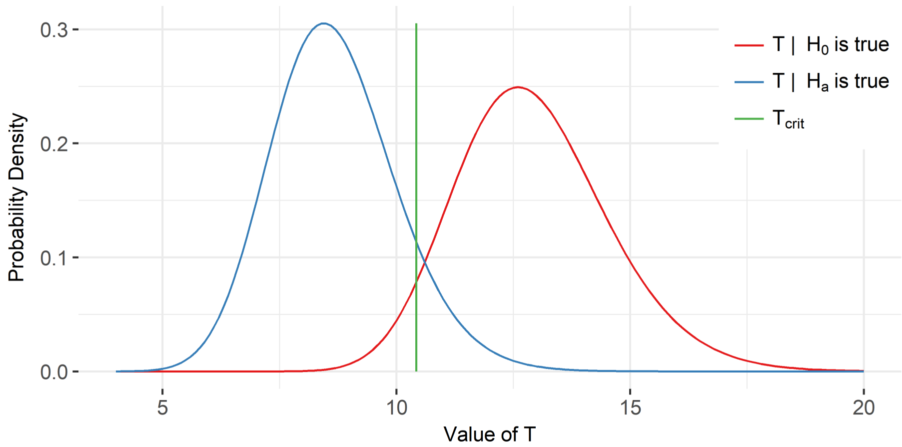

Figure 2: The distribution of the test statistic under H_0 and H_a for n=53.

Power, denoted as $1 - \beta$, is the probability of correctly rejecting H_0, or equivalently the probability of correctly rejecting the lot. Power is the area under the blue curve to the left of T_{crit}. The complement of power is the Type II error, denoted as β, which represents the consumer risk, because it conveys the risk of accepting a bad lot.

Confidence level, denoted as $1 - \alpha$, is the probability of correctly accepting H_0, or equivalently the probability of correctly accepting the lot. Confidence level is the area under the red curve to the right of T_{crit}. The complement of confidence level is the Type I error, denoted as α, which represents the producer's risk, because it conveys the risk of rejecting an acceptable lot.

7 Discussion

We highlighted three methods for sizing a stand-alone ASbV experiment, but other approaches can be found in literature. For instance, Lieberman and Resnikoff's (1955) landmark paper bases an acceptance criteria on a minimum variance unbiased estimator (MVUE). The power function for the MVUE in this case is quite complex and does not have the simple non-central t format that we saw from the other three methods in this paper, which may be why it is less commonly used in practice.

Similarly, Hamilton (1995) also bases the acceptance criteria on an estimator of the fraction defective, but this time the estimator is biased. Their sample size determination equation is similar in form to those used by the other three approaches in this paper, but Hamilton relies on OC curves as opposed to a hypothesis test and power function to formulate the sample size determination problem.

Alternative Bayesian approaches are also available in literature. Average coverage criteria, average length criterion, utility theory, or Bayes factors (see Adcock [1997]), could potentially be adapted to the ASbV problem. One method in particular (Easterling and Weeks 1970) adapted the FW approach for use in a Bayesian setting. Surely, other sample size determination methodologies exist that would also be appropriate for the ASbV problem.

We compared the OC curve approach, the FW approach, and the power analysis for a test on a quantile because of their prevalence of use for sizing ASbV tests. In this paper we believe we helped clarify how nearly identical these methods really are by reviewing their origins and by demonstrating the equivalence of their sample size determination equations (Equations 14, 21, and 30).

To demonstrate this equivalence, we had to make a couple of minor assumptions involving the inequality sign in the sample size equations. For the FW approach, the inequality sign organically occurs due to the definition of the "goodness criterion." For the OC curve approach, theoretically, the inequality sign should be an equals sign because the OC curve approach is typically defined by specifying two points that the curve passes through. Given that the integer solution for n prevents the curve from exactly passing through two points, it seemed reasonable that we impose the arbitrary constraint, $\alpha^* < \alpha$, for the sake of matching the OC curve sample size equation to the FW approach. For the power analysis, we used the same line of reasoning and imposed the same arbitrary constraint, $\alpha^* < \alpha$. We should note that Lenth's guideline for power analyses suggests that the sample size solution be the minimum value of n that provides a target value of power that is greater than the intended value of power. Consequently, this would flip the inequality sign. However, this is quite trivial in practice, as the direction of the inequality can only change the sample size result by one.

The FW approach and the OC curve approach use a nearly identical derivation that follows the "classical approach" (e.g., Krishnamoorthy [2009, 26]). That is, they start with the equation of a tolerance interval or the equation of the acceptance criteria, and re-express it in terms of the non-central t distribution. The FW approach and OC curve approach even share the same symbol, k. For tolerance intervals, k is the "k-factor", while in ASbV it is an "acceptability constant," but they have the same mathematical interpretation. In contrast, the derivation of the test statistic in the power analysis (e.g. Chakraborti and Li [2007]) more closely follows the "generalized variable approach" (e.g. Krishnamoorthy [2009, 26]) and does not use k or any equivalent symbol.

Complex methodologies can be presented in simpler terms to gain traction within a community of practitioners, but it obscures the underlying theory. The architects of ASbV worked in an applied setting, and it seems plausible that they conceived ASbV with statistical terminology and language in mind (e.g., Neyman Pearson hypothesis testing), but then packaged ASbV for their engineering audience by defining names, such as "consumer risk," and omitting statistical terms. It is challenging for a practitioner to customize a stand-alone ASbV plan when the underlying theory is obscured.

Framing the ASbV problem in terms of a power analysis can assist in creating customized test plans. For instance, the hypothesis test in this paper easily could have been changed to a regression model problem that focused on a hypothesis test on a conditional quantile corresponding to a particular point in a factorial experiment. Extensions like this could add to the body of ASbV theory, but we save such work for the future.

References

Adcock, CJ. 1997. "Sample size determination: a review." *Journal of the Royal Statistical Society: Series D (The Statistician)* 46 (2): 261–283.

Bowker, Albert Hosmer, and Henry Phillip Goode. 1952. *Sampling inspection by variables.* McGraw-Hill.

Casella, George, and Roger L Berger. 2002. *Statistical inference.* Vol. 2. Duxbury Pacific Grove, CA.

Chakraborti, S, and J Li. 2007. "Confidence interval estimation of a normal percentile." *The American Statistician* 61 (4): 331–336.

Duncan, Acheson J. 1975. "Sampling by variables to control the fraction defective: Part I." *Journal of Quality Technology* 7 (sup1): 61–69.

Duncan, Acheson Johnston. 1959. "Quality control and industrial statistics."

Easterling, Robert G, and David L Weeks. 1970. "An accuracy criterion for Bayesian tolerance intervals." *Journal of the Royal Statistical Society. Series B (Methodological):* 236–240.

Eisenhart, Wallis, Hastay. 1947. *Techniques of statistical analysis.* New York, NY, McGraw-Hill.

Faulkenberry, G David, and James C Daly. 1970. "Sample size for tolerance limits on a normal distribution." *Technometrics* 12 (4): 813–821.

Faulkenberry, G David, and David L Weeks. 1968. "Sample size determination for tolerance limits." *Technometrics* 10 (2): 343–348.

Gascoigne, JC, and ID Hill. 1976. "The Draft British Standard 6002:" Sampling inspection by variables"." *Journal of the Royal Statistical Society. Series A (General)* 139 (3): 299–317.

Hamilton, David C, and Mary L Lesperance. 1995. "A comparison of methods for univariate and multivariate acceptance sampling by variables." *Technometrics* 37 (3): 329–339.

Jennett, WJ, and BL Welch. 1939. "The control of proportion defective as judged by a single quality characteristic varying on a continuous scale." *Supplement to the Journal of the Royal Statistical Society* 6 (1): 80–88.

Jensen, Willis A, Douglas C Montgomery, Fugee Tsung, and Geoffery G Vining. 2018. "50 years of the Journal of Quality Technology." *Journal of Quality Technology* 50 (1): 2–16.

Johnson, NL, and BL Welch. 1940. "Applications of the non-central t-distribution." *Biometrika* 31 (3/4): 362–389.

Juran, Joseph, and A Blanton Godfrey. 1999. "Quality handbook." *Republished McGraw-Hill:* 173–178.

Krishnamoorthy, Kalimuthu, and Thomas Mathew. 2009. *Statistical tolerance regions: theory, applications, and computation.* Vol. 744. John Wiley & Sons.

Lawless, Jerald F. 2002. *Statistical models and methods for lifetime data.* Vol. 2nd Ed. John Wiley & Sons.

Lenth, Russell V. 2001. "Some practical guidelines for effective sample size determination." *The American Statistician* 55 (3): 187–193.

Lieberman, Gerald J, and George J Resnikoff. 1955. "Sampling plans for inspection by variables." *Journal of the American Statistical Association* 50 (270): 457–516.

Mathews, Paul. 2010. *Sample size calculations: Practical methods for engineers and scientists.* Mathews Malnar / Bailey.

Montgomery, Douglas C. 2009. *Statistical quality control.* Vol. 7. Wiley New York.

Neubauer, Dean V, and Stephen Luko. 2013. "Comparing Acceptance Sampling Standards, Part 2." *Quality Engineering* 25 (2): 181–187.

Owen, Donald B. 1963. *Factors for one-sided tolerance limits and for variables sampling plans.* Technical report. Sandia Corp., Albuquerque, N. Mex.

Resnikoff, George J, and Gerald J Lieberman. 1957. *Tables of the non-central t-distribution: density function, cumulative distribution function, and percentage points.* Stanford University Press.

Samohyl, Robert Wayne. 2017. "Acceptance sampling for attributes via hypothesis testing and the hypergeometric distribution." *Journal of Industrial Engineering International:* 1–20.

Schilling, Edward G, and Dean V Neubauer. 2017. *Acceptance sampling in quality control.* Crc Press.

Young, Derek S, Charles M Gordon, Shihong Zhu, and Bryan D Olin. 2016. "Sample size determination strategies for normal tolerance intervals using historical data." *Quality Engineering* 28 (3): 337–351.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
06-08-2019	OED Draft	1 July 2019 - 31 July 2019

4. TITLE AND SUBTITLE Sample Size Determination Methods Using Acceptance Sampling by Variables	**5a. CONTRACT NUMBER** HQ0034-14-D-0001
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Davis, Lindsey A.; Johnson, Thomas H.; Walzl, Kerry N.; Wojton, Heather M.	**5d. PROJECT NUMBER** BD-9-1833
	5e. TASK NUMBER 1833(10)
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Institute for Defense Analyses 4850 Mark Center Drive Alexandria, Virginia 22311-1882	**8. PERFORMING ORGANIZATION REPORT NUMBER** D-10666 H 2019-00359
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Director, Operational Test and Evaluation 1700 Defense Pentagon Washington, DC 20301	**10. SPONSOR/MONITOR'S ACRONYM(S)** DOT&E
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Walzl, Kerry N.

14. ABSTRACT

The sample size for an Acceptance Sampling by Variable experiment is often determined using a standard such as MIL-STD-414 or one of its many derivatives. These standards specify sample sizes for a series of experiments to be used on a stream of incoming lots. They are not intended to be used to plan a stand-alone experiment, which requires an alternative approach. In this paper we focus on three alternatives: the operating characteristic curve approach, the Faulkenberry and Weeks approach, and a power analysis for a hypothesis test on a quantile. These three methods originate from different scientific sub-disciplines, but their concepts are quite similar and are easily confounded. To illustrate this, we review their development and present them in a consistent notation that demonstrates their mathematical equivalence.

15. SUBJECT TERMS

Design of Experiments, Operational Testing, Statistics

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Kerry N. Walzl
Unclassified	Unclassified	Unclassified		28	**19b. TELEPHONE NUMBER** *(Include area code)* 703-845-6927