INL/EXT-05-00349

Statistical Methods Handbook For Advanced Gas Reactor Fuel Materials

J. J. Einerson

May 2005

Idaho National Laboratory

The INL is a U.S. Department of Energy National Laboratory
operated by Battelle Energy Alliance

Statistical Methods Handbook for Advanced Gas Reactor Fuel Materials

J. J. Einerson

May 2005

Idaho National Laboratory

Idaho Falls, Idaho 83415

Statistical Methods Handbook for Advanced Gas Reactor Fuel Materials

INL/EXT-05-00349

May 2005
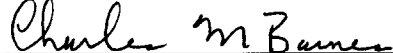
Approved by
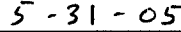
John R. Cox   _5-31-05_
John R. Cox   Date

Charles M Barnes   _5-31-05_
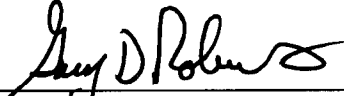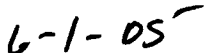Charles M. Barnes   Date

Gary D Roberts   _6-1-05_
Gary D. Roberts   Date

ABSTRACT

Fuel materials such as kernels, coated particles, and compacts are being manufactured for experiments simulating service in the next generation of high temperature gas reactors. These must meet predefined acceptance specifications. Many tests are performed for quality assurance, and many of these correspond to criteria that must be met with specified confidence, based on random samples. This report describes the statistical methods to be used. The properties of the tests are discussed, including the risk of false acceptance, the risk of false rejection, and the assumption of normality. Methods for calculating sample sizes are also described.

CONTENTS

FIGURES

TABLES

Statistical Methods Handbook for Advanced Gas Reactor Fuel Materials

1. INTRODUCTION

1.1 Overview

Kernels, coated particles, and compacts are being manufactured to predefined specifications for experiments simulating service in the Next Generation Nuclear Plant (NGNP). This handbook describes the statistical methods to be employed to determine the acceptability of the manufactured products with respect to specification requirements. The acceptability requires that properties of the product satisfy acceptance criteria set by the Advanced Gas Reactor (AGR) team. These properties include characteristics such as kernel diameter, coating thickness, and process impurities.

For most properties, the total population (that is, all particles manufactured) can not be measured, so acceptance therefore has to be based on a sample representative of the population. This report considers sampling criteria that have to be met with specified confidence levels.

The basic approach is to collect a representative sample, apply an acceptance test to the sample, and accept or reject the product based on this test. Even though a sample satisfies the acceptance test, it is not certain that the population meets the acceptance criteria. A chance of a wrong decision always exists when the decision is based on a random sample, but this can be quantified and made small, as follows.

There are two types of wrong decisions that can be made. The first is made when a product that does not meet the specifications is accepted (false acceptance). A test corresponds to c% confidence if any unacceptable product has at most a (100 – c)% chance of being accepted. For example, if the acceptance test has a 95% confidence level, there is no more than a 5% chance of accepting a product that should be rejected. The second wrong decision is made when an acceptable product is rejected (false rejection). The risk associated with making this decision must also be minimized. For a fixed sample size, the two kinds of wrong decision are inversely related. As the sample size increases, both risks decrease.

1.2 Terminology

Kernels are uncoated particles that are manufactured in batches of approximately three million kernels per batch. Each batch of kernels is tested to assure acceptable quality. Several batches (e.g., two or more) then form a composite, and the composite is also subjected to characterization and qualification. From the kernel composite, portions (coating batches) are split out to be coated. The kernels are coated with four layers: the buffer layer, the inner pyroltyic carbon (IPyC) layer, the silicon carbide (SiC) layer, and the outer pyrolytic carbon (OPyC) layer. The term particle refers to a coated kernel and these batches become particle batches and are tested for acceptable quality. Two or more particle batches are then combined to form a particle composite and subjected to acceptance tests. Finally, the particles from a composite are formed into compacts, consolidated objects bound together by a carbon matrix. For most characterization properties regarding compacts, compacts are selected at random and subjected to tests to assure acceptable quality of all the compacts. For dimensions qualification, all compacts are measured, not just a random selection.

2. ACCEPTANCE CRITERIA AND PRODUCT SPECIFICATIONS

2.1 Acceptance Criteria for Attribute Properties

An attribute property is a discrete property in the sense that the product is either defective or not defective with respect to the control limit defined for the property. For example, "missing OPyC layer" is an attribute property – either the OPyC layer is present or it is missing. Defective IPyC is also an attribute property; in this case a fuel particle is considered to have a defective IPyC if the uranium dispersion in the particle is greater than that shown in a visual standard (i.e., the control limit). The acceptance criterion for an attribute property is stated in terms of the allowable fraction of defective particles. To test whether the population satisfies the criterion for a given attribute property, a sample is drawn and each item in the sample is determined to be either acceptable or defective with respect to the control limit for that property. If the number of defective items is small enough, the population is accepted. The numbers defining the test (e.g., sample size and acceptable number of defective particles in the sample) are adjusted so that the probability of a false acceptance is at most 5% and the probability of a false rejection is reasonably low (i.e., ideally no more than 5%).

2.2 Acceptance Criteria for Variable Properties

A variable property is a property defined by a continuous distribution such as the normal distribution. The acceptance criterion for a variable property is stated in terms of the population mean and/or the population dispersion. For the population mean, the criterion is that the mean must lie within a specified interval. In some instances this interval is one-sided. The endpoints of the interval are the acceptance limits for the mean. For population dispersion, the criterion is that no more than a specified fraction of the population can exceed and/or be less than predetermined values. These values are called the critical limits for the population, or sometimes the critical limits for the dispersion. A confidence interval is used to test whether the population meets a criterion for the mean. There are, however, two ways to test whether the population meets a criterion for dispersion. The first way is to assume that the population is normally distributed, and to base the test on the sample mean, the sample standard deviation, and tabulated tolerance factors for the normal distribution. The second way, if the normality assumption is not justified, treats the property as an attribute property: a particle is called defective if it exceeds the critical limit, and is accepted otherwise. The test is based on the number of defective items found in a sample. The price for not assuming normality is that a much larger sample is required.

For the AGR experiments, the acceptance tests for most variable properties are based on the assumption that the populations are normally distributed. Substantial quantities of data are required to adequately test the assumption of normality and currently defined sample sizes will not be large enough. Therefore, in Section 4.1 of this report, the effect of possible non-normality is considered.

2.3 Product Specifications

Preliminary AGR fuel specifications are found in EDF-4198, "Preliminary AGR Fuel Specification" (Petti et al. 2004). Specifications for the AGR-1 fuel appear in EDF-4380, "AGR-1 Fuel Product Specification and Characterization Guidance" (Barnes 2005) and are duplicated here in Tables 1, 2, and 3 for kernels, coated particles, and compacts, respectively. There are 47 properties of the AGR-1 fuel, including 13 kernel properties, 19 coated particle properties, and 15 compact properties. For three compact properties, 100% inspection is specified. For the remaining properties, statistical specifications are given. All statistical specifications must be verified at the 95% confidence level.

Three general types of specifications appear in EDF-4380: mean specifications, dispersion specifications, and defect fractions. A mean specification sets an acceptable range for the mean of the sampled population. A dispersion specification defines a range within which a specified fraction of the population must lie. Dispersion specifications, often called critical limits, are used to control the variability in the population. A defect fraction specification defines a defective member of the population and sets an upper bound on the fraction of defectives allowed in the population. Dispersion specifications are essentially defect fraction specifications for properties measured on a continuous scale. The statistical methods used to address these types of specifications are discussed in Sections 3 and 4.

Table 1. AGR-1 product specifications for kernels.

Property	Specification on mean	Specification on dispersion[a]
Kernel batches		
Envelope density (Mg/m^3)	≥ 10.3	none
Diameter (μm)	350 ± 20	none
Sphericity (ellipticity)[b]	none	$\leq 0.10 \geq 1.06$
Kernel composite		
U-235 enrichment (wt %)[c]	19.80 ± 0.10	none
Carbon/uranium (atomic ratio)	0.50 ± 0.20	$\leq 0.01 \leq 0.20$ $\leq 0.01 \geq 0.80$
Oxygen/uranium (atomic ratio)	1.50 ± 0.20	none
(Carbon plus oxygen)/ uranium (atomic ratio)	≤ 2.0	none
Individual impurities: Li, Na, Ca, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Al, Cl, (ppm-wt)	≤ 100 ea. impurity	none
Process impurities: P, S (ppm-wt)	≤ 1500 ea. impurity	none
Envelope density (Mg/m^3)	≥ 10.4	none
Diameter (μm)	350 ± 10	$\leq 0.01 < 300$ $\leq 0.01 > 400$
Total uranium (wt %)	≥ 87.0	none
Sphericity (ellipticity)[b]	none	$\leq 0.10 \geq 1.05$

a. The first number is the allowable fraction outside of the critical limit and the second number is the critical limit.
b. Treated as an attribute property.
c. LEU kernels only.

Table 2. AGR-1 product specifications for coated particles.

Property	Specification on mean	Specification on dispersion[a]
Coated particle batches		
Buffer thickness (μm)	100 ± 15	none
IPyC thickness (μm)	40 ± 5	none
SiC thickness (μm)	35 ± 4	none
OPyC thickness (μm)	40 ± 5	none
Missing OPyC layer defect fraction[b]	$\leq 6 \times 10^{-4}$	none
Coated particle composites		
Buffer thickness (μm)	100 ± 10	$\leq 0.01 \ \leq 55$
IPyC thickness (μm)	40 ± 4	$\leq 0.01 \ \leq 30$ $\leq 0.01 \ \geq 56$
SiC thickness (μm)	35 ± 3	$\leq 0.01 \ \leq 25$
OPyC thickness (μm)	40 ± 4	$\leq 0.01 \ \leq 20$
Buffer bulk density (Mg/m^3)	0.95 ± 0.15	none
IPyC density (Mg/m^3)	1.90 ± 0.05	$\leq 0.01 \ \leq 1.80$ $\leq 0.01 \ \geq 2.00$
SiC density (Mg/m^3)	≥ 3.19	$\leq 0.01 \ \leq 3.17$
OPyC density (Mg/m^3)	1.90 ± 0.05	$\leq 0.01 \ \leq 1.80$ $\leq 0.01 \ \geq 2.00$
IPyC anisotropy	≤ 1.035	$\leq 0.01 \ \geq 1.06$
OPyC anisotropy	≤ 1.035	$\leq 0.01 \ \geq 1.06$
SiC gold spot defect fraction[b]	$\leq 1.0 \times 10^{-3}$	none
Aspect ratio (faceting)[b]	none	$\leq 0.01 \ \geq 1.084$
Defective SiC fraction[b]	$\leq 1.0 \times 10^{-4}$	none
Missing OPyC layer defect fraction[b]	$\leq 3.0 \times 10^{-4}$	none

a. The first number is the allowable fraction outside of the critical limit and the second number is the critical limit.
b. Treated as an attribute property.

Table 3. AGR-1 product specifications for compacts.

Property	Specification on mean	Specification on dispersion[a]
Mean uranium loading (g U/compact)	0.905 ± 0.04	none
Diameter (mm)	na	$0 < 12.22$ $0 > 12.46$
Length (mm)	na	$0 < 25.21$ $0 > 25.40$
Iron content (μg Fe outside SiC per compact)	≤ 25	$\leq 0.01 \; > 100$
Transition metal content (Cr, Mn, Co, and Ni) (μg outside SiC per compact)	≤ 75 each	$\leq 0.01 \; \geq 300$ total
Calcium content, (μg Ca outside SiC per compact)	≤ 25	none
Aluminum content, (μg Al outside SiC per compact)	≤ 25	none
Titanium and Vanadium content, (μg Ti + V outside SiC per compact)	≤ 400	none
Chlorine content, (wt ppm in heat treated matrix in compact)[e]	≤ 30	none
Compact integrity	(b)	none
Compact crush strength, MPa[c]	not specified	$\leq 0.05 \leq 3.5$
Heavy metal contamination fraction, (g exposed U/g U in compact)[d]	$\leq 1 \times 10^{-4}$	none
Defective SiC coating fraction (fraction of total particles)[d]	$\leq 2 \times 10^{-4}$	none
Defective IPyC coating fraction (fraction of total particles)[d]	$\leq 2 \times 10^{-4}$	none
Defective OPyC coating fraction (fraction of total particles)[d]	≤ 0.01	none

a. The first number is the allowable fraction outside of the critical limit and the second number is the critical limit.
b. No compacts shall exhibit surface defects in excess of visual standards shown in Figure 3-1 of Scheffel and Saurwein (2003).
c. Force needed to crush compact divided by compact diametral cross-sectional area.
d. Treated as an attribute property.
e. Chlorine specification applicable only if compacts are subjected to HCl cleaning.

3. STATISTICAL METHODS FOR ATTRIBUTE PROPERTIES

3.1 Assumptions

Attributes are properties measured on a discrete scale, such as acceptable/defective for defective SiC and present/absent for missing OPyC layer. Some properties are intrinsically attributes like missing OPyC layer, but any variable may be transformed into an attribute simply by dividing the range into an acceptable region and an unacceptable region, such as for sphericity.

Distribution-free statistical methods are methods where the statistical properties derive solely from random sampling and not from any assumption about the statistical distribution of the property of interest. The methods considered here are based on binomial or hypergeometric probability models and are most often used for attribute properties.

3.2 Acceptance Test

The test described here can be used when the property is inherently an attribute property (such as a missing OPyC layer), or when a variable property is treated as an attribute property because the normality assumption is not justified. One such variable property is the sphericity (aspect ratio). It is ideally 1.0, but it may be larger, and it cannot be smaller. Therefore it probably has an asymmetrical distribution, not a normal distribution. To treat the sphericity as an attribute property, count a particle as defective if the particle's aspect ratio is greater than the allowed limit, and count it as acceptable otherwise.

Single-stage and multi-stage sampling plans are used for AGR fuel materials. Multi-stage plans are attractive because they provide for instances of high quality product. Basically, the first stage of a multi-stage sampling plan is designed to allow a population to be accepted with a relatively small sample size when the actual quality level of the population is much better than the specification requirement. The subsequent stage(s) of the sampling plan allow for acceptance of populations having an actual quality level that is closer to the specification requirement. The advantage of such sampling plans is that they minimize waste of material in sampling/analysis when the product quality level is substantially better than the specified quality level.

Two-stage sampling plans are discussed here as they are the most commonly used for AGR fuel materials. Three-stage sampling plans are desirable when the necessary sample sizes get very large. They follow the same logic as two-stage plans. Single-stage plans are analogous to the final stage of a two- or three-stage plan.

Under the two-stage sampling plan, the first sample is drawn and then one of the following decisions is made. If the first sample is sufficiently good (in terms of the number of defective particles) the product is accepted without further testing. If it is sufficiently poor, the product is rejected. If the sample falls into neither category, a second sample is inspected before the decision is made.

More precisely, a two-stage sampling plan specifies an initial sample size n1, an acceptance number Ac1, and a rejection number Re1. If the first sample has Ac1 or fewer defective items, the product is accepted without further sampling. If the sample has Re1 or more defective items, the product is rejected. If the number of defective particles is between Ac1 and Re1, another sample of size n2 is taken. The cumulative acceptance and rejection numbers for the combined samples are denoted Ac2 and Re2. They satisfy Re2 = Ac2 + 1, so that the product will be either accepted (Ac2 or fewer defective items in the total of the two samples) or rejected (Re2 or more defectives) with no further sampling.

3.3   Sample Size Determination

For an attribute property, the sample size and acceptable number of defects are selected based on a balance between knowledge of process capability and the amount of testing that can realistically be performed. The knowledge of process capability is in terms of the suspected quality level of the population. If the quality level is such that the defect fraction is far from the specification, a relatively small sample would be sufficient to show conformance to the specification without an unacceptable risk of false rejection. If, however, the quality level of the population is close to the specification limit, the sample size that would be needed to have a reasonable probability of passing the acceptance test may be so large that the population fabrication process is not viable and must be improved.

3.3.1   Binomial, Hypergeometric, and Poisson Distributions

The acceptance and rejection numbers and the sample size for each sampling stage are determined based on the hypergeometric or binomial distribution (Bain and Engelhardt 1992), the required confidence level, and the acceptance criterion. The hypergeometric distribution is the theoretically correct distribution for this situation but the binomial distribution is an excellent approximation when the sample size is small relative to the population size. This is intuitively reasonable because the binomial distribution is applicable when we sample with replacement, while the hypergeometric distribution is applicable when we sample without replacement. If the size of the population sampled from is large, then it should not make a great deal of difference whether a particle is returned to the population before the next one is selected.

The probability density function (PDF) of the binomial distribution is given by

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, \dots, n$$

where:

x = number of defects,
n = number of samples,
p = probability of a defect, and
q = 1 - p.

The expression b(x; n, p) can also be written as $P(X = x / n, p)$ and is interpreted as the probability that exactly x defects are seen in a sample of size n from a population with a defect fraction of p.

The cumulative distribution function (CDF) of the binomial distribution is given by

$$B(x;n,p) = \sum_{x=0}^{n} b(x;n,p)$$

$$= \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, \dots, n.$$

This expression can be written as $P(X \leq x / n, p)$ and is interpreted as the probability that no more than x defects are seen in a sample of size n from a population with a defect fraction of p.

The Poisson distribution is the limiting form of the binomial distribution as the sample size increases and the defect fraction approaches zero. Hence, if the sample size is large and the defect fraction is near zero, the Poisson distribution can be used to approximate binomial probabilities.

The PDF of the Poisson distribution is given by

$$f(x;\mu) = \frac{e^{-\mu}\mu^{x}}{x!} \qquad x = 0, 1, 2, \ldots$$

where:

x = number of defects,
μ = np,
n = number of samples,
p = probability of a defect.

The CDF of the Poisson distribution is given by

$$F(x;\mu) = \sum_{k=0}^{x} \frac{e^{-\mu}\mu^{k}}{k!}.$$

Bain and Engelhardt (1992) state that, as a general rule, the approximation gives reasonable results provided the sample size is at least 100 and the defect fraction is less than or equal to 0.01. Use of the Poisson distribution to approximate binomial probabilities is attractive because of the simplicity of the calculations. However, with the widespread availability of statistical software capable of efficiently performing the binomial calculations, the advantage is not what it used to be.

3.3.2 Example of Two-Stage Sample Size Determination

To illustrate two-stage sampling, consider the product specification for SiC gold spot defects in coated particle composites. The specification is that the defect fraction in the population can be no more than 1E-3 with 95% confidence. A primary objective of the sample size determination is to identify a sample size that will give the population a reasonable chance of passing the acceptance test but not a sample size that is substantially larger than necessary. We use our knowledge of the process capability to construct this plan such that it provides no more than a 5% probability of false acceptance and a reasonably low probability of false rejection.

Consider Table 4 below. The fourth column presents the maximum number of defective particles, AC1 (for first stage) or AC2 (for first- and second-stage combined), that may be present in the sample to have 95% confidence that the population from which the sample was taken meets the specification, i.e., less than a 5% chance of false acceptance. The last column gives the probability of observing no more than AC1 (or AC2) defects in a sample of the size shown in column 2 given the true quality level of the population in column 1. Recall that the objective of the first stage of two-stage sampling is to sample such that very high quality particles are accepted and very poor quality particles are rejected. Suppose the quality level is expected, based on process knowledge, to be between 3E-04 and 7E-04. As shown in Table 4, if the actual quality level of the population is 3E-04, the sample size is 12,000, and 4 defects are observed in the sample, the specification is met with 95% confidence and the probability of false rejection is less than 8%. So 12,000 particles is a reasonable first-stage sample size.

Now suppose that the quality level of the population is at the upper end of the suspected range (at 7E-04) and that a sample size of 10,000 is selected for the second stage (for a cumulative sample size of

8

22,000). Per Table 4, the specification will be met with 95% confidence if there are 14 or fewer defects in the sample, but the probability of observing more than 14 defects is 0.5748, i.e., a false rejection rate of nearly 60%. Thus, a second-stage sample of 10,000 particles is too small; a much larger sample would be needed to give a reasonably high probability of passing the acceptance test. In this case, a second-stage sample size of 63,000 particles (for a cumulative sample size of 75,000) would be necessary to increase the probability of passing the acceptance test to 0.8645.   However, if the true defect fraction for gold spot defects is 5E-4, a second-stage sample size of 10,000 particles would provide for a reasonable false rejection probability of about 15% per Table 4. The data presented in Table 4 suggests that as a rule of thumb, the true defect fraction of a population of particles should be no more than about half of the specification limit in order to avoid excessively large samples or an unacceptably high probability of false rejection of in-specification product.

The acceptance number, AC1 (or AC2), and probability of $\leq$ AC1 (or AC2) defects given in the fourth and fifth columns below are binomial calculations easily performed in an Excel spreadsheet. The determination of first- and second-stage sample sizes is an iterative process and must be made in conjunction with expected quality level and prior feelings regarding reasonable sample sizes.

Table 4. Binomial probabilities for two-stage sampling of SiC gold spots in coated particles

Quality level of particles	Sample size	Expected number of defects in sample	Accept number for 1.0E-03 (95% conf.) (AC1 or AC2)	Probability of ≤ AC1 (or AC2) defects
3.00E-04	10000	3.0	4	0.8153
3.00E-04	12000	3.6	6	0.9268
3.00E-04	14000	4.2	7	0.9361
3.00E-04	16000	4.8	9	0.9749
3.00E-04	18000	5.4	10	0.9775
3.00E-04	20000	6.0	12	0.9912
3.00E-04	22000	6.6	14	0.9966
3.00E-04	24000	7.2	15	0.9969
5.00E-04	12000	6.0	6	0.6063
5.00E-04	14000	7.0	7	0.5987
5.00E-04	16000	8.0	9	0.7167
5.00E-04	18000	9.0	10	0.7060
5.00E-04	20000	10.0	12	0.7916
5.00E-04	22000	11.0	14	0.8541
5.00E-04	25000	12.5	16	0.8694
5.00E-04	28000	14.0	19	0.9235
5.00E-04	31000	15.5	21	0.9305
5.00E-04	34000	17.0	24	0.9594
7.00E-04	12000	8.4	6	0.2669
7.00E-04	14000	9.8	7	0.2387
7.00E-04	16000	11.2	9	0.3191
7.00E-04	18000	12.6	10	0.2875
7.00E-04	20000	14.0	12	0.3584
7.00E-04	22000	15.4	14	0.4252
7.00E-04	30000	21.0	20	0.4709
7.00E-04	50000	35.0	38	0.7291
7.00E-04	75000	52.5	60	0.8645
7.00E-04	100000	70.0	83	0.9436
9.00E-04	12000	10.8	6	0.0871
9.00E-04	22000	19.8	14	0.1127
9.00E-04	30000	27.0	20	0.1014
9.00E-04	50000	45.0	38	0.1663
9.00E-04	100000	90.0	83	0.2495
9.00E-04	200000	180.0	176	0.4015
9.00E-04	300000	270.0	271	0.5404
9.00E-04	500000	450.0	463	0.7394

4. STATISTICAL METHODS FOR VARIABLE PROPERTIES

4.1 Assumption of Normality

Variables are properties, such as coating thickness and density, which can be measured on a continuous scale. The methods used, confidence intervals and tolerance intervals, assume that the property under investigation follows a normal distribution. It may be the case that there are not sufficient data for testing normality (as in AGR-1) and therefore the assumption of normality must be made without validation. This should be clearly stated in the applicable fuel specification. If, however, there are sufficient data, then the following steps are recommended.

1. Perform the Shapiro-Wilk test for normality (Shapiro and Wilk 1965). If the data pass the Shapiro-Wilk test, continue with the acceptance test.

2. If the data fail the Shapiro-Wilk test, then investigate the departure from normality and assess its effect in terms of the robustness of the acceptance test. This is graphical and subjective and should involve the assistance of a statistician.

3. If the conclusions of steps 1 and 2 are that the assumption of normality is not met and the acceptance test is invalid due to the nature of the departure from normality, other methods of evaluating the data (such as treating the variable property as an attribute or performing a data transformation to make the data more normal) should be considered. This should also involve the assistance of a statistician.

4.1.1 Shapiro-Wilk Test for Normality

The Shapiro-Wilk test is a widely accepted formal test of normality (Shapiro, Wilk, and Chen 1968). The Shapiro-Wilk W statistic is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance (Shapiro and Wilk 1965). Large values of W indicate normality, small values indicate non-normality. A p-value (the probability of obtaining a W statistic less than or equal to the observed value) is calculated. If this p-value is less than 0.01, then the assumption of normality is rejected.

The Shapiro-Wilk W statistic is a difficult calculation when done by hand or in a spreadsheet. However, it is readily available in most commercial statistical software packages such as SAS (SAS 2004) and Statistica (StatSoft 2002).

4.1.2 Graphical Techniques for Assessing Robustness of Acceptance Tests

If the Shapiro-Wilk test rejects normality, it must still be decided whether the non-normality is extreme enough to make a practical difference; with a large data set even small departures from normality can be detected, but these small departures may have little effect on the properties of the acceptance test procedure.

Two commonly used graphical techniques for investigating departures from normality are normal probability plots and histograms. The standard normal probability plot is constructed as follows. First the values are rank ordered. From these ranks, z values (i.e., standardized values of the normal distribution) are computed based on the assumption that the data come from a normal distribution. Specifically, the normal probability value z_j for the j'th ordered value (rank) in a variable with n observations is computed as:

$z_j = \Phi^{-1}[(3*j-1)/(3*n+1)]$

where Φ^{-1} is the inverse standard normal cumulative distribution function (converting the normal probability p into the normal value z). Φ^{-1} can be calculated by the NORMSINV statistical function in an Excel spreadsheet.

These z values are plotted on the y-axis in the plot. If the observed values (plotted on the x-axis) are normally distributed, then all values should fall onto a straight line in the plot. If the values are not normally distributed they will deviate from the line. The deviations suggest the manner in which this population seems to depart from normality.

Simple histograms are helpful when there are 30 or more measurements. Histograms depict a variety of data characteristics. Both the construction and interpretation of histograms is subjective; however, they are a quick and easy way to assess the type and scale of suspected non-normality.

The use of normal probability plots and histograms are demonstrated below in an example.

A statistical procedure is called robust if moderate changes in the model assumptions produce only small changes in the properties of the procedure. Robust procedures are insensitive to the underlying model assumptions.

The robustness of the acceptance tests for kernel diameter is investigated here, by considering the assumption that the population is not normal but instead has the form of the distribution determined by 1000 randomly selected measurements of over 4700 measurements taken on BWXT NUCO kernels 69300R-38. These kernels have mean 346.46 µm and standard deviation 12.32 µm. A histogram of this distribution is plotted in Figure 1, overlaid with a normal density having the same mean and standard deviation. Figure 2 shows the actual measurements plotted against the standard normal values that would be expected from a normal population with the measured mean and standard deviation (normal probability plot). The Shapiro-Wilk test conclusively rejects normality of the population; this is in part because a sample size of 1000 is large enough to allow detection of even small departures from normality. As can be seen in Figures 1 and 2, the distribution resembles a normal distribution, although in both figures a substantial number of measurements are slightly to the right of the corresponding standard normal values, and two measurements are far to the left of the standard normal values.

The empirical distribution has $\mu = 346.46$ µm, $\sigma = 12.32$ µm, and 99th percentile = 376.88 µm. Therefore the empirical distribution satisfies

99th percentile = $\mu + 2.469\sigma$.

This is not very different from the relation for a normal distribution,

99th percentile = $\mu + 2.326\sigma$.

This is also illustrated by Figure 2. The 99th percentile can be found there by counting 10 data points from the right. The plotted point is very close to the diagonal line, indicating that the empirical 99th percentile is close to the normal 99th percentile. Therefore, tests for the dispersion that are based on the sample mean and sample standard deviation behave well, even though there are some obvious departures from normality.
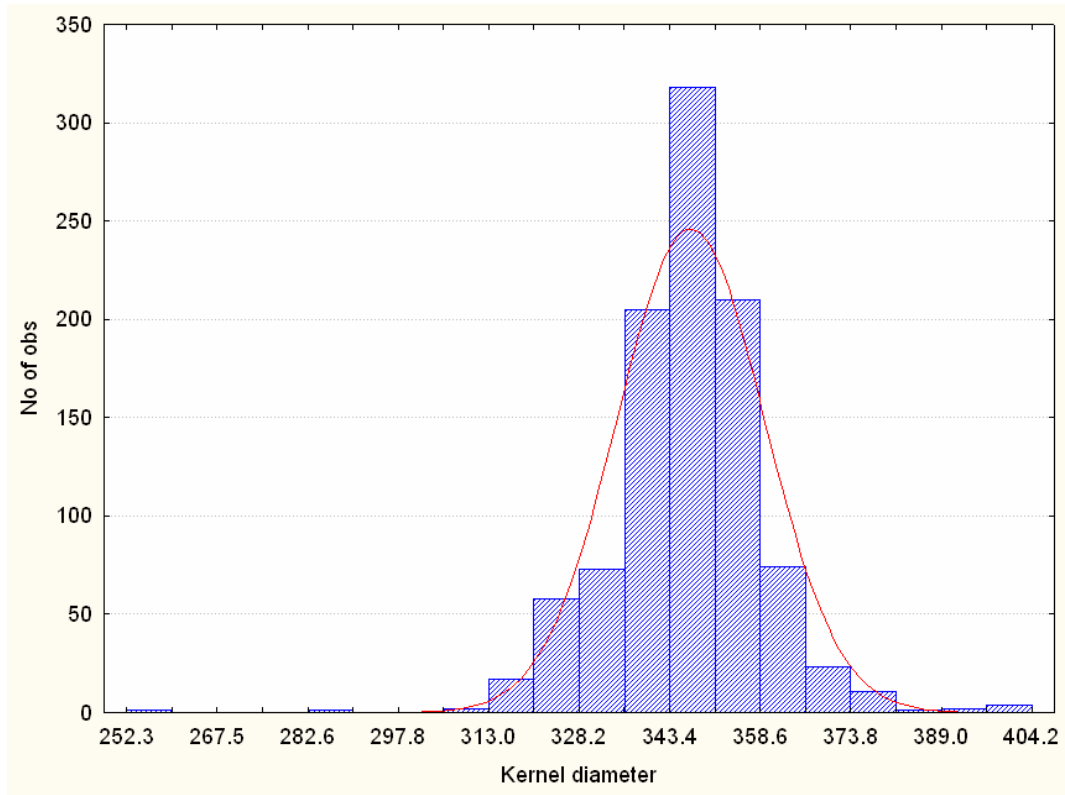
Figure 1. Histogram of 1000 kernel diameters with normal density superimposed.
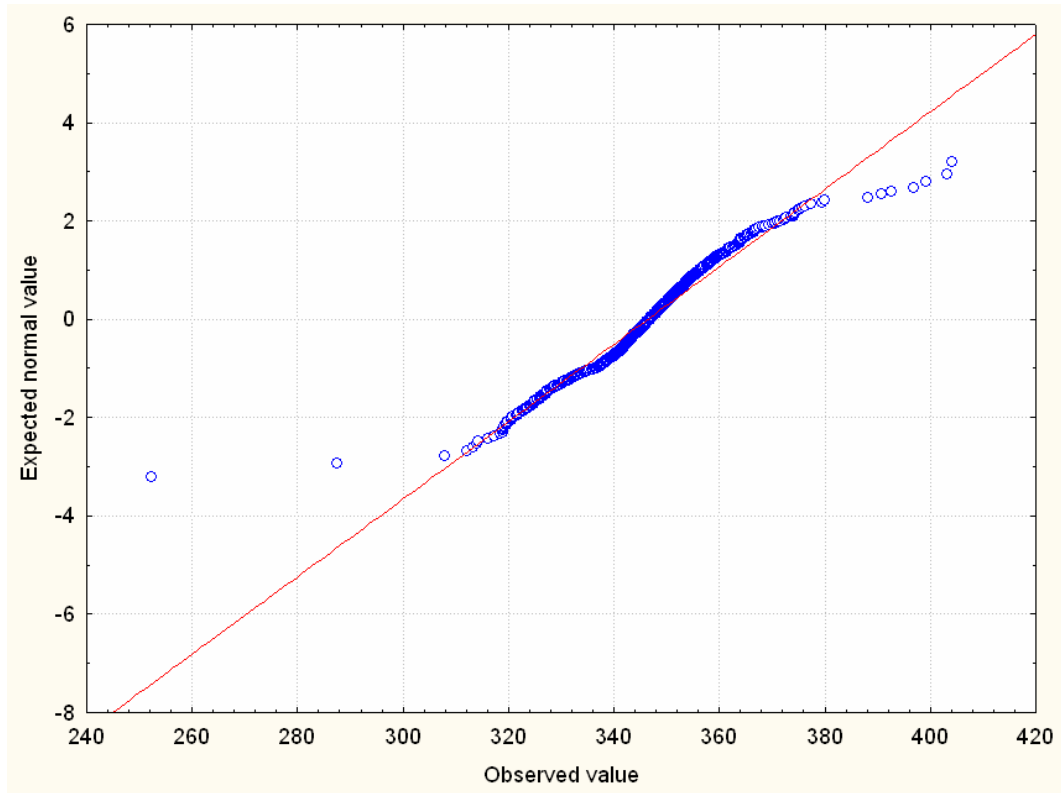
Figure 2. Normal probability plot of 1000 kernel diameters.

4.1.3   Discussion

Consider first criteria for the mean. In this case, the sample mean is used to estimate the population mean. The Central Limit Theorem of statistical theory states that the sample mean is approximately normally distributed, and therefore the test based on normality behaves well enough, unless the sample size is extremely small or the population distribution is very non-normal. This is an important point. Its implication to the AGR program is that if the specification is on the mean, then departures from normality that are not drastic do not affect the validity of the acceptance test.

There are rules of thumb for the minimum sample size needed to invoke the Central Limit Theorem. For example, Kennedy and Neville (1986) state: a sample size of at least 4 is needed if the population distribution is symmetrical and not radically different from a normal distribution; a sample of at least 12 is needed if the population distribution is fairly symmetrical with no prominent mode; a sample size of 100 or more is needed if the distribution is ill-behaved, with most density in the tails.

Kennedy and Neville's standards for an acceptable approximation to the normal distribution are not stated, but as a minimum they presumably want to compute a 95% confidence interval for the mean. The AGR criterion for the mean is less demanding, involving only a 90% confidence interval for the mean (5% for each tail, not 2.5%). Therefore, Kennedy and Neville's rules may well be sufficient, or even conservative, for the AGR applications. Of course, it is necessary to decide whether the underlying population is nearly normal, ill-behaved, or somewhere in between. The data give guidance when there are many measurements, for example when there are 200 measured kernel diameters. There also may be theoretical reasons for believing that the quantity being measured has an approximately normal distribution. For example, if the variation in the measured property is caused by many small and nearly independent variations in the manufacturing process, the Central Limit Theorem states that the resulting

14

measured quantity is approximately normally distributed. For such a property, a sample size as small as 20 is certainly sufficient for criteria for the mean, and an even smaller sample is good enough if the test is passed conclusively.

Consider now criteria for the dispersion. Here, the robustness issue is much more important than with criteria for the mean. The normal test uses the tail of the distribution to conclude that only a small fraction of the population is beyond the critical limit. This extrapolation can be sensitive to the shape of the distribution and since the acceptance criterion is on dispersion and not the mean, the Central Limit Theorem can not be invoked.

If enough data have been collected to give confidence that the underlying distribution is approximately normal, then tests based on the normal assumption may be used for future products manufactured by the same process. If instead few measurements have been taken, the assumption of normality is harder to justify. In the early stage of production, it may be necessary to use the test based on normality, but with caution and after assessment of any potential departures from normality. Thus, it is important to determine the validity of the assumption of normality and to evaluate the impact of departures from normality per the approach outlined in Section 4.1 when performing acceptance tests for population dispersion criteria.

4.2   Acceptance Tests

An acceptance test is simply a decision rule for determining acceptance or rejection of a population based on a sample. The following describes the acceptance testing procedure.

4.2.1   Test for the Mean

When testing for a variable property based on the population mean, and given a two-sided acceptance criterion, the quantities

$$A = \overline{X} - ts/\sqrt{n} \ \ \text{and} \ \ B = \overline{X} + ts/\sqrt{n}$$

are calculated, where

$\overline{X}$ = sample mean,
t = Student's t value based on sample size and confidence level,
s = sample standard deviation, and
n = sample size.

If

$A < L_m$, the lower acceptance limit for the mean, or
$B > U_m$, the upper acceptance limit for the mean,

then the population is rejected. Otherwise the population is accepted, based on the sample. The 95% confidence level is achieved by use of the 95th percentile of Student's t distribution, with n-1 degrees of freedom. For a one-sided interval, the method is the same except that only A or B applies. The 95th percentile of Student's t ensures that an unacceptable population is rejected with probability 0.95.

Note the 95th percentile of t is used whether the acceptance criterion is one- or two-sided. To see why this is correct, consider a two-sided criterion for the mean. Suppose that the true mean μ is at the

borderline between acceptable and unacceptable, that is, it equals either L_m or U_m. It cannot equal both, so assume for this example that $\mu = U_m$. Recall that B is defined using the 95th percentile of Student's t. Therefore, the usual algebra for confidence interval shows that $\Pr(B > U_m) = 0.95$, that is, this borderline-unacceptable population is rejected with probability 0.95. In addition, because the criterion is two-sided, there is a test based on the lower acceptance limit for the mean, L_m. Although it is unlikely that A will be less than L_m when μ is as large as U_m, there is still a small chance of rejecting the population for this reason. Therefore, the two tests together cause this borderline population to be rejected with probability somewhat greater than 0.95. It is worth noting that if there are also other tests, perhaps based on a criterion for the dispersion or on other properties altogether, the probability that the population is rejected is further increased. In summary, the 95th percentile of Student's t ensures that an unacceptable population is rejected with probability at least 0.95. With a one-sided acceptance criterion, a borderline population (that is, a population at the boundary between acceptable and unacceptable) is rejected with probability 0.95; when the criterion is two-sided, or when there are more acceptance criteria, a borderline population is rejected with probability somewhat greater than 0.95.

4.2.2 Test for Dispersion

The acceptance test of a variable property based on the population dispersion, using the assumption of normality, is as follows. Given a two-sided acceptance criterion, the quantities

$$C = \overline{X} - ks \text{ and } D = \overline{X} + ks$$

are calculated, where,

$\overline{X}$ = sample mean,
k = tolerance factor for normal distribution based on sample size,
    confidence level, and allowable fraction outside of critical limit, and
s = sample standard deviation.

If

$C < L_p$, the lower critical limit for the population, or
$D > U_p$, the upper critical limit for the population,

then the population is rejected. Otherwise the population is accepted. For one-sided acceptance criteria, the method is the same except that only C or D applies. The confidence level is achieved by the appropriate use of the tolerance factor, k. One-sided tolerance factors are used for both one- and two-sided acceptance criteria.

Student's t values, standard normal z values (see Section 4.3.2 for application), and one-sided tolerance factors (k) for sample sizes up to 100 are given in Table 5.

Table 5. Student's t values[a], standard normal values[b] (z), and tolerance factors[c] (k).

Sample size	$t_{0.95}$	$z_{0.95}$	$k_{0.95/0.99}$	$k_{0.95/0.95}$
2	6.314	1.645	37.094	26.260
3	2.920	1.645	10.553	7.656
4	2.353	1.645	7.042	5.144
5	2.132	1.645	5.741	4.203
6	2.015	1.645	5.062	3.708
7	1.943	1.645	4.642	3.399
8	1.895	1.645	4.354	3.187
9	1.860	1.645	4.143	3.031
10	1.833	1.645	3.981	2.911
11	1.812	1.645	3.852	2.815
12	1.796	1.645	3.747	2.736
13	1.782	1.645	3.659	2.671
14	1.771	1.645	3.585	2.614
15	1.761	1.645	3.520	2.566
16	1.753	1.645	3.464	2.524
17	1.746	1.645	3.414	2.486
18	1.740	1.645	3.370	2.453
19	1.734	1.645	3.331	2.423
20	1.729	1.645	3.295	2.396
21	1.725	1.645	3.263	2.371
22	1.721	1.645	3.233	2.349
23	1.717	1.645	3.206	2.328
24	1.714	1.645	3.181	2.309
25	1.711	1.645	3.158	2.292
26	1.708	1.645	3.136	2.275
27	1.706	1.645	3.116	2.260
28	1.703	1.645	3.098	2.246
29	1.701	1.645	3.080	2.232
30	1.699	1.645	3.064	2.220
31	1.697	1.645	3.048	2.208
32	1.696	1.645	3.034	2.197
33	1.694	1.645	3.020	2.186
34	1.692	1.645	3.007	2.176
35	1.691	1.645	2.995	2.167
36	1.690	1.645	2.983	2.158

Table 5 (continued).

Sample size	$t_{0.95}$	$Z_{0.95}$	$k_{0.95/0.99}$	$k_{0.95/0.95}$
37	1.688	1.645	2.972	2.149
38	1.687	1.645	2.961	2.141
39	1.686	1.645	2.951	2.133
40	1.685	1.645	2.941	2.125
41	1.684	1.645	2.932	2.118
42	1.683	1.645	2.923	2.111
43	1.682	1.645	2.914	2.105
44	1.681	1.645	2.906	2.098
45	1.680	1.645	2.898	2.092
46	1.679	1.645	2.890	2.086
47	1.679	1.645	2.883	2.081
48	1.678	1.645	2.876	2.075
49	1.677	1.645	2.869	2.070
50	1.677	1.645	2.862	2.065
60	1.671	1.645	2.807	2.022
70	1.667	1.645	2.765	1.990
80	1.664	1.645	2.733	1.964
90	1.662	1.645	2.706	1.944
100	1.660	1.645	2.684	1.927
∞	1.645	1.645	2.326	1.645

a.   Student's t values were generated using the tinv function of the SAS statistical software (SAS 2004).
b.   Standard normal z values are from Beyer (1968).
c.   Tolerance factors (k) are from Guttman (1970).

4.3   Sample Size Determination

Recall two kinds of error that have been mentioned: false acceptance, that is, acceptance of a product that does not, in fact, satisfy the required criteria; and false rejection, rejection of a product that satisfies the criteria. These errors can occur because each decision to accept or reject is based on a random sample, not on examination of the entire population. The probability of each kind of error can be found by calculating the probability of rejecting the population, assuming that the population is specified. Ideally, the probability of rejection should be large for an unacceptable population and small for an acceptable population. Sample size determination is based on the objective of minimizing the two types of errors with the constraint that the resulting sample size is reasonable.

Two approaches are considered here for determining sample sizes. The first is graphical through the use of power curves. The second is more objective and uses an equation which specifies the desired false acceptance and false rejection rates.

4.3.1   Power Curves for Sample Size

A power curve is a graphical depiction of the probability of rejection of a property as a function of the true mean of the property and fixed sample size and standard deviation. The probability of acceptance is simply 1.0 minus the probability of rejection. The power curve shows the probability of getting a sample that leads to rejection of the population, if in fact the population is normal with the stated mean and standard deviation. By varying the mean, a curve over the range of means is constructed. By varying the sample size, several curves are produced.

Consider the acceptance criteria for kernel diameter of a composite. In this example, there are two acceptance criteria for the population. The criterion on the mean requires that the population mean be less than the upper acceptance limit of 360 μm and greater than the lower acceptance limit of 340 μm. The criterion on the dispersion requires that at most 1% of the population be above the critical limit of 400 μm and at most 1% of the population be below the critical limit of 300 μm. For a normal population with mean μ and standard deviation σ, this criterion is equivalent to requiring that $\mu + 2.326\sigma \leq 400$ μm and $\mu - 2.326\sigma \geq 300$ μm, because 2.326 and -2.326 are the 99th and 1st percentiles of a standard normal distribution, respectively. Note that the criterion for dispersion is not a function of the sample size. Therefore, the sample size determination using power curves is based on the criterion for the mean. As the sample size increases, the probability of rejection when the true mean is within the specifications decreases.

Figure 3 depicts the probability of rejection (power curve) for kernel diameter assuming a standard deviation of 10 and varying the mean and sample size. The specifications on the mean are shown as vertical lines at 340 and 360 μm. The figure shows that for a small sample size (say n=10) the probability of rejection when the true mean is within the specification is unacceptably high. For example, if the true mean is 355 μm, the probability of rejecting the composite based on a sample of size 10 is more than 0.50. If the sample size is increased to 50, the probability of rejecting a composite with a true mean of 355 is less than 0.05.

4.3.2   Equation for Sample Size

The following equation can be used for determining sample size (EPA 1989). It requires a specification of the false acceptance rate, α, the false rejection rate, β, and estimates of true mean and true standard deviation.

$$n = \sigma^2 \left(\frac{z_{1-\alpha} + z_{1-\beta}}{spec - \mu} \right)^2 \text{, where}$$

σ = the true standard deviation,
μ = the true mean,
α = the false acceptance rate,
β = the false rejection rate,
$z_{1-\alpha}$ = the standard normal value corresponding to 1-α,
$z_{1-\beta}$ = the standard normal value corresponding to 1-β, and
spec = the property acceptance specification.

Consider the example used in the preceding section. Assume a true standard deviation of 10, and a true mean of 355. Since 355 is nearest to the upper specification on the mean, take "spec" to be 360. The
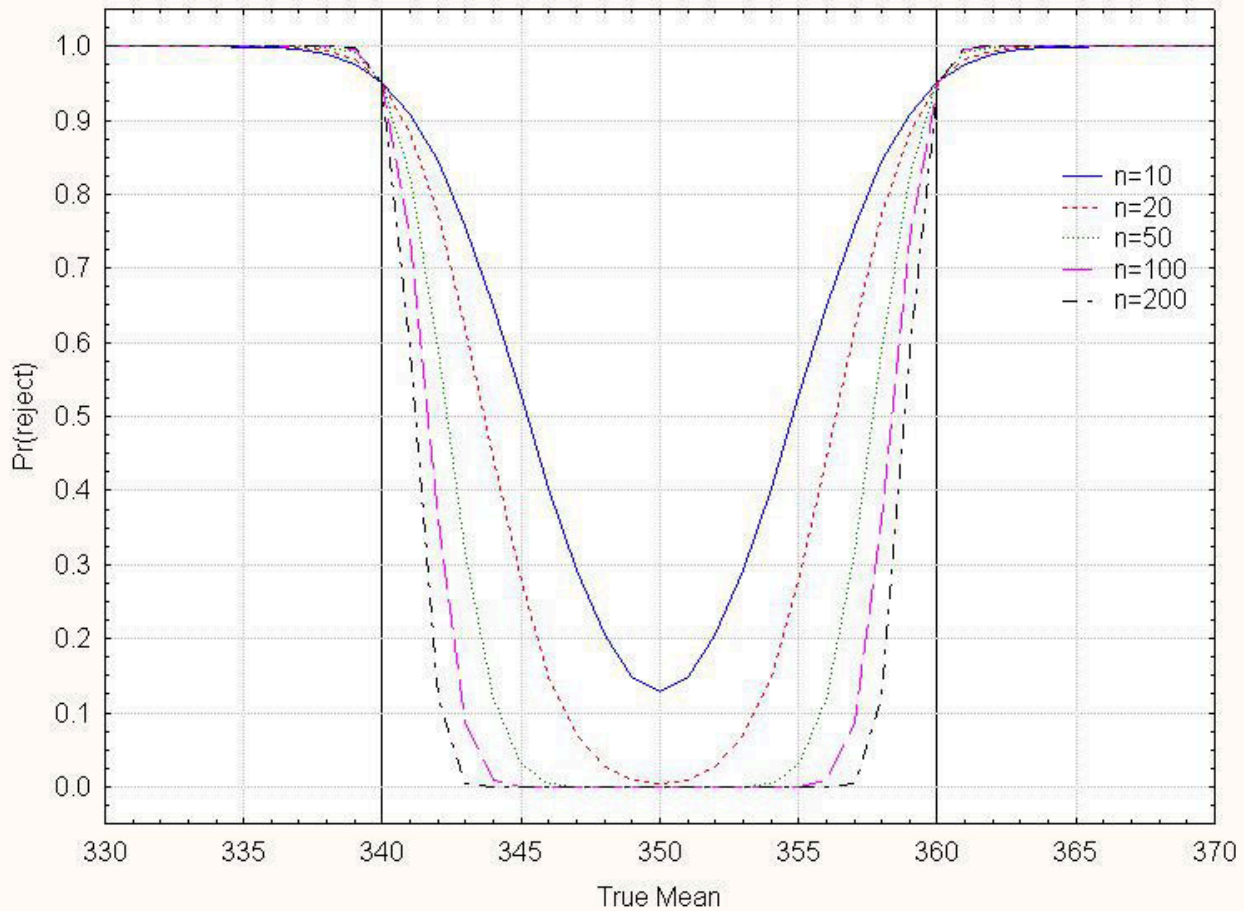
Figure 3. Probability of rejecting composite for mean kernel diameter when σ = 10.

required false acceptance rate is 0.05 (95% confidence). Suppose it is desired to maintain a false rejection rate of no more than 0.05. Then,

$$n = 10^2 \left(\frac{1.645 + 1.645}{360 - 355} \right)^2 = 44.$$

The result is a sample size of 44 (always round up). Note that this corresponds to the approximate sample size of 50 picked off the power curve but is more precise and objective. However, it still requires estimates of the true mean and true standard deviation, and the desirable false rejection rate. (The false acceptance rate is programmatically fixed at 0.05).

Required sample sizes for the AGR-1 experiment are given in EDF-4572, "Statistical Sampling Plan for AGR Fuel Materials" (Einerson 2005) and address the specifications listed in Tables 1, 2, and 3 of this report. As specifications for future experiments are finalized, sampling plans utilizing the methods described in this report will be developed.

5. REFERENCES

Bain, L. J., and M. Engelhardt, 1992, *Introduction of Probability and Mathematical Statistics, Second Edition*, Belmont, CA: Duxbury Press, 1992.

Barnes, C. M., 2005, "AGR-1 Fuel Product Specification and Characterization Guidance", EDF-4380, Revision 2, Idaho National Laboratory, April 14, 2005.

Beyer, W. H. (editor), 1968, *CRC Handbook of Tables for Probability and Statistics, Second Edition*, Boca Raton, FL: CRC Press Inc., 1968

Einerson, J. J., 2005, "Statistical Sampling Plan for AGR Fuel Materials", EDF-4542, Revision 4, Idaho National Laboratory, April 14, 2005.

EPA, 1989, *Methods for Evaluating the Attainment of Cleanup Standards, Volume 1, Soils and Solid Media*, EPA 230/2-89/042, Environmental Protection Agency, Washington, DC, February 1989.

Kennedy, J. B., and A. M. Neville, 1986, *Basic Statistical Methods for Engineers and Scientists, Third Edition,* New York: Harper & Row, 1986.

Petti, D. A., J. T. Maki, M. A. Ebner, G. K. Miller, 2004, "Preliminary AGR Fuel Specification", EDF-4198, Revision 1, Idaho National Engineering and Environmental Laboratory, September 22, 2003.

SAS Institute Inc. 2004. *SAS OnlineDoc® 9.1.3.* Cary, NC: SAS Institute Inc., 2004

Scheffel, W., and J. Saurwein, 2003, *Preliminary Fuel Product Specifications for the Baseline Advanced Gas Reactor Fuel Design*, 911034, Rev. 0 Draft, General Atomics, August 2003.

Shapiro, S. S. and M. B. Wilk, 1965, *An Analysis of Variance Test for Normality (complete samples)*, Biometrika, 52, 591 - 611.

Shapiro, S. S., M. B. Wilk and H. J. Chen (1968), *A comparative study of various tests of normality,* Journal of the American Statistical Association, 63, 1343-1372.

StatSoft, Inc., 2002. *STATISTICA for Windows [Computer program manual].* Tulsa, OK: StatSoft, Inc., 2002.