U. S. Department of Commerce
National Bureau of Standards

Research Paper RP1827
Volume 39, September 1947

Part of the Journal of Research of the National Bureau of Standards

Acceptance Sampling by Variables, With Special Reference to the Case in Which Quality Is Measured by Average or Dispersion[1]

By John H. Curtiss

This paper is devoted to a presentation of the theory and practice of certain types of acceptance sampling plan based on the statistical tests of hypotheses. The basic concepts of the statistical theory are discussed in detail, and are then applied to obtain a number of specific formulas for the single sampling case.

I. Introduction

Acceptance sampling is one of the most interesting and useful applications of modern mathematical statistics. Involving as it does the principles of experimental design and of the testing of statistical hypotheses, it furnishes a proving ground for these theories in which the validity of the basic assumptions is often more readily verifiable, and the correctness of the inference often more rapidly ascertainable, than in many of the standard biometric and sociometric applications.

This paper will be devoted to a presentation of the theory and practice of certain types of sampling plans which are based on statistical tests of hypotheses. Such a plan is intended to act as a sieve for separating unsatisfactory material from acceptable product, to be used with the understanding that there is no guarantee that an occasional poor lot will not accidentally pass through the sieve. The distinction between this type of acceptance plan and certain other popular types is made in the earlier sections of the paper, and the basic concepts of the statistical theory are then discussed in considerable detail. The later sections are devoted to the details of the application of the basic principles to the derivation of explicit formulas for sample size and acceptance numbers. These derivations are given because there seems to be no collection of these derivations and of the formulas themselves readily available in the literature.

II. What Is Acceptance Sampling?

Because statisticians who are not primarily interested in the engineering applications may wish to follow the discussion for a while, it might be well to start by explaining from the very beginning what acceptance sampling is all about.

Acceptance sampling is a branch of the science of engineering statistics, or industrial statistics, or statistical quality control; that is, the science of dealing with the variability encountered in data arising from engineering experiments and industrial processes. Consider a stream of nominally identical items produced by an industrial process. The items are supposed to be all exactly alike, but of course they never really are. Perhaps the most fundamental and natural way to study their variability is to express the several quality characteristics of each item in terms of a like number of variables which assume numerical values, and then to investigate the properties of the joint frequency distribution of these variables as built up for an aggregate of the items.

One obvious way to perform the investigation consists in actually setting up the frequency distribution in complete detail by making a 100-percent inspection of the items in the aggregate. But for most purposes a much better way to conduct the investigation is to inspect only a portion of the aggregate, and from the data so obtained, make estimates and draw inferences concerning the nature of the frequency distribution of the aggregate. When the portion, or *sample*, is properly chosen, and the methods of inference properly

[1] An invited address delivered to a joint session of the American Statistical Association and the Institute of Mathematical Statistics in Cleveland, Ohio, on January 27, 1946.

founded on sound theory, an exact balance can often be achieved between economy in cost of inspection and the needed accuracy of conclusions.

As the name implies, acceptance sampling is a special kind of sampling investigation of an aggregate in which the sample data are used not only to study the distribution of quality characteristics of the aggregate, but also to provide a criterion as to whether the aggregate should be accepted or rejected by a purchaser or consumer.

There are two basic types of acceptance sampling plans. In one type, the entire interest is centered on inferences drawn from the sample relating to the frequency distribution of the aggregate, and rejection or acceptance is specified in accordance with whether or not the inferences so drawn deny or affirm that the frequency distribution of the aggregate meets certain prestated requirements. This will hereinafter be called the *inferential* type of acceptance sampling plan. The other type of acceptance sampling plan is based on the assumption that a rejected aggregate will always be subjected to corrective action (such as 100-percent inspection) which will surely remove all, or a known fraction, of the deficiencies which caused rejection. The plan is then set up so that after the entire inspection process the distribution of the quality-defining variables in the *final outgoing* sequence of aggregates will meet certain prestated requirements. Since the requirements generally pertain to the mean values of the variables in the outgoing sequence of aggregates, this sort of plan might appropriately be termed an *average outgoing quality* type of acceptance sampling plan. Such plans were invented by Dodge and Romig [1].[2]

When there are only a few admissible values of a given quality-defining variable (and the case most frequently encountered is that in which there are only two values, conveniently taken to be 0 and 1, corresponding to whether or not the item meets a specified standard), or when the sampling acceptance-rejection requirements are based on the observed proportion of the total number of sample measurements which fall into a few (say 2 or 3) intervals in the range of a quality defining variable, the associated sampling inspection process is termed acceptance sampling "by attributes." More general cases are grouped together under the generic title of acceptance sampling "by variables." Theoretically, the distinction between "attributes" and "variables" is quite artificial and the boundary is indefinite; "attributes" sampling is really just a highly developed special case of "variables" sampling.

The average outgoing quality type of acceptance sampling plan has been developed only for sampling by attributes and only in the case in which 100-percent inspection (not partial inspection) of rejected lots is specified. The average outgoing quality theory for more general cases, and in particular for the case in which a rejected lot is only partially inspected, apparently has not yet been studied in any detail. This paper will be devoted henceforth entirely to the inferential type of acceptance sampling plan, and the average outgoing quality type will not be mentioned again.

III. An Example

To illustrate how inferential acceptance sampling plans are set up and used, a simple practical example based on a wartime procurement problem will now be presented. For expository purposes and for consideration of commercial confidence, the example has been considerably simplified and the actual dimensions have been altered.

During the war, a small metal cylinder for dispensing insecticide in the form of an aerosol was manufactured in large quantities for the U. S. Navy. The cylinders were manufactured by one set of contractors and filled by another set of contractors. It was necessary to control the volumetric capacity of the cylinder rather carefully, because the filling process used by some of the filling contractors was not automatically adjustable to the size of the individual cylinder. Overfills presented a hazard of explosions when the cylinders were stored in the sun.

A preliminary survey revealed that the frequency distribution of volumetric capacities of cylinders produced by each manufacturer when the manufacturer was aiming at one given nominal capacity, would be of the type illustrated in idealized form in figure 1. The preliminary survey consisted of determining the frequency distribution of a random sample of 1,800 cylinders selected from the first 100,000 produced by each manufacturer. The standard deviation of this observed frequency distribution in each case was about 2.0 cc. A normal (or Gaussian) theoretical distribu-

[2] Figures in brackets indicate the literature references at the end of this paper.

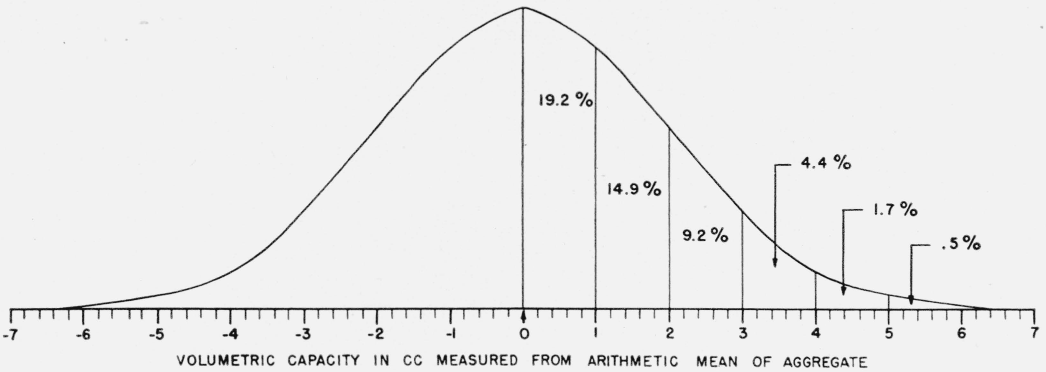

FIGURE 1. *Theoretical frequency distribution of volumetric capacities of cylinders to be used for insecticide dispensers.*

tion gave a satisfactory fit to the data; it is the frequency curve of this distribution which actually appears in figure 1.

The variability in successive determinations of capacity was probably due in no small part to testing error.[3] The chief manufacturing variable affecting volume was in the brazing operation by which the two halves of the cylinder were sealed together. Certain of the manufacturers were known to be experimenting with the nominal size of their cylinders, but it was believed that such changes would probably affect only the mean of the distribution of volumetric capacities, and not the standard deviation.

It seemed probable therefore that the proper control over volumetric capacity could be achieved by first formulating suitable restrictions on the size and pedigree of an aggregate or "lot" of cylinders so that it would have a distribution of the form shown in figure 1, and then imposing requirements on the arithmetic mean of this distribution which were enforced by an acceptance sampling plan.

The appropriate definition of lot will not be detailed here; this is purely an engineering matter. It was decided from the chemical and engineering considerations involved, that if the arithmetic mean of the distribution of volumetric capacities in a lot were to be reduced to less than 530.5 cc, the lot (in effect here, the process) [4] should almost surely be rejected, and if the lot mean remained above 534.0 cc, the lot should almost surely be accepted.

The problem of setting up the acceptance sampling plan that would provide the basis for deciding whether or not a lot meets these requirements can be solved in many different ways. The method which seems to be the most natural one from an intuitive point of view, and which indeed does theoretically have certain optimum properties (see section V below), consists in selecting a random sample of several cylinders from each lot, finding the average of the volumetric capacities of these cylinders, and accepting or rejecting the lot arbitrarily in accordance with whether or not this average is greater than some predetermined "acceptance number."

To determine the sample size and the value of the acceptance number, it is necessary to study the sampling distribution of the average volumetric capacity of a random sample of cylinders taken from a population with a distribution shown in figure 1. It is of course well known that if the population has a normal distribution with mean μ and standard deviation σ, the distribution of the average of a random sample of n observations will be normal with mean μ and standard deviation $\sigma/\sqrt{n}$. The distribution of the average of $n=4$ cylinders is shown in figure 2. From figure 3 it will be seen that if an acceptance number of 532.0 cc is specified, using a random sample of 4, the probability (or "risk") of accidentally rejecting a lot with a mean volumetric capacity of more than 534.0 cc would not exceed 0.023, and the probability of accidentally accepting a lot with mean volumetric capacity of less than 530.5 cc would not exceed 0.067.

[3] It should be noted that the population which is under study in any acceptance sampling plan for a given quality defining variable is the population of *apparent values* of the variable as determined by some test or inspection procedure. Thus the variance of the population is made up of two components: one due to the nonuniformity in the material and one due to lack of precision in the test.

[4] See section XII.

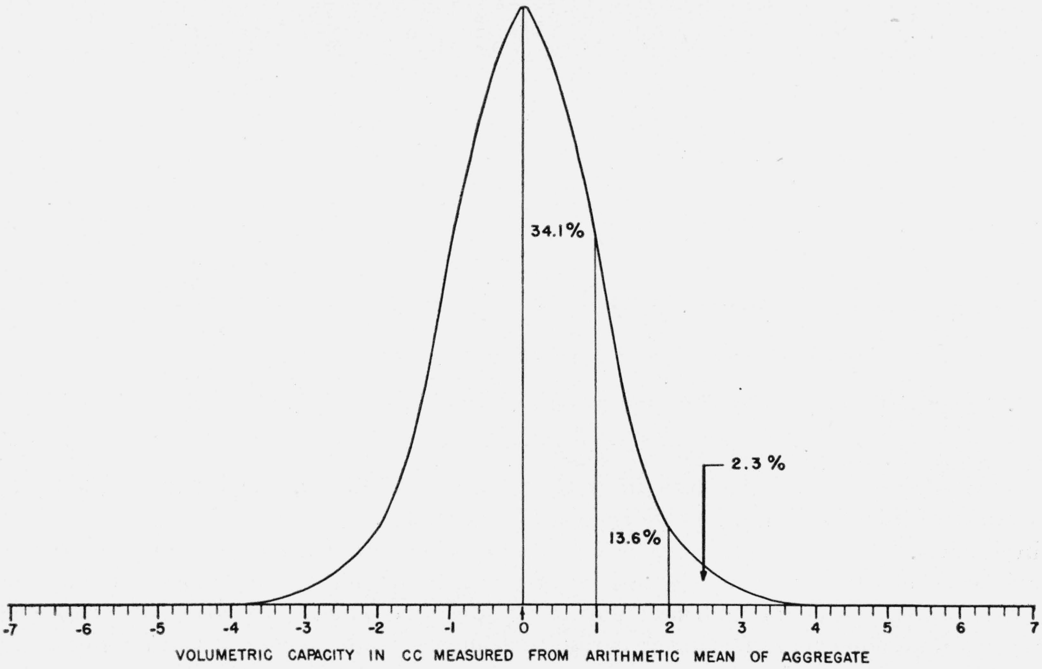

FIGURE 2. *Probability distribution of mean of random sample of four cylinders, derived from the theoretical distribution in figure 1.*

Either one of these two risks can obviously be decreased at the expense of increasing the other by merely changing the position of the acceptance number appropriately. By using a larger sample size, or by multiple or sequential sampling, it would be possible to reduce both of the risks simultaneously. The general order of magnitude of the risks in the sampling plan illustrated by figure 3 is considered to be appropriate for such work.

IV. Acceptance Sampling and Statistical Inference

Inferential acceptance sampling plans may be classified as statistical or nonstatistical, in accordance with whether or not they are based on the principles of statistical inference.

In nonstatistical inferential plans, sample items are purposively selected in accordance with prior information, and from the observations on the sample items a supposedly *certain* inference is made concerning the nature of the sampled aggregate. Such acceptance sampling plans will not be considered in this paper. They are useful and efficient in special cases, but their theory is a function of the application and does not lend itself readily to a general discussion.

Statistical inferential acceptance sampling plans are characterized by a selection of sample items from an aggregate in accordance with some method which insures that the observations on the respective items in the sample have a joint probability distribution which bears some direct and calculable relation to the frequency distribution of the aggregate. Through the use of this probability distribution, an *uncertain* inference into the nature of the distribution of the aggregate is drawn from the data of a sample. The example given in section III was a statistical acceptance sampling plan. The "random" selection of the sample cylinders implies a unique determination of the probability distribution of the sample in terms of the frequency distribution of volumetric capacities in the sampled aggregate.

The general theory of the uncertain inferences involved in statistical acceptance sampling plans may conveniently be identified with a branch of modern mathematical statistics called tests of statistical hypotheses.[5] A statistical hypothesis is a statement concerning the unknown frequency

[5] Simon [2] prefers to identify the theory of statistical acceptance plans with the theory of estimation of population parameters. The approach via the theory of statistical hypotheses, which was pioneered by Dodge and Romig [1], has the advantage that it provides a quantitative basis for a rational selection of the accuracy of estimation required in a given application; see section VI.

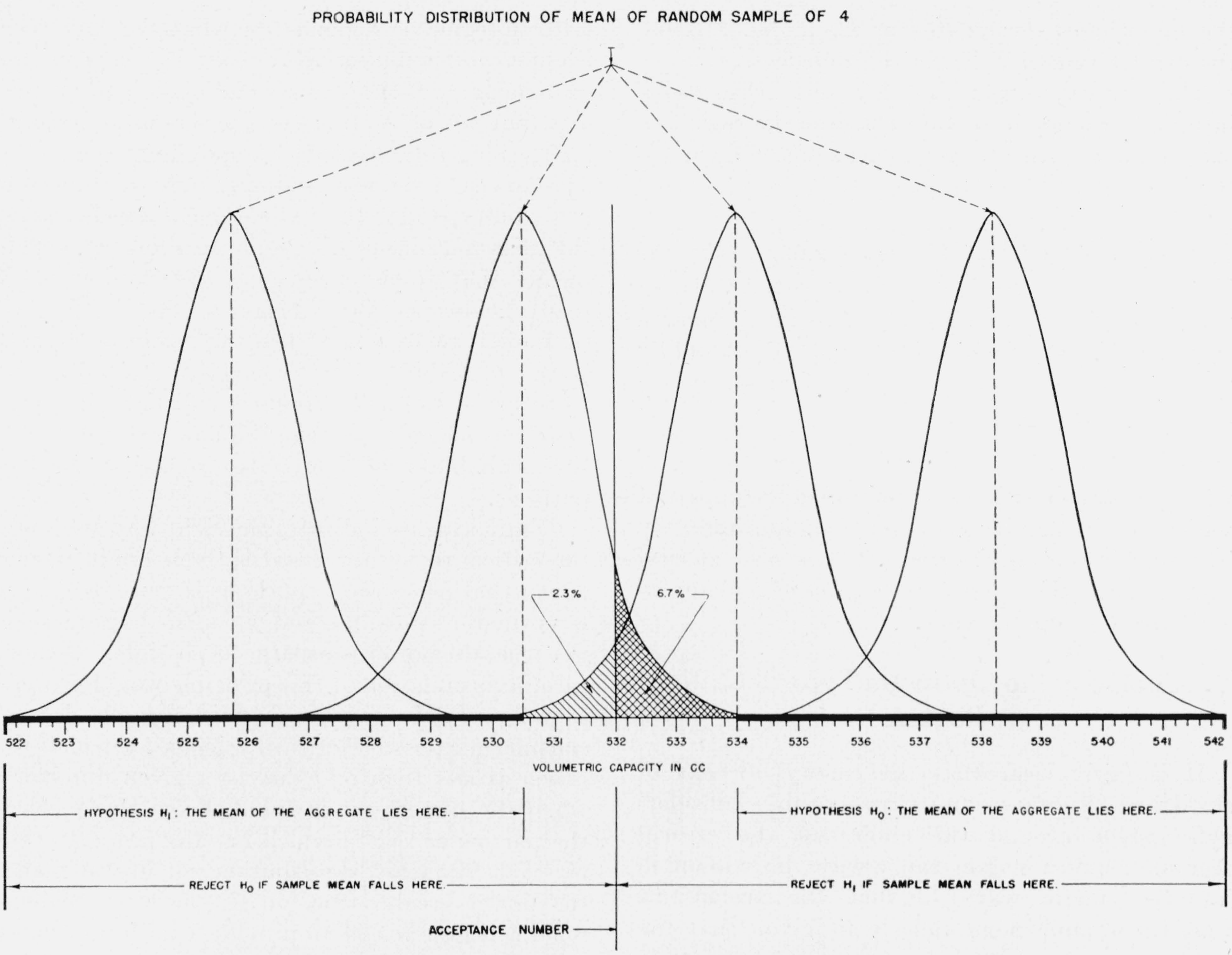

FIGURE 3. *Test of hypothesis concerning the arithmetic mean of the aggregate.*

distribution (or probability distribution) of a variate X (which may be a vector with several components), to the effect that this unknown frequency distribution belong to a certain well-defined subclass ω of a class Ω of admissible frequency distributions of X. The hypothesis is tested by first drawing a sample of observations on the variable X, say $(X_1, X_2, \ldots, X_n)$, in such a way that the probability distribution of the sample can be calculated from the distribution of X, and then noting just where the point defined by the coordinates $(X_1, X_2, \ldots, X_n)$ falls in n-dimensional Euclidean space. If this sample point falls on a certain predetermined region w of the n-dimensional space, the hypothesis is rejected. The subset w is called the "critical region"[6] for the test.

In the application to acceptance sampling, the variable X is a measure of the quality of the individual item, and the hypothesis to be tested is that the unknown frequency distribution of X in the sampled aggregate of items belongs to a certain subclass of distributions which is identified with definitely "good" or acceptable quality (alternatively, "bad" or unacceptable quality) with respect to the characteristics measured by X. Thus in the example of section III, X is volumetric capacity; the admissible class Ω of frequency distributions consists of all normal distributions with standard deviations equal to 2.0 cc; the subclass ω representing acceptable quality consists of all such frequency distributions with means not less than 534.0 cc; and the purpose of the sampling plan can be considered to be to test

[6] For the benefit of mathematicians reading this paper, it might be observed that this usage of the word "region" is at variance with the usage in classical mathematical analysis; the critical region is not necessarily an open connected set.

the hypotheses (designated as H_0 in fig. 3) that the distribution of X lies in this subclass.

The critical region for this test, chosen on intuitive grounds in section III, was the region in the sample space of four dimensions defined by the inequality

$$\frac{X_1+X_2+X_3+X_4}{4}<532.0, \qquad (1)$$

where X_1, X_2, X_3, and X_4 represent four observations on X. In general, in acceptance sampling whenever the critical region is defined by an inequality such as

$$c_1<f(X_1, X_2, \ldots, X_n)<c_2,$$

where f is some continuous function of the apparent variables, the numbers c_1 and c_2 can appropriately be called *acceptance numbers*. It is also natural to call the critical region the *region of rejection* in acceptance sampling work.

V. Choice of the Optimum Type of Critical Region

It is fairly clear that the theory of tests of statistical hypotheses must revolve to a considerable extent around the choice of the critical region. Such a region can always be chosen in infinitely many ways so that the probability that the sample point falls in it, given that the hypothesis under test is true, is not greater than a given arbitrarily small number. For example, the critical region used in section III is the region defined by the inequality (1), and the maximum probability of the sample point falling into it under the hypothesis of good quality is 0.023. But another critical region for which this maximum probability is also 0.023 is defined by the inequality

$$532.9<\frac{X_1+X_2+X_3+X_4}{4}<533.0, \qquad (2)$$

and the reader will have no difficulty in concocting other critical regions with the same maximum probability of rejection. Which one is the "best"?

To answer this sort of question, Neyman and Pearson [7] introduced the idea of the "power" of a test of a statistical hypothesis. This concept is based on the obvious fact that the probability that the sample point falls in a predetermined critical region w depends on what the unknown frequency distribution of X really is. That is, for each member of the class Ω of admissible frequency distributions of X, there is a probability, theoretically calculable, that the sample point falls in w. The power of the test is simply defined to be this probability, and is thus a functional depending on the distribution of X (and also on w). This means that if the frequency distribution of X really belongs to the subclass ω (that is, the hypothesis is really true), then the power of the test is the probability of (erroneously) rejecting the hypothesis, and if the frequency distribution of X does not belong to ω, then the power of the test is the probability of (correctly) rejecting the hypothesis.

Common sense indicates the right way to choose the critical region for a test is to choose it in such a way that when the hypothesis is true, the power is as small as possible, and when the hypothesis is not true, the power is as large as possible. A more careful formulation of this principle would proceed as follows: For any given fixed sample size n,[8] consider all possible critical regions for which the power is less than or equal to a given number α when the hypothesis is true. From these, select the particular region which has the property that for each frequency distribution not in ω it makes the power greater than do all the other regions. Such a region is said to provide a uniformly most powerful test. When this sort of optimum region does not exist, various compromises have been worked out which rationalize and systematize intuitional solutions of the problem In some cases where a critical region which is known to be in some sense optimum does exist, the boundary is so complicated that for practical reasons it is desirable to use a simpler critical region. An example of such a case will be discussed in section VIII.

To return to the example of section III, the power of the test for any member of the class Ω and for any critical region defined by an inequality like (1) or (2) is very easily calculated by simply noting that the probability distribution of $(X_1+X_2+X_3+X_4)/4$ has exactly the same mean as the parent frequency distribution in Ω, and has the standard deviation $2/\sqrt{4}=1$ cc. Therefore, the probability that the sample observations fall so

[7] See, for example, Wilks [3], chapter VII.

[8] If sequential sampling is used, this clause can be omitted (see Wald [4]). If multiple sampling is used, the clause should be replaced by a specification of the sample size at each step of the sampling procedure.

that they satisfy an inequality such as (1) or (2) can quickly be looked up in a table of a normal distribution. It appears reasonable from figure 3, which shows the distribution of $(X_1+X_2+X_3+X_4)/4$ drawn for four typical members of the class Ω, that among all critical regions of the type given by inequalities such as (1) and (2), the one given by (1) gives the uniformly most powerful test in the sense above defined. This can be proved rigorously with little trouble. It is not so easy to show that this same critical region gives the uniformly most powerful test among all possible imaginable critical regions in space of four dimensions, and not just among those critical regions defined by inequalities to be satisfied by the sample mean, but it can be proved that such is indeed the case [5].

VI. Adjustment of the Power of the Test

Presuming that an optimum (or at least, satisfactory) type of critical region for a given statistical hypothesis is known, the chief remaining question in setting up the test, or acceptance sampling plan, is that of arranging for the appropriate power. It was previously pointed out that the power of the test depended on the distribution of the variate X to which the hypothesis applied, and on the shape and size of the critical region w. In addition, it should now be observed that for a fixed specification of the distribution of X and for a fixed w, the power also depends on the probability distribution of the sample point $(X_1, X_2, \ldots, X_n)$, which is determined by such things as method of selection of the sample and the sample size.

The problem of deciding upon the appropriate power for a test is complicated by the fact that in most cases of practical importance, given any fixed critical region, the power functional does not have a sharp discontinuity as the frequency distribution of X crosses the boundary between ω and the remainder of Ω. In other words, if the hypothesis is almost, but not quite, true, the probability that the sample point falls into the critical region is about the same as if the hypothesis was almost, but not quite, untrue. This phenomenon can be seen very easily in figure 3. Obviously as the mean of the true distribution moves continuously leftward from the value 534.0 cc, which represents the boundary of good quality, the shaded area to the left of the point 532.0 cc (which represents the power of the test) continuously increases.

A convenient way of getting around this difficulty is to recognize that in the applications there is often no real need for distinguishing sharply between a case where the hypothesis is true and another case where it is "just a little bit" untrue. In many situations it is possible from practical considerations to select a second well-defined subclass ω_1 of the class Ω of admissible frequency distributions, such that there is a positive distance (as defined in some reasonable way) between ω and ω_1, and such that from the viewpoint of the application, ω_1, and ω are subclasses that are so radically different that if the true distribution of X really belongs to ω_1, then there should be a high probability of rejecting the hypothesis that it belongs to ω. It is natural in such a case to introduce a second statistical hypothesis into the picture, consisting of a statement that the unknown frequency distribution of X belongs to the subclass ω_1. The test is then conveniently thought of as affording a criterion of choice between the two hypotheses.

This approach, which will henceforth be called the method of alternative hypotheses, leads to simple quantitative rules for determining the appropriate power, as for example that the power shall not be greater than a preassigned number α when the distribution of X is truly a member of ω, and the power shall not be less than a preassigned number $1-\beta$ when the distribution of X is a member of ω_1. For a given method of sample selection and a given type of critical region, this sort of requirement will in many practical cases uniquely determine a minimum sample size and a corresponding position for the critical region, as will be seen in section VIII.

In the application of the theory of statistical tests to acceptance sampling, the method of alternative hypothesis is particularly useful and convenient. It is seldom indeed that the boundary between the characterization of a "good" lot and a "bad" one is sharply defined; but given a class of admissible frequency distribution of a quality-defining variable X, it is often possible to pick out two isolated subclasses which respectively represent definitely "good" or definitely "bad" quality with respect to the characteristics measured by X. When this has been done, the maximum [9] probability of erroneously rejecting an aggregate that is really in the good subclass is

[9] Maximum is here used in the sense of least upper bound.

called the *producer's risk*, and the maximum (see footnote 9) probability of erroneously accepting an aggregate that is really in the bad subclass is called the *consumer's risk*. In the terminology of the preceding paragraph, if β represents good quality and ω_1 represents bad quality, then α is the producer's risk and β is the consumer's risk—a notation that will be adhered to in the sequel.

The example of section III illustrates the method of alternative hypotheses. It was pointed out previously in section IV that the admissible class Ω of frequency distributions in this example consists of all normal distributions with standard deviations equal to 2.0 cc, and the subclass ω, identified with a good quality, consists of all frequency distributions in Ω with means not less than 534.0 cc. A second subclass ω_1, identified with unacceptable quality, is also defined in the example; this ω_1 subclass consists of the members of Ω with means less than or equal to 530.5 cc. The producer's risk of the sampling plan thereupon chosen is 0.023, and the consumer's risk is 0.067.

VII. Parametric Case and Operating Characteristic

The study of the power of a statistical test is considerably simplified when the members of the class Ω of admissible frequency distributions are uniquely determined by the values of one or more parameters. The discussion in the preceding sections has not been limited to the parametric case because the basic definitions and concepts can be stated just as easily for the general case as for the parametric case. Moreover, work is proceeding rapidly now on the theory of nonparametric tests, and it would seem to be worth while to lay down general definitions here with an eye to the day when nonparametric tests come into general use in acceptance sampling. But parametric hypotheses are still used far more commonly in the applications than nonparametric ones; this is especially true in acceptance sampling applications.

The case illustrated in the example of section III, where each member of the class Ω was uniquely determined by the value of its mean, is an example of the important case in which the class Ω is a one parameter family. Let the parameter in a one parameter family of admissible frequency distributions be denoted by θ. For a fixed critical region w, and method of sampling, the power of the test is a single-valued function of θ, which can be represented graphically in the usual way. In the theory of statistical inference, this graph is called the *power curve* of the test; in acceptance sampling work, it is called the *operating characteristic*. The power curve or operating characteristic of the sampling plan of section III will be found in figure 4.

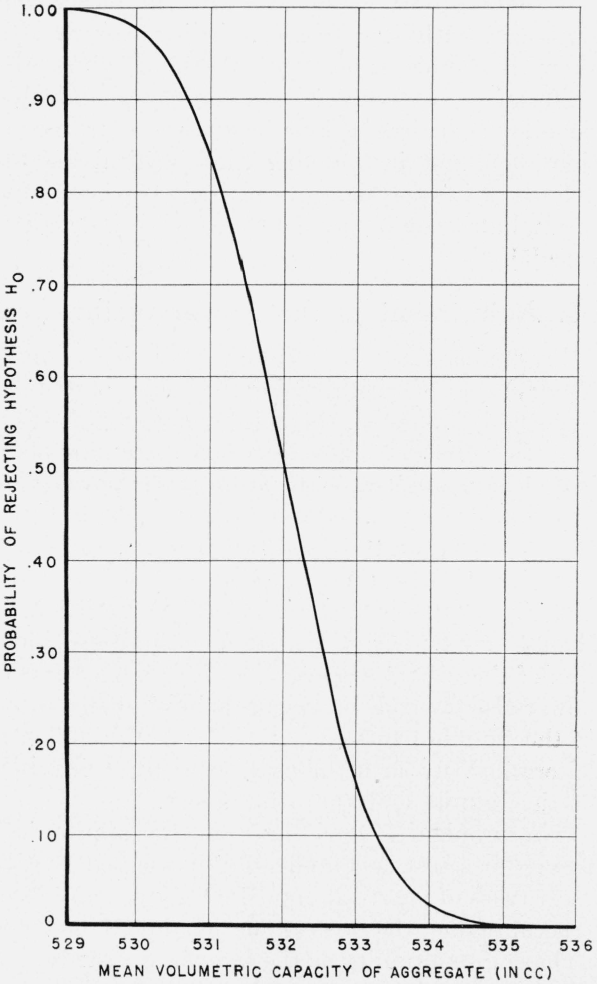

FIGURE 4. *Operating characteristic of the sampling plan of figure 3.*

By drawing the power curves, or operating characteristics of a number of proposed tests, corresponding to different positions of the critical region and different sample sizes (and perhaps even to different methods of sampling), a rational choice of the most suitable test can often be made.[10]

[10] Collections of operating characteristics for the attributes case will be found in [6] and in [7]. A collection of power curves for certain standard tests of statistical hypotheses for sampling by variables is given in [8].

The choice can be based on one or more of a number of rather obvious properties of the various curves, such as their relative slopes and their heights at specially chosen values of θ. The alternative hypothesis approach described in considerable generality above can also be used profitably in conjunction with a power curve. The subclasses ω and ω_1 are respectively represented by intervals (or more general point sets) on the θ-axis of the power curve, and from inspection of the curve the maximum probabilities of rejecting the alternative hypotheses when θ lies in these intervals can easily be estimated.

VIII. Sample Size and Region of Rejection: Normal Distribution

In the present section, and in the following one, the methods of alternative hypotheses will be applied to obtain some general formulas of practical importance for sample sizes and regions of rejection of acceptance sampling plans involving parametric hypotheses. The parameters to which the hypotheses pertain will be the mean and the standard deviation of the frequency distribution of a quality-defining variable X in the sampled aggregate. Single sampling only will be considered; this means that a sample of n items is to be drawn from the aggregate, the value of X for each of the n items is to be observed, and on the basis of these observations the aggregate is to be sentenced. It will further be assumed that the sample is so selected that the observations on the items are statistically independent, and the probability distribution of each observation is identical with the unknown frequency distribution of X. This can be accomplished, for example, by what is generally known as "random" selection from a finite aggregate with replacement after each drawing. The assumption will not be strictly valid if the sample items are selected "at random" *without* replacements, but in the practical applications of sampling by variables the size of the aggregate is usually so much larger than that of the sample that the requirement of replacements can be dropped without causing any appreciable change in the probability distribution of the sample. As a convenient rule of thumb, it might be stated that the formulas to be derived are valid for sampling without replacements if the ratio of aggregate size to sample size is greater than 10 to 1.

Consider first the case in which the parameter to which the hypotheses are to be applied is the mean μ of the frequency distribution of the aggregate, as in section III. Suppose further that this distribution is normal, that the value of its standard deviation σ is known, say $\sigma=\sigma_0$, and that two values of μ, μ_0, and μ_1 can be found, so that acceptable quality is identifiable with μ lying in the interval $\mu_0 \leqq \mu < \infty$, and unacceptable quality is identified with $-\infty < \mu \leqq \mu_1$. (This is the case exemplified in sec. III.) As pointed out in section V, a critical region of the type $\bar{x} < c$ will provide a uniformly most powerful test; here c is some constant and

$$\bar{x}=\frac{X_1+X_2+\ .\ .\ .\ +X_n}{n},$$

where X_1, X_2, . . ., X_n, denote the observations on the sample items. Let α be the producer's risk and β be the consumer's risk. Then the definition of these risks yields the following pair of simultaneous inequalities

$$\text{Prob.}\ [\bar{x} < c | \mu_0 \leqq \mu < \infty,\ \sigma=\sigma_0] \leqq \alpha, \tag{3}$$

$$\text{Prob.}\ [\bar{x} \geqq c | -\infty < \mu \leqq \mu,\ \sigma=\sigma_0] \leqq \beta. \tag{4}$$

(The first formula is to be interpreted as stating that the probability of $\bar{x} < c$, given that $\mu_0 \leqq \mu < \infty$ and $\sigma=\sigma_0$ is less than or equal to α, and the second formula has an analogous meaning.)

Now it can be seen, say from a figure like figure 3, that for a fixed c, if μ is allowed to approach μ_0 from the right, while the distribution of $\bar{x}$ is otherwise kept unchanged, then Prob. $[\bar{x} < c]$ continually increases, and achieves a maximum at $\mu=\mu_0$. Similarly, Prob. $[\bar{x} > c]$ achieves a maximum for all $\mu \leqq \mu_1$ when $\mu=\mu_1$. Thus (3) and (4) can be revised to read

$$\text{Prob.}\ [\bar{x} < c | \mu=\mu_0,\ \sigma=\sigma_0]=\alpha. \tag{5}$$

$$\text{Prob.}\ [x \geqq c | \mu=\mu_1,\ \sigma=\sigma_0]=\beta. \tag{6}$$

As $\bar{x}$ has a normal distribution with mean μ and standard deviation $\sigma_0/\sqrt{n}$ it follows that $(\bar{x}-\mu)/(\sigma_0/\sqrt{n})$ has a normal distribution with mean zero and unit standard deviation. Tables of this sort of normal distribution (henceforth to be called a standard normal distribution) are widely available. Consider (6) first. It follows that

$$\text{Prob. } [\bar{x} \geqq c | \mu=\mu_1, \sigma=\sigma_0]$$

$$= \text{Prob. } \left[\frac{\bar{x}-\mu_1}{\sigma_0/\sqrt{n}} \geqq \frac{c-\mu_1}{\sigma_0/\sqrt{n}} \middle| \mu=\mu_1, \sigma=\sigma_0\right]$$

$$= \left[\text{Area under standard normal curve from ordinate at } \frac{c-\mu_1}{\sigma_0/\sqrt{n}} \text{ to} +\infty\right]. \tag{7}$$

Let k_p be the $100p$ percent point on the standard normal scale, that is, the point at which the ordinate of the standard normal curve bounds a tail area exactly equal to p. Then combining (6) and (7),

$$\frac{c-\mu_1}{\sigma_0/\sqrt{n}} = k_\beta. \tag{8}$$

A similar argument applied to (5) leads to the equation

$$\frac{c-\mu_0}{\sigma_0/\sqrt{n}} = -k_\alpha. \tag{9}$$

The solution of (8) and (9) simultaneously for n and c yields formulas for these unknowns in terms of σ_0, k_α, and k_β, which provide a general solution to the problem of setting up the sampling plan. These formulas will be found under case 1 in the list of formulas in the appendix. Analogous formulas for the case in which the two hypotheses are interchanged are given there under case 2.

A slight generalization is obtained by remarking that the argument that led from (3) and (4) to (5) and (6) is valid if the class of admissible distributions is extended from just those normal distributions with $\sigma=\sigma_0$ for all values of μ, to include all normal distributions with $\sigma \leqq \sigma_0$ when μ lies in the interval identified with good quality, and with $\sigma \leqq \sigma_1$ when μ lies in the interval identified with bad quality (where σ_1 is any arbitrary positive number), and with no restriction at all on σ when μ lies in the interval identified with neither bad nor good quality.

The region of rejection $\bar{x}<c$ is then no longer a uniformly most powerful critical region, but it does still have certain optimum properties. This generalization is incorporated in the formulas given in the appendix.

The significance in practice of this extension of the class of admissible distributions is that while in a given instance it is seldom known exactly what the standard deviation of the distribution of the aggregate is, frequently a good guess can be made at a reasonable upper bound for this standard deviation. The practical side of making estimates of bounds for dispersion will be considered in more detail in section XIII.

A case which is often encountered is one in which the standard deviation of the distribution of a quality-defining variable X is directly proportional to the mean of X. Minor adjustments in the above argument yield the formulas listed under cases 3 and 4 in the appendix. The region of rejection used in these formulas is not a uniformly most powerful region, and in fact a theoretically "better" region can be found fairly easily.[11] However, this better region is a hypersphere in the sample space of n dimensions with radius and center which depend in a complicated way on μ_0, μ_1, and n. Although the matter has not been carefully investigated, there would seem to be very little loss in power in using a critical region of the type $\bar{x}<c$, as is done in the formulas in the appendix, and the gain in simplicity is very considerable. It would be of interest to make a more thorough study of this point.

Turning now to the case in which the hypotheses are to apply to the standard deviation σ of the sampled aggregate, an optimum type of critical region for testing the hypothesis $\sigma \leqq \sigma_0$ against the hypothesis $\sigma \geqq \sigma_1$ in the case of a normal aggregate $\sigma_0 < \sigma_1$ is known to be given by the inequality $s^2 > c$, where c is a positive constant, and s^2 is the usual unbiased sample estimate of σ^2; that is,

$$s^2 = \frac{\sum_{j=1}^{n}(X_j-\bar{x})^2}{n-1}.$$

Inequalities (3) and (4) become

$$\text{Prob. } [s^2 > c | \sigma \leqq \sigma_0] \geqq \alpha \tag{10}$$

$$\text{Prob. } [s^2 \leqq c | \sigma \geqq \sigma_1] \leqq \beta \tag{11}$$

Now the probability distribution of $(n-1)s^2/\sigma^2$ depends only on n and not on μ or σ, and is a standard tabulated distribution known as a chi-square distribution with $n-1$ degrees of freedom.[12]

[11] This better critical region is the region for the "likelihood ratio test"; see p. 150 of [3].

[12] See [9], p. 150. For tables of the chi-square distribution, see [9], p. 169.

This means that since

$$\text{Prob. } [s^2 > c] =$$

$$\text{Prob.} \left[\frac{(n-1)s^2}{\sigma^2} > \frac{(n-1)c}{\sigma^2} \right], \tag{12}$$

for a fixed value of n and c, the probabilities in this identity increase as the quantity $(n-1)c/\sigma^2$ decreases; that is, as σ increases. This remark leads to the following revisions of (10) and (11):

$$\text{Prob. } [s^2 > c | \sigma = \sigma_0] = \alpha. \tag{13}$$

$$\text{Prob. } [s^2 \leqq c | \sigma = \sigma_1] = \beta. \tag{14}$$

Now let $\chi^2_{n,\,p}$ denote in general the $100p$ percent point of the chi-square distribution with n degrees of freedom; that is, the point at which the ordinate of the frequency curve of this distribution bounds a *right-hand* tail area exactly equal to p. By applying the identity (12) to (13) and (14), the following simultaneous equations are obtained:

$$\frac{(n-1)c}{\sigma_0^2} = \chi^2_{\alpha,\ n-1},$$

$$\frac{(n-1)c}{\sigma_1^2} = \chi^2_{1-\beta,\ n-1}.$$

The solution of these simultaneous equations for n and c yields the formulas that define the sampling plan. The best way to obtain the solution seems to be as follows:

Replace the equations by the equivalent pair

$$\frac{\chi^2_{\alpha,n-1}}{\chi^2_{1-\beta,n-1}} = \frac{\sigma_1^2}{\sigma_0^2}. \tag{15}$$

$$c = \frac{\sigma_0^2 \chi^2_{\alpha,n-1}}{n-1}. \tag{16}$$

For fixed α and β, the quantity on the left side of (15) decreases as n increases. By inspection of a table of the percentage points of the chi-square distribution,[13] determine the first value of $(n-1)$ which makes the left member of (15) greater than (or equal to) the right member. Then substitute into (16) to obtain c.

Most tables of the χ^2 distribution stop at $n=30$. Beyond this value of n, a satisfactory approximation is obtained by treating s as if it had a normal distribution with mean equal to σ and with standard deviation equal to $\sigma/\sqrt{2n-2}$.

The derivation of formulas for n and c then becomes essentially the same as that for case 4 in the appendix; the results are given under case 7.

IX. Sample Size and Region of Rejection: Nonnormal Distribution

The principal limitation to the use of the formulas so far developed is that they involve the assumption that the distribution of the quality-defining variable X in the sampled aggregate is of the normal or Gaussian type. If in any given case, the assumption of normality appears to be inapplicable, it may happen that some other fairly tractable form of distribution can be specified. The general principles outlined in the preceding sections for determining sample sizes and regions of rejection will again be applicable, but rather troublesome computational difficulties may be encountered in determining the distribution of the sample mean and in solving for the sample size and boundary of the optimum critical region.

Moreover, in practice, cases are frequently encountered in which, due to lack of "statistical control" (see section XII) or lack of any historical data concerning the production process, very little *a priori* information is available concerning the possible form of the distribution of the variable X in a given aggregate. A method for dealing with such cases will be described in the remaining paragraphs of this section. It naturally leads to somewhat higher sample sizes than are obtained when for equivalent hypotheses and risks of error the form of the distribution is completely specified in advance. However, the increase in sample size over that obtained even with the assumption of normality is often not unreasonably great, and it is suggested that the method about to be described might be used whenever there is serious doubt about normality, unless there exists *a priori* knowledge of the physical nature of the production process which leads quite definitely to the conclusion that X should have some special form of nonnormal distribution, such as a Poisson distribution.

The method consists in making use of the various inequalities of the so-called Tchebycheff type, which place upper bounds on the proportion of a distribution that can lie at more than a preassigned distance from a central point in the distribution.[14]

[13] See footnote 12.

[14] See [10] p. 182 to 183.

The best-known and most frequently employed of these inequalities is the Bienaymé-Tchebycheff inequality,[15] which presupposes no restrictions at all on the distribution except that there is a finite standard deviation. As might be expected from the absence of prior information assumed, the use of this inequality to derive formulas analogous to those of section VIII leads to very high sample sizes compared to those obtainable under the assumption of normality. It is possible to improve the Bienaymé-Tchebycheff inequality somewhat by imposing certain additional restrictions on the distribution, such as that it be continuous and unimodal.[16] This type of restriction is not unrealistic in acceptance sampling work.[17] But another and more fundamental difficulty in the application of the Bienaymé-Tchebycheff inequality to acceptance sampling by variables is that the inequality seems to fail to take full advantage of the usually rapid approach of the sampling distributions of $\bar{x}$ and s to the normal form as n increases.

Thus in spite of the popularity of the Bienaymé-Tchebycheff inequality and its various modifications, in the case of hypotheses concerning μ, it would seem that the appropriate type of Tchebycheff inequality to use is one which is known as the inequality of S. Bernstein. This inequality has the double virtue of appearing to avoid the difficulty just now mentioned, and of involving restrictions on the distribution of X that are automatically satisfied in most acceptance sampling work. The Bernstein inequality may be written in the form of a pair of inequalities as follows: [18]

$$\text{Prob. } [\bar{x}-\mu \geqq K] \leqq e^{-nK^2/(2\sigma^2+2hK)}, \qquad (17)$$

$$\text{Prob. } [\bar{x}-\mu \leqq -K] \leqq e^{-nK^2/(2\sigma^2+2hK)}, \qquad (18)$$

and the restrictions on the distribution of the aggregate are that the absolute moments

$$\nu_j = E(|X-\mu|^j),\ j=1,\ 2,\ .\ .\ .,$$

all exist and satisfy the inequality

$$\nu_j \leqq \frac{\sigma^2}{2} j! h^{j-2},\ j=2,\ 3,\ .\ .\ .,$$

where h is a positive number. These conditions are satisfied if, for example, the distribution of X is bounded, which is usually the case in practice. If the distribution of X lies entirely in the interval $(\mu-\Delta,\ \mu+\Delta)$, then it can be shown that h can be taken as equal to $\Delta 3$.[19]

The use of the inequalities (17) and (18) to derive formulas for sample size and critical regions for hypotheses relating to μ, follows closely in the pattern set up in section VIII. The same type of regions of rejection will be used as before; these regions at least have optimum properties in the limit as the sample size tends to infinity.

The argument used in section VIII is applicable without change down through (5) and (6). At that point the inequalities (17) and (18) are introduced. The analogue of (7), obtained by using (17), is

$$\text{Prob. } [x \geqq c | \mu=\mu_1,\ \sigma=\sigma_0] =$$

$$\text{Prob. } [\bar{x}-\mu_1 \geqq c-\mu_1 | \mu=\mu_1,\ \sigma=\sigma_0]$$

$$\leqq e^{-\frac{n(c-\mu_1)^2}{2\sigma_0^2+2h(c-\mu_1)}} = \beta,$$

where $e=2.71828$ . . . . That is,

$$\frac{n(c-\mu_1)^2}{\sigma_0+h(c-\mu_1)} = -2 \log_e \beta. \qquad (19)$$

Similarly, applying (18) to (5), the following equation is obtained:

$$\frac{n(c-\mu_0)^2}{\sigma_0{}^2-h(c-\mu_0)} = -2 \log_e \alpha. \qquad (20)$$

The two simultaneous equations (19) and (20) define n and c. The extension to the case in which the *a priori* information $\sigma=\sigma_0$ is replaced by $\sigma \leqq \sigma_0$ for H_0 and $\sigma \leqq \sigma_1$ for H_1 is again valid. The only change in the equations (19) and (20) necessitated by this generalization consists in replacing σ_0 by σ_1 in (19).

Unfortunately, (19) and (20) happen to be equations of the third degree in the unknowns, and the general solution is somewhat complicated. The situation is greatly simplified from the algebraic point of view by taking $\alpha=\beta$. The solution for this case, and with $\sigma_0=\sigma_1$ for additional simplicity, is given in cases 5 and 6 in the appendix. The reader will have no great difficulty in obtaining the solution of (19) and (20) in more general cases, either by graphical methods or by straightforward algebra.

[15] See [10] p. 183.

[16] See [10] p. 183, also [11].

[17] Simon [2] discusses in some detail the use of the sharpened Bienaymé-Tchebycheff inequality in industrial sampling work.

[18] See [12] p. 204 to 205; also [13].

[19] See [12] p. 205.

X. The Case of Unknown σ

It has doubtless been noticed by the reader that the formulas so far derived for acceptance sampling plans involving hypotheses relating primarily to μ, have unfortunately always explicitly involved upper bounds for σ. This situation stems from the fact that the power curves for any efficient test of a statistical hypothesis concerning μ will in general always depend on higher moments of the distribution of μ, except in the case of certain very special distributions, where σ is a known function of μ. (An example of such a special distribution is the Poisson distribution, where $\sigma=\sqrt{\mu}$.)

If it can be assumed without too much error that the distribution of X is normal, then it is possible to devise a test of a hypothesis relating to μ, which has the property that the power curves corresponding to different values of σ all go through one point, say the point (μ_0, α) or the point (μ_1, β) in the notation of section VIII and the appendix. Moreover, the power curve of the test corresponding to each given value of σ either always rises or always falls as μ increases, so that if the hypothesis to be tested is of the form $\mu_0 \leqq \mu < \infty$, say, and if the test is arranged so that each power curve passes through the point (μ_0, α), then it is possible further to arrange the test so that the power for $\mu > \mu_0$ is always less than α no matter what σ may be.

Translated into the language of acceptance sampling, this all means that if the distribution of X in the sampled aggregate can be assumed to be normal, and if good quality is defined, say, by $\mu_0 \leqq \mu < \infty$ and bad quality by $-\infty < \mu \leqq \mu_1$ (as in case 1 in the appendix), then there is a way to set up a sampling plan that will have a preassigned producer's risk, or alternatively a preassigned consumer's risk, regardless of the standard deviation of the aggregate. But only a chosen one of the two kinds of risks can be preassigned; when a series of aggregates having differing values of σ are sampled, the value of the opposite risk will be different for each aggregate in the series, and will increase with σ.

Assuming (as always) the method of sampling described in section VIII, the critical region (or region of rejection) for the test in question is defined by inequalities of the type

$$\frac{\bar{x}-\mu'}{s/\sqrt{n}} < c, \tag{21}$$

$$\frac{\bar{x}-\mu'}{s/\sqrt{n}} \geqq c, \tag{22}$$

where μ' is the particular value of μ for which it is desired that the power of the test shall be independent of σ. In case the producer's risk is to be fixed for a hypothesis of good quality given by $\mu_0 \leqq \mu < \infty$, then (21) would be used with $\mu=\mu_0$. The left member of (21) and of (22) is a function of the observations $X_1, X_2 \ldots, X_n$ known as Student's t[20] with $n-1$ degrees of freedom. Tables giving certain of the percentiles of the distribution of this function for various values of n are widely available. Let $t_{n,p}$ denote the $100p$ percent point of the distribution of Student's t with n degrees of freedom; that is, the point at which the ordinate of the frequency curve of this distribution bounds a right-hand tail area exactly equal to p. Then if the producer's risk is to be fixed, the formula for determining c is simply

$$c = \pm t_{n-1,\alpha}$$

with the sign chosen so as to place the region of rejection properly in accordance with the obvious requirements of the hypothesis. If the consumer's risk is to be fixed, the formula becomes

$$c = \pm t_{n-1,\beta}$$

Numerical examples and further details will be found in [9].

The procedure given above is applicable for any proposed sample size n and does not fix the value of n. Perhaps the most satisfactory method of selecting n is by inspection of a family of power curves for the Student's t test drawn for various different values of σ and n.[21] If the use of alternative hypotheses seem to be appropriate, then these curves will be found to give in convenient form, for each n, the various values of the consumer's risk (assuming that the producer's risk was fixed in setting up the plan) which correspond to the particular values of σ, which are likely to be encountered in the applications.

It is perhaps a little discouraging to notice in this connection that if a sampling plan is set up by the formulas given in case 1, or case 2 in the appendix, and if the Student's t plan with, say, fixed producer's risk is set up for the same defi-

[20] See, for example, [9], p. 14. A table of the distribution of t is given on p. 167 of [9].

[21] See [8].

nition of good quality, the same producer's risk, and the same sample size, then the consumer's risk for a given definition of bad quality and a given upper bound for σ is considerably greater in the case of the Student's t plan than in the case of the other plan. (Of course, this consumer's risk in the Student's t plan can be reduced by increasing the sample size beyond that of the other plan). This fact suggests that the use of the Student's t approach in acceptance sampling work should be limited to cases in which it is essential at all costs to maintain either a constant producer's risk or a constant consumer's risk. But in practice, the exact validity of the underlying assumption of normality is always in doubt, and the method of sampling is seldom truly random, so the concept of a constant risk is in reality quite a nebulous one. *There is thus some basis for a recommendation that the Student's* t *approach should never be used at all in acceptance sampling work.*

The dependence on σ of the power curve of tests relating to μ in the normal case can be eliminated by the device of using a two-sample test, or double-sampling plan, in which the first sample serves the purpose of exploring the situation as regards σ, and the sample size of the second sample is adjusted accordingly. If the assumption of normality could be removed, such double-sampling plans might be of considerable utility in acceptance sampling work in which a variety of unique aggregates of uncertain pedigree and composition are to be sentenced by sampling in the complete absense of any history on which to base estimated upper bounds for σ. But as far as the writer is aware, there has been no attempt to eliminate the normality assumption in the literature of such tests to date. The assumption of normality, or even approximate normality, in the situation just now described would seem to be rather dangerous.

If the normality assumption is admitted, then there are several ways of going about the problem of setting up the test. A particularly interesting and straightforward method is given by Stein in [14]. Although it is beyond the scope of this paper to discuss multiple or sequential sampling plans at any length, it seems worth while to give here the instructions for setting up an acceptance sampling plan for alternative hypotheses on the basis of the Stein theory. For the hypotheses of case 1 of the appendix, the steps are as follows in terms of the notation introduced in this section and previously:

(1) Select a random sample of size m from the aggregate, and calculate s^2 for this sample. (The sample size m is not prescribed by the conditions of the problem, and is left to the judgment of the person doing the sampling. A small m might result in a large total sample size if σ is large.)

(2) Find the numerical value of

$$n'=1+\frac{s^2(t_{m-1,\,\alpha}+t_{m-1,\,\beta})^2}{(\mu_0-\mu_1)^2}.$$

Let n be the total number of observations in the first and second samples. The value of n is to be taken as $[n']$ [22] or m, whichever is the larger. (This means that if $[n']>m$, it will be necessary to take a second sample of $[n']-m$ items, but if $[n]\leqq m$, no second sample is necessary.)

(3) The region of rejection is now defined to be

$$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}<c, \tag{23}$$

where

$$c=-t_{m-1,\,\alpha} \text{ and } \bar{x}=\frac{\sum_1^n x_j}{n}.$$

(Note that s is computed only from the first sample of m items, but x is computed from the first and second samples combined, if there is a second sample.)

For the hypotheses of case 2 of the appendix, the sense of the inequality (23) and the sign of c is reversed; the instructions are otherwise unchanged.

The critical region defined by (23) is obviously a very complicated one, as n is a function of the first m observations. It is not known to the writer whether or in what sense the test so defined has optimum properties.

It is important to observe that the producer's and consumer's risks of a double-sampling plan such as the above one apply to *the entire double-sampling process* and not just to the second sample. In sampling a sequence of aggregates, it would not be legitimate to select the first sample of m only in the case of the first aggregate and then to use the value of s and n so derived for all the succeeding aggregates. The value of s must be determined separately for each aggregate of the sequence.

[22] The symbol $[n']$ means largest integer in n'.

XI. Quality Measured Simultaneously by Several Parameters

It occurs frequently that the quality of an item must be measured by a number of variables that may be correlated with varying degrees of intensity.

If the quality of the sampled aggregate is to be measured by the mean value of two or more variables, and if these variables may be considered on engineering grounds to be statistically independent, then it is very easy to generalize the theory of the single-parameter case. For example, suppose that aggregate quality is to be measured by the three arithmetic mean values μ_1, μ_2, μ_3, and that the hypothesis H_0 of good quality has the form

$$a_1 \leqq \mu_1 < \infty, \; a_2 \leqq \mu_2 < \infty, \; a_3 \leqq \mu_3 < \infty.$$

Let α_1, α_2, and α_3 be the producer's risks associated individually and respectively with these three inequalities; that is, the producer's risks for sampling plans designed to control the means, one at a time. If $\overline{x}_1$, $\overline{x}_2$, and $\overline{x}_3$ are the three respective sample means, then an efficient test of H_0 simply consists in requiring rejection if one or more of the $\overline{x}$'s fall below acceptance numbers determined individually for the three quality characteristics by the methods used for the single-variate case. From this it can be inferred that the total producer's risk of the entire sampling plan (that is, the maximum probability of rejecting H_0 if H_0 is true) is given by

$$1-(1-\alpha_1)(1-\alpha_2)(1-\alpha_3)=\alpha_1+\alpha_2+\alpha_3-\alpha_1\alpha_2-\alpha_1\alpha_3-\alpha_2\alpha_3+\alpha_1\alpha_2\alpha_3.$$

If each of these risks is small, say less than 0.05, as is usually the case in practice, the product terms can safely be ignored, and the total producer's risk becomes simply the sum of the three individual risks.

On the other hand, the problem of the consumer's risk requires a different approach. The aggregate would usually be considered to be of bad quality if *any one* of the three parameters μ_1, μ_2, μ_3 lies in an interval that is associated with bad quality for that parameter; that is, an aggregate is generally considered unacceptable if it is bad in any one respect. Under these conditions the total consumer's risk of the sampling plan would be simply equal to the largest of the individual consumer's risks associated with sampling plans designed to control the means one at a time. Of course this point of view is somewhat arbitrary, and perhaps there would be circumstances in which an aggregate would not be rejected unless an inference could be made that it was bad on two or more counts. In this case the total consumer's risk would be the largest of the products of pairs of individual consumer's risks.

The special case of three parameters so far considered can immediately be generalized to a higher number of independent quality characteristics. The general result is that the producer's risks for individual quality characteristics are always added to get the total producer's risk, but the consumer's risks for each of the quality characteristics must ordinarily be considered separately.

A similar situation occurs when aggregate quality is simultaneously measured by the arithmetic mean of a quality characteristic and by the standard deviation. If a normally distributed aggregate can be assumed, the customary sample estimates of these parameters are statistically independent. (This is true asymptotically for any aggregate.) Thus if sampling plans are set up individually for these quality measures in accordance with the formulas in the appendix, the combined producer's risk will be approximately the sum of the two producer's risks, but the consumer's risks cannot be combined.

When the quality characteristics are correlated, the problem becomes far more complicated, both from the engineering and from the computational points of view. The definition of what is good and what is bad quality is often harder to decide upon in this case. For instance, a lot of plywood that is low on shear strength but high on percentage wood failure in sheared specimens might be just as acceptable as a lot that is very high on shear strength but very low on percentage wood failure in sheared specimens. Thus in the case of correlation, the definition of good quality may correspond in some cases to a spherical or ellipsoidal region of the space whose coordinates are the population parameters representing quality characteristics.[23] The formulation of the "best" test of the hypothesis of good quality in such cases and the computation of the risks may involve a

[23] It is conceivable that such a definition of good quality might occasionally be considered appropriate even in the case of independent characteristics.

number of difficulties. Even for the hypothesis H_0 considered above, the computation of the producer's risk in a given case would involve a rather laborious mechanical quadrature. However, the formula given above for the case of independent variables provides an upper bound that often suffices in practical applications.

XII. Lot Acceptance and Process Acceptance

It has been emphasized several times in the foregoing discussion that the theory of inferential statistical sampling plans always involves the assumption that the sample is selected in such a way that the distribution of the sample observations bears a known, calculable relation to the unknown frequency distribution of the aggregate. In particular, it will be recalled that the validity of the formulas developed in sections VIII to X is dependent upon a method of sampling that will insure that the observations are statistically independent and that each has a probability distribution identical with the frequency distribution of the aggregate.

When such a requirement is considered from the operational point of view, it becomes necessary to distinguish between the case in which the sampled aggregate is all physically present and complete in itself at the time of sampling, and that in which the aggregate is viewed as a whole stream of items, some of which have yet to be manufactured. In other words, is it the purpose of the sampling plan to sentence a finite "lot" or delivery already produced, or the entire process? It has been customary to call the plan a "lot acceptance sampling plan "in the first case. The second case might be called "process acceptance sampling," and is one of the fundamental techniques of statistical quality control.

In lot acceptance sampling work, it is always possible, though often not practicable, by the use of a table of random numbers or some such device, to obtain more or less complete theoretical conformance with the requirement of "randomness," which is basic to the formulas of sections VIII to X. In process acceptance work, the difficulties in securing randomness with respect to the entire process are much more formidable. In addition, the viewpoint is broader, and economic and psychological considerations may dictate that it is more important to avoid erroneous rejection of a good process (and consequent shutdown of machinery) than it is to avoid temporary erroneous acceptance of a substandard process.

These considerations suggest that a rather frequently repeated statistical test of low nominal power and with a very low nominal producer's risk might be more suitable for process control than a less frequently repeated test of considerably higher power. Such an approach was advocated by Shewhart [15] and has now become the accepted sampling technic for statistical process control. An essential feature of the Shewhart system consists in plotting the sample data on a "control chart." This graphical device has been found useful in bringing about a condition in which the sequence of observations on successive items produced by the process has the characteristics of a sequence of observations on items successively drawn "at random" (in the sense of section VIII) from a finite aggregate. Shewhart calls this condition that of "statistical control" and has pointed out a number of benefits that it entails.

When a process is known to be in statistical control, a fairly extensive history of the production process is necessarily at. hand. If this history indicates that the distribution of the aggregate has remained for some time within a subclass of distributions identifiable with good quality, then it is the viewpoint of many experts in the field of statistical quality control that the weight of such historical data should be ordinarily sufficient to permit the continual acceptance of lots without the use of a more powerful (and more expensive) lot acceptance sampling plan.

Such a viewpoint clearly rests on experience rather than on mathematical theory. If the purchaser does not have his own inspectors in the plant of the manufacturer, he may very well feel that he needs some sort of definite lot-by-lot protection against lack of real control or shifts in the distribution of a statistically controlled process. Moreover, each lot produced by a process in statistical control is itself a random sample from the aggregate consisting of the process, so a poor lot will occasionally be produced by a good process through sampling fluctuations.

Thus it is by no means irrational under some circumstances to employ a lot acceptance sampling plan on lots produced by a process supposedly in a state of statistical control. If the decision is made to do this, the *historical data* concerning the control of the process, when available, will often provide safe estimates of nuisance parameters such as σ in the case of hypotheses concerning the mean and also will provide appraisals of the applicability of assumptions concerning the general form of the distribution of the lot.

It is interesting to notice that if a lot acceptance sampling plan is set up in connection with a controlled process, with the ultimate aim of accepting or rejecting the process, then the chances of making an erroneous decision are likely to be about the same for the individual lot and for the process. More precisely, if a sampling plan involving a hypothesis concerning, say, the mean μ of the sampled aggregate (assumed to be normally distributed), is set up for a pair of hypotheses similar to those of case 1 of the appendix, with the process playing the role of the sampled aggregate, then the upper limit of the conditional probability of rejection for a lot with mean $\overline{X}$ satisfying the condition $\overline{X} \geqq \mu_0$ is less than (but only slightly less than) α, and the upper limit of the conditional probability of acceptance for a lot with $\overline{X} \leqq \mu_1$ is less than (but only slightly less than) β. This situation arises from the fact that for any given μ, the "unconditional" marginal distribution of $\overline{x}$ (the sample mean) is normal with mean μ and standard deviation $\sigma/\sqrt{n}$, while for a given fixed value of $\overline{X}$, the conditional distribution of $\overline{x}$ (which is incidentally independent of μ) is normal with mean $\overline{X}$ and standard deviation $(\sigma/\sqrt{n})$ $(\sqrt{(N-1)N})$ where N is the lot size), which is only very slightly less than $\sigma/\sqrt{n}$ for reasonably large values of N. Similar considerations apply to sampling plans involving the standard deviation.

Thus under the hypothesis of control, if the mean of the process is in a region associated with good quality, the occasional "bad" lot that is produced by an accident of chance itself has a high probability of being rejected by the sampling plan. The probability of accidentally accepting a bad lot under such circumstances is a little less than the probability, multiplied by β, that a bad lot will be produced. This probability of accidentally accepting a bad lot, for $N > n$, is infinitesimal for the pairs of hypothesis considered in this paper; for example, in the case of hypotheses concerning the mean, the region for $\overline{X}$ identified with bad lies well beyond the acceptance number c, which itself lies in the extreme tail of the distribution of a sample of n.

This suggests that when there is a certain amount of assurance that a state of statistical control exists, a "cheap" lot acceptance sampling plan with low power might be the reasonable one to use. For given definitions of good and bad quality, when statistical control is *not* assumed, it is customary to use consumer's and producer's risks of the order of 1 to 10 percent. It is suggested that when statistical control can be assumed, a satisfactory plan will result if a consumer's risk of 50 percent is used in conjunction with the value of the producer's risk, and the definitions of good and bad quality, that would have been chosen if no control were assumed.

XIII. Random Remarks on the Applications

In practice, the control of dispersion in lot acceptance sampling work is still usually handled by the method of imposing upper and lower tolerances on the variable in question, which are then enforced by sampling by attributes. However, the writer has seen a number of successful applications of variable technics involving the sample range or standard deviation. It seems probable that as statistical quality control methodology becomes more familiar to inspectors and factory personnel, the variables approach to the dispersion problem will be given proper emphasis in cases where it is not the position of the distribution, but its "width" that must be controlled. A case in point is to be found in recent issues of Navy Department specifications for certain types of high-pressure cylinders. The range of tensile strength in the steel in individual cylinders, which affects the weldability properties of the metal, is controlled in these specifications by means of a variables sampling plan.

In the case of quality measured by the mean, the situation as regards the present popularity of the variables approach is quite different. Acceptance sampling plans involving the use of the sample mean are very frequently met with in practice. Needless to say, they are usually set up on the assumption that the sample is an exact replica of the lot, and are then defended by their authors on the basis that "years of experience" have proved them to be satisfactory.

It is easy to see why this type of approach actually does not often lead to trouble. The fact is that variables sampling plans involving the mean are usually applied in cases where there is considerable flexibility in the establishing of the definition of good and bad quality. By far the most frequent applications of this sampling technic are to cases in which the main problem is to see that a purchaser receives full measure under a contract or order. The replica-of-the-lot theory of sampling places the acceptance number for the sample mean squarely on the value of the mean agreed upon for the total amount of material under contract, and the contractor usually has sense enough to avoid rejections caused by sampling fluctuations by keeping his lot mean a little above the acceptance number. The purchaser, in his turn, has unconsciously established for himself a definition of a bad lot; but he seldom really knows exactly what is really in a given delivery, and a single underweight lot accidentally accepted through sampling errors will probably never be detected.

Closely related to the problem of full measure is the problem of the average analysis. In the purchase of certain commercial chemicals where complete homogeneity is not necessary, an average analysis is all that is needed to insure that the purchaser is getting his money's worth. This situation also obtains in the purchase of many metal items, such as pots and pans. In the same category are service life lot acceptance problems, in which the item is easily replaced from stockroom supplies if it wears out in use. Lamp bulbs, welding rods, searchlight carbons, and abrasives are all examples of materials which, under mass procurement conditions, give rise to inspection problems in which the estimation of the average service life of the items in each lot (perhaps by means of an accelerated life test) is the really important problem. In all these cases, the definition of what is good and what is bad quality is never really clear-cut. It is safe to say in general that empirically set up sampling plans using variables are not as likely to be really disadvantageous or pernicious from the purchaser's point of view as are certain empirical attributes sampling plans, such as the familiar "take one in each lot of 250" plan.

To be sure, the writer has seen some exceptions to this statement, which have occurred when specification writers in attempting to extrapolate from "years of experience" unconsciously have run afoul of the laws of probability. An example that comes to mind is a case in which the *mean* of a fairly large preliminary sample of observations on the breaking strength of glass portlights was taken by a specification writer as the *acceptance number* for the mean of a sample of four portlights selected from each lot. In another case, the mean of the buoyancy values of a few experimental kapok-stuffed life preserver pads was reduced arbitrarily by 1 pound and then used in the ensuing specification as a lower lot tolerance limit to be strictly enforced by attributes sampling. Because the standard deviation of most contractors for pad-stuffing was about ¾ pound (as could have been roughly inferred from the original data), trouble struck quickly when the specification hit the street.

These examples are admittedly somewhat extreme. It is not on the basis of such incidents that the scientific approach to variable acceptance sampling plans should be recommended, but rather on the basis of the fact that such an approach pays off in the long run. As in so many other fields of application of the modern theory of statistical inference, one can always do better on the average by using the theory than by not using it. And once in a while, a spectacular saving results or a serious error is averted, which alone justifies the extra trouble involved in basing the work on sound theory.

The chief obstacles (other than the difficulty of training personnel involved) to overcome in applying statistical theory to acceptance sampling by variables are that the sample selection should be truly random and that prior data on the form of the distribution, or at least on its higher moments, must generally be obtained. The problem of randomness has already been mentioned briefly in section XII and will not be further elaborated on here. The question of estimating bounds for higher moments is worth discussing here in a little more detail, at least in the important case in which the hypothesis concerns the mean and limits must be assigned to the dispersion of the sampled aggregate.

In some cases, the dispersion of the population is not due so much to nonhomogeneity of material as to lack of precision or reproducibility of a test, and once the testing error has been determined from a preliminary sample, it is likely to remain rather constant. Examples of cases in which the lot, if properly defined, is likely to be rather uniform but the test (or tests) may exhibit considerable variability, are soap, impregnite, and various commercial chemicals produced by wet mix processes. In the case of chemical compounds blended by dry mix processes and in the case of metals produced by continuous melting furnaces, the composition of a lot is likely to change slowly and uniformly from the first portion poured to the last portion, and if it were not for the error of the test, purposively selected samples from the first and last portion should "bracket" the lot. In these cases, it is not the number of sample items, but rather the number of repetitions of a test or an analysis that must be calculated on the basis of the known precision of the analysis.

But in the majority of cases, it is the material rather than the test that is variable, and predictions as to future dispersions must be made on the basis of familiarity with the engineering process, intuition, and a certain amount of historical evidence.

In the first place, purely engineering considerations that limit the range of the variable under observation will sometimes yield useful upper bounds for the standard deviation of the sampled aggregate. In the gaging of the inside diameter of a bearing, for example, it may be known that the range of measurements in a lot will almost surely be ±0.001 inch, from which a reasonable bound for the standard deviation can be inferred. In the production of fiber glass, the range of diameters of the fibers may be pretty well established by the process; in the winding of metallic-asbestos spiral-wound gaskets, the length of the strips of asbestos used will insure that the number of plies will fall within definite limits. An interesting case is furnished by resistance wire, which is subjected to average requirements on diameter mainly so that a piece of wire with a given resistance will not take up too much or too little room when wound on a mandrel. Unless different sizes of wire accidentally get mixed in a lot (which does sometimes happen but which would probably be detected by the resistance test), it can safely be inferred that the range of diameters in a lot will be restricted by the dies so that it lies between certain known limits.

XIV References

[1] H. F. Dodge and H. G. Romig, The Bell System Tech. J. **8,** 613 (1929).

[2] L. E. Simon, An engineers' manual of statistical methods (John Wiley & Sons, Inc., New York, N. Y., 1941).

[3] S. S. Wilks, Mathematical statistics (Princeton University Press, Princeton, N. J., 1944).

[4] A. Wald, J. Am. Stat. Assn. **40,** 277 (1940).

[5] J. Neyman and E. S. Pearson, Phil. Trans. Roy. Soc. (London) [A], **231,** 289 (1933).

[6] J. H. Curtiss, Ann. Math. Stat. **17,** 62 (1946).

[7] The Navy sampling inspection manual, chapter D4.00.00 in the administrative manual for the Materials Inspection Service, U. S. N., 1946.

[8] Charles D. Ferris, Frank E. Grubbs, and Chalmers L. Weaver, Ann. Math. Stat. **17,** 178 (1946).

[9] Harry Freeman, Industrial statistics (John Wiley & Sons, Inc., New York, N. Y., 1942).

[10] H. Cramér, Mathematical methods of statistics (Princeton University Press, Princeton, N. J., 1946).

[11] B. H. Camp, Bul. Am. Math. Soc. **28,** 427 (1922).

[12] J. V. Uspensky, Introduction to mathematical probability (McGraw-Hill Book Co., Inc., New York, N. Y., 1937).

[13] C. C. Craig, Ann. Math. Stat. **4,** 94 (1933).

[14] C. Stein, Ann. of Math. Stat. **16,** 243 (1945).

[15] W. A. Shewhart, Economic control of quality of manufactured product (D. Van Nostrand Co., Inc., New York, N. Y., 1931).

XV. Appendix

Single-Sample Formulas for Acceptance Sampling by Variables

1. Notation

X=a variable which measures a quality characteristic of an individual item

μ=the arithmetic mean of the frequency distribution of X in an aggregate of the items

σ=the standard deviation of the distribution of X in an aggregate of the items

n=sample size (number of items in a sample)

$$\bar{x}=\frac{X_1+X_2+\ .\ .\ .\ +X_n}{n}, \text{ where}$$

$X_1, X_2, \ldots, X_n$ are the observations on the sample items

$$s^2=\frac{(X_1-\bar{x})^2+(X_2-\bar{x})^2+\ .\ .\ .\ +(X_n-\bar{x})^2}{n-1}$$

c=acceptance number

α=producer's risk (maximum probability of rejecting an aggregate of good or acceptable quality)

β=consumer's risk (maximum probability of accepting an aggregate of bad or unacceptable quality)

k_p=defined by the equation

$$p=\int_{kp}^{\infty}\frac{1}{\sqrt{2\pi}}\,e^{-t^2/2_{dt}}.$$

2. Pairs of Alternative Hypotheses

Case number	H_0		H_1	
	Definition of good quality	A priori information concerning parameters	Definition of bad quality	A priori information concerning parameters
1	$\mu \geqq \mu_0$	$\sigma \leqq \sigma_0$	$\mu \leqq \mu_1 < \mu_0$	$\sigma \leqq \sigma_1$
2	$\mu \leqq \mu_0$	$\sigma \leqq \sigma_0$	$\mu \geqq \mu_1 > \mu_0$	$\sigma \leqq \sigma_1$
3	$\mu \geqq \mu_0$	$\frac{\sigma}{\mu} \leqq \alpha_0$	$\mu \leqq \mu_1 < \mu_0$	$\frac{\sigma}{\mu} \leqq a_1$
4	$\mu \leqq \mu_0$	$\frac{\sigma}{\mu} \leqq \alpha_0$	$\mu \geqq \mu_1 > \mu_0$	$\frac{\sigma}{\mu} \leqq a_1$
5	$\mu \geqq \mu_0$	$\sigma \leqq \sigma_0$	$\mu \leqq \mu_1 < \mu_0$	$\sigma \leqq \sigma_0$
6	$\mu \leqq \mu_0$	$\sigma \leqq \sigma_0$	$\mu \geqq \mu_1 > \mu_0$	$\sigma \leqq \sigma_0$
7	$\sigma \leqq \sigma_0$	----------------	$\sigma \geqq \sigma_1 > \sigma_0$	
8	$\mu_0 \leqq \mu \leqq \mu_0'$	$\sigma \leqq \sigma_0$	$\mu \leqq \mu_1 < \mu_0$ $\mu \geqq \mu_1' > \mu_0'$	$\sigma \leqq \sigma_1$

3. Single-Sample Formulas

(Random sampling)

Case number	Assumed form of distribution of aggregate	n	c	Region of rejection	Reference to notes
1	Normal	$\left(\frac{k_\beta\sigma_1+k_\alpha\sigma_0}{\mu_1-\mu_0}\right)^2$	$\frac{k_\beta\mu_0\sigma_1+k_\alpha\mu_1\sigma_0}{k_\beta\sigma_1+k_\alpha\sigma_0}$	$\bar{x}<c$	(a)
2	do	$\left(\frac{k_\beta\sigma_1+k_\alpha\sigma_0}{\mu_1-\mu_0}\right)^2$	$\frac{k_\beta\mu_0\sigma_1+k_\alpha\mu_1\sigma_0}{k_\beta\sigma_1+k_\alpha\sigma_0}$	$\bar{x}>c$	(a)
3	do	$\left(\frac{k_\beta a_1\mu_1+k_\alpha a_0\mu_0}{\mu_1-\mu_0}\right)^2$	$\mu_0\mu_1\frac{k_\beta a_1+k_\alpha a_0}{k_\beta a_1\mu_1+k_\alpha a_0\mu_0}$	$\bar{x}<c$	(a)
4	do	$\left(\frac{k_\beta a_1\mu_1+k_\alpha a_0\mu_0}{\mu_1-\mu_0}\right)^2$	$\mu_0\mu_1\frac{k_\beta a_1+k_\alpha a_0}{k_\beta a_1\mu_1+k_\alpha a_0\mu_0}$	$\bar{x}>c$	(a)
5	Lies entirely in finite interval $(\mu-\Delta, \mu+\Delta)$	$-18.42(\log_{10}\alpha)\frac{\sigma_0^2+\frac{\Delta}{6}\lvert\mu_0-\mu_1\rvert}{(\mu_0-\mu_1)^2}$	$\frac{\mu_0+\mu_1}{2}$	$\bar{x}<c$	(a), (b)
6	do	$-18.42(\log_{10}\alpha)\frac{\sigma_0^2+\frac{\Delta}{6}\lvert\mu_0-\mu_1\rvert}{(\mu_0-\mu_1)^2}$	$\frac{\mu_0+\mu_1}{2}$	$\bar{x}>c$	(a), (b)
7	Normal	$1+\frac{1}{2}\left(\frac{k_\beta\sigma_1+k_\alpha\sigma_0}{\sigma_1-\sigma_0}\right)^2$	$\sigma_0\sigma_1\left(\frac{k_\beta+k_\alpha}{k_\beta\sigma_1+k_\alpha\sigma_0}\right)$	$s>c$	(c)

4. Notes

(a) In the case of the "two-sided alternative" hypothesis, identified as case 8 in the list of pairs of alternative hypotheses, calculate sampling plan for each side separately by use of cases 1 and 2, or 3 and 4, or 5 and 6, as applicable, and use larger value of n for the final sampling plan. Final producer's risk will be less than or equal to the sum of the two producer's risks for each side.

(b) These formulas apply only when $\alpha=\beta$. In practice it is often easier to estimate Δ than σ_0, and it may safely be assumed that $\sigma<\Delta/3$. This leads to the following useful alternative formula for n:

$$-18.42\ (\log_{10}\alpha)\ \frac{\left(\frac{\Delta}{3}\right)^2+\frac{\Delta}{6}\lvert\mu_0-\mu_1\rvert}{(\mu_0-\mu_1)^2}.$$

(c) Use these formulas only for $n>30$. For $n\leqq 30$, set up sampling plan with the use of the χ^2 distribution.

Washington, November 22, 1946.