OPERATING CHARACTERISTICS FOR THE COMMON STATISTICAL TESTS OF SIGNIFICANCE

By Charles D. Ferris, Frank E. Grubbs, Chalmers L. Weaver

Ballistic Research Laboratory, Aberdeen Proving Ground

1. Summary. Methods making possible quick calculation of operating characteristics or power curves of common tests of significance involving the χ^2, F, t, and normal distributions are presented. In addition, a comprehensive set of curves illustrating graphically the power of each test for the 5% significance level are included. We are interested in the power of: (1) the χ^2-test to determine whether an unknown population standard deviation is greater or less than a standard value, (2) the F test to determine whether one unknown population standard deviation is greater than another (one-sided alternative), and (3) the t-test and normal test to determine whether an unknown population mean differs from a standard or two unknown population means differ from each other. Such operating characteristics have application for the quality control engineer and statistician in the design of sampling inspection plans using variables where they may be used to determine the sample size that will guarantee a specified consumer's and producer's risk. On the other hand they are of use in displaying the power of a test if the sample size has already been set. Finally, they are a necessary adjunct to the proper interpretation of the common tests of significance.

2. Introduction. In the application of the common statistical tests of significance there has been a great need for readily accessible information on the power of the test employed to distinguish between the null hypothesis and pertinent alternative hypotheses for given sample size. In this connection, two important applications arise. On one hand it becomes important for the sampler to know, for a given sample size and critical region, something about the power of the test in rejecting the stated hypothesis when some alternative hypothesis is true. On the other hand, if the sampler wants a given degree of assurance in rejecting the null hypothesis when a particular alternative is true, he would like to know the minimum sample size which would accomplish this when the probability of rejecting the null hypothesis when true is given. In particular, the need for such information arises most frequently in setting sample sizes to distinguish effectively, on the basis of single sample results, between (1) population standard deviations and (2) population means. If the sample size has already been set, as is the case with most specifications, quick information on whether or not it is large enough to keep the risk of accepting poor material down to a reasonable figure is highly desirable. Such probabilities will be recognized, of course, as the Type I and Type II errors of the Neyman-Pearson theory. Such risks must be given proper consideration in the interpretation of a significance test or in designing the provisions of an acceptance test.

Needless to say, the appropriate expressions for the power functions of the χ^2-test, F-test, normal-test, and t-test have been derived at one time or another in the literatuer. However, insofar as the practical statistician or quality control engineer is concerned, such information has not been employed to advantage widely since no informative graphs or extensive tables of power functions for the common statistical tests of significance have been presented. Due to the practical importance of questions of this type, the authors believe there is need for operating characteristics or graphical power functions of the common statistical tests of significance. This paper supplies such a need over a useful range of sample sizes and alternative hypotheses for the 5% significance level.

3. Definitions. In the following account, we will refer to one or both of the normal populations, π_1 and π_2. We will let x_1 be a variate from π_1 whose expected value or mean is μ_1 and standard deviation σ_1. By n_1 we will mean the number of observations drawn at random from π_1 and our sample statistics will be defined in the usual fashion:

$$\bar{x}_1 = \sum_1^{n_1} x_1/n_1, \qquad s_1^2 = \sum_1^{n_1} (x_1 - \bar{x}_1)^2/(n_1 - 1).$$

Similar definitions apply to the normal population π_2 with the appropriate subscript for sample statistics and population values. In dealing with a single population we will drop the subscripts from the sample statistics.

We also define

σ = a standard or arbitrary value of the standard deviation,

α = a standard or given level,

$$s_{12}^2 = \frac{\sum_1^{n_1} (x_1 - \bar{x}_1)^2 + \sum_1^{n_2} (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$$ when two normal populations are encountered.

H_0 will be used to denote the null hypothesis and H_1 any one of a set of alternative hypotheses. The probability of rejecting the null hypothesis H_0 when it is true (Type I error) will be denoted by α, and the probability of accepting the null hypothesis when some alternative hypothesis H_1 is true (Type II error) will be denoted by β.

4. Power function of the χ^2-test. The statistic $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$ (dropping subscripts of sample statistics) is used to accept or reject the hypothesis that the standard deviation, σ_1, of the normal population sampled is some specified or given value, σ.

Our hypotheses are

$$H_0: \sigma_1 = \sigma$$
$$H_1: \sigma_1 = \lambda\sigma, \ (\lambda > 0).$$

A. To determine whether or not $\sigma_1 > \sigma$. We choose a significance level, α, and compute $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$. If $\chi^2 > \chi^2_\alpha$, where the percentage point α is determined by

$$\frac{\left(\frac{1}{2}\right)^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}\right)} \int_{\chi^2_\alpha}^{\infty} u^{(n-3)/2} e^{-u/2}\, du = \alpha \tag{1}$$

we reject H_0 and conclude that $\sigma_1 > \sigma$.

To set up the power function we note that:
If H_0 is true

$$Pr\left\{\frac{(n-1)s^2}{\sigma^2}\right\} > \chi^2_\alpha = \alpha$$

If H_1 is true

$$Pr\left\{\frac{(n-1)s^2}{\sigma^2}\right\} > \chi^2_\alpha = 1 - \beta, \qquad (1 - \beta = \alpha, \text{ if } \lambda = 1).$$

However, since

$$Pr\left\{\frac{(n-1)s^2}{\sigma_1^2} > \chi^2_{1-\beta}\right\} = 1 - \beta$$

or

$$Pr\left\{\frac{(n-1)s^2}{\sigma^2} > \lambda^2 \chi^2_{1-\beta}\right\} = 1 - \beta$$

we have the relation

$$\lambda^2 \chi^2_{1-\beta} = \chi^2_\alpha \quad \text{or} \quad \lambda = \sqrt{\frac{\chi^2_\alpha}{\chi^2_{1-\beta}}}.$$

Therefore, for a given significance level, α (Type I error), and various Type II errors, β, we can make use of the Tables of Percentage Points of the χ^2-distribution [1] and compute enough of the points (λ, β) to plot the power curves depicted in Fig. 1. The Type I error, α, has been set at the practical level of .05 for Fig. 1.

B. To detect $\sigma_1 < \sigma$. We compute

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

and if $\chi^2 < \chi^2_{1-\alpha}$ we reject H_0, concluding that $\sigma_1 < \sigma$.

By reasoning similar to that in A. we arrive at the relationship

$$\chi^2_{1-\alpha} = \lambda^2 \chi^2_\beta \quad \text{or} \quad \lambda = \sqrt{\frac{\chi^2_{1-\alpha}}{\chi^2_\beta}}.$$

Again, by use of the Table of Percentage Points of the χ^2-Distribution the operating characteristics of Fig. 2 are obtained. We have chosen the practical level of $\alpha = .05$ for Fig. 2.

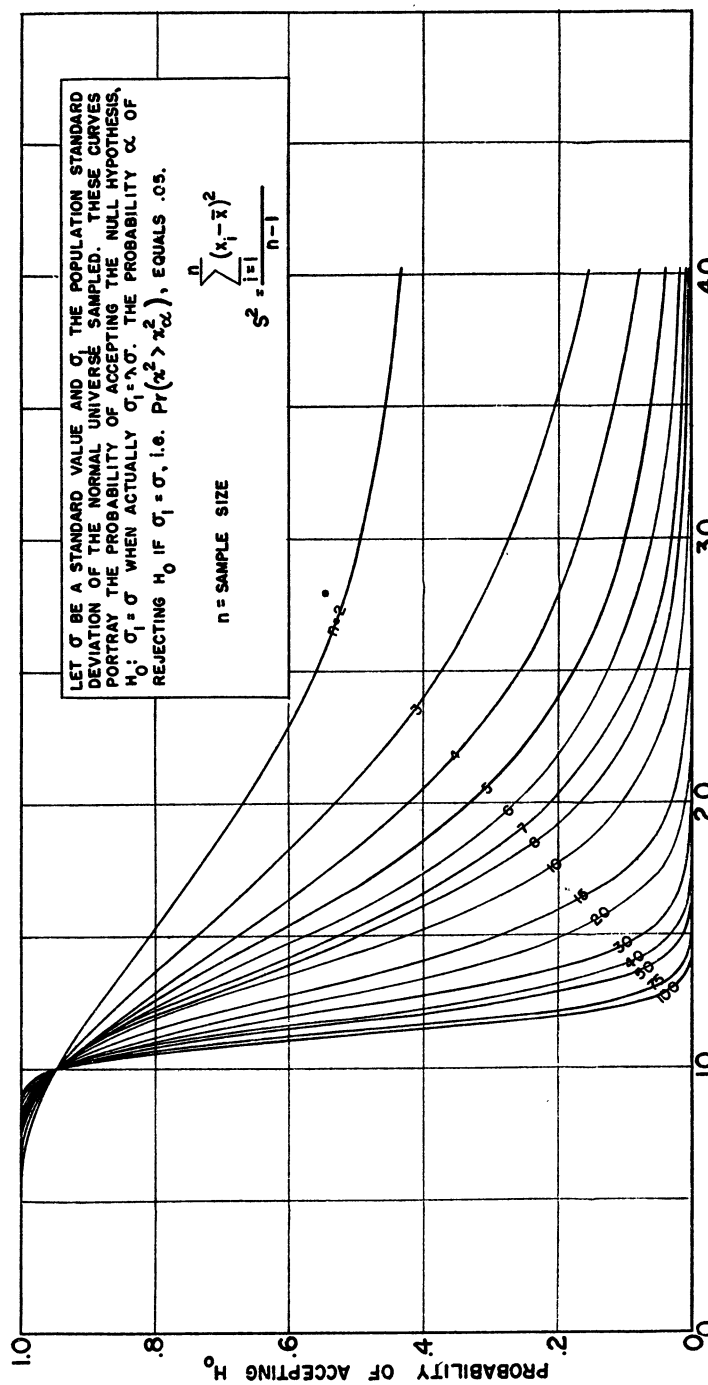

FIG. 1. OPERATING CHARACTERISTICS OF THE χ^2-TEST $\left[\chi^2 = \frac{(n-1)s^2}{\sigma^2}\right]$ FOR TESTING $\sigma_1 = \sigma$ AGAINST $\sigma_1 > \sigma$

Example:[1] A Rifle Association is purchasing small arms ammunition for match purposes. It is the desire of the rifle club that the dispersion in muzzle velocity of a lot of ammunition intended for match purposes be kept down to a practical minimum. Acceptance or rejection of an ammunition lot must, of course, be made on a sampling basis since the ballistic acceptance test is destructive in nature. Moreover, for practical reasons acceptance of a given lot is to be on the basis of a single sample. The Association specifies that they are not willing to accept more than 5% of the lots whose standard deviation in muzzle velocity is 6 ft./sec. The ammunition manufacturer agrees that he will accept these terms provided not more than 5% of the lots whose standard deviation in muzzle velocity is 4 ft./sec. will be rejected. Under these agreements, it is desired to know what sample size is necessary to provide the stated assurances for the Rifle Association and the ammunition manufacturer.

In this problem, $\alpha = .05$, $\beta = .05$, and $\lambda = 1.5$. Referring to Fig. 1, we find the required sample size is approximately 35.

On the other hand, if a sample size had already been set, the appropriate curve in Fig. 1 could be examined to determine whether it provided sufficient protection against the acceptance of inferior ammunition.

5. Power function of the *F*-test. In discussing the power function of the *F*-test we will focus our attention on the problem of comparing the standard deviations of two normal populations.

A. To determine whether or not the standard deviation, σ_1, of one normal population is greater than the standard deviation, σ_2, of another normal population. We choose a significance level, α, and compute $F = s_1^2/s_2^2$. If $F > F_\alpha$, where the percentage point F_α is determined by

$$\frac{\Gamma[\frac{1}{2}(n_1 + n_2 - 2)]}{\Gamma[\frac{1}{2}(n_1 - 1)]\Gamma[\frac{1}{2}(n_2 - 1)]} (n_1 - 1)^{\frac{1}{2}(n_1-1)} (n_2 - 1)^{\frac{1}{2}(n_2-1)} \cdot \int_{F_\alpha}^{\infty} \frac{u^{\frac{1}{2}(n_1-3)}}{[(n_1 - 1)u + n_2 - 1]^{\frac{1}{2}(n_1+n_2-2)}} du = \alpha, \tag{2}$$

we conclude that $\sigma_1 > \sigma_2$.

Our hypotheses are

$$H_0: \sigma_1 = \sigma_2$$

$$H_1: \sigma_1 = \lambda\sigma_2, (\lambda > 1).$$

To set up the power function of the *F*-test we note that:
If H_0 is true

$$Pr\{s_1^2/s_2^2 > F_\alpha\} = \alpha.$$

[1] This example is used to illustrate the use of the power of the χ^2-test and is not advocated as a most powerful sampling technique. (See ref. [10]).

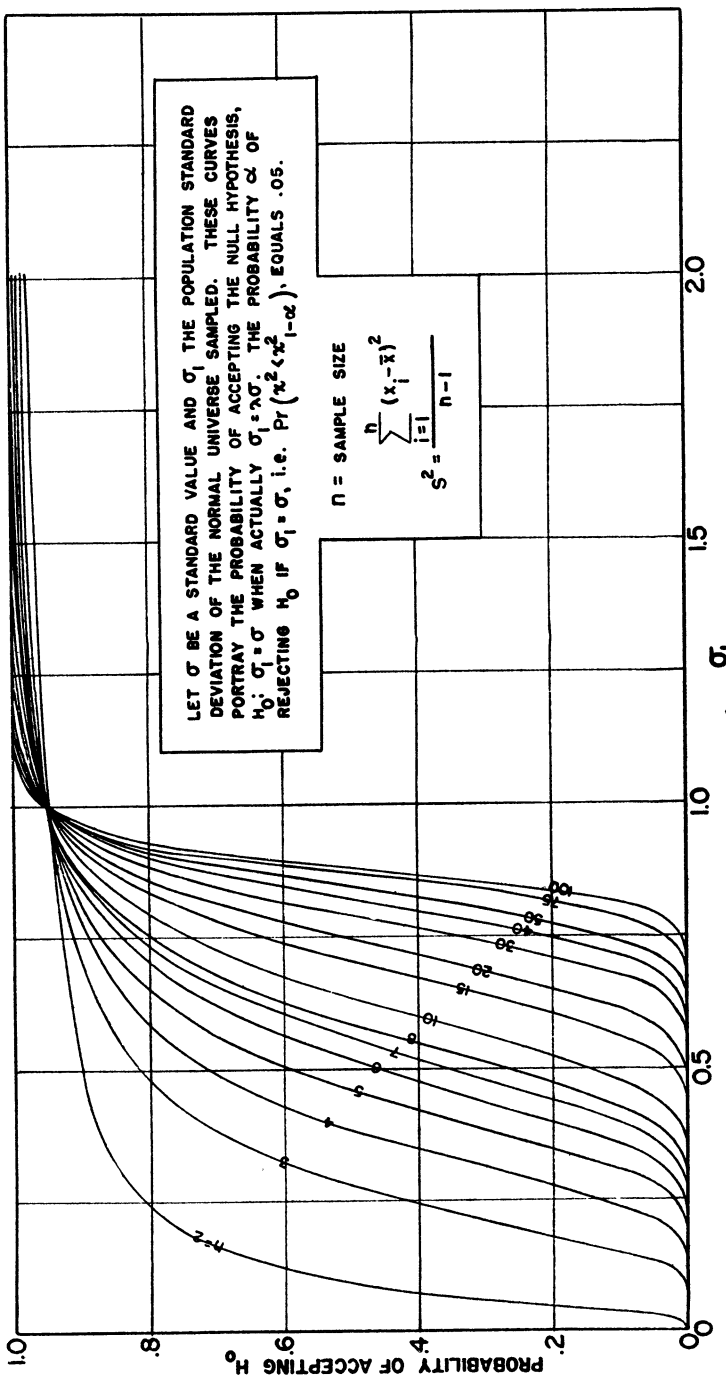

FIG. 2. OPERATING CHARACTERISTICS OF THE χ^2-TEST $\left[\chi^2 = \frac{(n-1)s^2}{\sigma^2}\right]$ FOR TESTING $\sigma_1 = \sigma$ AGAINST $\sigma_1 < \sigma$

If H_1 is true

$$Pr\{s_1^2/s_2^2 > F_\alpha\} = 1 - \beta, \qquad (1 - \beta = \alpha \text{ if } \lambda = 1).$$

However, since

$$Pr\left\{\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} > F_{1-\beta}\right\} = 1 - \beta$$

or

$$Pr\{s_1^2/s_2^2 > \lambda^2 F_{1-\beta}\} = 1 - \beta,$$

we have the relation $\lambda^2 F_{1-\beta} = F_\alpha$ or $\lambda = \sqrt{\frac{F_\alpha}{F_{1-\beta}}}$.

Therefore, for a given Type I error, α, and various Type II errors, β, we can make use of the Table of Percentage Points of the F-Distribution [2] and compute sufficient points (λ, β) to plot the operating characteristics depicted in Figs. 3, 4, and 5. In these figures, α has been set at the practical level of .05.

It should be emphasized that the operating characteristics presented in this paper are applicable only when one is interested in the one-sided alternative that $\sigma_1 > \sigma_2$ and not $\sigma_1 < \sigma_2$. Under these circumstances, the exact formation of the F ratio will be set beforehand and will not depend upon test results (for example, placing the greatest mean square in the numerator). In those cases where one is interested in the two-sided alternative, a two-tail F-test such as described by H. Scheffé [3] should be used. It is hoped that at a later date operating characteristics of such a test calculated in a manner similar to the example in [3] will be presented.

Example: It became necessary for a manufacturer to make a choice between a new type casting and one produced under standard design practices. One of the bases of comparison was dispersion in tensile strength. It was considered that if the standard deviation of the standard casting were larger than the new type, definite preference should be given to the latter. When the question of a practical criterion for rejecting the standard casting was considered, it was decided that if its true standard deviation in tensile strength were actually $1\frac{1}{2}$ times that of the new type there should be a 90% chance of rejection. It would be of little practical importance to detect any ratio less than $1\frac{1}{2}$ in this particular case. It was also decided that the 5% significance level would suffice insofar as rejection of equal quality was concerned. A preliminary sample size of 20 was selected, and the question arose as to how well a sample of this size gave the protection desired.

The question can be answered immediately by reference to Fig. 3 (here s_1^2 is computed from the standard casting data, of course) where it is seen that a sample size of 20 will fail to detect the stated difference 47% of the time. In order to achieve the desired protection, it is seen at once from Fig. 3 that a sample size of over 50 will be necessary. The exact sample size, determined with the aid of the formulas above, is found to be 54.

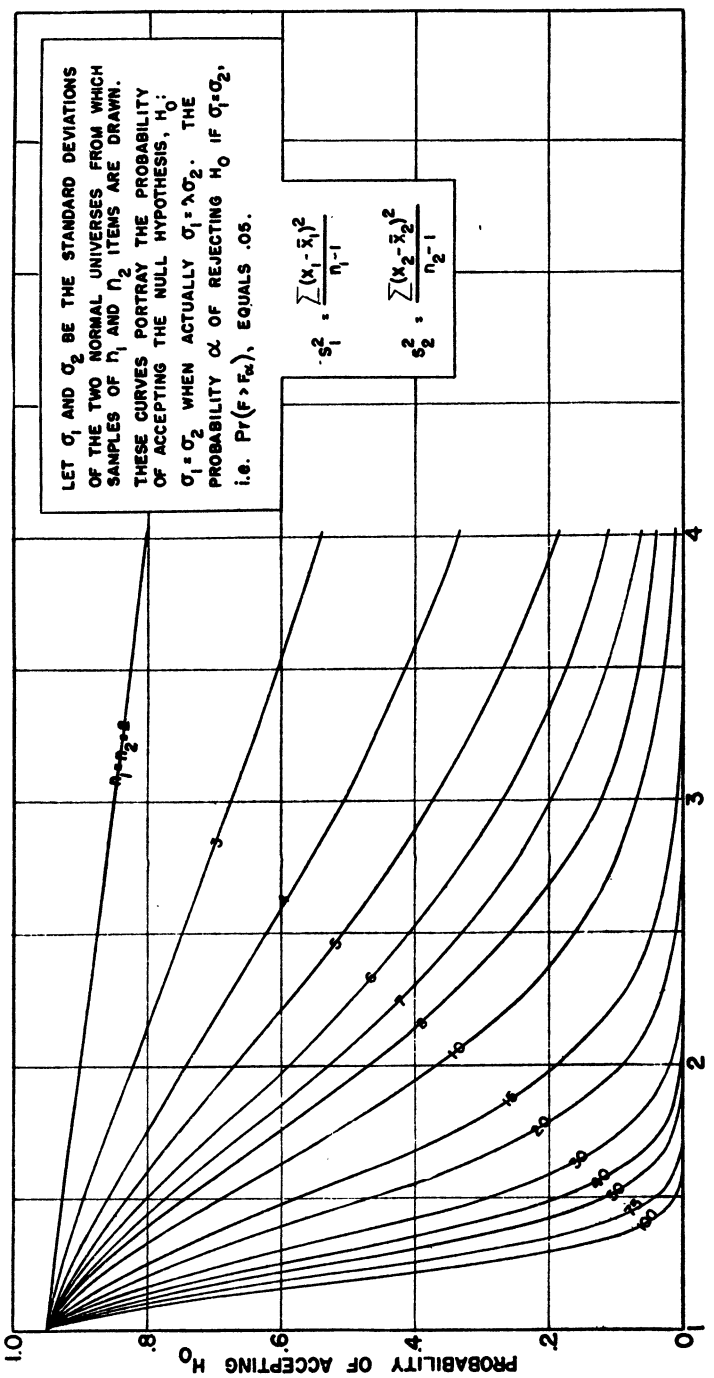

FIG. 3. OPERATING CHARACTERISTICS OF THE F-TEST $\left[F = \frac{s_1^2}{s_2^2}\right]$ FOR TESTING $\sigma_1 = \sigma_2$ AGAINST $\sigma_1 > \sigma_2$
$(n_1 = n_2)$

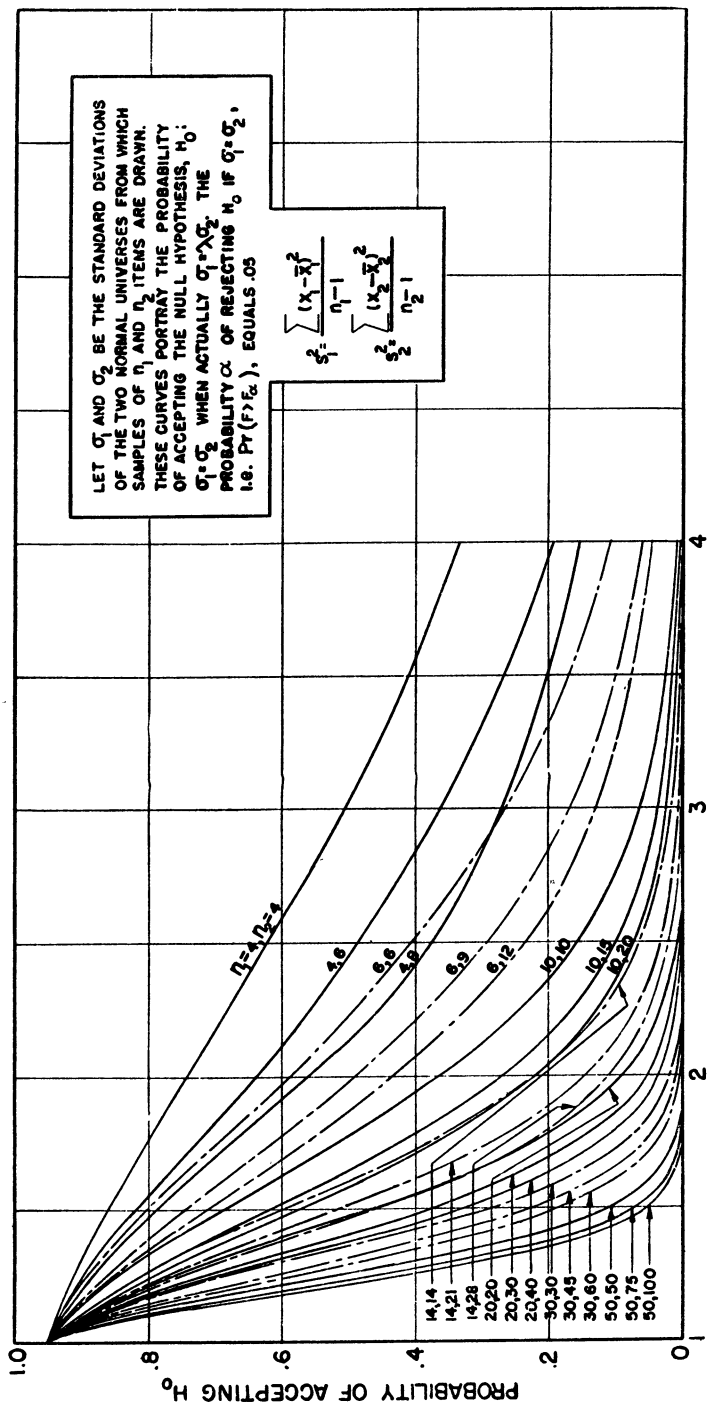

FIG. 4. OPERATING CHARACTERISTICS OF THE F-TEST $\left[F=\frac{s_1^2}{s_2^2}\right]$ FOR TESTING $\sigma_1=\sigma_2$ AGAINST $\sigma_1>\sigma_2$
$(n_1=n_2, 3n_1=2n_2, 2n_1=n_2)$

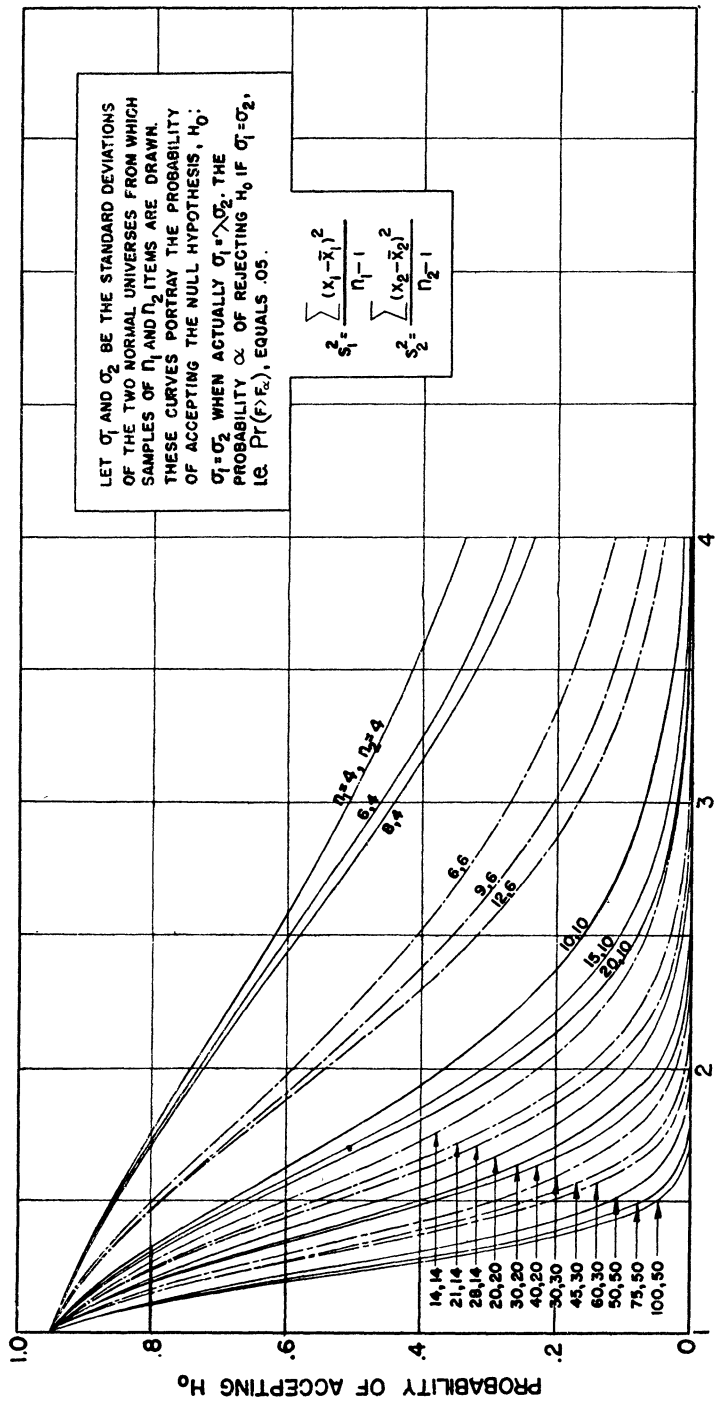

FIG. 5. OPERATING CHARACTERISTICS OF THE F-TEST $\left[F = \frac{s_1^2}{s_2^2}\right]$ FOR TESTING $\sigma_1 = \sigma_2$ AGAINST $\sigma_1 > \sigma_2$
$(n_1 = n_2, 2n_1 = 3n_2, n_1 = 2n_2)$

B. *Analysis of variance.* We shall consider the analysis of variance layout where a sample of n items is drawn from each of m normal populations with common variance σ^2. It is required to decide on the basis of the sample results whether or not there is any variation among the true means of the m normal populations sampled.

Let x_{ij} be the jth item drawn at random from the ith population,

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}, \quad \text{and} \quad \bar{x} = \frac{1}{m}\sum_{i=1}^{n} \bar{x}_i .$$

The F-test utilizes the comparison of the variation among the sample means (external variance) with that among the items within the samples (internal variance) in order to test the equality of population means by making use of the ratio

$$F = \frac{n\sum_{i=1}^{m} (\bar{x}_i - \bar{x})^2 m(n-1)}{\sum_{i,j} (x_{ij} - \bar{x}_i)^2 (m-1)} .$$

If $F > F_\alpha$, where F_α is defined as in 5.A., we conclude that the population means are not equal.

In our approach we will assume that the m true lot means represent a sample from a super-population, also normal, with variance equal to $\theta^2\sigma^2$. Since the sampling variance of the means is σ^2/n, the total variance among the sample means equals

$$\sigma^2/n + \theta^2\sigma^2 = \lambda^2\sigma^2/n, \qquad (\lambda^2 = 1 + n\theta^2).$$

Hence, our hypotheses are

$$H_0\colon \theta = 0$$

$$H_1\colon \theta > 0.$$

Since F/λ^2 follows the F-distribution with $m - 1$ and $m(n - 1)$ degrees of freedom the operating characteristic, i.e. the probability for various θ of accepting H_0, may be obtained from the curves already graphed by setting $n_1 = m$, $n_2 = nm - m + 1$, and $\lambda^2 = 1 + n\theta^2$.

In the design of experiments when the number of populations is indefinite (for example, daily tests) and the total sample size mn is limited, the above procedure will enable one to determine what values of m and n give the most powerful operating characteristic for the given amount of sampling. For example, for $mn = 24$ operating characteristics for all possible pairings were computed and charted. They were observed to cross one another, each combination in turn becoming most powerful for a given interval of θ. The following table gives the best pairings for various intervals of θ:

m	n	θ
2	12	.00– .32
3	8	.32– .60
4	6	.60– .91
6	4	.91–1.37
8	3	1.37–2.50
12	2	2.50–

In contrast to the above discussion, mention should be made of P. C. Tang's approach [4] to the power function of the analysis of variance. The basic difference lies in the method of expressing the alternative hypothesis. Tang expresses it in terms of the variance of a finite number of population means. We express it in terms of normally distributed population means. We believe our approach has considerable practical value in control chart analyses where we are interested in the quality of the flow of production of a large number of lots. In addition, our approach obviates the difficulties imposed by the non-central χ^2-distribution.

6. Power function of the normal test.

A. The statistic $u = \dfrac{\sqrt{n}(\bar{x} - a)}{\sigma_1}$ is used to accept or reject the hypothesis that the mean, μ, of the normal population sampled, is some specified standard level, a, when the population standard deviation is known (for example, from past data).

Our hypotheses are

$$H_0: \mu = a$$
$$H_1: |\mu - a| = \lambda\sigma_1, \ (\lambda > 0).$$

To test the hypothesis $\mu = a$, we choose a significance level, α, and compute u. If $|u| > u_\alpha$, where the percentage point, u_α, is determined by

$$\frac{1}{\sqrt{2\pi}} \int_{-u_\alpha}^{+u_\alpha} e^{-\frac{1}{2}\chi^2} d\chi = 1 - \alpha, \tag{3}$$

we reject H_0 and conclude that $\mu \neq a$.

To set up the power function we note that:

If H_0 is true

$$Pr\{-u_\alpha < u < +u_\alpha\} = 1 - \alpha$$

If H_1 is true

$$Pr\left\{-u_\alpha < \frac{\sqrt{n}(\bar{x} - a)}{\sigma_1} < u_\alpha\right\} = \beta, \qquad (1 - \beta = \alpha \text{ if } \lambda = 0),$$

$$= Pr\left\{-u_\alpha + \lambda\sqrt{n} < \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma_1} < u_\alpha + \lambda\sqrt{n}\right\}$$

where $\lambda = \dfrac{|\mu - a|}{\sigma_1}$.

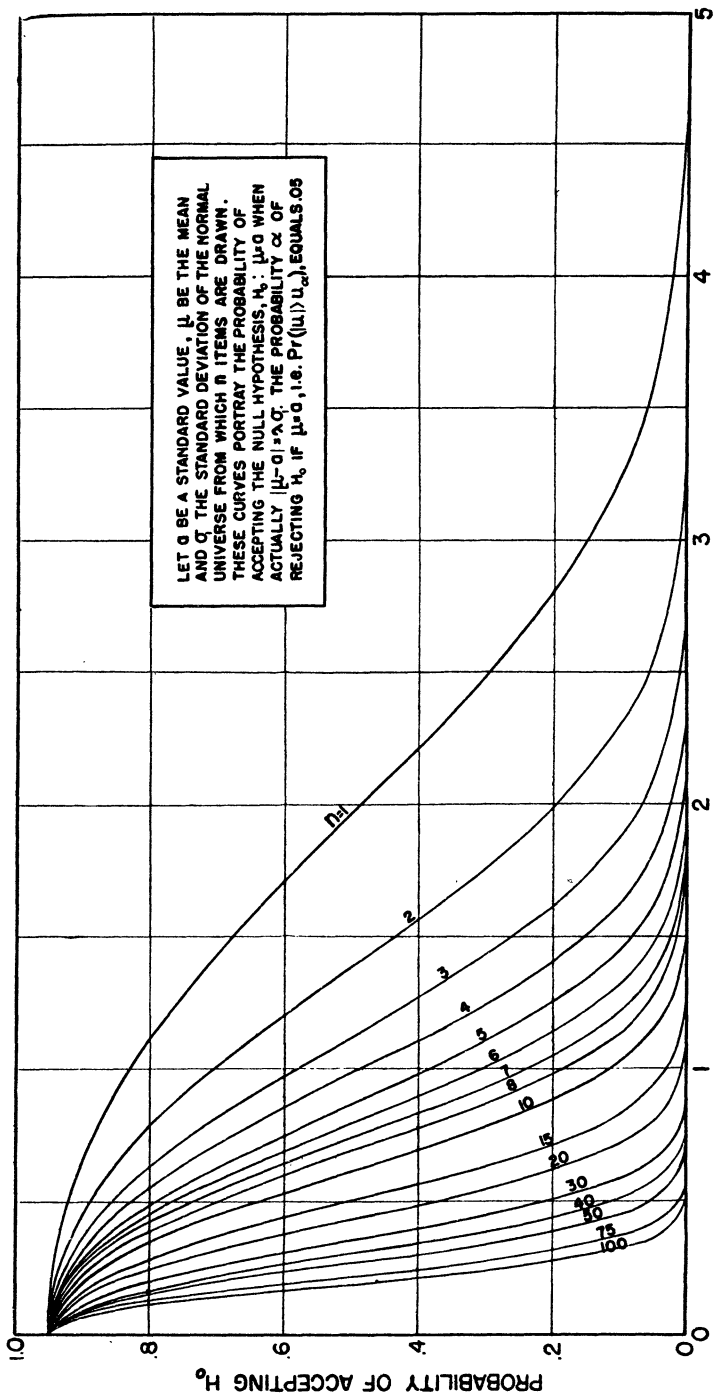

FIG. 6. OPERATING CHARACTERISTICS OF THE NORMAL TEST $\left[u = \frac{\sqrt{n}(\bar{x}-a)}{\sigma_1}\right]$ FOR TESTING $\mu = a$ AGAINST $\mu \neq a$

In the latter expression the statistic $\frac{\sqrt{n}(\bar{x} - \mu)}{\sigma_1}$ is normally distributed with zero mean and unit variance. The required probabilities are found easily from tables of areas under the normal frequency curve. By computing enough points (λ, β) the operating characteristics depicted in Fig. 6 were constructed.

It should be noted that the β corresponding to a pair of values n' and λ' may be obtained from any other operating characteristic by use of the relation $\lambda = \lambda'\sqrt{n'/n}$. For example, if it is desired to find the Type II error for a sample size of $n' = 12$ and $\lambda' = 1$, select any operating characteristic, say for $n = 3$, as the reference curve, compute $\lambda = 1\sqrt{12/3} = 2$, and find from the curve for $n = 3$ that $\beta = .07$. In Fig. 6, however, individual operating characteristics are plotted for convenience and to provide a picture of the comparative efficiency of various sample sizes.

Example: Pressure-measuring instruments are being tested against a standard level. It has been decided that instruments whose true mean reading is as much as 10 pounds per square inch away from the standard level should be rejected 95% of the time. On the other hand only 5% of instruments whose true mean reading equals that of the standard should be rejected. From past data, it is known that all test instruments of the type being considered have a stable standard deviation of 5 psi. If rejection or acceptance is to occur on the basis of a single sample and the normal criterion of significance, what sample size should be chosen to accomplish this purpose? Referring to Fig. 6 with $\lambda = 10/5 = 2$ it is seen that a sample size of 4 provides the required assurance.

B. In sampling two normal populations π_1 and π_2, the statistic

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is used to accept or reject the hypothesis that $\mu_1 = \mu_2$. For generality it will be assumed that the population standard deviations σ_1 and σ_2 may not be equal, although they are known accurately.

Our hypotheses are

$$H_0: \mu_1 = \mu_2$$
$$H_1: |\mu_1 - \mu_2| = \lambda\sigma_1.$$

Significance is determined in the same manner as in 5.A., and the power function is set up in identical fashion. The value β is found to be the area under the standardized normal curve between the abscissas.

$$\pm u_\alpha + \lambda\sqrt{\frac{n_1 n_2}{k^2 n_1 + n_2}}$$

where $\sigma_2 = k\sigma_1$. The value of β may easily be read from Fig. 6 for any λ', n_1, n_2, and k by selecting the curve for a convenient sample size, n, on Fig. 6 and taking

$$\lambda = \frac{\lambda'}{\sqrt{n}}\sqrt{\frac{n_1 n_2}{k^2 n_1 + n_2}}.$$

7. Power function of the t-test.

A. The statistic $t = \dfrac{\sqrt{n}(\bar{x} - a)}{s}$ is used to accept or reject the hypothesis that the mean, μ, of the normal population sampled, is equal to some specified level, a, when the population standard deviation, σ_1, is unknown.

Our hypotheses are

$$H_0 : \mu = a$$

$$H_1 : |\mu - a| = \lambda\sigma_1, \ (\lambda > 0).$$

In order to test the hypothesis $\mu = a$ we choose a significance level, α, and compute the statistic $t = \dfrac{\sqrt{n}(\bar{x} - a)}{s}$. If $|t| > t_\alpha$, where the percentage point, t_α, is determined by

$$\frac{\Gamma(n/2)}{\Gamma[\frac{1}{2}(n-1)]\sqrt{n-1}\sqrt{\pi}} \int_{-t_\alpha}^{+t_\alpha} \left(1 + \frac{x^2}{n-1}\right)^{-n/2} dx = 1 - \alpha, \tag{4}$$

we reject H_0 and conclude that $\mu \neq a$.

To set up the power function we note that:

If H_0 is true

$$\Pr\{-t_\alpha < t < +t_\alpha\} = 1 - \alpha.$$

If H_1 is true,

$$\Pr\{-t_\alpha < t < t_\alpha\} = \beta, \qquad (1 - \beta = \alpha \text{ if } \lambda = 0).$$

However, we have the identity

$$Pr\left\{-t_\alpha \frac{s}{\sigma_1} + \lambda\sqrt{n} < \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma_1} < +t_\alpha \frac{s}{\sigma_1} + \lambda\sqrt{n}\right\} = Pr\{-t_\alpha < t < +t_\alpha\}$$

where $\lambda = \dfrac{|\mu - a|}{\sigma_1}$. Hence, for any fixed $\dfrac{s}{\sigma_1}$, the above probability may be denoted by say $h(s/\sigma_1)$ or, using the notation of section 4, $h\left(\sqrt{\dfrac{\chi^2}{n-1}}\right)$, and evaluated as the area under the standardized normal curve between the abscissas indicated. Then

$$\beta = \int_0^\infty h\left(\sqrt{\frac{\chi^2}{n-1}}\right) f(\chi^2)\, d(\chi^2)$$

where $f(\chi^2)$ is the probability density function of χ^2 for $n - 1$ degrees of freedom. This is one method of evaluating β and it was used for calculating the operating characteristics for $n < 5$.

It has been noted that such a formula had been employed by Neyman and Tokarska [6] in calculating Type II errors where only one tail of the t-curve is used as the region of rejection. Probabilities calculated in this manner are

provided by Neyman and Tokarska for degrees of freedom $n = 1$ to 30 and Type I errors of .01 and .05. As soon as the area in one tail of the non-central t-distribution becomes negligible these curves are equivalent to the test treated herein with an α of .02 and .10 respectively. An idea of the critical values of λ at which this occurs may be obtained from a table in a succeeding footnote in which they are quoted for $\alpha = .05$. The values are surprisingly small, such that almost all of Neyman's figures can be interpreted for a two-tail region of rejection.

Using C. C. Craig's development of the non-central t [7] we obtain[2]

$$\beta = Pr\left\{-t_\alpha < \frac{\sqrt{n}(\bar{x} - \mu)/\sigma_1 + \sqrt{n\lambda}}{s/\sigma_1} < +t_\alpha\right\}$$

$$= e^{-\frac{1}{2}n\lambda^2} \sum_{r=0}^{\infty} \frac{(\frac{1}{2}n\lambda^2)^r}{r!} I\left[(r + 1/2), \tfrac{1}{2}(n - 1); \frac{t_\alpha^2}{n - 1 + t_\alpha^2}\right]$$

where $I(p, q;x)$ represents the Incomplete-Beta Function Ratio [7]. This may be conveniently used for those values of n where the necessary values are obtainable from Tables of the Incomplete-Beta Function ratio [8] and for small values of λ where the above series converges rapidly.

The method actually used for $n > 4$, however, made use of the tables prepared by Johnson and Welch [9]. Replacing their λ by π to avoid confusion with our notation, these tables give values of π tabulated against f, t, and ϵ such that

$$Pr\left\{t = \frac{z + \delta}{\sqrt{w}} > t_0\right\} = \epsilon$$

where z is a normally distributed variate with zero mean and unit variance, fw is distributed according to the χ^2-distribution with f degrees of freedom, and $\delta = t_0 - \pi\sqrt{1 + t_0^2/2f}$. We want

$$\beta = 1 - \Pr\{t < -t_\alpha\} - \Pr\{t > t_\alpha\}.$$

For those values of λ and n for which $\Pr\{t < -t_\alpha\}$ is negligible[3] we can, for any given ϵ, take $t_0 = t_\alpha$ and $f = n - 1$ and read π from the tables, then deter-

[2] It should be noted that Craig's formula as published is in error in having $\frac{1}{2}(r + 1)$ as the parameter in the incomplete beta function instead of $r + \frac{1}{2}$.

[3] Values of λ for which $Pr\{t < -t_{.05}\} = .005$ are listed below.

$f = n - 1$	λ
4	.34
5	.30
6	.27
7	.25
8	.23
9	.216
16	.159
36	.103
144	.051
∞	.000

mine δ and finally λ from the relation $\lambda = \delta/\sqrt{n}$. After computing $\beta = 1 - \epsilon$, the point (λ, β) on the operating characteristic may be graphed. At the few places where $\Pr\{t < -t_\alpha\}$ is not negligible and β is needed we can for a given λ take

$$\pi = \frac{t_0 - \delta}{\sqrt{1 + t_0^2/2f}}$$

and then by reading π for various values of ϵ, f, t_0 make an inverse interpolation for ϵ thus setting values for $\Pr\{t > -t_\alpha\}$ and $\Pr\{t > t_\alpha\}$. Finally

$$\beta = \Pr\{t > -t_\alpha\} - \Pr\{t > +t_\alpha\}.$$

It was found that for $n > 10$ a good approximation for computing operating characteristics is given by

$$\beta = \Pr\{-t_\alpha + \lambda\sqrt{n} < t < +t_\alpha + \lambda\sqrt{n}\}$$

in which the variable t is distributed as central t with $n - 1$ degrees of freedom. This formula proved to be quite useful in preparation of the operating characteristics for the t-test.

Fig. 7 presents operating characteristics of the t-test calculated by these methods. It should be noted that in using the t-test, alternative hypotheses are expressed as so many multiples of the unknown population standard deviation away from the level stated in the null hypothesis. In some applications the alternatives may be naturally so expressed. In many applications, however, it may be desired to control the distance $\mu - a$ regardless of the standard deviation of the lot sampled. In this case, one could place confidence limits on the estimate of σ, determine the λ value corresponding to each estimate, and finally obtain limits on the sample sizes or risks involved.[4]

B. For the case of two normal populations, the statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{12}\sqrt{1/n_1 + 1/n_2}}$$

is used to accept or reject the hypothesis that $\mu_1 = \mu_2$ when the two normal population standard deviations are unknown but equal to say, σ_1.

Our hypotheses are

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : |\mu_1 - \mu_2| = \lambda\sigma_1.$$

Significance is determined in the same manner as in par. 6.A., and, by reasoning similar to that in the preceding section, it is found that β for a given λ' can be read from Fig. 7 by taking

$$\lambda = \frac{\lambda'}{\sqrt{n}}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

[4] For a test of this nature in which the power of the test depends only on the absolute value of the distance $\mu - a$ see [10].

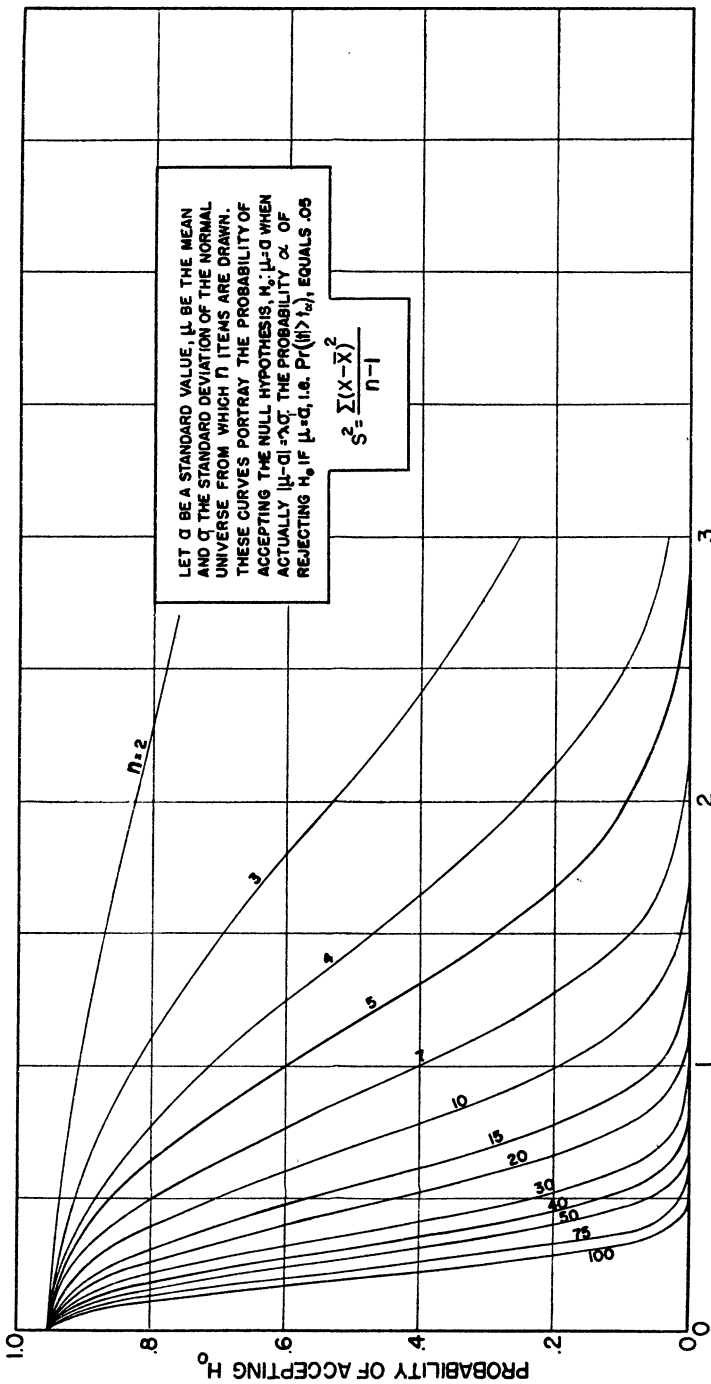

FIG. 7. OPERATING CHARACTERISTICS OF THE t-TEST $\left[t=\frac{\sqrt{n}(x-a)}{s}\right]$ FOR TESTING $\mu = a$ AGAINST $\mu \neq a$

and $n = n_1 + n_2 - 1$. Before a statistical test of this nature is applied the data should be examined to verify consistency with the assumption that $\sigma_1 = \sigma_2$.

Example: An analysis of the difference in tensile strength between two types of castings is being conducted. A sample of 10 items is selected from each type of casting and the t-test employed to establish superiority of one over the other. Experience has shown that the variability in tensile strength for one type of casting is comparable to that of the other type. If α is set equal to .05, what percentage of the time would our significance test fail to detect a superiority of one standard deviation in tensile strength? $n = 10 + 10 - 1 = 19$ and $\lambda = .513$. Referring to Fig. 7 for this λ and n, it is seen that the percentage β is approximately 45.

In this paper we have presented power curves or operating characteristics of the common significance tests employed but a single sample of items. The power of the tests obtained here does not represent the limit that can be obtained for the average amount of inspection performed, say, over many consecutive lots. Tests, sequential in character [11], have been shown to be much more efficient. Nevertheless, single sampling is often the only practical procedure available. Again, the data may be brought to the analyst as single sample results collected supplementary to other purposes or prescribed by a standard procedure. Finally, in performing a significance test, it is quite important to be able to give constructive advice when the data indicate practical differences although no statistical significance is found.[5]

Although sequential tests using variables have been devised, no investigation of double sampling schemes for variables similar to the Dodge-Romig [12] plans for attributes has, as yet, been designed with the exception of [9]. It is believed, however, that such plans would have considerable application for industry in combining efficiency with practicability.

The graphs of the operating characteristics in this report have been made by calculating a sufficient number of points to draw them in by use of French curves. Considering this method of plotting slight error should be allowed for in reading probabilities of acceptance from the graphs, especially where the curves are steep.

REFERENCES

[1] CATHERINE M. THOMPSON, "Tables of percentage points of the χ^2-distribution," *Biometrika*, Vol. 32 (1941).

[2] MAXINE MERRINGTON AND CATHERINE THOMPSON, "Tables of percentage points of the inverted (F) distribution," *Biometrika*, Vol. 33 (1943).

[3] HENRY SCHEFFÉ, "On the ratio of the variances of two normal populations," *Annals of Math. Stat.*, Vol. 13 (1942).

[4] P. C. TANG, "The power function of the analysis of variance tests with tables and illustrations of their use," *Stat. Res. Mem.*, Univ. of London, Vol. 2 (1938), pp. 126–149.

[5] Acknowledgment is made to Colonel Leslie E. Simon, Director, and Mr. R. H. Kent, Associate Director, of the Ballistic Research Laboratory who proposed practical problems related to the work of this report.

[5] MAXINE MERRINGTON, "Table of percentage points of the t-distribution," *Biometrika*, Vol. 32 (1942).

[6] J. NEYMAN AND B. TOKARSKA, "Errors of the second kind in testing 'Student's hypothesis'," *Amer. Stat. Assoc. Jour.*, Vol. 31 (1936).

[7] C. C. CRAIG, "Note on the distribution of non-central t with an application," *Annals of Math. Stat.*, Vol. 12 (1941).

[8] KARL PEARSON, *Tables of the Incomplete Beta Function*, London, 1934.

[9] N. L. JOHNSON AND B. L. WELCH, "Applications of the non-central t-distribution," *Biometrika*, Vol. 31 (1940).

[10] CHARLES STEIN, "A two-sample test for a linear hypothesis whose power is independent of the variance," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 243–258.

[11] A. WALD, "Sequential tests of statistical hypotheses," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 117–186.

[12] H. F. DODGE AND H. G. ROMIG, "Single sampling and double sampling inspection tables," *Bell System Tech. Jour.*, Vol. 20 (1941).