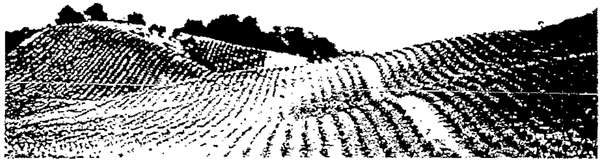
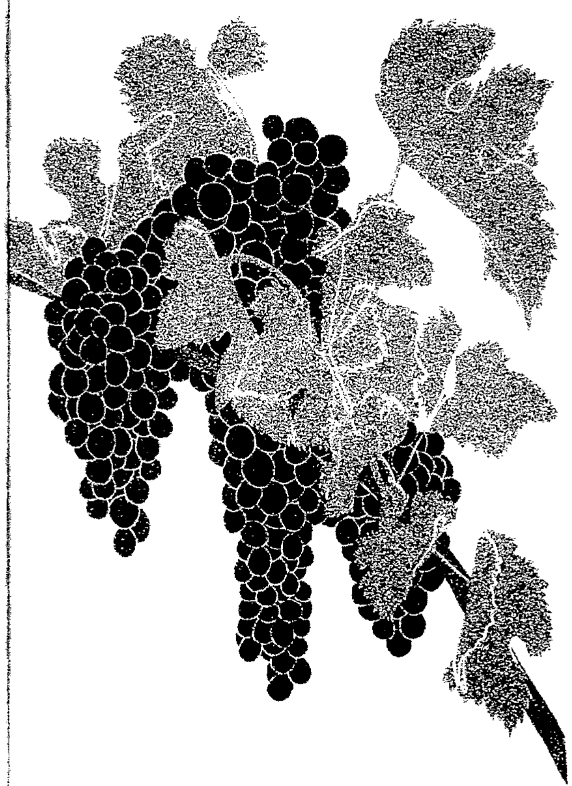
WINES
Their Sensory Evaluation

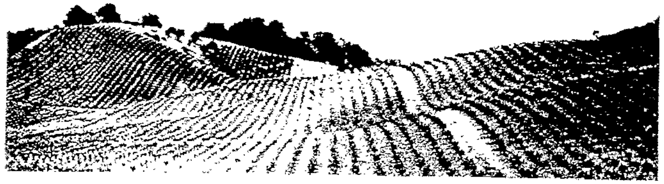


Maynard A. Amerine   Edward B. Roessler
University of California, Davis

W. H. Freeman and Company
San Francisco

II
STATISTICAL PROCEDURES





No food product has a longer history of quality evaluation than wine. Homer, Pliny, and Horace wrote of wines that were famous long before, or at the beginning of, the Christian era. The fame of these wines was undoubtedly based on subjective comparisons, or perhaps even on some sort of deliberate sensory examinations.

Prior to 1940 many quality evaluations in the wine industry were performed by only one or two professionals. Even today, considerable quantities of wine are purchased by skilled wine brokers who base their selections solely on their own evaluations. However, with increasing consumer demand for better wines, greater competition among wine producers, and the development of appropriate statistical procedures for the analysis of sensory data, many wine professionals have concluded that it is unsound to rely on the quality and standards-of-identity judgments of only one or two individuals.

Today it is standard practice in many wineries and wine-distributing companies (and, indeed, throughout the entire food industry) to have regular panel evaluations, not only for quality control of their own products, but also for comparisons with competing products. The data obtained in such evaluations should be subjected to appropriate statistical analysis. Unfortunately, reported differences among wines often imply significance when there is, in fact, no statistical justification for such a conclusion. It is the purpose of Part II of this book to encourage the use of statistical procedures for the analysis of sensory data.

Fundamentals

In Part I of this book we have referred to the importance of statistical procedures in providing tests of significance. A discussion of significance of experimental data is usually based on a comparison of the actual results with those that would be obtained if chance alone were the determining factor. Since the interpretation of such tests depends upon the probabilities of the events in question, some understanding of the concept of probability is essential.

Probability. Briefly, the probability of an event can be defined as the relative frequency of that event in a large number of trials. From this definition it is clear that probability is a number between 0 and 1. An event with probability $p = 0$ cannot occur, and one with probability $p = 1$ is certain to occur. When we say that the probability of getting heads on the toss of a well-balanced coin is $\frac{1}{2}$, we mean that one of every two tosses, *on the average*, will give heads. In other words, it is probable that in a large number of tosses 50% heads and 50% tails will be obtained. This does not mean that in 10 tosses of a coin we will get *exactly* 5 heads and 5 tails, nor that in 100 tosses we will get exactly 50 heads and 50 tails. However, if we continue tossing the coin indefinitely, the ratio of the number of heads (or tails) to the total number of tosses will approach the value $\frac{1}{2}$ (0.5) ever more closely.

Imagine that a judge is presented with three glasses, two of which contain the same wine and the third a different but very similar wine. If he cannot detect a difference among the three, chance alone will determine his ability to pick the odd wine. The probability that he will be successful in doing this is $\frac{1}{3}$; the probability that he will fail is $\frac{2}{3}$.

In a sequence of trials in each of which a certain result may or may not occur, the occurrence of the result is called a *success* and its nonoccurrence a *failure*. In a sequence of coin tosses, for example, getting heads might be designated a success; getting tails would therefore constitute a failure. This terminology is purely conventional, and the result called success need not necessarily be the desired one. The sum of the probabilities of success and failure for a given result is always equal to 1. Therefore, if the probability of success is p, the probability of failure is $1 - p$.

Problems requiring a statistical treatment of events (or results) often entail decisions based on a limited number of observations, the conclusions from which are to apply to a much larger category of events, of which those actually observed are only a part. The larger category about which we wish information is called the *population* (or universe) and the actual observations constitute the *sample*. If the sample is selected in such a way that all components of the population have an equal chance of being included, the sample is called a *random sample*. A quantity calculated from a sample, e.g., its standard deviation (see page 130), is called a *sample statistic*, or simply a *statistic*. Using a statistic to draw conclusions concerning a population from a sample of that population is called *statistical inference*. For such conclusions to be valid the sample must be randomly selected.

Null Hypothesis. The statistical method used in any scientific investigation originates with an investigator's idea, which leads to a tentative hypothesis about the population to be studied. This hypothesis, commonly called the *null hypothesis*, must be a specific assumption, made about some statistical measure of the population, with which to compare the experimental results. For example, in the toss of a fair coin the null hypothesis, $p = \frac{1}{2}$,

states that in a single toss the chances are one in two (50:50) that a head will show.

In a consideration of a judge's ability to differentiate between two wine samples of differing quality, the null hypothesis, $p = \frac{1}{2}$, states that the chances are 50:50 that the judge will make the correct decision, i.e., it states that he does not have the sensory ability to detect a difference. In the previous example of the judge trying to select the odd wine sample from three, two of which are alike, the null hypothesis, $p = \frac{1}{3}$, states that the chances are one in three that the judge will correctly select the odd sample, i.e., it states that he does not have the sensory ability required for this task. In a comparison of the average quality ratings (scores) of two different wines, the null hypothesis, $\mu_1 - \mu_2 = 0$, states that the difference between the mean scores μ_1 and μ_2 for the two populations is zero, i.e., there are no quality differences between the two wine populations from which the samples were selected.

Statistical methods allow us to predict whether or not a null hypothesis is likely to be true or false. A *statistical test*, which is a decision rule or procedure, is then applied to the observed results to decide whether they agree sufficiently well with the expected values to support the null hypothesis or to suggest its rejection in favor of an *alternative hypothesis*. An alternative to the null hypothesis ($p = \frac{1}{2}$) of no sensory ability to differentiate between two wine samples might be $p > \frac{1}{2}$. This alternative hypothesis states that in a single trial the probability of the judge's making the correct decision is greater than $\frac{1}{2}$, i.e., it states that he does have some sensory ability to perform the task. If this hypothesis is true, the chances of his being successful in detecting a difference are therefore better than 50:50. Analogously, an alternative to the null hypothesis of $p = \frac{1}{3}$ might be $p > \frac{1}{3}$. The null hypothesis is usually designated H_0 and the alternative hypothesis H_1.

An alternative hypothesis is called a *one-sided alternative* and the corresponding test a *one-tailed test* if the hypothesis specifies a value on only one side of the value stated in the null hypothesis. The alternative hypotheses $p > \frac{1}{2}$ and $p > \frac{1}{3}$ are therefore both one-sided. If, however, an alternative hypothesis specifies values on both sides of the value stated in the null hypothesis, it is called a *two-sided alternative* and the corresponding test is called a *two-tailed test*. One- and two-tailed tests are illustrated in Figures 1 and 2. We will discuss these illustrations in detail shortly.

Types of Errors. Decision rules are seldom infallible and may lead to rejection of a true hypothesis, which is called an error of the first kind, or a *type I error*. Or, they may lead to acceptance of a false hypothesis, which is called an error of the second kind, or a *type II error*. The probabilities of occurrence of these errors can be minimized but never reduced to zero.

Experimental results rarely lead to obvious conclusions, and the question immediately arises as to the dividing line between acceptance and rejection of the null hypothesis. By a commonly accepted convention the null hypothesis is rejected if, under the

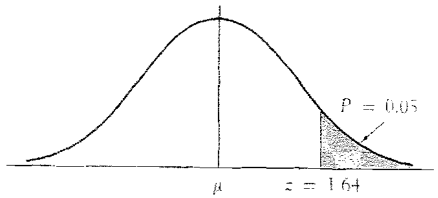


FIGURE 1
One-tailed test, 5% level.

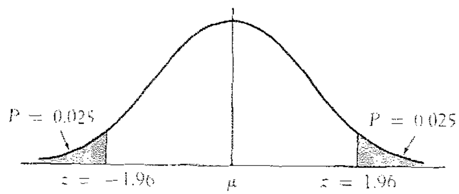


FIGURE 2
Two-tailed test, 5% level.

hypothesis, the result observed in the sample would occur by chance alone *at most* once in 20 trials ($P \leq 0.05$).* Such a result is called *significant*. If, under the null hypothesis and by chance alone, the result would occur at most once in 100 trials ($P \leq 0.01$), it is called *highly significant*, and if it would occur at most once in 1000 trials ($P \leq 0.001$), it is called *very highly significant*. These are known as the 5%, 1%, and 0.1% levels of significance, respectively. It should be understood, however, that, although we accept or reject the null hypothesis on the basis of these levels, we have not proved or disproved it, because there is always the possibility, however remote, that the difference between the observed result and that expected under the null hypothesis could have arisen by chance alone. At the 5% level of significance ($P = 0.05$) we wrongly reject the null hypothesis 5% of the time; at the 1% level ($P = 0.01$) we wrongly reject it 1% of the time; and at the 0.1% level ($P = 0.001$) we wrongly reject it 0.1% of the time, or once every 1000 times, on the average.

Frequency Distributions

For large sets of data comprising many values of a given variable, some form of summarization is needed so that the main features can be readily observed. The simplest method of arranging the data is to divide the whole range of values into a number of equal intervals called class intervals and to count the number of values falling within each such interval. The number of values within a class interval is called the class frequency, or simply the *frequency*. This set of frequencies is called a *frequency distribution*. If the actual frequencies are expressed as fractions of the total frequency, the resulting distribution is called a *probability distribution*. Before considering specific testing procedures we will briefly discuss the usefulness of two frequency distributions—the normal

*The small p introduced earlier is used to denote the probability of a simple event, such as getting heads in a single toss of a coin. The capital P is used to denote the probability of a composite of simple events, such as getting 3 heads in 5 tosses of a coin.

and chi-square (χ^2) distributions—in problems concerning sensory evaluation.

Normal Distribution. The normal distribution can be used to estimate the probabilities of chance results in a judge's performance, but *only* in a task in which there are only *two* possible events, such as picking the odd sample correctly (success) or picking it incorrectly (failure). Probabilities in the distribution are represented by areas under the *normal probability curve*, which is bell-shaped and symmetrical about the mean, μ, of the distribution. Because the value of any normally distributed variable *must* fall somewhere, i.e., because the probability of its falling *anywhere* is 1, the total area (probability) under the curve is equal to 1. Tables for the normal probability curve list the values of the areas (probabilities) corresponding to various values of z, the *normal deviate*, which is defined as the deviation $X - \mu$ measured in terms of the standard deviation, σ:

$$z = \frac{X - \mu}{\sigma} \tag{1}$$

Here X is the value of any normally distributed variable with mean μ, and σ, the standard deviation, is a measure of the dispersion (scatter) in the distribution of X-values about the mean. The smaller the value of σ, the more tightly the X-values cluster about the mean; approximately $\frac{2}{3}$ of them fall between $\mu - \sigma$ and $\mu + \sigma$. The probability of a chance result is a maximum (midpoint on the curve) when $z = 0$, i.e., when $X = \mu$.

In sensory evaluations in which the null hypothesis, H_0, specifies the probability p of success (correct choice) in a single trial, the mean μ (expected number of successes) in n trials is equal to np, and the standard deviation σ can be shown to be $\sqrt{np(1 - p)}$. The *observed* number X of successes is obtained by counting and is therefore always a whole number (integer). When it is used in finding areas under the normal probability curve—which is continuous and therefore permits fractional as well as integral values—X, if it is greater than μ, must be reduced by the number 0.5. This is called a *correction for continuity*. For

Table 1. Values of z and χ^2 at three levels of significance.

Level of significance	Difference (one-tailed)		Preference (two-tailed)	
	$\pm z$	χ^2	z	χ^2
5% (significant)	±1.64	2.71	1.96	3.84
1% (highly significant)	±2.33	5.41	2.58	6.64
0.1% (very highly significant)	±3.09	9.55	3.30	10.82

example, 5 or more on a counting scale is recorded as 4.5 or more on a continuous scale. Then the normal deviate becomes

$$z = \frac{(X - 0.5) - \mu}{\sigma} = \frac{(X - 0.5) - np}{\sqrt{np(1 - p)}} \tag{2}$$

Appendix A gives areas under the normal probability curve to the right of positive values of z or to the left of the corresponding negative values of z. Because the curve is symmetrical the two areas are the same, so only the area to the right of a positive value of z is shown in the graph there and only positive values of z are listed. For a one-tailed test the notation $+z_{.05}$ is used to denote that value of z to the right of which 5% of the total area lies, as shown in Figure 1. Analogously, $-z_{.05}$ would be the value to the left of which 5% of the area lies. From Table 1 we see that, in a one-tailed test, $+z_{.05} = +1.64$ and $-z_{.01} = -2.33$. In a two-tailed test the notation $z_{.05}$ denotes that value of z that defines *two* tail areas, each of which contains 2.5% of the total area, as shown in Figure 2. From Table 1 we see that, in a two-tailed test, $z_{.05} = 1.96$ and $z_{.01} = 2.58$.

Example 1. A judge is presented with three glasses of wine. Two glasses contain the same wine and the third glass a different but similar wine. He is asked to pick the odd sample. What is the probability that, by chance alone, he will be successful 9 or more times in 18 trials?

The formulation of the question ($9 \leq X \leq 18$) implies that we need to find the area under the normal probability curve between the z values corresponding to $X = 9$ and $X = 18$. Because the probability of the judge's being successful in all 18 trials is vanishingly small, however, the area under the curve to the right of the z value corresponding to $X = 18$ is so small that we can include it without introducing any significant error. We can therefore simply let $X = 9$ and find the entire area (to within the accuracy of 4 significant figures, as given in Appendix A) to the right of the corresponding z value.

The probability p of a correct choice in a single trial is $1/3$ and the number of trials n is 18. Therefore $\mu = np = 6$, $\sigma = \sqrt{np(1-p)} = \sqrt{18(1/3)(2/3)} = \sqrt{4} = 2$, and, since $X = 9$,

$$z = \frac{(X - 0.5) - \mu}{\sigma} = \frac{(9 - 0.5) - 6}{2} = \frac{2.5}{2} = 1.25$$

From Appendix A we see that, for $z = 1.25$, $P = 0.1056$. This is the probability (i.e., the chance is about 10 to 11%) that, by chance alone, the judge will correctly identify the odd sample 9 or more times in 18 trials.

Chi-Square (χ^2) Distribution. The chi-square distribution is useful in comparing a set of k observed frequencies (o) with a corresponding set of k expected or hypothesized frequencies (e), particularly when k is greater than 2. The appropriate statistic, which is called *chi-square*, is defined as

$$\chi^2 = \sum \frac{(o - e)^2}{e} \tag{3}$$

where the Greek letter Σ denotes the sum of the k terms $(o_1 - e_1)^2/e_1 + (o_2 - e_2)^2/e_2 + \cdots + (o_k - e_k)^2/e_k$. If the events in question are those of success and failure, as in the examples we have been considering, then $k = 2$, so there are two observed frequencies and two expected frequencies. Chi-square is never negative because in each term the numerator is squared and the denominator is positive. If the observed and expected frequencies agree exactly in every one of the k terms, then $\chi^2 = 0$. It has a positive value if there is any difference between an observed and expected frequency, and it increases as the difference becomes greater.

The distribution of χ^2 depends upon the number of independent differences, called *degrees of freedom* (df). Since the sum of all the expected frequencies, $\sum e$, must agree with the sum of all the observed frequencies, $\sum o$, the sum of all the differences is $\sum(o - e) = 0$. Therefore only $k - 1$ of the expected values are independent, and the remaining one can be calculated from the relation $\sum(o - e) = 0$. The number of degrees of freedom is therefore, $k - 1$. Values of χ^2 for various combinations of probabilities and numbers of degrees of freedom are given in Appendix B.

Imagine a series of n trials, with X observed successes and $n - X$ failures. If the null hypothesis specifies the probability of success in a single trial as p, and therefore that of failure as $1 - p$, χ^2 takes the form

$$\begin{aligned}\chi^2 &= \frac{(|X - np| - 0.5)^2}{np} + \frac{[|(n - X) - n(1 - p)| - 0.5]^2}{n(1 - p)} \\ &= (|X - np| - 0.5)^2[1/np + 1/n(1 - p)] \\ &= \frac{(|X - np| - 0.5)^2}{np(1 - p)}\end{aligned} \tag{4}$$

where $|X - np|$ is the *absolute value* of the expression $X - np$, i.e., it is the value without regard to algebraic sign (it can therefore be interpreted as a positive quantity). As in the normal distribution, the number -0.5 is a correction for continuity because the χ^2 curve is also continuous, whereas the observed frequencies can only be integers. This correction is applicable only for 1 df, which holds for the examples we have been considering, because $k = 2$ (success and failure). In this case the one-tailed probability associated with a value of χ^2 equals the two-tailed probability associated with the corresponding value of z, the normal deviate.

Example 2. Use χ^2 to estimate the probability in Example 1.

$$\chi^2 = \frac{(|X - np| - 0.5)^2}{np(1 - p)} = \frac{(|9 - 6| - 0.5)^2}{18(1/3)(2/3)} = \frac{(2.5)^2}{4} = 1.56$$

From Appendix B we see that, for 1 df, $\chi^2 = 1.56$ is very close to the value 1.64, which corresponds to a probability of 0.20.

Since this equals the total probability for both tails of the normal distribution, the one-tailed probability is close to 0.10, which agrees with the result obtained in Example 1.

The applications and appropriateness of the statistical terms and reasoning outlined above will be evident in the discussions and examples that follow.

Difference Tests

Difference tests are used in the comparison of two wines to evaluate objectively the differences between them, to test the ability of judges to make comparisons of chemical constituents or sensory characteristics, and, on the basis of preference ratings, to establish quality differences.

Sensory evaluations are usually conducted by a small laboratory panel of judges or by members of the consuming public. The number of panelists in laboratory testing varies with conditions, such as the number of qualified persons available. Many investigators recommend panels of 5 to 10 members; we agree. Large panels are customary in preference tests in which the only criterion for the selection of members is representativeness of some consumer population. Laboratory panels can suggest probable consumer reactions but any resulting conclusions relating to the consuming public should be very carefully evaluated. We view such conclusions with considerable skepticism because the relation of the laboratory panel to the consuming public is generally not clear.

The results of a sensory evaluation have little meaning unless the panelists have demonstrated the ability to detect differences that *can* be detected, and to do so consistently. These differences are often very subtle and difficult to detect. Obviously the panel should consist of individuals with the greatest sensitivity and experience. *When no difference can be established, the question of preference is obviously irrelevant.*

Although in the usual statistical analysis the assumptions and test procedures used for one judge making n comparisons are the same as those used for n judges making a single comparison each,

these two experiments are not the same. In all difference tests it is customary to assume an unchanging fundamental probability. Tests based on this assumption are more reliable when performed by one "competent" judge, but even then their validity is doubtful owing to the possibility of fatigue and the effects of various psychological factors (see page 50). The problems encountered in panel or consumer tests are even more complicated because of varying thresholds and differing directions of preference. To conform to basic assumptions in detecting possible differences it is clearly important to use the best judge or judges available.

It has already been pointed out (page 62) that in all trials in wine evaluations the samples should be presented as uniformly as possible—at the same temperature, in identical glasses, but in different orders. Three testing procedures in common use are the paired-sample, duo-trio, and triangle tests.

Paired-Sample Test. In this test the judge is presented with two samples and asked to identify the one with the greater intensity of a specific constituent or well-defined characteristic (see Figure 3). Or, he may be asked to express a preference. This procedure may be carried out by one judge several times or by a panel of judges one or more times. Based on the null hypothesis of no difference, about one-half of the responses should be correct by chance alone, i.e., H_0: $p = \frac{1}{2}$.

Type of test
(e.g., sweetness of wine)

Taste both samples. Circle the sweeter of the two.

Test	Samples	
1	______	______
2	______	______
3	______	______

Name ______________________ Date ______________

FIGURE 3
Record form for paired-sample test.

The paired-sample test is useful not only in quality control and preference evaluation but also in the selection of judges. The presence of more or less of some constituent in one of the samples may already be known to the experimenter, or it can be determined by a specific chemical test. If, in several trials, the judge makes the differentiation correctly significantly more often than would be expected by chance ($p = \frac{1}{2}$), the experimenter can infer that the judge does possess some ability to detect that particular constituent. In this case a one-tailed test is applicable and the alternative hypothesis is H_1: $p > \frac{1}{2}$ because the judge shows ability only if he can make the correct choice more often than he could by guessing. The one-tailed *region of significance* in the normal distribution is shown in Figure 4 for the 5% level. Calculated values of z that exceed $+1.64$, the value at the 5% level ($+z_{.05}$), indicate a significant differentiation ability.

In preference testing the judge is asked to express a preference between two wines. A statistically significant preponderance of selections of one wine over the other then indicates a significant preference difference and, therefore (assuming the judge's tastes are conventional), a significant, objective quality difference. Since either wine may be the preferred one (i.e., since the selection of a given wine very infrequently is just as meaningful as its selection very frequently), the alternative hypothesis here is $H_1: p \neq \frac{1}{2}$ and the two-tailed test is applicable. The two-tailed region of significance in the normal distribution is shown in Figure 5 for the 5% level. Calculated values of z that numerically exceed 1.96, the value at the 5% level ($z_{.05}$), indicate a significant preference or quality difference.

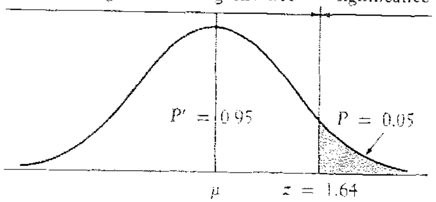


FIGURE 4
One-tailed test. 5% level. H_0: $p = \frac{1}{2}$. H_1: $p > \frac{1}{2}$.

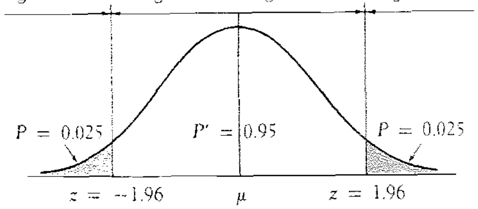


FIGURE 5
Two-tailed test. 5% level. H_0: $p = \frac{1}{2}$. H_1: $p \neq \frac{1}{2}$.

Duo-Trio Test. This test is a modified paired-sample test, in which a reference sample is identified and presented first, followed by two coded samples, one of which is identical to the reference sample. The judge is asked to decide which of the two coded samples is the same as the reference sample (see Figure 6). As in the paired-sample test, the null hypothesis is H_0: $p = \frac{1}{2}$ because, by chance alone, the judge will pick the correct sample about one-

Type of test
(e.g., comparison of old and new blends)

Taste or smell (or both) the reference sample and the two coded samples. Decide which of the latter is the same as the reference sample.

Test	Coded samples		Sample same as reference sample
1	______	______	______
2	______	______	______
3	______	______	______

Name ____________________ Date ____________

FIGURE 6
Record form for duo-trio test.

half of the time. Since this is a difference test, it is one-tailed. It is especially applicable in quality control, in which a sample is to be compared with a reference standard.

Triangle Test. In the triangle test the judge is presented with three samples, two of which are identical. He is asked to select the odd sample (see Figure 7). The probability of a correct choice by chance alone is one-third, i.e., the null hypothesis is H_0: $p = \frac{1}{3}$. The test is easy to administer and is also useful in quality control.

The duo-trio and triangle procedures should be used only for difference (one-tailed) testing, as described above, because it has been shown that having two samples of one wine and one sample of the other tends to cause bias in preference judgments.

For various numbers of trials in the paired-sample and duo-trio tests, Appendix C gives the minimum numbers of correct judgments required to establish a significant difference (one-tailed test) at the 5%, 1%, and 0.1% levels. Also given, for the paired-sample test, are the minimum numbers of agreeing judgments required to establish a significant preference (two-tailed test). Appendix D gives analogous information for establishing a significant difference in the triangle test. Values for $X > \mu$ that are not in the tables can be found by solving the following equations:

Type of test
(e.g., difference in wine flavored by two agents)

Taste or smell (or both) all three samples. Decide which of the three is unlike the other two.

Test	Samples			Sample unlike the other two
1	______	______	______	______
2	______	______	______	______
3	______	______	______	______

Name ______________________ Date ______________

FIGURE 7
Record form for triangle test.

$$X = \frac{n + z\sqrt{n} + 1}{2} \quad \text{or} \quad X = \frac{n + \sqrt{n\chi^2} + 1}{2}$$

$$\text{for } p = \tfrac{1}{2} \text{ (one- or two-tailed)} \qquad (5)$$

and

$$X = \frac{2n + 2.83z\sqrt{n} + 3}{6} \quad \text{or} \quad X = \frac{2n + 2.83\sqrt{n\chi^2} + 3}{6}$$

$$\text{for } p = \tfrac{1}{3} \text{ (one-tailed only)} \qquad (6)$$

In n trials (number of judges or judgments) the minimum number of correct or agreeing judgments required for significance is the next greater integer above the value of X obtained from the appropriate equation above, for the value of z or χ^2 found in Table 1. Values of z for other levels of significance can be found in Appendix A, and values of χ^2 in Appendix B.

Example 3. In a paired-sample test a judge is given two glasses containing a dry white table wine, to one of which a small amount of ethyl acetate has been added. Fourteen times in 20 trials he correctly identifies the adulterated sample. From Appendix C we see that in 20 trials at least 15 correct judgments are required for significance at the 5% level. On the basis of this test, therefore, the judge is not able to detect the ethyl acetate that has been added.

Example 4. In a paired-sample test 50 judges are asked to express their preference for one of two wines. Thirty-six preferences are expressed for wine S_1 and 14 for wine S_2. From Appendix C we see that the minimum number of agreeing judgments required for significance at the 5% level in a two-tailed test is 33, and at the 1% level, 35. On the basis of this test, wine S_1 is judged better than wine S_2 at both the 5% (significant) and 1% (highly significant) levels. Therefore the chances of being wrong in rejecting the null hypothesis (H_0: $p = \frac{1}{2}$) of there being no difference between the wines are less than one in 100.

Example 5. In a duo-trio test of 24 trials, how many correct identifications of the identical samples are required for significance at the 5% and 1% levels? From Appendix C we see that, for a one-tailed test, at least 17 and 19 correct identifications are required for significance at the 5% and 1% levels, respectively.

Example 6. In a triangle test a judge correctly identifies the odd sample in 13 of 23 trials. He therefore indicates ability at the 5% level of significance because, from Appendix D, at least 13 correct identifications are required at this level.

Example 7. In a paired-sample preference test with 64 trials, how many agreeing judgments are required for significance at the 5% level? Since $n = 64$ does not appear in Appendix C, we use Equation 5 to determine X, the number of agreeing judgments required.

$$X = \frac{64 + 1.96\sqrt{64} + 1}{2} = \frac{64 + 1.96(8) + 1}{2} = 40.3$$

Therefore at least 41 agreeing judgments are required at the 5% level of significance.

In testing procedures entailing two or more wines, differences among wine samples can be established by quantitative measures obtained from score cards or other means of scoring, by ranking, or by hedonic rating. We will discuss each of these procedures, but first we must examine in more detail the procedures for selecting judges.

Sequential Procedure for Selection of Judges

When paired-sample, duo-trio, and triangle tests are used in the selection of judges, a predetermined number of trials is employed and those candidates showing the greatest ability are selected. Questions have been raised regarding the number of trials needed and the quality of the judges thus obtained. Often too

little testing is done because of limitations of time and suitable experimental material.

Sequential procedures can provide considerable improvement over other selection procedures and can save valuable time and materials. In a sequential testing plan the number of trials is not predetermined, and the decision to terminate the experiment at any time depends upon the previous results. The sequential procedure described here is a modification of that developed by Wald (1947) and adapted by Bradley (1953).

Let p be the true proportion of correct decisions that would be obtained in paired-sample, duo-trio, or triangle tests if the potential judge were to continue testing indefinitely. This is a measure of his inherent ability in the test in question. Values of p_0 and p_1 are specified such that individuals having abilities equal to or greater than p_1 will be accepted as judges, and those with abilities equal to or less than p_0 will be rejected. The testing plan depends upon the values assigned to p_0 and p_1 and also upon the values of α and β, the probabilities of committing errors of the first and second kind, respectively (α is the probability of rejecting a qualified judge and β is the probability of accepting an unqualified one). Potential judges are accepted or rejected on the basis of their performance with respect to a chart of two parallel straight lines L_0 and L_1, which are uniquely determined by the assigned values of p_0, p_1, α, and β. These lines divide the plane into three regions: one of acceptance, one of rejection, and one of indecision, as shown in Figure 8.

The equations of the lines are

$$L_0: d_0 = a_0 + bn \quad \text{and} \quad L_1: d_1 = a_1 + bn \qquad (7)$$

where n is the total number of trials, d (either one) is the accumulated number of correct decisions, b is the common slope of the two lines, and a_0 and a_1 are the intercepts on the vertical axis. The common slope b of L_0 and L_1 is

$$b = k_2/(k_1 + k_2) \qquad (8)$$

and the intercepts a_0 and a_1 are

$$a_0 = -e_1/(k_1 + k_2) \quad \text{and} \quad a_1 = e_2/(k_1 + k_2) \qquad (9)$$

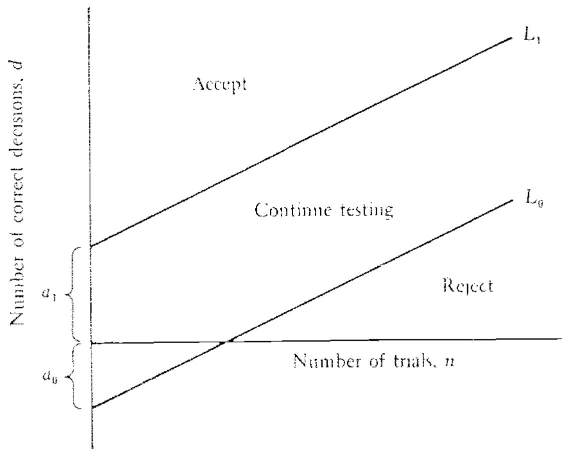


FIGURE 8
Sequential test chart.

where

$$k_1 = \log(p_1/p_0) = \log p_1 - \log p_0$$
$$k_2 = \log[(1 - p_0)/(1 - p_1)] = \log(1 - p_0) - \log(1 - p_1)$$
$$e_1 = \log[(1 - \beta)/\alpha] = \log(1 - \beta) - \log \alpha$$
$$e_2 = \log[(1 - \alpha)/\beta] = \log(1 - \alpha) - \log \beta$$

After each trial the experimenter plots the point (d, n), representing the accumulated number of correct decisions (vertical scale) versus the total number of trials (horizontal scale). Each plotted point is therefore one n unit to the right of the preceding point, and is either one d unit above the preceding point or on the same horizontal level, depending on whether the decision was correct or incorrect, respectively. Testing continues until a plotted point falls on or above the upper line, resulting in acceptance of the candidate as a judge, or on or below the lower line, resulting in his rejection.

The number of trials required depends upon the ability of the potential judge and on the assigned values of p_0, p_1, α, and β, which are determined by the experimenter. Before committing

himself to a given set of values the experimenter may wish to know the average number of trials that can be expected for that set of values. The number of trials required can be decreased by increasing the difference between p_0 and p_1 or by increasing α or β, or both. If competent judges are in good supply the experimenter may wish to increase α and accept a greater risk of rejecting a competent judge.

The average number of trials to be expected, $\bar{n}$, can be obtained from among four calculated values corresponding to special values of p, as shown below.

$p = 0$ (no ability)

$$\bar{n}_0 = e_1/k_2$$

$p = p_0$ (maximum unacceptable ability)

$$\bar{n}_{p_0} = \frac{(1 - \beta)e_1 - \beta e_2}{(1 - p_0)k_2 - p_0 k_1}$$

$p = p_1$ (minimum acceptable ability)

$$\bar{n}_{p_1} = \frac{(1 - \alpha)e_2 - \alpha e_1}{p_1 k_1 - (1 - p_1)k_2}$$

$p = 1$ (infallible ability)

$$\bar{n}_1 = e_2/k_1$$

The average number of trials to be expected is the largest of these four values.

Example 8. Suppose that a triangle test is used as a basis for selecting judges in a sequential procedure. For the assigned values $p_0 = 0.45$, $p_1 = 0.70$, $\alpha = 0.10$, and $\beta = 0.05$, find the average number of trials to be expected. (Competent judges are in good supply, so α is being taken as 0.10.)

We begin by finding the values of k and e:

$$k_1 = \log(0.70/0.45) = 0.1919$$
$$k_2 = \log(0.55/0.30) = 0.2632$$
$$e_1 = \log(0.95/0.10) = 0.9777$$
$$e_2 = \log(0.90/0.05) = 1.2553$$

We then use these values in the four equations for $\bar{n}$:

$$\bar{n}_0 = 0.9777/0.2632 = 3.7$$

$$\bar{n}_{p_0} = \frac{(0.95)(0.9777) - (0.05)(1.2553)}{(0.55)(0.2632) - (0.45)(0.1919)} = \frac{0.866}{0.058} = 14.9$$

$$\bar{n}_{p_1} = \frac{(0.90)(1.2553) - (0.10)(0.9777)}{(0.70)(0.1919) - (0.30)(0.2632)} = \frac{1.032}{0.055} = 18.8$$

$$\bar{n}_1 = 1.2553/0.1919 = 6.5$$

We see that the test will require an average of 19 trials. The number required for each candidate will, of course, depend upon his inherent ability, p.

Example 9. Using the values of k and e calculated in Example 8, find the equations of the lines L_0 and L_1.

From Equations 8 and 9 we obtain

$$b = 0.2632/0.4551 = 0.578$$

$$a_0 = -0.9777/0.4551 = -2.15$$

$$a_1 = 1.2553/0.4551 = 2.76$$

The equations of the lines are therefore

$$L_0: \quad d_0 = -2.15 + 0.578n$$

$$L_1: \quad d_1 = 2.76 + 0.578n$$

Example 10. The performances of two candidates for wine judge, A and B, are shown in the table below, where a 1 indicates a correct decision and a 0 an incorrect decision. Evaluate their performances with respect to the lines L_0 and L_1 and determine the number of trials after which each candidate is either accepted or rejected.

No. of trials	n:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Decisions	A:	1	0	1	0	0	0	1	0	1	1	0	0	0						
	B:	1	1	1	0	1	0	0	1	1	1	0	1	1	1	0	1	1	1	1
No. of correct	A:	1	1	2	2	2	2	3	3	4	5	5	5	5						
decisions, d	B:	1	2	3	3	4	4	4	5	6	7	7	8	9	10	10	11	12	13	14

The performances of A and B are shown in Figure 9, in which the number of correct decisions is plotted against the

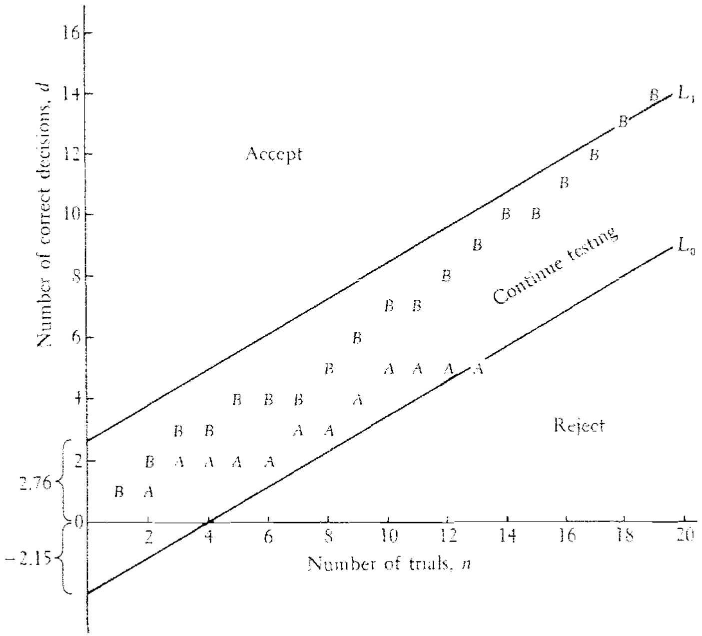


FIGURE 9

Performances of candidates A and B in a sequential test procedure.

total number of trials. By the criteria specified for the sequential procedure, we see that A is rejected as a judge after 13 trials and B is accepted after 19 trials.

Scoring

With experienced judges scoring is usually the most acceptable procedure for establishing differences among wine samples because it measures the magnitudes of the differences. The scoring scale to be used must be clearly defined and understood by all the judges. A 9-point quality scale (see Figure 10) has been widely used. It is an example of an ordinal scale (Stone et al., 1974). The judge

Wine sample ____________

Check the appropriate quality

- ______ Excellent
- ______ Very good
- ______ Good
- ______ Below good, above fair
- ______ Fair
- ______ Below fair, above poor
- ______ Poor
- ______ Very poor
- ______ Extremely poor

Name ____________________ Date ____________

FIGURE 10
A 9-point quality scale.

checks the appropriate quality, which is converted to a numerical score: 1 for extremely poor to 9 for excellent.

In the evaluation of overall wine quality, score cards usually provide for 10-point or 20-point rating scales. On the basis of a 20-point scale the following groupings are suggested: (a) *superior* (17–20 points)—wines of fine quality, well-balanced, no pronounced defects, and free of excess "young" character; (b) *standard* (13–16 points)—the wines of commerce (including ordinary bottled wines), not deficient in any important characteristic, but lacking proper age or the balance required for fine quality; (c) *below standard* (9–12 points)—wines lacking some required characteristic or suffering from some malady (wines with off odors or off taste or high volatile acidity); (d) *unacceptable*, or *spoiled* (1–8 points)—wines so spoiled that they must be discarded. See pages 130–161 for methods of analyzing the results.

Davis Score Card. The original, so-called Davis score card (see Figure 11) was developed by the staff of the Department of Viticulture and Enology at the University of California, Davis, as a method of rating the large number of experimental wines that were produced there. Later it was used as a training device for students who were beginning their education in the sensory evaluation of wines. This score card overemphasized some factors

Wine sample ____________

Characteristic	Weight
Appearance	2
Color	2
Aroma and bouquet	4
Volatile acidity	2
Total acidity	2
Sweetness	1
Body	1
Flavor	2
Bitterness	2
General quality	2

Ratings: *superior* (17–20); *standard* (13–16); *below standard* (9–12); *unacceptable*, or *spoiled* (1–8)

Name ____________________ Date ____________

FIGURE 11
The Davis score card. (The meanings specified for the total scores serve to assure relative uniformity of the judges' interpretations of these terms.)

(acescence, for example) and underemphasized others (aroma and bouquet being the worst examples). Among its other defects was that it did not differentiate between bitterness and astringency (page 42). The concepts of flavor (now generally regarded as odor perceived via the mouth) and general quality were not clearly defined. It also became apparent that the definitions of superior (17–20 points), standard (13–16 points), below standard (9–12 points), and unacceptable, or spoiled (1–8 points) varied from judge to judge, depending on the judge's experience and the severity of his judgment.

Despite these deficiencies the Davis score card has been successfully used by highly skilled judges at Davis without serious difficulty. In fact, the staff has learned to use it with remarkable precision of the results and their interpretation. As a pedagogical tool it has proved useful for both regularly enrolled students and those taking adult wine-appreciation courses. The above-mentioned problems are always explained to the students. A modified Davis score card is shown in Figure 12.

Wine sample ______________

Characteristic	Weight
Appearance	2
Color	2
Aroma and bouquet	6
Total acidity	2
Sweetness	1
Body	1
Flavor	2
Bitterness	1
Astringency	1
General quality	2

Ratings: *superior* (17–20); *standard* (13–16); *below standard* (9–12); *unacceptable, or spoiled* (1–8).

Name ______________________ Date ______________

FIGURE 12
A modified Davis score card.

In recent years the Davis score card has been used (or misused) by professional and amateur groups with less success. Most of the difficulty arises from varying interpretations of the score card. Some amateurs assign high scores to all the wines, whereas professionals usually spread their scores over a larger range. Disaster occurs when amateurs and professionals judge together and the average scores for the individual wines are used to rank the wines. This cannot be done safely without appropriate statistical analysis of the data, and the latter is hardly ever done.

One solution to this problem would be to hold one or more practice sessions of the group and discuss the meaning of the scores. Another possible solution would be to use the shorter, 10-point score card devised by Ough and Baker (1961). However, bunching of the scores in the 8-to-10-point range would then be even more acute than bunching in the 17-to-20-point range of the 20-point scale.

We recommend using only professional judges if the objective is to rank a group of wines in order of merit *by their scores.* The judges, though experienced, will still require one or more practice

sessions in which their scores are compared. Although it may embarrass a judge to be found scoring too high or too low, it is essential that this be revealed if the average scores are to be meaningful. Also, judges may have very different standards of excellence for different types of wines. With samples before them the judges should discuss the various types of wines to be evaluated. Questions such as the following must be discussed: What range of color will be tolerated in a given type of wine? What is the typical varietal aroma? How much fermentation bouquet can be allowed (especially in young white wines)? Are dry and sweet wines to be judged together? How much credit should be given for bottle bouquet (as in a well-aged red wine)? With respect to these and similar questions the differences between superior and standard wines must be clear to all the judges.

Other Score Cards. A 20-point score card that avoids the detailed evaluation required for the Davis score card is shown in Figure 13 and appears very useful. Two noteworthy features of this score card are the provision for listing specific defects and the specification of the minimum acceptable number of points for each of the three categories. One disadvantage is the heavy weight given to taste in evaluating the wine.

Klenk (1972) has used the following, very similar 20-point score card: color, 2; appearance, 2; odor, 4; taste, 12. Again the taste contribution to quality seems to us to be greatly overemphasized. In competitions in which this scale was used, the gold medal was given to wines that scored 19.6 to 20, the silver medal to wines scoring 18.6 to 19.5, and the bronze medal to wines scoring 17.6 to 18.5. For example, Klenk gives data for 8 wines, each of which was judged by 4 judges. The averages were 20 and 19.9 (gold medal), 19.0, 19.0, and 18.8 (silver medal), and 18.5, 18.5, and 17.8 (bronze medal). Statistical analysis of Klenk's data shows that differences of less than 0.51 between average scores were not significant. Therefore the silver-medal wines and the first two bronze-medal wines were not significantly different from one another.

Klenk has also used a 40-point score card scaled as follows: color, 3; appearance, 3; odor, 10; taste, 24. We believe that this is too

Wine Evaluation
Seagram Wine Quality Laboratory

To: ____________ Date: ________ Sample no. ______
From: ____________ Location ____________
Brand: ____________ Type: ____________
Producer ____________ Country: ____________
Age: ______ % Ethanol: ________ No. of bottles ____________
Samples recd from: ____________ Date evaluated ____________
Identification of case, bottle, date, etc.: ____________

Appearance — *Points*

Ordinary sound wine ____________ 3
Outstanding color and clarity (+1) ____________
____
Defects (−1 to −3) ____________
____

Total appearance points [ ]
(Maximum 4; minimum acceptable 2)

Odor — *Points*

Ordinary sound wine ____________ 2
Positive attributes (+1 to +2) ____________
____
Defects (−1 to −2) ____________
____

Total odor points [ ]
(Maximum 4; minimum acceptable 1)

Taste — *Points*

Ordinary sound wine ____________ 7
Positive attributes (+1 to +5) ____________
____
Defects (−1 to −7) ____________
____

Total taste points [ ]
(Maximum 12; minimum acceptable 5)

Total wine rating [ ]
(Maximum 20; minimum acceptable 9)

Ratings: *great* (18–20); *fine* (15–17); *good* (12–14); *fair* (9–11); *poor*, or *unacceptable* (below 9). Wine is also unacceptable if it does not meet the minimum in all three categories. Unless it also *exceeds* the minimum in at least one category, it cannot meet the overall minimum of 9 points.

FIGURE 13
A 20-point score card.
(Courtesy of Joseph E. Seagram & Sons, Inc.)

great a range for normal use because judges cannot differentiate 40 levels of quality.

The typical 20-point score card is well suited for the evaluation of still table wines, but it requires modification for other types of wines. For example, the persistence of the sparkle in sparkling wines must be taken into account; either flavor or general quality may be invoked as a means of subtracting points for lack of persistence. In dessert wines (except muscatels) aroma is not a prominent characteristic: greater emphasis must be given to bouquet.

Another score card that has been used in international judgings is that of the *Office International de la Vigne et du Vin*, in Paris (see Figure 14). The perfect score is 0. Defects are marked on an increasing scale for each category as a multiplying factor (×0, ×1, ×4, ×9, ×16). As with all score cards, some degree of familiarity with the terms is necessary. Odor intensity and odor quality seem clear enough. The difference between taste intensity and taste quality is by no means so clear. Taste intensity would *seem* to pertain to the positive aspects of sweetness, sourness, and bitterness, i.e., the ideal intensity of each. Taste quality would then pertain to the balance (or lack of it) in the overall taste character.

		Multiplying factor for increasing defects				
Characteristic	Weight	×0	×1	×4	×9	×16
Appearance	1	____	____	____	____	____
Color	1	____	____	____	____	____
Odor intensity	1	____	____	____	____	____
Odor quality	2	____	____	____	____	____
Taste intensity	2	____	____	____	____	____
Taste quality	3	____	____	____	____	____
Harmony or balance	2	____	____	____	____	____

Multiplying factors: *outstanding* (0); *very good* (1); *good* (4); *acceptable* (9); *unacceptable* (16).

Name ____________ Date ____________

FIGURE 14
Score card of the Office International de la Vigne et du Vin, *Paris.*

The *Associazione Enotecnici Italiani* (1975), in Milan, has proposed a 100-point score card (see Figure 15). This system will probably work as well as most others, although it has several disadvantages: a 100-point scale is too large, the words *finesse* and *harmony* are difficult to define in sensory terms, and old red wines and most dessert wines would score low in freshness. It does have the advantage, however, of forcing the judge to quantify his judgments, from *bad* to *excellent*, on several wine attributes.

When other evaluation methods are used, such as ranking or hedonic rating, it is still necessary that the judges understand the problems discussed above and that they agree as closely as possible on the definitions and interpretations of the terms to be used in describing the wines.

	Weight	Excellent 4	Good 3	Average 2	Mediocre 1	Bad 0
Visual						
Appearance	2	____	____	____	____	____
Color	2	____	____	____	____	____
Olfactory						
Finesse	2	____	____	____	____	____
Intensity	2	____	____	____	____	____
Freshness	2	____	____	____	____	____
Taste						
Body	2	____	____	____	____	____
Harmony	2	____	____	____	____	____
Intensity	2	____	____	____	____	____
Final taste-odor sensation	3	____	____	____	____	____
Typicalness	3	____	____	____	____	____
General impression	3	____	____	____	____	____

Name ______________________ Date ____________

FIGURE 15
Score card adapted from that published by the Associazione Enotecnici Italiani (1975), Milan.

Rank the 6 samples in order of increasing ethanol content.

Highest ______
______
______
______
______
Lowest ______

Name ______________________ Date ____________

FIGURE 16
Ranking wines in order of percent ethanol.

Ranking

In the ranking procedure the judges are asked to arrange a series of two or more samples in increasing or decreasing order with respect either to the intensity of a particular characteristic or to their own preference (see Figure 16). The test is simple to administer, may not require highly skilled judges, and makes possible a distribution-free analysis. It does, however, disregard degrees of difference among the wines and is therefore usually less sensitive to the effects of such differences than tests based on scoring. See pages 161–167 for methods of analyzing the results.

Hedonic Rating

Hedonic rating is what the name implies: quality evaluation based on the pleasure that the judge finds in the wine. The evaluations are usually made on 5- to 9-point balanced scales ranging from extreme disapproval to extreme approval, such as the one shown in Figure 10. The results are converted to numerical scores, which are then treated by rank analysis or the analysis of variance (these topics are discussed later). The procedure is used by both experts and untrained consumers, but is more appropriate for the latter group.

What do the results of hedonic rating mean? Are they merely a subjective preference opinion? If so, averaging the scores is not very meaningful. However, if they denote a degree of quality relative to some theoretical, agreed-upon standard of perfection, then the average score may have objective value. In fact, if tested by appropriate statistical procedures, the differences among the average scores of the various wines may reveal significant differences among the wines, or they may indicate no significant differences. See pages 145–147 for methods of analyzing the results.

Tests of Significance of Scores

Regardless of the type of evaluation procedure used, the overall results for each wine in the test are usually expressed in terms of a single numerical score. These scores can be analyzed statistically to determine if significant differences exist. Although the usual statistical procedures presuppose a normal distribution of scores, moderate deviations from such a distribution do not invalidate the results. Studies have shown that the distribution of scores in most tests is only moderately asymmetrical, and the usual test procedures are valid. Sometimes the scores fit a *bimodal distribution* (one with two peaks in its graph), which means that we may be dealing with two types of judges who differ significantly in their quality standards or preferences. It may then be desirable to separate and compare the scores for the two groups making up the bimodal distribution.

Variability. Tests of significance entailing means (averages) of scores are based on estimates of the variability of that population of which the scores constitute a random sample (see page 102). The customarily used estimates of the variability are the variance, v, of a sample distribution of scores and its square root, $s = \sqrt{v}$. The latter represents what is called the best estimate of the standard deviation of the population, as determined from a sample of that

population.* The variance is thus a measure of the dispersion of the observed values of a variable (here, the score) about the mean value. If $X_1, X_2, X_3, \cdots, X_n$ represent n sample scores, their mean value is

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum X}{n} \qquad (10)$$

where, in analogy with our previous usage, the Greek letter Σ denotes the sum of the n values of X.

The best estimate of the variance of the population of which the n scores are a random sample is defined as

$$v = s^2 = \frac{\sum(X - \overline{X})^2}{n-1} = \frac{\sum X^2 - (\sum X)^2/n}{n-1} = \frac{\sum X^2 - C}{n-1} \qquad (11)$$

where $C = (\sum X)^2/n$ is a correction term that converts the sum of the squares of the deviations of the scores from 0, $\sum(X - 0)^2 = \sum X^2$, into the sum of the squares of the deviations of the scores from their mean value, $\overline{X}$, $\sum(X - \overline{X})^2$. It is customary to refer to the numerator of the expression for v as the *sum of squares* (*SS*) and to the denominator as the corresponding number of *degrees of freedom* (*df*). The latter is $n - 1$ because $\sum(X - \overline{X}) = 0$ and therefore only $n - 1$ of the differences $X - \overline{X}$ are independent. A sum of squares divided by the number of degrees of freedom gives an unbiased estimate of the variance of the population.

Example 11. From the 8 sample scores X = 8, 7, 6, 5, 5, 6, 8, and 7, verify numerically that $\sum(X - \overline{X}) = 0$ and that $\sum(X - \overline{X})^2 = \sum X^2 - C$. Find the value of s, the best estimate of the standard deviation of the population from which the sample was selected.

Partial calculations are shown in Table 2, from which we see immediately that $\sum(X - \overline{X}) = 0$. Using the other sums shown there, we obtain $C = (52)^2/8 = 338$, so

*Note that the standard deviation of the population is denoted by σ (see page 106) but the best estimate of it, based on the actual sample, is denoted by s.

Table 2. Partial calculations for the scores given in Example 11.

	X	$X - \overline{X}$	$(X - \overline{X})^2$	X^2
	8	1.5	2.25	64
	7	0.5	0.25	49
	6	−0.5	0.25	36
	5	−1.5	2.25	25
	5	−1.5	2.25	25
	6	−0.5	0.25	36
	8	1.5	2.25	64
	7	0.5	0.25	49
Total	52	0	10.00	348
Mean	6.5			

$$\sum X^2 - C = 348 - 338 = 10 = \sum(X - \overline{X})^2$$

From Equation 11 we obtain $v = 10/7 = 1.43$, so

$$s = \sqrt{v} = \sqrt{1.43} = 1.20$$

We will encounter calculations of this kind again (see page 137) in the discussion of analysis of variance.

The *t*-Distribution

When the standard deviation σ of the population is known, the normal distribution is applicable in "either-or" decision problems, such as: is there a significant difference between these two mean scores or not? If σ is unknown and must be estimated from a sample by calculating s, the sampling distribution of the resulting statistic (see page 130) is no longer a normal one. The appropriate test statistic in this case is denoted by t. Like χ^2, t has a different distribution for each value of the number of degrees of freedom. When the population is normal, the t-curve is symmetrical and bell-shaped, but non-normal. As the size of the sample from which

s is calculated increases, the t-curve approaches the normal curve as a limiting form.

Values of t for various combinations of probabilities and numbers of degrees of freedom are given in Appendix E. The probabilities shown at the top of the table pertain to a two-tailed test, and those shown at the bottom of the table are the corresponding values for a one-tailed test.

Two Sets of Scores (Unpaired). Statistical tests for significant difference are based on the null hypothesis that no difference exists. This assumption applies both to population mean scores and standard deviations. The statistic t is useful in determining significance in such tests. If, for two sets of scores, no score from one set corresponds to any particular score from the other set (as, e.g., in the sets of scores obtained for one wine by two different panels of judges), the scores are independent, or *unpaired*, and the t-distribution furnishes the appropriate test of significance for comparing the mean scores of the two sets. Suppose there are n_1 X-scores and n_2 Y-scores (n_1 may or may not equal n_2); t is then defined as

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\left(\dfrac{n_1 + n_2}{n_1 n_2}\right)\left[\dfrac{\sum X^2 + \sum Y^2 - (\sum X)^2/n_1 - (\sum Y)^2/n_2}{n_1 + n_2 - 2}\right]}}$$

$$(df = n_1 + n_2 - 2) \qquad (12)$$

The significance of the result is determined by comparing the calculated value of t with the two-tailed values given in Appendix E, for the appropriate number of degrees of freedom. Calculated values of t that exceed those in the table indicate significant differences between the mean scores $\overline{X}$ and $\overline{Y}$, at the level of significance in question. In other words, such values of t lead to rejection of the null hypothesis of no difference.

Example 12. A panel of 6 judges scores a wine on a 10-point scale (see X-scores in Table 3) and a second panel of 8 judges scores the same wine, using the same scale (see Y-scores in

Table 3. A wine scored by two panels of judges (see Example 12).

	Panel			
	X	Y	X^2	Y^2
	9	8	81	64
	8	7	64	49
	7	6	49	36
	9	5	81	25
	7	5	49	25
	8	6	64	36
		8		64
		7		49
Total	48	52	388	348
Mean	8.0	6.5		

Table 3). Is there a significant difference at the 5% level between the mean scores for the two panels?

Using the total and mean values obtained in Table 3, we solve Equation 12:

$$t = \frac{8.0 - 6.5}{\sqrt{\left[\frac{6+8}{6(8)}\right]\left[\frac{388 + 348 - (48)^2/6 - (52)^2/8}{6+8-2}\right]}}$$

$$= \frac{1.5}{\sqrt{0.340}} = \frac{1.5}{0.583} = 2.57$$

From Appendix E we see that $t_{.05}(12\ df) = 2.179$. Since the calculated value $t = 2.57$ is greater than the tabular value 2.179, the null hypothesis of no difference must be rejected, and the analysis indicates that the mean scores for the two panels are significantly different. The two panels are therefore not using the same standards of judgment in evaluating the wine.

Two Sets of Scores (Paired). If one judge compares the same two wines on several different occassions, or if each member of a panel

of judges compares the same two wines, a set of *paired* scores results. For the n paired scores X and Y, the differences $D = X - Y$ are then computed, and the mean difference $\overline{D} = \sum D/n$ between the mean scores $\overline{X}$ and $\overline{Y}$ is tested with the t-distribution. The expression for t in this case is

$$t = \frac{\overline{D}}{\left(\frac{1}{n}\right)\sqrt{\frac{n\sum D^2 - (\sum D)^2}{n-1}}} = \frac{\sum D}{\sqrt{\frac{n\sum D^2 - (\sum D)^2}{n-1}}} \qquad (df = n - 1) \qquad (13)$$

Again the calculated value of t is compared with the two-tailed values given in Appendix E to determine the significance of the result.

***Example* 13.** A panel of 7 judges scores two wines on a 20-point scale, as shown in Table 4. Is there a significant difference at the 5% level between the mean scores of the wines?

Using the total values for D and D^2 obtained in Table 4, we solve Equation 13:

Table 4. Two wines scored by 7 judges (see Example 13).

	Wine			
Judge	X	Y	D	D^2
A	15	14	1	1
B	12	14	−2	4
C	14	15	−1	1
D	17	14	3	9
E	11	11	0	0
F	16	14	2	4
G	15	13	2	4
Total	100	95	5	23
Mean	14.3	13.6	0.714	

$$t = \frac{5}{\sqrt{\frac{7(23) - (5)^2}{7 - 1}}} = \frac{5}{\sqrt{22.7}} = \frac{5}{4.76} = 1.05$$

From Appendix E we see that $t_{.05}$ (6 *df*) = 2.447. Since the calculated value t = 1.05 is less than the tabular value 2.447, there is no reason to reject the null hypothesis. Therefore the mean scores of the wines are not significantly different, i.e., this panel of judges cannot distinguish between the two wines.

Analysis of Variance

Scores for Several Wines. In comparing the mean scores of more than two wines, the *t*-distribution is no longer appropriate. Instead, the statistical technique called *analysis of variance* is used to determine whether there are significant differences in the mean scores of the wines. The analysis of variance is essentially an arithmetic process for partitioning a total sum of squares (page 131) into components associated with various sources of variation.

To analyze a number, say k, of wines, for each of which n scores are available, a so-called *one-way*, or single-classification, analysis of variance is appropriate. Such a classification is shown in Table 5,

Table 5. One-way analysis of variance.

	Wine					
	1	2	3	···	k	
	X_{11}	X_{21}	X_{31}	···	X_{k1}	
	X_{12}	X_{22}	X_{32}	···	X_{k2}	
	X_{13}	X_{23}	X_{33}	···	X_{k3}	
	⋮	⋮	⋮	X_{ij}	⋮	
	X_{1n}	X_{2n}	X_{3n}	···	X_{kn}	
Total	W_1	W_2	W_3	···	W_k	Grand total $G = \Sigma W_i$
Mean	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	···	$\bar{X}_k$	Total no. of scores = kn

where X_{ij} represents the j-th score of the i-th wine sample (i can have any value from 1 to k, and j can have any value from 1 to n).

The variance of this classification of scores can be estimated in three ways, from three sums of squares (two of which include a relevant correction term, C) and their corresponding numbers of degrees of freedom. The three sums of squares in question are the *total* sum of squares, the *sample* sum of squares (i.e., the sum of squares between means of wine samples), and the *error* sum of squares (i.e., the sum of squares within samples). The correction term and the three sums of squares are defined as follows:

$$C = (\text{Grand total})^2/kn = G^2/kn \qquad (14)$$

$$\text{Total } SS = \sum_{ij} X_{ij}^2 - C \qquad (df = kn - 1) \qquad (15)$$

$$\text{Sample } SS = n\left(\sum_i \bar{X}_i^2 - G^2/kn^2\right)$$
$$= (W_1^2 + W_2^2 + \cdots + W_k^2)/n - C$$
$$= \sum_i W_i^2/n - C \qquad (df = k - 1) \qquad (16)$$

$$\text{Error } SS = \left(\sum_j X_{1j}^2 - W_1^2/n\right) + \left(\sum_j X_{2j}^2 - W_2^2/n\right) + \cdots$$
$$+ \left(\sum_j X_{kj}^2 - W_k^2/n\right)$$
$$= \sum_{ij} X_{ij}^2 - \sum_i W_i^2/n \qquad [df = k(n - 1)] \qquad (17)$$

From these relations it follows that

$$\text{Total } SS = \text{Sample } SS + \text{Error } SS \qquad (18)$$

and

$$\text{Total } df = \text{Sample } df + \text{Error } df \qquad (19)$$

The within-sample sum of squares (Error *SS*) is usually calculated by subtracting the between-sample sum of squares (Sample *SS*) from the total sum of squares (Total *SS*). The value of the *error mean square* (the error variance) is given by v = Error *SS*/ Error *df*. It is often referred to as a *generalized error term* because it is a measure of the error variation contributed by all the samples.

It is independent of any differences that might exist among the sample means. The value of the *sample mean square* (Sample SS/Sample *df*), on the other hand, is a measure of the differences among the sample means; the larger the differences, the larger the sample mean square. The null hypothesis is that the samples come from k populations, all having the same means μ and the same variances v. This implies equality among the sample means.

The sample mean square and the error mean square provide two independent estimates of the common population variance. They are compared by calculating their ratio, which is a statistic called F:

$$F = \frac{\text{Sample mean square}}{\text{Error mean square}} \qquad (20)$$

This calculated F-value is compared with the tabular values given in Appendixes F-1, F-2, or F-3. The F-distribution is represented by double-entry tables with respect to the degrees of freedom. The degrees of freedom for the numerator are shown in the top rows of the tables, and the degrees of freedom for the denominator are shown in the left-hand columns. Calculated F-values that exceed the tabular values for the appropriate values of *df* indicate rejection of the null hypothesis of no differences among the sample means, i.e., there are significant differences. (If the sample mean square is less than the error mean square, $F < 1$ and the result is nonsignificant by definition. The null hypothesis is then accepted without the need to refer to the table.) A significant F-value implies that the evidence is sufficiently strong to indicate differences among the sample means, but it does not reveal *which* of the various differences among the sample means may be statistically significant. To determine these differences is the next step in the analysis.

Least Significant Difference. One procedure for determining which wine-sample means are significantly different, following the demonstration of a significant F-value, is to calculate the *least significant difference* (*LSD*), which is the smallest difference that could exist between two significantly different sample means:

$$LSD = t_\alpha \sqrt{2v/n} \qquad [df = k(n-1)] \qquad (21)$$

where t_α is the t-value, with $k(n-1)$ degrees of freedom, at the significance level α, v is the error variance, and n is the number of scores on which each mean is based. For the difference between two means to be significant at the level of significance selected, the observed difference must exceed the *LSD*-value.

Example 14. Given 5 scores for each of 4 wines, as shown in Table 6, analyze the results for significance.

$$C = (142)^2/20 = 1008.2$$

$$\text{Total } SS = (10)^2 + (8)^2 + \cdots + (6)^2 - C$$

$$= 1066 - 1008.2 = 57.8 \qquad (19\ df)$$

$$\text{Wine } SS = \frac{(42)^2 + (43)^2 + (31)^2 + (26)^2}{5} - C$$

$$= 5250/5 - 1008.2 = 41.8 \qquad (3\ df)$$

$$\text{Error } SS = 57.8 - 41.8 = 16.0 \qquad (16\ df)$$

It is customary to combine these results into a so-called *analysis of variance table*, as shown in Table 7, where $ms = SS/df$ is the mean square (the error *ms* is also denoted by v, as we have seen above).

Table 6. Five scores for each of 4 wines (see Example 14).

	Wine				
	S_1	S_2	S_3	S_4	
	10	9	7	6	
	8	9	5	5	
	7	8	6	4	
	9	10	7	5	
	8	7	6	6	
Total	42	43	31	26	G = 142
Mean	8.4	8.6	6.2	5.2	

Table 7. Analysis of variance table for the data in Example 14.

Source	SS	df	ms	F	$F_{.05}$	$F_{.01}$	$F_{.001}$
Total	57.8	19					
Wines	41.8	3	13.9	13.9***	3.24	5.29	9.00
Error	16.0	16	1.0				

Since the calculated F-value is larger than any of the three tabular values from Appendixes F, significant differences among the means of the wine scores are indicated at all three levels. The level of significance of a calculated F-value is often denoted by one or more asterisks: one for the 5% level, two for the 1% level, and three for the 0.1% level. In this example the significance of the calculated F-value is denoted by 13.9***. Significance at any given level obviously implies significance at all lower levels.

For the 1% level we use the t-value from Appendix E to calculate the LSD by Equation 21:

$$LSD = t_{.01}(16\ df)\sqrt{2(1.0)/5} = 2.921\sqrt{0.40} = 1.85$$

Significance is usually shown by ranking the mean scores and underlining those that are *not* significantly different. The difference between *any two* scores that are not connected by an underline is therefore significant. For the mean scores in the present example we would write

	Wine			
	S_2	S_1	S_3	S_4
Mean	8.6	8.4	6.2	5.2

Thus there is no significant difference between wines S_1 and S_2 because the difference between their mean scores, 0.2, is less than 1.85, the calculated LSD. However, each of these wines is significantly better than wines S_3 and S_4. Wines S_3 and S_4 are not significantly different from each other.

Duncan's New Multiple-Range Test. Some experimenters prefer one of the newer tests for establishing significance among the sample means. These tests do not require the preliminary F-test but are applied directly to the mean scores. One such test is *Duncan's new multiple-range test,* in which, after ranking, each sample mean is compared with every other sample mean, using a set of significant differences that depend upon, and increase with, the increase in the range between the ranked means. The smallest value is obtained for adjacent means, and the largest value for the extremes. In Duncan's test the shortest significant range R_p for comparing the largest and smallest of p mean scores, after they have been ranked, is given by

$$R_p = Q_p\sqrt{v/n} \qquad [df = k(n-1)] \tag{22}$$

where the number of degrees of freedom is that for the error variance v. The appropriate value of Q_p can be obtained from Appendixes G-1, G-2, or G-3.

Example 15. Use Duncan's new multiple-range test to establish significance for the data in Example 14.

For the 1% level, $\sqrt{v/n} = \sqrt{1.0/5} = \sqrt{0.2} = 0.447$, and the values of Q_p for $p = 2, 3,$ and 4 are obtained from Appendix G-2. The results are summarized in Table 8. We see that the R_p-values are appropriate for making the following comparisons:

$R_2 = 1.85$ S_2 with S_1, S_1 with S_3, and S_3 with S_4
$R_3 = 1.93$ S_2 with S_3, and S_1 with S_4
$R_4 = 1.98$ S_2 with S_4

Table 8. Duncan's new multiple-range test (1% level) for the data in Example 14 (see Example 15).

Shortest significant range				Comparison				
p	2	3	4					
Q_p	4.13	4.31	4.42	Wine	S_2	S_1	S_3	S_4
R_p	1.85	1.93	1.98	Mean	8.6	8.4	6.2	5.2

The results are the same as those obtained in Example 14. There is no significant difference between wines S_1 and S_2, but each of these wines is significantly better than wines S_3 and S_4. Wines S_3 and S_4 are not significantly different from each other.

If the mean scores of the wines are based on different numbers of individual scores, that is, n_1 scores for the first wine, n_2 scores for the second wine, . . . , n_k scores for the k-th wine, the analysis is very similar but the following modifications must be made:

1. Sample $SS = W_1^2/n_1 + W_2^2/n_2 + \cdots + W_k^2/n_k - C$
2. Effective number of replications n_{eff} replaces n:

$$n_{\text{eff}} = \left(\frac{1}{k-1}\right)\left(\frac{\sum n_j - \sum n_j^2}{\sum n_j}\right)$$

where $\sum n_j$ is the total number of wine samples in the experiment.

3. $LSD = t_\alpha\sqrt{2v/n_{\text{eff}}}$ and $R_p = Q_p\sqrt{v/n_{\text{eff}}}$ where t_α and Q_p are based on $\sum n_j - k$ degrees of freedom.

Table 9. Two-way analysis of variance (randomized complete-block design).

Judge	Wine 1	2	3	$\cdots$	k	Total
1	X_{11}	X_{21}	X_{31}	$\cdots$	X_{k1}	T_1
2	X_{12}	X_{22}	X_{32}	$\cdots$	X_{k2}	T_2
3	X_{13}	X_{23}	X_{33}	$\cdots$	X_{k3}	T_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	X_{ij}	$\vdots$	$\vdots$
n	X_{1n}	X_{2n}	X_{3n}	$\cdots$	X_{kn}	T_n
Total	W_1	W_2	W_3	$\cdots$	W_k	$G = \Sigma T_i = \Sigma W_i$
Mean	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\cdots$	$\bar{X}_k$	Total no. of scores $= kn$

Scoring of Several Wines by Several Judges. In the customary sensory evaluation in which a panel of n judges scores each of k wines, the so-called *two-way*, or double-classification, analysis of variance is appropriate in testing for significance. In this classification the total sum of squares, calculated as the variation among all scores, is subdivided into three parts: a sum of squares based on the variation among wines, a sum of squares based on the variation among judges, and a remainder sum of squares. The latter is not the result of variation among wines or judges but is a measure of the unexplained variation, or error variation. The degrees of freedom are subdivided in the same way. This is known as a *randomized complete-block design*; its pattern is shown in Table 9. The definitions are as follows (compare them with Equations 14–19):

		df
(a)	$C = G^2/kn$	
(b)	Total $SS = \sum X_{ij}^2 - C$	$kn - 1$
(c)	Wine $SS = \sum W_i^2/n - C$	$k - 1$
(d)	Judge $SS = \sum T_j^2/k - C$	$n - 1$
(e)	Error $SS =$ (b) $-$ (c) $-$ (d)	$(kn-1) - (k-1) - (n-1) = (k-1)(n-1)$

From these sums of squares and the corresponding numbers of degrees of freedom, three independent estimates of the population variance are computed. On the assumption that the groups making up the total set of measurements (scores) are random samples from populations with the same means, the three estimates of the population variance can be expected to differ only within the limits of chance fluctuation. There are two null hypotheses here, namely, that the population means for the wines are all the same and that those for the judges are all the same. These hypotheses are tested by comparing the among-wine and the among-judge variance, respectively, with the error variance. The comparisons consist of calculating the variance ratios

$$F = \frac{\text{variance for wines}}{\text{error variance}} \quad \text{and} \quad F = \frac{\text{variance for judges}}{\text{error variance}} \tag{23}$$

To establish significance, as before, the calculated values of F are compared with the tabular values at the three levels of significance.

Example 16. Five judges score 4 wines on a 20-point scale, as shown in Table 10. Are there significant differences among the sample means at the 1% level?

Substituting the data into the equations given above, we obtain

$$C = (267)^2/20 = 3564.45$$

$$\text{Total } SS = (13)^2 + \cdots + (12)^2 - C = 142.55 \qquad (19\ df)$$

$$\text{Wine } SS = \frac{(67)^2 + \cdots + (52)^2}{5} - C = 112.95 \qquad (4\ df)$$

$$\text{Judge } SS = \frac{(56)^2 + \cdots + (56)^2}{4} - C = 8.80 \qquad (3\ df)$$

$$\text{Error } SS = 142.55 - 112.95 - 8.80 = 20.80$$

$$(19 - 4 - 3 = 12\ df)$$

These results and the remaining calculations are shown in Table 11.

Since the calculated F-value for wines is greater than the tabular value, significant differences among the means of the

Table 10. Five judges score 4 wines on a 20-point scale (see Example 16).

	Wine				
Judge	S_1	S_2	S_3	S_4	Total
1	13	18	15	10	56
2	15	16	12	11	54
3	14	15	11	9	49
4	12	17	13	10	52
5	13	19	12	12	56
Total	67	85	63	52	267 = G
Mean	13.4	17.0	12.6	10.4	

Table 11. Analysis of variance table for the data in Example 16.

Source	SS	df	ms	F	$F_{.01}$	$F_{.001}$
Total	142.55	19				
Wines	112.95	4	28.24	16.32***	5.41	9.63
Judges	8.80	3	2.93	1.69	5.95	
Error	20.80	12	1.73			

wine scores do exist at the 1% level. (In fact, they exist at the 0.1% level, as implied by the three asterisks on the calculated F-value.) The calculated F-value for judges is less than the tabular value, so there are no significant differences among the judges, i.e., they have been consistent in their scoring.

Specific differences among the wines can be tested by calculating the least significant difference:

$$LSD = t_\alpha\sqrt{2v/n} = t_{.01}(12\ df)\sqrt{2(1.73)/5} = 3.055\sqrt{0.692}$$
$$= 2.54$$

Therefore 2.54 is the smallest difference that can exist between two significantly different sample means. Again using the method of underlining mean scores that are not significantly different, we write

	Wine			
	S_2	S_1	S_3	S_4
Mean	17.0	13.4	12.6	10.4

We see that wine S_2 is significantly better than wines S_1, S_3, and S_4. Wine S_1 is significantly better than wine S_4. Wines S_1 and S_3 are not significantly different, and wines S_3 and S_4 are not significantly different.

Hedonic Rating. Hedonic rating of wines is usually done with a scale of 5, 7, or 9 points. The usual 9-point scale comprises the following categories: *like extremely* (4); *like very much* (3); *like* *moderately* (2); *like slightly* (1); *neither like nor dislike* (0); *dislike slightly* (−1); *dislike moderately* (−2); *dislike very much* (−3); *dislike extremely* (−4). (See also Figure 10.) To analyze the results the numerical values shown in parentheses are used and the analysis of variance is applied. Any set of consecutive integers could be used instead of these numbers, but those used here result in the smallest intermediate values.

Example 17. Fifty judges rate 4 wines on a 7-point hedonic scale, as shown in Table 12. Are there significant differences in the judges' preference among the wines?

$$C = (227)^2/200 = 257.64$$

$$\text{Total } SS = 729 - 257.64 = 471.36 \qquad (199\ df)$$

$$\text{Wine } SS = \frac{(109)^2 + (89)^2 + (28)^2 + (1)^2}{50} - C$$

$$= 411.74 - 257.64 = 154.10 \qquad (3\ df)$$

$$\text{Error } SS = 471.36 - 154.10 = 317.26 \qquad (196\ df)$$

Table 12. Fifty judges assign hedonic ratings to 4 wines (see Example 17).

			Frequency of response, f						
			Wine						
Rating		X	S_1	S_2	S_3	S_4	Σf	$(\Sigma f)X$	$(\Sigma f)X^2$
Like very much		3	22	8	2	5	37	111	333
Like moderately		2	17	25	13	8	63	126	252
Like slightly		1	10	15	18	3	46	46	46
Neither like nor dislike		0	0	2	5	10	17	0	0
Dislike slightly		−1	1	0	4	15	20	−20	20
Dislike moderately		−2	0	0	6	9	15	−30	60
Dislike very much		−3	0	0	2	0	2	−6	18
Total	Σf		50	50	50	50	200		
	ΣfX		109	89	28	1		227 = G	
	ΣfX^2								729
Mean	$\Sigma fX/\Sigma f$		2.18	1.78	0.56	0.02			

Table 13. Analysis of variance table for the data in Example 17.

Source	SS	df	ms	F	$F_{.05}$	$F_{.01}$	$F_{.001}$
Total	471.36	199					
Wines	154.10	3	51.4	31.7***	2.60	3.78	5.42
Error	317.26	196	1.62				

Table 14. Duncan's new multiple-range test (0.1% level) for the data in Example 17.

Shortest significant range				Comparison				
p	2	3	4					
Q_p	4.65	4.80	4.90	Wine	S_1	S_2	S_3	S_4
R_p	0.837	0.864	0.882	Mean	2.18	1.78	0.56	0.02

These results and the remaining calculations are shown in Table 13. [Since F-values for 196 degrees of freedom (denominator) are not given in Appendixes F, the values for $df = \infty$ are used.]

Since $F = 31.7$ (calculated) exceeds $F_{.001} = 5.42$ (tabular), very highly significant differences among the mean scores of the wines are indicated. If Duncan's new multiple-range test is applied, we have

$$R_p = Q_p\sqrt{v/n} = Q_p\sqrt{1.62/50} = Q_p(0.18)$$

The results are summarized in Table 14. (Again the numbers for $df = \infty$ are used.)

In this example we see that wines S_1 and S_2 are significantly better than wines S_3 and S_4. Wine S_1 is not significantly different from wine S_2, and wine S_3 is not significantly different from wine S_4.

Interaction. The term *interaction* is used in statistics to describe a differential response to two variables, usually referred to as *factors*, which may or may not act independently of each other. In the analysis of variance, interaction is expressed by a so-called residual term, which provides another estimate of variance. It reflects the relations between experimental factors or the failure of one factor to vary in accord with variations in the second factor. For example, judges differ in their susceptibility to physical and mental fatigue and in their reactions to the foods they consume. Such differences can lead to interaction effects when the same judges evaluate the same wines at two different times. (Time is always one of the factors in interaction effects in wine evaluation.)

Some possible situations are shown in Figure 17, which relates the scoring of two wines by two judges to the time of day. If the lines joining the morning and afternoon scores for each judge are parallel, there is no interaction. The greater the departure from parallelism, the greater the interaction, owing to the differential

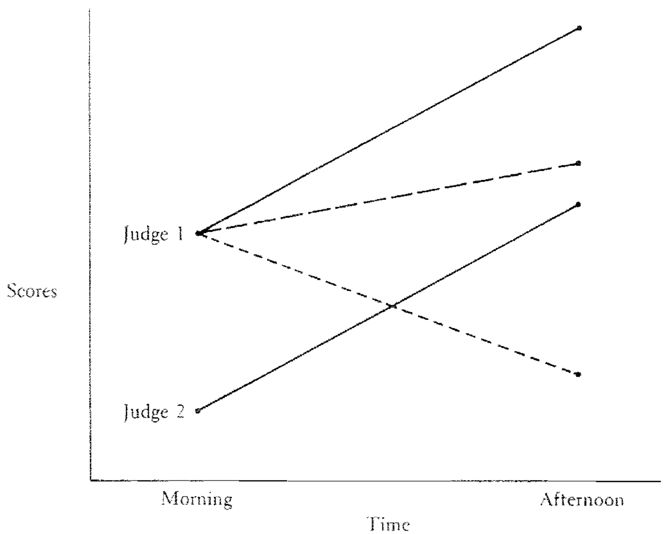


FIGURE 17
Changes in scores with time. The two solid lines show no interaction between the judges' scores and time. The lower solid line and the two dashed lines show different degrees of interaction.

response of the judges to the factors time and, say, fatigue. Small departures from parallelism may be caused by variation in, or treatment of, wine samples or as a result of random sampling errors. The problem is to test statistically whether an observed departure from parallelism is greater than could reasonably be expected to occur by chance alone.

The significance of an interaction is determined by comparing its estimate of variance with that of experimental error. A significant interaction is one that is too large to be explained on the basis of chance alone, under the null hypothesis of no interaction. A nonsignificant interaction leads to the conclusion that the factors in question act independently of each other. The existence or nonexistence of interactions can only be determined when scores are replicated.

Example 18. Five judges score 4 wines on two successive days, called time I and time II. The results are shown in Table 15. Analyze the results for significance, to determine whether there is interaction.

For the 40 individual scores we have

$$C = (310)^2/40 = 2402.5$$

$$\text{Total } SS = (10)^2 + (9)^2 + \cdots + (5)^2 - C$$

$$= 2504 - 2402.5 = 101.5 \qquad (39\ df)$$

Table 15. Five judges score 4 wines on two successive days (see Example 18).

Time I						Time II					
	Wine						Wine				
Judge	S_1	S_2	S_3	S_4	Total	Judge	S_1	S_2	S_3	S_4	Total
1	10	10	8	6	34	1	8	9	6	7	30
2	9	9	6	8	32	2	7	8	6	6	27
3	10	10	9	8	37	3	9	8	7	9	33
4	8	8	8	5	29	4	10	9	8	5	32
5	8	7	6	4	25	5	9	10	7	5	31
Total	45	44	37	31	157	Total	43	44	34	32	153

If the individual scores for the two times are added, as shown in Table 16, the result is a classification of wines and judges called a *two-way pattern*. Since the entries in the table are the totals of two scores, the denominators of the equations for the sums of squares are twice as great as in the usual analysis, and the means are obtained by dividing the totals by 10 (5 judges × 2 times). The correction term remains the same because it always pertains to the same totals. The total sum of squares for this pattern is called a *subtotal sum of squares* to distinguish it from the total sum of squares for the independent scores. The calculations follow.

$$\text{Subtotal } SS = \frac{(18)^2 + (16)^2 + \cdots + (9)^2}{2} - C$$

$$= 2481 - 2402.5 = 78.5 \qquad (19\ df)$$

$$\text{Wine } SS = \frac{(88)^2 + (88)^2 + (71)^2 + (63)^2}{10} - C$$

$$= 2449.8 - 2402.5 = 47.3 \qquad (3\ df)$$

$$\text{Judge } SS = \frac{(64)^2 + (59)^2 + \cdots + (56)^2}{8} - C$$

$$= 2416.75 - 2402.5 = 14.25 \qquad (4\ df)$$

Table 16. Combined (two-way) scores for times for the data in Table 15.

Wines × Judges (disregard times)					
	Wine				
Judge	S_1	S_2	S_3	S_4	Total
1	18	19	14	13	64
2	16	17	12	14	59
3	19	18	16	17	70
4	18	17	16	10	61
5	17	17	13	9	56
Total	88	88	71	63	310 = G
Mean	8.80	8.80	7.10	6.30	

$$\text{Interaction } SS = 78.5 - 47.3 - 14.25 = 16.95$$

$$(\text{Wine} \times \text{Judge}) \qquad (19 - 3 - 4 = 12\ df)$$

The next step in the analysis is to combine the total scores for the 5 judges, which results in a two-way pattern of wines and times, as shown in Table 17. Since the entries in the table are the totals of 5 individual scores, the denominators of the equations are 5 times as great as in the usual analysis. The calculations follow.

$$\text{Subtotal } SS = \frac{(45)^2 + (43)^2 + \cdots + (32)^2}{5} - C$$

$$= 2451.2 - 2402.5 = 48.7 \qquad (7\ df)$$

$$\text{Wine } SS = 47.3 \quad (\text{from preceding pattern}) \qquad (3\ df)$$

$$\text{Time } SS = \frac{(157)^2 + (153)^2}{20} - C$$

$$= 2402.9 - 2402.5 = 0.4 \qquad (1\ df)$$

$$\text{Interaction } SS = 48.7 - 47.3 - 0.4 = 1.0$$

$$(\text{Wine} \times \text{Time}) \qquad (7 - 3 - 1 = 3\ df)$$

Next the total scores for the 4 wines are combined to give a two-way pattern of judges and times, as shown in Table 18. Since the entries in the table are the totals of 4 individual scores, the denominators of the equations are 4 times as great as in the usual analysis. The calculations follow.

Table 17. Combined (two-way) scores for judges for the data in Table 15.

Wines × Times (disregard judges)					
	Wine				
Time	S_1	S_2	S_3	S_4	Total
I	45	44	37	31	157
II	43	44	34	32	153
Total	88	88	71	63	310 = C
Mean	8.80	8.80	7.10	6.30	

Table 18. Combined (two-way) scores for wines for the data in Table 15.

Judges × Times (disregard wines)						
	Judge					
Time	1	2	3	4	5	Total
I	34	32	37	29	25	157
II	30	27	33	32	31	153
Total	64	59	70	61	56	310 = G
Mean	8.00	7.38	8.75	7.62	7.00	

$$\text{Subtotal } SS = \frac{(34)^2 + (30)^2 + \cdots + (31)^2}{4} - C$$

$$= 2429.5 - 2402.5 = 27.0 \qquad (9\ df)$$

$$\text{Judge } SS = 14.25 \qquad (4\ df)$$

$$\text{Time } SS = 0.4 \qquad (1\ df)$$

$$\text{Interaction } SS = 27.0 - 14.25 - 0.4 = 12.35$$

$$(\text{Judge} \times \text{Time}) \qquad (9 - 4 - 1 = 4\ df)$$

These results and the remaining calculations are shown in Table 19. (Recall the meaning of the asterisks on the calculated F-values, mentioned in Example 14.)

Table 19. Analysis of variance table for the data in Example 18.

Source	SS	df	ms	F	$F_{.05}$	$F_{.01}$	$F_{.001}$
Total	101.50	39					
Wines	47.30	3	15.77	20.48***	3.49	5.95	10.80
Judges	14.25	4	3.56	4.62*	3.26	5.41	
Times	0.40	1	0.40				
Interactions							
W × J	16.95	12	1.41	1.83	2.69		
W × T	1.00	3	0.33				
J × T	12.35	4	3.09	4.01*	3.26	5.41	
Error	9.25	12	0.77				

We see that the wines are significantly different at all three levels, and that the values for the judges and the judge × time interaction are significant at the 5% level. The significant interaction indicates that the judges have reacted differently at the two times, as can be seen from their total scores at the two times. The total scores for the first three judges are less at time II than at time I, but the last two judges have total scores greater at time II than at time I. This might mean that we are dealing with two different types of judges. It could be the result of different foods consumed on the two days, varying mental or physical conditions, temperature differences, or other causes.

The least significant differences can now be used to make specific comparisons of the mean scores for the wines and for the judges.

$$\text{Wines: } LSD = t_{.001}(12\ df)\sqrt{2(0.77)/10} = 4.318\sqrt{0.154}$$
$$= 1.69$$

	Wine			
	S_1	S_2	S_3	S_4
Mean	8.80	8.80	7.10	6.30

$$\text{Judges: } LSD = t_{.05}(12\ df)\sqrt{2(0.77)/8} = 2.179\sqrt{0.192}$$
$$= 0.96$$

	Judge				
	3	1	4	2	5
Mean	8.75	8.00	7.62	7.38	7.00

Some experimenters combine the sum of squares and number of degrees of freedom for nonsignificant interactions with the sum of squares and number of degrees of freedom, respectively, for the error, and use the resulting value as a revised error term. This increases the number of degrees of freedom upon which the error is based. The results of these calculations for the data in Example 18 are shown in Table 20. The corresponding LSD values are shown below.

Table 20. Analysis of variance table for the data in Example 18, with nonsignificant interactions combined with error.

Source	SS	df	ms	F	$F_{.05}$	$F_{.01}$	$F_{.001}$
Total	101.50	39					
Wines	47.30	3	15.77	15.46***	2.96	4.60	7.27
Judges	14.25	4	3.56	3.49*	2.73	4.11	
Times	0.40	1	0.40				
J × T	12.35	4	3.09	3.06*	2.73	4.11	
Error	27.20	27	1.01				

Wines: $LSD = t_{.001}(27\ df)\sqrt{2(1.01)/10} = 3.690\sqrt{0.202} = 1.66$

Judges: $LSD = t_{.05}(27\ df)\sqrt{2(1.01)/8} = 2.052\sqrt{0.252} = 1.03$

This procedure results in slight changes in the significance patterns when the F-values are close to the borderline between significance and nonsignificance. It often yields a smaller value for the error variance, although in Example 18 it yields a larger value.

Incomplete Blocks

In wine judging, if each judge scores all the samples at the same session, the randomized complete-block design discussed previously (page 142) is appropriate. However, the judge finds it increasingly difficult to make satisfactory ratings as the number of wines presented to him at one time becomes larger. The number of samples that can be reliably scored at any one session depends upon several factors, including the type of wine being evaluated. If the judge at any one session scores only some of the wines under study, the result is an *incomplete-block design*, and the scores in question constitute an incomplete block. Sometimes the judge rates only one incomplete block and sometimes several, with intervening rest periods. Incomplete-block designs reduce the need for the judge to have long-term memory because he need be consistent in his level of judgment only within the incomplete-block limit.

An incomplete-block design in which each block contains the same number of samples, k, and in which each pair of samples appears together in the same block the same number of times, λ, is called a balanced incomplete-block design. In such designs all pairs of samples are compared with approximately the same precision.

Since only some of the wines are judged at the same time, and since each wine is compared with every other wine, only certain arrangements of blocks, samples within blocks, and replications are possible. The relevant procedures and possible incomplete-block designs for specific numbers of samples and judges can be found in Fisher and Yates (1974) and Cochran and Cox (1957).

The customary notation and method of analysis is outlined below.

t = number of samples (wines)

r = number of replications

b = number of blocks (judges)

k = number of samples per block

N = total number of scores in the design = $tr = bk$

λ = number of times each pair of samples appears in the same block = $r(k-1)/(t-1)$

W_i = total score for sample i

B_i = sum of totals for blocks in which sample i appears

$A_i = kW_i - B_i$ represents, for sample i, the sample effect adjusted for and free of the effects of the blocks in which it appears ($\sum A_i = 0$)

The calculations and the analysis of variance follow the usual patterns except for the sample sum of squares adjusted for blocks, which is defined as

$$\text{Wine } SS \text{ (adj.)} = \frac{\sum A_i^2}{kt\lambda} \qquad (24)$$

Since each A_i is free of block effects, it represents, for sample i, an estimated sample effect w_i that provides an adjustment to the general mean score, namely, an adjusted mean score for the sample. The adjusted mean for each sample is $\mu + w_i$, where $w_i = A_i/t\lambda$; $(\sum w_i = 0)$. In using the *LSD* or Duncan's new multiple-range test to compare adjusted mean scores for samples, the value of the effective error variance to be used instead of v is

$$v_{\text{eff}} = v\left[\frac{k(t-1)}{t(k-1)}\right] \qquad (25)$$

Example 19. Six wines are scored on a 10-point scale by judges in 10 blocks of 3 samples each. There are 5 scores for each wine sample, each of which is compared twice with every other sample in the same block. The pattern is shown in Table 21. Analyze the data for significance.

In this design $t = 6$, $b = 10$, $k = 3$, $r = 5$, and $\lambda = 2$. The calculations are shown below.

Table 21. Six wines scored on a 10-point scale by judges in 10 blocks of 3 samples each (incomplete-block design; see Example 19).

Block (judge)		Wine						Total
		S_1	S_2	S_3	S_4	S_5	S_6	
1		4	5			5		14
2		6	7				6	19
3		6		7	5			18
4		7		5			7	19
5		4			6	4		14
6				6		4	10	20
7			8	7	5			20
8			10	4		6		20
9			6		4		9	19
10					4	5	8	17
Total	W_i	27	36	29	24	24	40	$180 = G$
	kW_i	81	108	87	72	72	120	$\mu = 180/30$
	B_i	84	92	97	88	85	94	$= 6.00$
	A_i	−3	16	−10	−16	−13	26	
	w_i	−0.25	1.33	−0.83	−1.33	−1.08	2.17	
	$\mu + w_i$	5.75	7.33	5.17	4.67	4.92	8.17	

$$C = (180)^2/30 = 1080$$

$$\text{Total } SS = (4)^2 + (5)^2 + \cdots + (8)^2 - C$$

$$= 1168 - 1080 = 88 \qquad (29\ df)$$

$$\text{Block } SS = \frac{(14)^2 + (19)^2 + \cdots + (17)^2}{3} - C$$

$$= 1096 - 1080 = 16 \qquad (9\ df)$$

$$\text{Wine } SS \text{ (adj.)} = \frac{\sum A_i^2}{kt\lambda} = \frac{(-3)^2 + (16)^2 + \cdots + (26)^2}{3(6)(2)}$$

$$= 1466/36 = 40.72 \qquad (5\ df)$$

$$\text{Error } SS = 88 - 16 - 40.72$$

$$= 31.28 \quad \text{(intra-block error)} \qquad (15\ df)$$

These results and the remaining calculations are shown in Table 22.

The analysis indicates significant differences among the sample means at the 5% level because the calculated value $F = 3.89$ exceeds the tabular value $F_{.05} = 2.90$. If the *LSD* is used to test for specific differences among the wines, we have

$$LSD = t_{.05}(15\ df)\sqrt{\frac{2v}{r}\left[\frac{k(t-1)}{t(k-1)}\right]}$$

$$= 2.131\sqrt{\left[\frac{2(2.09)}{5}\right]\left[\frac{3(5)}{6(2)}\right]} = 2.131\sqrt{(0.836)(1.25)}$$

$$= 2.131\sqrt{1.04} = 2.17$$

	Wine					
	S_6	S_2	S_1	S_3	S_5	S_4
Mean	8.17	7.33	5.75	5.17	4.92	4.67

We see that there is no significant difference between wines S_6 and S_2. Wine S_6 is significantly better than wines S_1, S_3, S_5, and S_4. Wine S_2 is not significantly different from wines S_1 and S_3 but is significantly better than wines S_5 and S_4. There are no significant differences among wines S_1, S_3, S_5, and S_4.

Table 22. Analysis of variance table for the data in Example 19.

Source	SS	df	ms	F	$F_{.05}$	$F_{.01}$
Total	88.00	29				
Blocks	16.00	9	1.78			
Wines (adj.)	40.72	5	8.14	3.89*	2.90	4.56
Error	31.28	15	2.09			

Sometimes it is possible to have the judges score each of the wines in an incomplete-block design, scoring a part of the total number at different times. For each judge the incomplete blocks are grouped to form a replication. This design permits the removal of variations in replications from the block sum of squares. *Balanced lattices* are of this type of design. They are useful and the calculations are simple. The number of such designs is limited because the number of samples must be a perfect square, k^2, grouped in blocks of k samples with $k + 1$ replications.

Example 20. Nine wines are scored on a 10-point scale by 4 judges, each judge scoring all 9 samples in 3 incomplete blocks of 3 samples each, as shown in Table 23. Test the wine scores for significance.

In this design $k = 3$, $t = k^2 = 9$, $r = k + 1 = 4$, and $\lambda = 1$. The calculations are shown below.

$$C = (211)^2/36 = 1236.69$$

$$\text{Total } SS = (9)^2 + (3)^2 + \cdots + (3)^2 - C$$

$$= 1399 - 1236.69 = 162.31 \qquad (35\ df)$$

$$\text{Block } SS = \frac{(19)^2 + (16)^2 + \cdots + (14)^2}{3} - C$$

$$= 1255 - 1236.69 = 18.31 \qquad (11\ df)$$

$$\text{Replication } SS = \frac{(53)^2 + (50)^2 + (55)^2 + (53)^2}{9} - C$$

$$= 1238.11 - 1236.69 = 1.42 \qquad (3\ df)$$

$$\text{Block (in repl.) } SS = 18.31 - 1.42 = 16.89 \qquad (8\ df)$$

Table 23. Nine wines scored on a 10-point scale by 4 judges in blocks of 3 samples each (balanced lattice; see Example 20).

Replication (Judge)	Block	Wine S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	Total	
	1	9	3	7							19	
I	2				5	4	7				16	
	3							9	7	2	18	53
	4	8			4			8			20	
II	5		5			3			7		15	
	6			8			4			3	15	50
	7	9				4				3	16	
III	8			7	4				8		19	
	9		5				6	9			20	55
	10			7		3		9			19	
IV	11	7					7		6		20	
	12		5		6					3	14	53
Total W_i		33	18	29	19	14	24	35	28	11	$211 = G$	
kW_i		99	54	87	57	42	72	105	84	33	633	
B_i		75	68	72	69	66	71	77	72	63	633	
A_i		24	−14	15	−12	−24	1	28	12	−30	0	
w_i		2.67	−1.56	1.67	−1.33	−2.67	0.11	3.11	1.33	−3.33	0	
$\mu + w_i$		8.53	4.30	7.53	4.53	3.19	5.97	8.97	7.19	2.53	$\mu = 211/36 = 5.86$	

$$
\begin{aligned}
\text{Wine } SS \text{ (adj.)} &= \frac{\sum A_i^2}{kt\lambda} \\
&= \frac{(24)^2 + (-14)^2 + \cdots + (-30)^2}{3(9)(1)} \\
&= 131.33 \qquad (8\ df) \\
\text{Error } SS &= 162.31 - 18.31 - 131.33 \\
&= 12.67 \quad \text{(intra-block error)} \qquad (16\ df)
\end{aligned}
$$

These results and the remaining calculations are shown in Table 24.

We will use Duncan's new multiple-range test to compare the adjusted mean scores of the wines. The standard error of an adjusted mean score is

$$\frac{R_p}{Q_p} = \sqrt{\frac{v}{r}\left[\frac{k(t-1)}{t(k-1)}\right]} = \sqrt{\frac{0.79}{4}\left[\frac{3(8)}{9(2)}\right]} = \sqrt{0.26} = 0.51$$

The results are summarized in Table 25.

The incomplete-block designs that we have described involve only what is known as the *intra-block error* and are based on the assumption that the blocks are fixed. If the block effects are assumed to be random, however, more efficient estimates of the treatment means can sometimes be obtained by a procedure called *recovery of inter-block information.* This procedure is described in Cochran and Cox (1957). It is recommended only for large experiments in which the numbers of degrees of freedom for blocks and error exceed 25.

Table 24. Analysis of variance table for the data in Example 20.

Source	SS	df	ms	F	$F_{.001}$
Total	162.31	35			
Blocks	18.31	11			
Replications	1.42	3			
Blocks (in repl.)	16.89	8	2.11	2.67	
Wines (adj.)	131.33	8	16.42	20.78***	6.19
Error	12.67	16	0.79		

Table 25. Duncan's new multiple-range test (1% level) for the data in Example 20.

Shortest significant range								
p	2	3	4	5	6	7	8	9
Q_p	4.13	4.31	4.42	4.51	4.57	4.62	4.66	4.70
R_p	2.11	2.20	2.25	2.30	2.33	2.36	2.38	2.40

Comparison									
Wine	S_7	S_1	S_3	S_8	S_6	S_4	S_2	S_5	S_9
Mean	8.97	8.53	7.53	7.19	5.97	4.53	4.30	3.19	2.53

Ranking Procedures

In evaluating wines, judges may find it difficult to express preferences in terms of a quantitative measure. They usually find it much easier to rank the wines. Since ranking gives no indication of the magnitudes of the differences among the wines under study, it does not supply as much information as scoring. On the other hand, it not only simplifies the procedure for the judging panel, but also often represents as satisfactory a method of detecting the differences as is required.

Pairs of Ranks. When only two wines are being compared, pairs of ranks are obtained. One test that is then used is based on the signs of the differences between the paired values. The procedure is identical to that used in preference testing of paired samples. The null hypothesis of equal numbers of positive and negative differences (H_0: $p = 0.5$) is tested approximately by calculating

$$\chi^2 = \frac{(|n_1 - n_2| - 1)^2}{n_1 + n_2} \tag{26}$$

where n_1 and n_2 are the numbers of positive and negative differences, respectively, $|n_1 - n_2|$ represents the numerical (nonnegative) value of the difference between them, and χ^2 is based on one degree of freedom.

Example 21. Two wines, S_1 and S_2, are ranked 15 times, as shown below. Is there a significant difference between them?

S_1	1	1	2	1	2	1	1	2	1	1	1	2	1	1	1
S_2	2	2	1	2	1	1	1	1	2	2	2	1	2	2	2
Sign	+	+	−	+	−	0	0	−	+	+	+	−	+	+	+

The + sign means that wine S_1 was ranked above wine S_2 and the − sign means that wine S_2 was ranked above wine S_1. Ties (denoted by 0) are disregarded in the analysis. The + sign appears 9 times and the − sign 4 times. Therefore

$$\chi^2 = \frac{(|9 - 4| - 1)^2}{13} = 16/13 = 1.23$$

Appendix B shows that $\chi^2_{.05}(1\ df) = 3.84$, which is larger than the calculated value. There is therefore no reason to reject the null hypothesis, and no significant difference between the two wines is indicated.

The advantages of this test are simplicity, no requirement of equal variances, and relative insensitivity to recording errors. The disadvantage, however, is that it disregards the magnitude of the difference, if any, between the wines. This problem is inherent in ranking procedures.

Ranking of Several Wines by Two Judges. To determine whether two judges are significantly different in their rankings of several wines, *Spearman's rank correlation coefficient* can be used to test the agreement between the rankings. This correlation coefficient is defined as

$$R = 1 - \frac{6\sum d^2}{k(k^2 - 1)} \tag{27}$$

where $\sum d^2$ is the sum of the squares of the differences between the rank values given by the two judges to each of k wine samples. (If any wines in one ranking are tied, each is assigned the mean of the rank values they would otherwise have had.) The value of R can vary from −1 (totally opposite rankings by the two judges) to +1 (perfect agreement between the judges). The intermediate value $R = 0$ indicates that the two rankings are totally unrelated, i.e., they are the result of chance alone. This, in fact, is the null hypothesis, which can be written $H_0: \rho = 0$, where ρ is the *population rank correlation*.

Little reliability can be placed on a value of R obtained from the rankings of fewer than 10 samples. The significance of a calculated value of R can be determined by comparing the value of

$$t = R\sqrt{\frac{k - 2}{1 - R^2}} \tag{28}$$

with the appropriate t-value, based on $k - 2$ degrees of freedom, in Appendix E. For significance the calculated t-value must exceed the tabular value. A significant positive t-value indicates that the judges agree in their rankings. The significance of calculated R-values can also be determined by the use of Appendix H. Calculated values that exceed those in the table are significantly different from zero and indicate agreement in the rankings.

Example 22. Two judges rank 10 wines, as shown below. Is there a significant difference in their rankings?

	Wine									
Judge	1	2	3	4	5	6	7	8	9	10
J_1	2	1	10	7	8	6	3	4	5	9
J_2	3	1	8	9	10	7	4	2	5	6
Difference										
d	−1	0	2	−2	−2	−1	−1	2	0	3
d^2	1	0	4	4	4	1	1	4	0	9
										$\sum d^2 = 28$

The null hypothesis (H_0: $\rho = 0$) is that there is no correlation between the rankings. Solving Equations 27 and 28, we obtain

$$R = 1 - \frac{6(28)}{10(99)} = 0.830$$

$$t = 0.830\sqrt{\frac{8}{1 - 0.689}} = 4.21$$

From Appendix E we see that $t_{.01}(8\ df) = 3.355$. Since the calculated value $t = 4.21$ exceeds the tabular value, we reject (at the 1% level) the null hypothesis and conclude that the value $R = 0.830$ is highly significantly different from 0. The agreement between the rankings of the two judges is therefore highly significant. If we use Appendix H (recalling that $df = 10 - 2 = 8$) we see that any value of R greater than 0.7646 is significant at the 1% level. Therefore $R = 0.830$ is highly significant. Using Appendix H eliminates the need to calculate t.

This procedure can also be applied in the evaluation of judging ability. Adding increasing amounts of some constituent to a wine provides a set of samples of known order. If a panelist is asked to rank the set for increasing amounts of the constituent, we have an accurate standard with which to compare his ranking, and Spearman's rank correlation coefficient is appropriate for rating his competence.

Ranking of Several Wines by Two or More Judges. The ranking of k wines by n judges is a very common procedure. Two methods of analyzing the data are presented here.

Method 1. A quick appraisal of possible significant differences among a set of rankings can be made by the use of Appendixes I-1 and I-2. These tables list ranges of *rank totals*, which are the sums of the n individual rank values for a given wine. Rank totals that lie *outside* the ranges shown in the tables indicate results significantly different from those that would be obtained by chance alone.

Example 23. Twelve judges rank 5 wines, yielding the following rank totals: S_1 (34), S_2 (20), S_3 (52), S_4 (26), S_5 (48). Use Appendixes I to determine whether there are significant differences among these rankings.

Appendix I-1 shows that for 12 rankings of 5 samples there are significant differences at the 5% level for rank totals not within the range 25–47. Thus we see that wine S_2 is ranked significantly low, and wines S_3 and S_5 are ranked significantly high. At the 1% level the range is 22–50, so at this level wine S_2 is ranked significantly low and wine S_3 is ranked significantly high.

For small values of k and n, there may be more significance than is indicated by the tables of rank totals. In such situations the following method of analyzing the data is more effective.

Method 2. Rankings can be replaced by a set of quantities called *normal scores*, which are listed in Appendix J. Then the usual procedures for analyzing normally distributed data are appropriate. For example, Appendix J shows that for 6 ranked wines the normal scores that replace the rank values 1, 2, 3, 4, 5, and 6 are 1.267, 0.642, 0.202, −0.202, −0.642, and −1.267, respectively. This transformation converts the ranking into a normal population, and the usual analysis of variance procedure is applied. Since the positive and negative values of the normal scores are distributed symmetrically about their mean value, 0, the total for each judge is zero and therefore the grand total, G, is also zero. This greatly simplifies the computations.

Example 24. Five judges rank 6 wines, as shown in Table 26. Use Appendix J to analyze the results for significance.

The rankings are converted to normal scores as shown in Table 27. The calculations follow.

Table 26. Six wines ranked by 5 judges (see Example 24).

	Wine					
Judge	S_1	S_2	S_3	S_4	S_5	S_6
1	6	4	2	3	5	1
2	3	6	4	1	5	2
3	1	2	5	3	6	4
4	5	6	3	1	4	2
5	6	5	4	2	3	1
Rank total	21	23	18	10	23	10

Table 27. Normal scores for the rankings in Table 26.

	Wine						
Judge	S_1	S_2	S_3	S_4	S_5	S_6	Total
1	−1.267	−0.202	0.642	0.202	−0.642	1.267	0
2	0.202	−1.267	−0.202	1.267	−0.642	0.642	0
3	1.267	0.642	−0.642	0.202	−1.267	−0.202	0
4	−0.642	−1.267	0.202	1.267	−0.202	0.642	0
5	−1.267	−0.642	−0.202	0.642	0.202	1.267	0
Total	−1.707	−2.736	−0.202	3.580	−2.551	3.616	0 = G
Mean	−0.341	−0.547	−0.040	0.716	−0.510	0.723	

$$C = 0$$

$$\text{Total } SS = 10[(1.267)^2 + (0.642)^2 + (0.202)^2] = 20.583 \qquad (29\ df)$$

$$\text{Wine } SS = [(-1.707)^2 + (-2.736)^2 + \cdots + (3.606)^2]/5 = 8.568 \qquad (5\ df)$$

$$\text{Error } SS = 20.583 - 8.568 = 12.015 \qquad (24\ df)$$

These results and the remaining calculations are shown in Table 28.

Since the calculated F value of 3.41 exceeds the tabular value of 2.62, significant differences at the 5% level are indicated, and the *LSD* can be used to determine which wines are significantly different from each other.

$$LSD = t_{.05}(24\ df)\sqrt{2(0.501)/5} = 2.064\sqrt{0.200} = 0.923$$

Table 28. Analysis of variance table for the data in Example 24.

Source	*SS*	*df*	*ms*	*F*	$F_{.05}$	$F_{.01}$
Total	20.583	29				
Wines	8.568	5	1.71	3.41*	2.62	3.90
Error	12.015	24	0.501			

Using the mean normal scores, the differences can be summarized as follows:

	Wine					
	S_6	S_4	S_3	S_1	S_5	S_2
Mean	0.723	0.716	−0.040	−0.341	−0.510	−0.547

We see that at the 5% level there are no significant differences among wines S_3, S_4, and S_6, but wines S_4 and S_6 are significantly better than wines S_1, S_2, and S_5. There are no significant differences among wines S_1, S_2, S_3, and S_5. (As in all such analyses Duncan's new multiple-range test, which does not require the calculation of F, could be used instead of the *LSD* procedure.)

The two methods that have been presented here for analyzing ranked data have the advantage over other methods that they provide ways of establishing significant differences among individual wines. Other methods merely indicate whether significant differences do or do not exist among the wines taken as a group.

Descriptive Sensory Analysis

The best-known method of descriptive sensory analysis is the *flavor profile* developed by the Arthur D. Little Company, Cambridge, Massachusetts. It has been used in product development, quality control, and laboratory research by numerous food and drug companies (Amerine et al., 1965*a*). In this method a panel of highly trained judges is used to identify the individual and overall odor and flavor characteristics of a food, in terms of the sensory impressions they create. Properly trained panels achieve considerable agreement, after group discussion, on overall sensory impressions and the intensities and order of detection of the various sensory factors. Disadvantages of the flavor profile method are the expense of training the judges, the possible bias introduced by a dominant (assertive) member of the panel during the group discussion, and the difficulty of statistical analysis of the results.

For an example of a record form for the descriptive sensory analysis of wines, see Figure 18. As in the flavor profile method, many winery staff members and private groups make their decisions on the quality of a wine after group discussion of the results obtained in the individual sensory examinations. Is group discussion beneficial or does it entail too great a risk of prejudicial influences? Meyers and Lamm (1975) have studied this problem; the answer is by no means as unequivocal as one would wish. There is first of all the danger of a dominant individual's imposing his judgment on the group, by either his reputation or force of personality. If this occurs, group discussion is useless except as an ego-cultivating exercise for the dominant individual (e.g., the winery owner). Jones (1958) and Foster et al. (1955) have noted that a group judgment is not the same as a group of judgments, because an individual can sway the group judgment. (The obvious analogy with trial juries here is inescapable.)

Even if there is no dominant individual, the group influence itself may be detrimental. As Meyers and Lamm say, "What people learn from discussion is mostly in the direction supporting the majority's initial preference." The problem is that, probably subconsciously, members of the group usually show a disproportionate interest in facts and opinions that support their initial preferences and tend to ignore those facts and opinions that do not. This appears to be true for both verbal and written opinions. If knowledge of the positions of other members of the panel has a polarizing effect (and how can it help but do so if the owner or winemaker is present?), we recommend that all the panelists withhold information on their initial preferences.

Stone et al. (1974) have introduced a quantitative method of descriptive sensory analysis. The various sensory attributes of the product are evaluated separately. For each attribute a scale of 6 inches is provided, with two labeled anchor points ½ inch from the ends of the scale and one at the center. For example, the scale for sweetness would look like this:

Weak — Moderate — Strong

	Identification	Intensity 0 to 10	Quality −5 to +5
Prior to Tasting			
1. *Visual*			
Appearance: cloudy, dull (hazy), clear, brilliant			
Color: straw yellow, greenish yellow, yellow, gold, amber; pink, violet-pink, eye-of-the-partridge (light brownish red), ruby red, violet-red, brownish red (tawny)			
Intensity: light, strong			
Gas release: none, fine bubbles, medium bubbles, large bubbles			
2. *Olfactory**			
Complex: vinous, distinct, varietal, flowery, musty(?), oxidized			
Specific: ethyl acetate, fusel oils, hydrogen sulfide, mercaptan, sulfur dioxide			
In-Mouth			
3. *Gustatory*			
Balanced: thin, full-bodied			
Specific: sweet, sour, bitter, salty			
4. *Olfactory (flavor)**			
Complex: earthy, fruity, herbaceous, woody			
Specific (identify)			
5. *Texture*: astringent, burning, prickly, foreign*			

*The chemical origin of the sensory impression should be specified if possible

Name ______________________ Date ____________

FIGURE 18
Record form for descriptive sensory analysis of wines. (Adapted from J. Puisais et al., 1974.)

After tasting the product the judge marks a cross at the point representing the magnitude of the sensation in question. The distance from the end of the scale to the cross is a measure of this magnitude. Stone et al. believe that the scale is linear, i.e., that with several data points a straight-line plot of measured distance versus true sweetness (or other sensory attribute) is obtained.

The procedure requires extensive training with the product (about 20 hours) and individual testing. The individual and panel data are evaluated by analysis of variance. Correlation coefficients are calculated to determine the degree of correlation between the scales. Primary sensory values are measured by principal component analysis, factor analysis, etc. Finally, a multidimensional model can be developed and its relation to consumer response or other external factors can be established.

From the data one should be able to identify inconsistent responses (indicating the need for more training) and the adequacy of the judge's discrimination between different levels of a given sensory attribute. One can also determine whether individual scales are producing consistent results and whether the scales are adequately discriminating between products. Finally, the extent to which products differ in the specific attributes can be measured, and the most accurate and consistent judges can be identified.

Computer programs for one-way and two-way analyses of variance are used to measure the agreement between a judge and the panel as a whole. The interaction sum of squares is estimated for each judge and the F-value is calculated. A high F-value for an individual judge indicates his disagreement with the panel, i.e., there is interaction between the product and the judge.

Our conclusion is that descriptive sensory analysis, in the hands of highly trained personnel, should prove useful in solving certain industrial and research sensory evaluation problems.

Some Suggested Exercises

The serious amateur wine judge usually wishes to improve his judging ability, but how does he go about it? Obviously he prac-

tices. His main problem is finding a fixed frame of reference for each of the major odor and taste components of wines. What, for example, is low or high sourness? How does a low concentration of acetaldehyde smell compared with a high concentration? Can one distinguish low, moderate, and high concentrations of sulfur dioxide in wines?

The following exercises are intended to help answer these and similar questions. They should also prove useful in selecting the best judges for many sensory evaluation panels. However, there is certainly no direct relation between one's inherent taste or odor sensitivity and one's ability to evaluate wine quality. For each specific sensory characteristic, one must also know the level of intensity that is appropriate in the wine in question, and one must be able to recognize the proper balance among the various sensory characteristics. Experience is what really counts.

Obviously most people do not have a supply of citric acid or glycerol or ethyl acetate, nor do they have the equipment for measuring or weighing such chemicals. We suggest that you solicit the interest and help of an enologically-inclined chemist or pharmacist. They do have the necessary chemicals and equipment, or can get them without difficulty. (See also Marcus, 1974.)

Thresholds. A suggested series of concentrations for testing sensitivity to sucrose (sweetness) in aqueous solution is 0.1, 0.3, 0.7, and 1.2% by weight. The "A-not-A" type of test may be used, although other methods work equally well. In this test a water blank (the standard) is tasted first. Then one of the sucrose solutions (in a random order) or another blank is tasted. The judge decides whether the sample presented is the same as or different from the standard. (For a record form see Figure 19.) The test is repeated 6 times for each concentration, including the blank (30 times in all). Typical results for such a test might be the following:

	Sample				
	Blank	0.1%	0.3%	0.7%	1.2%
Correct decision	3	3	4	5	6
% Correct	50	50	66.7	83.3	100

Nature of difference: ____________________

Taste (or smell) the standard (S) and the sample. Decide whether the sample is the same as or different from the standard.

Sample no.	Same as S	Different from S

Name ____________________ Date ____________

FIGURE 19

Record form for an A-not-A test.

What is this judge's threshold for sucrose in water? Obviously 50% of his decisions should be correct by chance alone. The percentage of correct decisions *above* chance is defined as $P_c = 2 \times (P_o - 50)$, where P_o is the percentage of correct decisions observed. In practice the threshold is usually taken to be that concentration at which the judge makes 50% correct decisions above chance ($P_c = 50$), i.e., 75% correct decisions observed ($P_o = 75$), since $2(75 - 50) = 50$. In the present example the sucrose threshold is therefore somewhere between 0.3 and 0.7%. A more exact threshold could be established by repeating the test with solutions between 0.3 and 0.7% sucrose, e.g., 0.3, 0.35, 0.43, 0.53, and 0.68%.

The results can be plotted on log-probability paper, with P_c on the probability axis (ordinate) and concentration on the log axis (abscissa). Draw a straight line as close to the data points as possible. The intersection of this line with a horizontal line drawn from $P_c = 50$ defines the concentration threshold. For a still more accurate value the line can be plotted by the method of least squares, either manually or with an electronic calculator or computer. For purposes of demonstration we suggest that the group results be pooled and the average threshold calculated. However, it is instructive to compare the thresholds of various members of the group. For this purpose the test should probably be repeated until there are at least 15 correct decisions for each individual.

This type of test can also be used to determine the thresholds for many other substances, in either wine or water. For example, the following amounts of various chemicals could be added to the base wine or water (the standards), which constitutes the first of five samples in the series: acetaldehyde (40, 80, 140, and 200 mg per liter); acetic acid (3, 5, 9, and 14 grams per liter); biacetyl (4, 8, 12, and 20 mg per liter); citric acid (0.2, 0.4, 0.8, and 1.6 grams per liter); ethyl acetate (30, 60, 100, and 150 mg per liter); sorbic acid (50, 100, 175, and 275 mg per liter); sulfur dioxide (40, 90, 150, and 250 mg per liter); tartaric acid (0.03, 0.07, 0.10, and 0.15 gram per liter). The sulfur dioxide test should be the last one attempted, and should be made no more than once per day.

When water is used as the standard rather than a base wine, these tests establish the absolute thresholds of the judges (see page 73). When wines are used the thresholds should be interpreted as difference thresholds (except for sorbic acid), because the concentration of the component in the base wine may already exceed that corresponding to the absolute threshold. Care should be exercised in selecting a fairly neutral wine of normal composition as the base wine. If testing time is limited one may use four concentrations instead of five (omitting the lowest).

Thresholds can also be determined by the methods of just noticeable difference (*jnd*) and just not noticeable difference (*jnnd*). In the former test the samples are presented in order of increasing concentration, from below threshold to well above threshold. The judge indicates the first sample that he finds just noticeably different (sweeter, sourer, etc.) from the preceding sample. (For a record form see Figure 20.) This test can be used for determining absolute as well as difference thresholds. Because the errors of expectation and habituation may occur, the test should be done in both directions, i.e., *jnd* and *jnnd*. In the latter test the samples are presented in order of *decreasing* concentration; the judge indicates the first sample that he finds just *not* noticeably different from the preceding sample.

For example, the test for a *jnd* is done 5 times with a series of wines containing citric acid. The base wine (nothing added) contains 0.50 gram per 100 ml; the amounts of citric acid added to

Nature of difference: ______________________________

Sample order: C H J K N R T

Taste (or smell) the samples, from the lowest concentration (left) to the highest (right). Indicate the first sample that is just noticeably different in taste (or smell) from the preceding sample

Difference first noticed in sample ______________________

Name ______________________ Date ______________

FIGURE 20

Record form for a just-noticeable-difference test.

make the remaining four samples are 0.02, 0.05, 0.10, and 0.25 gram per 100 ml, giving samples with 0.52, 0.55, 0.60, and 0.75 gram per 100 ml, respectively. In the *jnd* series the actual *jnd* is 0.55 three times and 0.60 twice; in the corresponding *jnnd* series (also done 5 times) the actual *jnnd* is 0.55 three times and 0.52 twice. The weighted means of these two sets of data are given by

$$\frac{0.55(3) + 0.60(2)}{5} = 0.57 \qquad (jnd)$$

$$\frac{0.55(3) + 0.52(2)}{5} = 0.54 \qquad (jnnd)$$

and the overall mean value is therefore 0.55. Thus this judge's difference threshold for citric acid in wine is $0.55 - 0.50 = 0.05$ gram per 100 ml. The usual measures of central tendency, significance, probable error, etc., can be applied.

Off Odors. The threshold tests for acetaldehyde, biacetyl, ethyl acetate, sorbic acid, and sulfur dioxide (listed above) can also be used for familiarizing the student with common off odors. Other off odors can be produced by adding a small amount of the substance in question to a neutral wine. For example, about 5 to 10 parts per billion of hydrogen sulfide will be detectable. For the higher alcohols, 400 mg per liter of 3-methyl-1-butanol (isoamyl alcohol) will be adequate to give a fusel oil odor to the wine. Securing wines with typical and easily detectable off odors of corkiness,

moldiness, or woodiness may be difficult. One should inquire of wine merchants or wineries for help in locating such wines.

Other Exercises. Most of the procedures discussed previously can also be used in the training and selection of judges. For detecting differences of a nonspecific character (an unidentified off odor, for example), the duo-trio test (page 113) and triangle test (page 114) are most useful. Judges who cannot distinguish the off odor can be screened out. When potential judges are being trained, those who fail to detect the odor will know that they must practice to reach the requisite proficiency, or be disqualified. The duo-trio and triangle tests can also be used in blending wines to match a standard—an important winery operation. They are useful not only in winery operations but also in the training and selection of blenders.

Paired-sample tests (page 111) can be used for establishing quality differences. However, ranking (page 129) and scoring (page 121) are often the preferred procedures. Can an individual correctly rank a series of wines in increasing order of Cabernet aroma, sweetness, sourness, ethanol content, etc.? Those who are deficient in one or more such skills need further training and practice, or should simply not be used on a sensory evaluation panel for which the skill in question is a requirement.

Because individuals differ in their understanding of the tests, some preliminary training is desirable so that all the potential judges start the test series on an approximately equal basis. In all training tests, the statistical significance of the results must be calculated unless it is obvious from inspection of the data that the results are insignificant.

Quality. For judging the quality of wines we recommend the scoring of groups of 5 to 7 wines of a closely related type, e.g., wines of the same variety but from different wineries or of different vintages, wines of a given region or district, etc. Should the wines be served "blind" or with the labels showing? For beginning students we favor the latter method because it gives the student the best chance to associate the label with the odor, taste, and flavor of the wine. However, this assumes that the students, and especially the instructor, are completely unprejudiced—a very big assumption. For more advanced students, "blind" judgings are much to be preferred. At home the wines should be served with the labels showing unless some consensus opinion is desired. In this case the wines should be served "blind." Ranking procedures are then usually preferred, but if the group has had experience in using a particular score card, scoring can be employed. (See also pages 59–62.)

When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.
—Lord Kelvin

PLATE 1
Wine glass painted black on the outside to prevent observation of appearance or color.

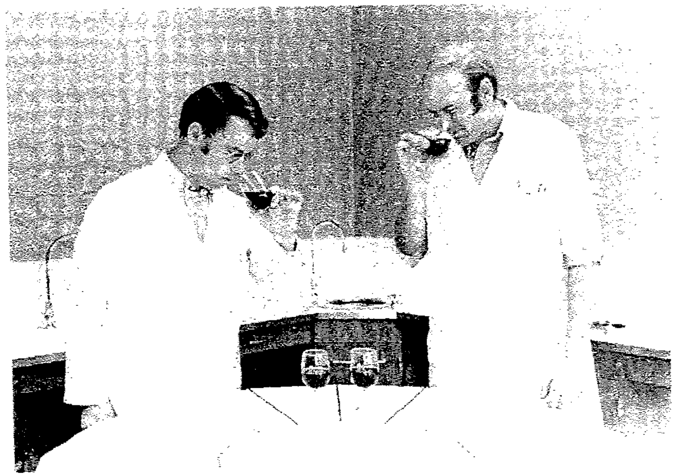
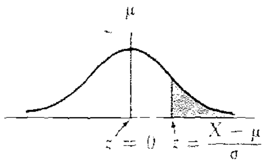


PLATE 2
Lazy susan serving table. Note sections for separating samples. (Courtesy of E. and J. Gallo, Modesto, Cal.)

Appendix A Normal Distribution

The entries in this table are the areas under the normal probability curve to the right of the marginal value of the normal deviate z (or to the left of $-z$), i.e., they are the probabilities that a random value of z will equal or exceed the marginal value.



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
3.9	.0000									

Appendix B Chi-Square Distribution

The entries in this table are the χ^2-values for distributions with from 1 to 30 degrees of freedom, at 10 values of the probability.

	Probability of a Larger Value of χ^2									
df	0.99	0.95	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.0002	0.004	0.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	0.020	0.103	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	0.115	0.35	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	0.30	0.71	3.36	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	0.55	1.14	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	0.87	1.64	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	2.17	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.73	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	3.32	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.94	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	4.58	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	5.23	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	5.89	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	6.57	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	7.26	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	7.96	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.25
17	6.41	8.67	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.79
18	7.02	9.39	17.34	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	7.63	10.12	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	10.85	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	11.59	20.34	23.86	26.17	29.62	32.67	36.34	38.93	46.80
22	9.54	12.34	21.34	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	13.09	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	13.85	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	14.61	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	15.38	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	16.15	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	16.93	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	17.71	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	14.95	18.49	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

Source: Abridged from Table IV of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 6th ed., 1974. Longman Group Ltd., London (previously published by Oliver and Boyd Ltd., Edinburgh). By permission of the authors and publisher.

Appendix C Significance in Paired-Sample and Duo-Trio Tests. H_0: $p = \frac{1}{2}$.

The number n is the number of trials, i.e., the number of judges or judgments in the test.

	Minimum correct judgments to establish significant difference (one-tailed test)			Minimum agreeing judgments necessary to establish significant preference (two-tailed test)		
n	$p = 0.05$	$p = 0.01$	$p = 0.001$	$p = 0.05$	$p = 0.01$	$p = 0.001$
7	7	7	—	7	—	—
8	7	8	—	8	8	—
9	8	9	—	8	9	—
10	9	10	10	9	10	—
11	9	10	11	10	11	11
12	10	11	12	10	11	12
13	10	12	13	11	12	13
14	11	12	13	12	13	14
15	12	13	14	12	13	14
16	12	14	15	13	14	15
17	13	14	16	13	15	16
18	13	15	16	14	15	17
19	14	15	17	15	16	17
20	15	16	18	15	17	18
21	15	17	18	16	17	19
22	16	17	19	17	18	19
23	16	18	20	17	19	20
24	17	19	20	18	19	21
25	18	19	21	18	20	21
30	20	22	24	21	23	25
35	23	25	27	24	26	28
40	26	28	31	27	29	31
45	29	31	34	30	32	34
50	32	34	37	33	35	37
60	37	40	43	39	41	44
70	43	46	49	44	47	50
80	48	51	55	50	52	56
90	54	57	61	55	58	61
100	59	63	66	61	64	67

SOURCE: Adapted from a table by E. B. Roessler, G. A. Baker, and M. A. Amerine, *Food Research* 21, 117–121 (1956).

Appendix D Significance in Triangle Tests. H_0: $p = \frac{1}{3}$.

The number n is the number of trials, i.e., the number of judges or judgments in the test.

	Minimum correct judgments to establish significant difference				Minimum correct judgments to establish significant difference		
n	$p = 0.05$	$p = 0.01$	$p = 0.001$	n	$p = 0.05$	$p = 0.01$	$p = 0.001$
5	4	5	5	56	26	28	31
6	5	6	6	57	27	29	31
7	5	6	7	58	27	29	32
8	6	7	8	59	27	30	32
9	6	7	8	60	28	30	33
10	7	8	9	61	28	30	33
11	7	8	9	62	28	31	33
12	8	9	10	63	29	31	34
13	8	9	10	64	29	32	34
14	9	10	11	65	30	32	35
15	9	10	12	66	30	32	35
16	10	11	12	67	30	33	36
17	10	11	13	68	31	33	36
18	10	12	13	69	31	34	36
19	11	12	14	70	32	34	37
20	11	13	14	71	32	34	37
21	12	13	15	72	32	35	38
22	12	14	15	73	33	35	38
23	13	14	16	74	33	36	39
24	13	14	16	75	34	36	39
25	13	15	17	76	34	36	39
26	14	15	17	77	34	37	40
27	14	16	18	78	35	37	40
28	15	16	18	79	35	38	41
29	15	17	19	80	35	38	41
30	16	17	19	81	36	38	41
31	16	18	19	82	36	39	42
32	16	18	20	83	37	39	42
33	17	19	20	84	37	40	43
34	17	19	21	85	37	40	43
35	18	19	21	86	38	40	44
36	18	20	22	87	38	41	44
37	18	20	22	88	39	41	44
38	19	21	23	89	39	42	45
39	19	21	23	90	39	42	45
40	20	22	24	91	40	42	46
41	20	22	24	92	40	43	46
42	21	22	25	93	40	43	46
43	21	23	25	94	41	44	47
44	21	23	25	95	41	44	47
45	22	24	26	96	42	44	48
46	22	24	26	97	42	45	48
47	23	25	27	98	42	45	49
48	23	25	27	99	43	46	49
49	23	25	28	100	43	46	49
50	24	26	28	200	80	84	89
51	24	26	29	300	117	122	127
52	25	27	29	400	152	158	165
53	25	27	29	500	188	194	202
54	25	27	30	1000	363	372	383
55	26	28	30	2000	709	722	737

SOURCE: Adapted from a table by E. B. Roessler, J. Warren, and J. F. Guymon, *Food Research* 13, 503–505 (1948).

Appendix E *t*-Distribution

The entries in this table are the *t*-values for distributions with from 1 to ∞ degrees of freedom, at 10 values of the two tailed probability (sum of the two tail areas) and the 10 corresponding values of the one-tailed probability (one tail area).

	Probability of a larger value of *t*, sign ignored (two-tailed test)									
df	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.002	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	318.31	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.214	12.941
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.859
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.405
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
df	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
	Probability of a larger value of *t*, sign considered (one-tailed test)									

Source: Abridged from Table III of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 6th ed., 1974 Longman Group Ltd., London (previously published by Oliver and Boyd Ltd., Edinburgh). By permission of the authors and publisher.

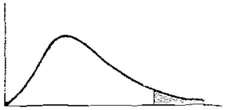
Appendix F-1 *F*-Distribution, 5% Level

0.05

The entries in this table are the *F*-values for which the tail area equals 0.05.

df for denominator	*df* for numerator: 1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

Appendix F-2 *F*-Distribution, 1% Level

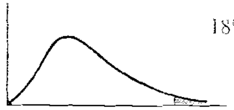


0.01

The entries in this table are the *F*-values for which the tail area equals 0.01.

df FOR DENOMINATOR	*df* FOR NUMERATOR 1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5982	6106	6234	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.46	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Appendix F-3 *F*-Distribution, 0.1% Level



0.001

The entries in this table are the *F*-values for which the tail area equals 0.001.

df FOR DENOMINATOR	*df* FOR NUMERATOR 1	2	3	4	5	6	8	12	24	∞
1	405284	500000	540379	562500	576405	585937	598144	610667	623497	636619
2	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.5	999.5
3	167.0	148.5	141.1	137.1	134.6	132.8	130.6	128.3	125.9	123.5
4	74.14	61.25	56.18	53.44	51.71	50.53	49.00	47.41	45.77	44.05
5	47.18	37.12	33.20	31.09	29.75	28.84	27.64	26.42	25.14	23.78
6	35.51	27.00	23.70	21.92	20.81	20.03	19.03	17.99	16.89	15.75
7	29.25	21.69	18.77	17.19	16.21	15.52	14.63	13.71	12.73	11.69
8	25.42	18.49	15.83	14.39	13.49	12.86	12.04	11.19	10.30	9.34
9	22.86	16.39	13.90	12.56	11.71	11.13	10.37	9.57	8.72	7.81
10	21.04	14.91	12.55	11.28	10.48	9.92	9.20	8.45	7.64	6.76
11	19.69	13.81	11.56	10.35	9.58	9.05	8.35	7.63	6.85	6.00
12	18.64	12.97	10.80	9.63	8.89	8.38	7.71	7.00	6.25	5.42
13	17.81	12.31	10.21	9.07	8.35	7.86	7.21	6.52	5.78	4.97
14	17.14	11.78	9.73	8.62	7.92	7.43	6.80	6.13	5.41	4.60
15	16.59	11.34	9.34	8.25	7.57	7.09	6.47	5.81	5.10	4.31
16	16.12	10.97	9.00	7.94	7.27	6.81	6.19	5.55	4.85	4.06
17	15.72	10.66	8.73	7.68	7.02	6.56	5.96	5.32	4.63	3.85
18	15.38	10.39	8.49	7.46	6.81	6.35	5.76	5.13	4.45	3.67
19	15.08	10.16	8.28	7.26	6.62	6.18	5.59	4.97	4.29	3.52
20	14.82	9.95	8.10	7.10	6.46	6.02	5.44	4.82	4.15	3.38
21	14.59	9.77	7.94	6.95	6.32	5.88	5.31	4.70	4.03	3.26
22	14.38	9.61	7.80	6.81	6.19	5.76	5.19	4.58	3.92	3.15
23	14.19	9.47	7.67	6.69	6.08	5.65	5.09	4.48	3.82	3.05
24	14.03	9.34	7.55	6.59	5.98	5.55	4.99	4.39	3.74	2.97
25	13.88	9.22	7.45	6.49	5.88	5.46	4.91	4.31	3.66	2.89
26	13.74	9.12	7.36	6.41	5.80	5.38	4.83	4.24	3.59	2.82
27	13.61	9.02	7.27	6.33	5.73	5.31	4.76	4.17	3.52	2.75
28	13.50	8.93	7.19	6.25	5.66	5.24	4.69	4.11	3.46	2.70
29	13.39	8.85	7.12	6.19	5.59	5.18	4.64	4.05	3.41	2.64
30	13.29	8.77	7.05	6.12	5.53	5.12	4.58	4.00	3.36	2.59
40	12.61	8.25	6.60	5.70	5.13	4.73	4.21	3.64	3.01	2.23
60	11.97	7.76	6.17	5.31	4.76	4.37	3.87	3.31	2.69	1.90
120	11.38	7.32	5.79	4.95	4.42	4.04	3.55	3.02	2.40	1.54
∞	10.83	6.91	5.42	4.62	4.10	3.74	3.27	2.74	2.13	1.00

SOURCE: Abridged from Table V of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 6th ed., 1974. Longman Group Ltd., London (previously published by Oliver and Boyd Ltd., Edinburgh). By permission of the authors and publisher.

Appendix G-1 Duncan's New Multiple Ranges, 5% Level

The entries in this table are the Q_p-values used to find R_p, the shortest significant range, at the 5% level.

df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	Number of means p within range being tested																	
2	6.085																	
3	4.501	4.516																
4	3.927	4.013	4.033															
5	3.635	3.749	3.797	3.814														
6	3.461	3.587	3.649	3.680	3.694													
7	3.344	3.477	3.548	3.588	3.611	3.622												
8	3.261	3.399	3.475	3.521	3.549	3.566	3.575											
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544										
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522									
11	3.113	3.256	3.342	3.397	3.435	3.462	3.480	3.493	3.501	3.506								
12	3.082	3.225	3.313	3.370	3.410	3.439	3.459	3.474	3.484	3.491	3.496							
13	3.055	3.200	3.289	3.348	3.389	3.419	3.442	3.458	3.470	3.478	3.484	3.488						
14	3.033	3.178	3.268	3.329	3.372	3.403	3.426	3.444	3.457	3.467	3.474	3.479	3.482					
15	3.014	3.160	3.250	3.312	3.356	3.389	3.413	3.432	3.446	3.457	3.465	3.471	3.476	3.478				
16	2.998	3.144	3.235	3.298	3.343	3.376	3.402	3.422	3.437	3.449	3.458	3.465	3.470	3.473	3.477			
17	2.984	3.130	3.222	3.285	3.331	3.366	3.392	3.412	3.429	3.441	3.451	3.459	3.465	3.469	3.473	3.475		
18	2.971	3.118	3.210	3.274	3.321	3.356	3.383	3.405	3.421	3.435	3.445	3.454	3.460	3.465	3.470	3.472	3.474	
19	2.960	3.107	2.199	3.264	3.311	3.347	3.375	3.397	3.415	3.429	3.440	3.449	3.456	3.462	3.467	3.470	3.472	3.473
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.391	3.409	3.424	3.436	3.445	3.453	3.459	3.464	3.467	3.470	3.472
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390	3.406	3.420	3.432	3.441	3.449	3.456	3.461	3.465	3.469
30	2.888	3.035	3.131	3.199	3.250	3.290	3.322	3.349	3.371	3.389	3.405	3.418	3.430	3.439	3.447	3.454	3.460	3.466
40	2.858	3.006	3.102	3.171	3.224	3.266	3.300	3.328	3.352	3.373	3.390	3.405	3.418	3.429	3.439	3.448	3.456	3.463
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333	3.355	3.374	3.391	3.406	3.419	3.431	3.442	3.451	3.460
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.287	3.314	3.337	3.359	3.377	3.394	3.409	3.423	3.435	3.446	3.457
∞	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294	3.320	3.343	3.363	3.382	3.399	3.414	3.428	3.442	2.454

(The last entry in each row remains the same for all succeeding values of p.)

Appendix G-2 Duncan's New Multiple Ranges, 1% Level

The entries in this table are the Q_p-values used to find R_p, the shortest significant range, at the 1% level.

df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	Number of means p within range being tested																	
2	14.04																	
3	8.261	8.321																
4	6.512	6.677	6.740															
5	5.702	5.893	5.989	6.040														
6	5.243	5.439	5.549	5.614	5.655													
7	4.949	5.145	5.260	5.334	5.383	5.416												
8	4.746	4.939	5.057	5.135	5.189	5.227	5.256											
9	4.596	4.787	4.906	4.986	5.043	5.086	5.118	5.142										
10	4.482	4.671	4.790	4.871	4.931	4.975	5.010	5.037	5.058									
11	4.392	4.579	4.697	4.780	4.841	4.887	4.924	4.952	4.975	4.994								
12	4.320	4.504	4.622	4.706	4.767	4.815	4.852	4.883	4.907	4.927	4.944							
13	4.260	4.442	4.560	4.644	4.706	4.755	4.793	4.824	4.850	4.872	4.889	4.904						
14	4.210	4.391	4.508	4.591	4.654	4.704	4.743	4.775	4.802	4.824	4.843	4.859	4.872					
15	4.168	4.347	4.463	4.547	4.610	4.660	4.700	4.733	4.760	4.783	4.803	4.820	4.834	4.846				
16	4.131	4.309	4.425	4.509	4.572	4.622	4.663	4.696	4.724	4.748	4.768	4.786	4.800	4.813	4.825			
17	4.099	4.275	4.391	4.475	4.539	4.589	4.630	4.664	4.693	4.717	4.738	4.756	4.771	4.785	4.797	4.807		
18	4.071	4.246	4.362	4.445	4.509	4.560	4.601	4.635	4.664	4.689	4.711	4.729	4.745	4.759	4.772	4.783	4.792	
19	4.046	4.220	4.335	4.419	4.483	4.534	4.575	4.610	4.639	4.665	4.686	4.705	4.722	4.736	4.749	4.761	4.771	4.780
20	4.024	4.197	4.312	4.395	4.459	4.510	4.552	4.587	4.617	4.642	4.664	4.684	4.701	4.716	4.729	4.741	4.751	4.761
24	3.956	4.126	4.239	4.322	4.386	4.437	4.480	4.516	4.546	4.573	4.596	4.616	4.634	4.651	4.665	4.678	4.690	4.700
30	3.889	4.056	4.168	4.250	4.314	4.366	4.409	4.445	4.477	4.504	4.528	4.550	4.569	4.586	4.601	4.615	4.628	4.640
40	3.825	3.988	4.098	4.180	4.244	4.296	4.339	4.376	4.408	4.436	4.461	4.483	4.503	4.521	4.537	4.553	4.566	4.579
60	3.762	3.922	4.031	4.111	4.174	4.226	4.270	4.307	4.340	4.368	4.394	4.417	4.438	4.456	4.474	4.490	4.504	4.518
120	3.702	3.858	3.965	4.044	4.107	4.158	4.202	4.239	4.272	4.301	4.327	4.351	4.372	4.392	4.410	4.426	4.442	4.456
∞	3.643	3.796	3.900	3.978	4.040	4.091	4.135	4.172	4.205	4.235	4.261	4.285	4.307	4.327	4.345	4.363	4.379	4.394

(The last entry in each row remains the same for all succeeding values of p.)

Appendix C-3 Duncan's New Multiple Ranges, 0.1% Level

The entries in this table are the Q_p-values used to find R_p, the shortest significant range, at the 0.1% level.

df	Number of means p within range being tested: 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
2	44.69																	
3	18.28	18.45																
4	12.18	12.52	12.67															
5	9.714	10.05	10.24	10.35														
6	8.427	8.743	8.932	9.055	9.139													
7	7.648	7.943	8.127	8.252	8.342	8.409												
8	7.130	7.407	7.584	7.708	7.799	7.869	7.924											
9	6.762	7.024	7.195	7.316	7.407	7.478	7.535	7.582										
10	6.487	6.738	6.902	7.021	7.111	7.182	7.240	7.287	7.327									
11	6.275	6.516	6.676	6.791	6.880	6.950	7.008	7.056	7.097	7.132								
12	6.106	6.340	6.494	6.607	6.695	6.765	6.822	6.870	6.911	6.947	6.978							
13	5.970	6.195	6.346	6.457	6.543	6.612	6.670	6.718	6.759	6.795	6.826	6.854						
14	5.856	6.075	6.223	6.332	6.416	6.485	6.542	6.590	6.631	6.667	6.699	6.727	6.752					
15	5.760	5.974	6.119	6.225	6.309	6.377	6.433	6.481	6.522	6.558	6.590	6.619	6.644	6.666				
16	5.678	5.888	6.030	6.135	6.217	6.284	6.340	6.388	6.429	6.465	6.497	6.525	6.551	6.574	6.595			
17	5.608	5.813	5.953	6.056	6.138	6.204	6.260	6.307	6.348	6.384	6.416	6.444	6.470	6.493	6.514	6.533		
18	5.546	5.748	5.886	5.988	6.068	6.134	6.189	6.236	6.277	6.313	6.345	6.373	6.399	6.422	6.443	6.462	6.480	
19	5.492	5.691	5.826	5.927	6.007	6.072	6.127	6.174	6.214	6.250	6.281	6.310	6.336	6.359	6.380	6.400	6.418	6.434
20	5.444	5.640	5.774	5.873	5.952	6.017	6.071	6.117	6.158	6.193	6.225	6.254	6.279	6.303	6.324	6.344	6.362	6.379
24	5.297	5.484	5.612	5.708	5.784	5.846	5.899	5.945	5.984	6.020	6.051	6.079	6.105	6.129	6.150	6.170	6.188	6.205
30	5.156	5.335	5.457	5.549	5.622	5.682	5.734	5.778	5.817	5.851	5.882	5.910	5.935	5.958	5.980	6.000	6.018	6.036
40	5.022	5.191	5.308	5.396	5.466	5.524	5.574	5.617	5.654	5.688	5.718	5.745	5.770	5.793	5.814	5.834	5.852	5.869
60	4.894	5.055	5.166	5.249	5.317	5.372	5.420	5.461	5.498	5.530	5.559	5.586	5.610	5.632	5.653	5.672	5.690	5.707
120	4.771	4.924	5.029	5.109	5.173	5.226	5.271	5.311	5.346	5.377	5.405	5.431	5.454	5.476	5.496	5.515	5.532	5.549
∞	4.654	4.798	4.898	4.974	5.034	5.085	5.128	5.166	5.199	5.229	5.256	5.280	5.303	5.324	5.343	5.361	5.378	5.394

(The last entry in each row remains the same for all succeeding values of p.)

SOURCE: Adapted from tables compiled by D. B. Duncan, *Biometrics* 11, 1–42 (1955), and modified by H. L. Harter, *ibid.*, 16, 671–685 (1960) and 17, 321–324 (1961).

Appendix H Correlation Coefficients

The entries in this table are the R-values for distributions with from 1 to 100 degrees of freedom, at 5 values of the probability.

df $(k-2)$	Probability of a larger value of R: 0.10	0.05	0.02	0.01	0.001
1	.98769	.99692	.999507	.999877	.9999988
2	.90000	.95000	.98000	.990000	.99900
3	.8054	.8783	.93433	.95873	.99116
4	.7293	.8114	.8822	.91720	.97406
5	.6694	.7545	.8329	.8745	.95074
6	.6215	.7067	.7887	.8343	.92493
7	.5822	.6664	.7498	.7977	.8982
8	.5494	.6319	.7155	.7646	.8721
9	.5214	.6021	.6851	.7348	.8471
10	.4973	.5760	.6581	.7079	.8233
11	.4762	.5529	.6339	.6835	.8010
12	.4575	.5324	.6120	.6614	.7800
13	.4409	.5139	.5923	.6411	.7603
14	.4259	.4973	.5742	.6226	.7420
15	.4124	.4821	.5577	.6055	.7246
16	.4000	.4683	.5425	.5897	.7084
17	.3887	.4555	.5285	.5751	.6932
18	.3783	.4438	.5155	.5614	.6787
19	.3687	.4329	.5034	.5487	.6652
20	.3598	.4227	.4921	.5368	.6524
25	.3233	.3809	.4451	.4869	.5974
30	.2960	.3494	.4093	.4487	.5541
35	.2746	.3246	.3810	.4182	.5189
40	.2573	.3044	.3578	.3932	.4896
45	.2428	.2875	.3384	.3721	.4648
50	.2306	.2732	.3218	.3541	.4433
60	.2108	.2500	.2948	.3248	.4078
70	.1954	.2319	.2737	.3017	.3799
80	.1829	.2172	.2565	.2830	.3568
90	.1726	.2050	.2422	.2673	.3375
100	.1638	.1946	.2301	.2540	.3211

SOURCE: Abridged from Table VII of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 6th ed., 1974. Longman Group Ltd., London (previously published by Oliver and Boyd Ltd., Edinburgh). By permission of the authors and publisher.

Appendix I-1 Rank Totals Excluded for Significant Differences (5% Level)

Any rank total outside the given range is significant

Number of judges	Number of wines: 2	3	4	5	6	7	8	9	10	11	12
3				4–14	4–17	4–20	4–23	5–25	5–28	5–31	5–34
4		5–11	5–15	6–18	6–22	7–25	7–29	8–32	8–36	8–39	9–43
5		6–14	7–18	8–22	9–26	9–31	10–35	11–39	12–43	12–48	13–52
6	7–11	8–16	9–21	10–26	11–31	12–36	13–41	14–46	15–51	17–55	18–60
7	8–13	10–18	11–24	12–30	14–35	15–41	17–46	18–52	19–58	21–63	22–69
8	9–15	11–21	13–27	15–33	17–39	18–46	20–52	22–58	24–64	25–71	27–77
9	11–16	13–23	15–30	17–37	19–44	22–50	24–57	26–64	28–71	30–78	32–85
10	12–18	14–26	17–33	20–40	22–48	25–55	27–63	30–70	32–78	35–85	37–93
11	13–20	16–28	19–36	22–44	25–52	28–60	31–68	34–76	36–85	39–93	42–101
12	15–21	18–30	21–39	25–47	28–56	31–65	34–74	38–82	41–91	44–100	47–109
13	16–23	20–32	24–41	27–51	31–60	35–69	38–79	42–88	45–98	49–107	52–117
14	17–25	22–34	26–44	30–54	34–64	38–74	42–84	46–94	50–104	54–114	57–125
15	19–26	23–37	28–47	32–58	37–68	41–79	46–89	50–100	54–111	58–122	63–132
16	20–28	25–39	30–50	35–61	40–72	45–83	49–95	54–106	59–117	63–129	68–140
17	22–29	27–41	32–53	38–64	43–76	48–88	53–100	58–112	63–124	68–136	73–148
18	23–31	29–43	34–56	40–68	46–80	52–92	57–105	61–118	68–130	73–143	79–155
19	24–33	30–46	37–58	43–71	49–84	55–97	61–110	67–123	73–136	78–150	84–163
20	26–34	32–48	39–61	45–75	52–88	58–102	65–115	71–129	77–143	83–157	90–170

Appendix I-2 Rank Totals Excluded for Significant Differences (1% Level)

Any rank total outside the given range is significant.

Number of judges	Number of wines: 2	3	4	5	6	7	8	9	10	11	12
3									4–29	4–32	4–35
4				5–19	5–23	5–27	6–30	6–34	6–38	6–42	7–45
5			6–19	7–23	7–28	8–32	8–37	9–41	9–46	10–50	10–55
6		7–17	8–22	9–27	9–33	10–38	11–43	12–48	13–53	13–59	14–64
7		8–20	10–25	11–31	12–37	13–43	14–49	15–55	16–61	17–67	18–73
8	9–15	10–22	11–29	13–35	14–42	16–48	17–55	19–61	20–68	21–75	23–81
9	10–17	12–24	13–32	15–39	17–46	19–53	21–60	22–68	24–75	26–82	27–90
10	11–19	13–27	15–35	18–42	20–50	22–58	24–66	26–74	28–82	30–90	32–98
11	12–21	15–29	17–38	20–46	22–55	25–63	27–72	30–80	32–89	34–98	37–106
12	14–22	17–31	19–41	22–50	25–59	28–68	31–77	33–87	36–96	39–105	42–114
13	15–24	18–34	21–44	25–53	28–63	31–73	34–83	37–93	40–103	43–113	46–123
14	16–26	20–36	24–46	27–57	31–67	34–78	38–88	41–98	45–109	48–120	51–131
15	18–27	22–38	26–49	30–60	34–71	37–83	41–94	45–105	49–116	53–127	56–139
16	19–29	23–41	28–52	32–64	36–76	41–87	45–99	49–111	53–123	57–135	62–146
17	20–31	25–43	30–55	35–67	39–80	44–92	49–104	53–117	58–129	62–142	67–155
18	22–32	27–45	32–58	37–71	42–84	47–97	52–110	57–123	62–136	67–149	72–162
19	23–34	29–47	34–61	40–74	45–88	50–102	56–115	61–129	67–142	72–156	77–170
20	24–36	30–50	36–64	42–78	48–92	54–106	60–120	65–135	71–149	77–163	82–178

Source. Adapted from tables compiled by A. Kramer and published in revised form in *Food Technology* 17(12), 124–125 (1963).

Appendix J Normal Scores

The entries in this table show the conversion of rankings to normal scores. Negative values are omitted for samples larger than 10.

Rank order	Size of sample: 2	3	4	5	6	7	8	9	10
1	0.564	0.864	1.029	1.163	1.267	1.352	1.424	1.485	1.539
2	−0.564	0.000	0.297	0.495	0.642	0.757	0.852	0.932	1.001
3		−0.864	−0.297	0.000	0.202	0.353	0.473	0.572	0.656
4			−1.029	−0.495	−0.202	0.000	0.153	0.275	0.376
5				−1.163	−0.642	−0.353	−0.153	0.000	0.123
6					−1.267	−0.757	−0.473	−0.275	−0.123
7						−1.352	−0.852	−0.572	−0.376
8							−1.424	−0.932	−0.656
9								−1.485	−1.001
10									−1.539

Rank order	11	12	13	14	15	16	17	18	19	20
1	1.586	1.629	1.668	1.703	1.736	1.766	1.794	1.820	1.844	1.867
2	1.062	1.116	1.164	1.208	1.248	1.285	1.319	1.350	1.380	1.408
3	0.729	0.793	0.850	0.901	0.948	0.990	1.029	1.066	1.099	1.131
4	0.462	0.537	0.603	0.662	0.715	0.763	0.807	0.848	0.886	0.921
5	0.225	0.312	0.388	0.456	0.516	0.570	0.619	0.665	0.707	0.745
6	0.000	0.103	0.191	0.267	0.335	0.396	0.451	0.502	0.548	0.590
7			0.000	0.088	0.165	0.234	0.295	0.351	0.402	0.448
8					0.000	0.077	0.146	0.208	0.264	0.315
9							0.000	0.069	0.131	0.187
10									0.000	0.062

Rank order	21	22	23	24	25	26	27	28	29	30
1	1.889	1.910	1.929	1.948	1.965	1.982	1.998	2.014	2.029	2.043
2	1.434	1.458	1.481	1.503	1.524	1.544	1.563	1.581	1.599	1.616
3	1.160	1.188	1.214	1.239	1.263	1.285	1.306	1.327	1.346	1.365
4	0.954	0.985	1.014	1.041	1.067	1.091	1.115	1.137	1.158	1.179
5	0.782	0.815	0.847	0.877	0.905	0.932	0.957	0.981	1.004	1.026
6	0.630	0.667	0.701	0.734	0.764	0.793	0.820	0.846	0.871	0.894
7	0.491	0.532	0.569	0.604	0.637	0.668	0.697	0.725	0.752	0.777
8	0.362	0.406	0.446	0.484	0.519	0.553	0.584	0.614	0.642	0.669
9	0.238	0.286	0.330	0.370	0.409	0.444	0.478	0.510	0.540	0.568
10	0.118	0.170	0.218	0.262	0.303	0.341	0.377	0.411	0.443	0.473
11	0.000	0.056	0.108	0.156	0.200	0.241	0.280	0.316	0.350	0.382
12			0.000	0.052	0.100	0.144	0.185	0.224	0.260	0.294
13					0.000	0.048	0.092	0.134	0.172	0.209
14							0.000	0.044	0.086	0.125
15									0.000	0.041

Source: Abridged from tables compiled by H. L. Harter, *Biometrika* 48, 151–165 (1961).