yield data from fertilizer experiments, *Agronomy Journal, 61,* 829–834.
[15] Lindley, D. V., and Smith, A. F. M. (1972): Bayes estimates for the linear model (with discussion), *J. Royal Statistical Society, Series B, 34,* 1–41.
[16] Mallows, C. L. (1973): Some comments on Cp, *Technometrics, 15,* 661–675.
[17] Marquardt, D. W. (1963): An algorithm for least-squares estimation of nonlinear parameters, *J. Soc. Indust. Appl. Math., 11,* 431–441.
[18] Marquardt, D. W. (1970): Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, *Technometrics, 12,* 591–612.
[19] Marquardt, D. W. (1974): Discussion of "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," by A. E. Beaton and J. W. Tukey, *Technometrics, 16,* 189–192.
[20] Mayer, L. S. and Willke, T. A. (1973): On biased estimation in linear models, *Technometrics, 15,* 497–508.
[21] McDonald, G. C. and Schwing, R. C. (1973): Instabilities of regression estimates relating air pollution to mortality, *Technometrics, 15,* 463–481.
[22] Snee, R. D. (1973): Some aspects of nonorthogonal data analysis. Part I. Developing prediction equations, *J. Qual. Technol., 5,* 67–79.
[23] Theobald, C. M. (1974): Generalizations of mean square error applied to ridge regression. *J. Royal Statistical Society, Series B, 36,* 103–106.

P-values: Interpretation and Methodology*

JEAN D. GIBBONS** AND JOHN W. PRATT***

1. *Introduction*

The most common traditional method of carrying out any hypothesis test is to select a region for rejection and form a rejection rule such that the probability of committing a Type I error does not exceed some preselected number called the level of the test. Then the investigator reports whether or not the observations are "significant" at the chosen level. This procedure probably stems from the use of the Neyman-Pearson theory in classical statistics, where the decision function for the test is determined such that the probability of a Type II error is a minimum subject to the conditions imposed by the level selected. This method of test construction circumvents the problem of interrelationship between the probabilities of the Type I and Type II error. However, in many cases the choice of a significance level is completely arbitrary. In nonparametric statistics particularly, but also in parametric statistics when the null distribution is discrete, the chosen level may not even be attainable. Further, in nonparametric statistics, there is usually not sufficient information about alternative distributions so that the probability of a Type II error can even be discussed in general. Rather, the decision function is selected by logical reasoning, or according to the research hypothesis, or sometimes even by the data.

Another approach to hypothesis testing is currently attaining wide acceptance. This is the practice of reporting the smallest level at which the observations are significant in a particular direction. This quantity, which is herein called the *P-value*, is sometimes called the "critical level" or "significance level" (e.g., in Birnbaum, [3, p. 289]), the "observed level of significance" (e.g., in Kraft and Van Eeden, [8, p. 63]), the "prob-value" (e.g., in Wonnacott and Wonnacott, [12, p. 190]), or the "associated probability" (e.g. in Siegel, [11, p. 11]). Many elementary textbooks are now introducing this procedure, in addition to or instead of the more traditional one, for one sided tests based on both parametric and nonparametric methods. However, little attention has been paid to the proper interpretation of a *P*-value, nor to the inherent problem of defining *P*-values for two sided tests, particularly when the null distribution is not symmetric. These questions will be discussed in this paper, along with some comments about the need for making a clear distinction between statistical significance and practical significance in decision making.

2. *Methodology and Advantages of One Sided P-values*

Consider any hypothesis testing situation where the appropriate critical region for the test clearly lies in one particular tail of the sampling distribution of the test statistic. Then the observed value of the test criterion can be used to compute a tail probability which we call the *P*-value. The *P*-value is defined as the probability under null distributions of a sample outcome equal to or more extreme than that observed. In well-behaved problems, which include almost all one sided tests commonly used, the possible outcomes can be ordered according to how "extreme" they are in one direction relative to the outcome expected under the null hypothesis, and the values of the test statistic are also ordered in a corresponding manner. Then the *P*-value is a well defined quantity, because the meaning of extreme is clear.

* This paper was written by Gibbons, but its content overlaps parts of Chapter 1 of a forthcoming book, *Concepts of Nonparametric Theory*, written by both authors. The first draft of Chapter 1 was prepared by Pratt.

** Dept. of Statistics, Univ. of Alabama, University, AL 35486.
*** Grad. School of Bus. Admin., Harvard Univ., Boston, MA 02163.

Now suppose that the P-value is well defined, and that the goal of the experiment is to reach a statistical decision through a significance test. The P-value can then be interpreted as the smallest level of significance, that is, the "borderline level", since the outcome observed would be judged significant at all levels greater than or equal to the P-value but not significant at any smaller levels. Thus it is sometimes called the "level attained" by the sample. A decision based on the P-value will always be the same as a decision based on a critical value for a *conservative* test.

Sometimes tables of critical values for test statistics which have discrete null distributions are constructed such that the exact probability of a Type I error is as near the preselected level as possible, whether above or below (e.g., Siegel, [11, Table G]). This is equivalent to judging an outcome as significant at all levels greater than, and not significant at any level smaller than, a quantity called the *mid P-value*. The mid P-value is defined as the arithmetic average of the ordinary P-value (as defined above) and the probability of an outcome more extreme than that observed. (See Lancaster [9] for discussion.)

For the purpose of statistical decision making, it is clear that reporting a P-value conveys as much information as reporting whether or not the observations are statistically significant at some preselected level as long as the reader is also informed as to what maximum probability of a Type I error is considered tolerable. However, if we consider the ultimate goal of statistical analysis as the reduction of data to a brief condensation which contains the gist of some experimental results, reporting a P-value as, say, .039 is considerably more informative than a "bare bones" statement like "significant at level .05" or "reject at level .05," especially when .05 was chosen rather arbitrarily or by habit.

Assuming that the problem is well-behaved and that tables of exact tail probabilities are available, the exact P-value is easily found. When the available tables give only critical values at selected levels, the P-value can be specified as within a certain interval. Reporting a P-value, whether exact or within an interval, in effect permits each individual to choose his own level of significance as the maximum tolerable probability of a Type I error. This is especially important when the investigator has no real justifiable reason for choosing a particular level of statistical significance, or does not have much feeling about the cost and consequences of statistical error for this particular experimental situation. Further, in many investigations the decision to be reached ultimately is not a statistical one but a practical one. Then the statistical result should be considered no more than an objective aid to the formation of a subjective decision. Statistical significance does not necessarily imply practical significance. Rather, the decision-making process is frequently influenced by many factors in addition to the P-value. Some of these factors, like reliability of sampling procedure, or validity of test procedure, are statistical, while others, like economic or practical implications of the decision, are purely non-statistical factors relevant for this particular decision. When the decision of importance relates to a target population which is different from the population sampled, either because it is impractical or impossible to sample the target population, the decision-making process becomes even less objective. Any or all of these factors may be even more important than the P-value in reaching a practical decision. However, in all cases the P-value provides an objective measure which can be helpful to the ultimate decision maker, whether it is his sole basis for judgment or one of several input factors.

3. *Further Interpretations of One Sided P-values*

By definition, the P-value is properly interpreted as an aid to decision making since it measures the level attained by the sample outcome. Can a one sided P-value be given any other interpretations which are relevant to inference? In particular, can the P-value be interpreted as a measure of the degree to which the observations support the null hypothesis, or contradict it in a particular direction?

It certainly is true that in any single experiment, the P-value measures the degree of agreement or disagreement in a specific direction between the particular observed value of the statistic and its expected value under the null hypothesis. Suppose, for instance, that n is large or moderate, so that the test is reasonably powerful. Then if the P-value is large or moderate, the test has not merely failed to disprove the statement in the null hypothesis; it has also provided substantial evidence that the null hypothesis is true or almost true. In addition, the larger the P-value, the more affirmative is the evidence by this experiment. On the other hand, the smaller the P-value, the more "extreme" is the outcome by this experiment. Hence, as long as extreme is properly defined, the P-value does measure the degree of disagreement (or agreement) with the null hypothesis. However, any strictly increasing or decreasing function of the P-value, in particular the value of the test statistic itself (or its negative), also measures the degree of disagreement.

The real question is, can one compare P-values across sample sizes, or across experiments, as can be done with power functions? Unfortunately, such comparisons of P-values have little meaning. Arguments both within and outside the frequency theory of probability are convincing that the extent of contradiction of the null hypothesis *in general* is *not* a function of the P-value, but rather of the likelihood function (see Birnbaum [3]). Finding an event which is rare under a null hypothesis H_0 can be taken as some evidence in favor of an alternative

hypothesis H_1 as long as it is contradictory in the proper direction. However, the important point to remember is that the P-value is calculated under the assumption that H_0 is true, while the power is calculated under H_1, and it does not necessarily follow that an event which is rare under H_0 is relatively frequent under H_1 (and vice versa). Furthermore, the extent of contradiction implied by a given P-value depends on the power. If a test is very powerful, then it is very likely to reject H_0 and hence the P-value is likely to be small even when the departure from the null situation is small. Similarly, if a test is not very powerful, it is not very likely to reject H_0, and hence the P-value is not likely to be small even when H_0 is moderately far from true. Thus, the relationship between the extent of contradiction of H_0 and the magnitude of the P-value depends on the power of the test.

In the Bayesian framework, where probabilities are used not only with an "objective" meaning, but also to represent "degrees of belief", the probability that the null hypothesis is true, given the observations (that is, the "posterior" probability of the null hypothesis), may vary widely, depending on the sample size and the problem, for a fixed P-value and a fixed probability of the null hypothesis before observation (that is, a fixed prior probability of the null hypothesis). Even in practical problems, if the null hypothesis is *a priori* as likely true as false, its posterior probability after observation may well be as small as six times or as large as twelve times the P-value for P-values between .001 and .05, although it is seldom less than three times or more than thirty times the P-value (Good, [5]). (These figures are rough, and are based on less than one might desire. See also Jeffreys [7] and Lindley [10]. For an interesting example with discussion, see Good [6] and Efron [4].) In this framework then, if the value of a test statistic is just significant at the .05 level, there is still a substantial chance (at least .15) that the null hypothesis is true. This suggests that bare significance at the .05 level is at best not a very strong justification for assuming that the null hypothesis is false. Of course, significance substantially beyond the .05 level is another matter.

Often a null hypothesis is almost certainly not exactly true but is perhaps nearly true; then it is frequently convenient to treat the null hypothesis as true even though it is only nearly true. The foregoing discussion should be read in this light. For instance, the next-to-last sentence of the previous paragraph would then mean that if something is just significant at the .05 level, then there is still a substantial chance (at least .15) that the null hypothesis is nearly true, where "nearly" is defined so that, before observation, one would have considered the null hypothesis "perhaps nearly true".

In summary, even though it is not appropriate to interpret a P-value as more than a measure of the extent to which the observations contradict or support the null hypothesis in a single experiment, the method is well justified and advised on the grounds that it contains information about the experimental results which is not reflected in a simple statement of significance at a preselected level.

4. *P-values for Two Sided Tests*

If P-values are to be widely adopted, some convention is needed to define them for two sided tests. Some people claim that P-values are not appropriate in the two sided situation, but that seems an inappropriate dismissal of a problem which is not trivial and should be examined. Several different procedures will be described here.

One approach is to report the one tailed P-value even in a two sided test and remark that the two tailed P-value, while depending on what kind of two sided critical region would have been formed, is presumably nearly twice as large as the one tailed P-value reported.

If this practice is not followed, the logical definition of a P-value is the sum of the probability of a value equal to or more extreme than that observed in the same tail and *some* probability from the opposite tail. However, then a single observed result could give various P-values depending on what probability is added to represent the other tail. The most common procedure is to report a two tailed P-value as twice the one tailed P-value. This seems a very reasonable practice when the null distribution is symmetric, in that it corresponds in principle to the standard two sided test which at level α is a combination of two one sided tests, each at level $\alpha/2$. For asymmetric null distributions, the practice of attributing an equal maximum probability to each tail seems less reasonable for P-values than for selecting critical regions, since in the latter case it is possible to resort to randomized tests to achieve equality of probabilities. (Randomization could theoretically be used for P-values.) In general, there seems to be no serious objection to doubling the one tailed P-value except that for discrete distributions the P-value reported may exceed one, or otherwise may not correspond to any probability which is attainable under that distribution. The interpretation of such a P-value is then clouded even in a single experiment. A logical modification of procedure which avoids this problem is to define the P-value as the sum of the one tailed P-value and an attainable probability in the other tail which is as close as possible to the one tailed P-value obtained. "As close" could be defined as meaning in either direction, only in the conservative direction, or only in the liberal direction.

Another possibility is to make the two tails complementary in terms of the distance from some specified location parameter in the null distribution, e.g., the mean, median, midrange or mode. Then if the test criterion is X and m is the chosen

parameter, the P-value for an observed x is $P(X \geq x) + P[X \leq m - (x - m)]$ if x is in the upper tail and $P(X \leq x) + P[X \geq m + (m - x)]$ if x is in the lower tail. This procedure is especially logical when one interprets the P-value as a measure of the degree of agreement or disagreement between the particular observed value and its average or central value under the null hypothesis. It could be modified with some sort of skewness correction for severely asymmetric distributions.

Table 1

Binomial Probabilities for $p = .6$, $n = 10$

s	0	1	2	3	4	5	6	7	8	9	10
$P_{p=.6}(S = s)$	.000	.002	.010	.043	.111	.201	.251	.215	.121	.040	.006

We illustrate these procedures in the binomial case with $n = 10$ and H_0: $p = .6$. The point probabilities for S, the observed number of successes, are given in Table 1. Suppose that $S = 3$ is observed. The appropriate one tailed P-value is lower tailed, and $P_{p=.6}(S \leq 3) = .055$. This could be reported, with the comment that the two tailed P-value is presumably around .110. Since .110 is not an attainable probability under this null distribution, we could use the modification suggested above for asymmetric distributions. Then we add $P_{p=.6}(S \geq 9) = .046$, the closest attainable level in the upper tail, and report a P-value of $.055 + .046 = .101$. Since the distribution in Table 1 has mean, median and mode each equal to 6, the method of complementary distances from any of these location parameters also gives a two tailed P-value of .101 when $S = 3$. If the distance is measured from the midrange however, the complementary value of $S = 3$ is $5 + (5 - 3) = 7$ and the P-value is $P_{p=.6}(S \leq 3) + P_{p=.6}(S \geq 7) = .055 + .382 = .437$.

Now suppose that $S = 6$ is observed. In essence, 6 does not lie in either tail since $P_{p=.6}(S \geq 6) = .633$ and $P_{p=.6}(S \leq 6) = .618$. If either of these P-values is doubled, or if they are added according to the method of values equally distant from the mean, median or mode, the result exceeds one. While it is clear that H_0 is strongly supported by the outcome $S = 6$, these two methods, when applied strictly, both lead to an absurd result even though the distribution is only moderately skewed.

There are several other procedures which are sometimes used to define two tailed P-values based on discrete null distributions. Two will be described here. The first one might be called the method of placing an equal number of extreme values in each tail. In the binomial case, suppose that the observed number of successes is s, and s falls in the upper tail of the null distribution. Since there are $n - s + 1$ different values of S which are at least as large as s and occur with positive probability, the two tailed P-value could be defined as the sum of the probabilities of these values of S and the $n - s + 1$ smallest values of S, that is, $P(S \geq s) + P(S \leq n - s)$. Of course, when the possible values of the test criterion are evenly spaced (as in the binomial case), this method is equivalent to making the tails complementary in terms of distance from the midrange. In general this procedure makes the two tails complementary in terms of the number of possible values of the test criterion, rather than the distance from m or the amount of probability in each tail.

Another approach to computing a two tailed P-value might be called the "principle of minimum likelihood". If the value $S = s$ is observed, the P-value at s is found by summing the probabilities of all values of S in either tail which do not exceed the probability $P(S = s)$. In other words, the possible sample points contribute to the P-value in order of their null probability, going from the least favorable case up to the observed value, or vice versa.

These two procedures are also illustrated in the binomial case with $n = 10$ and H_0: $p = .6$ when $S = 3$ is observed. For the first method, since $S = 3$ is the fourth most extreme value in the left tail, the corresponding extreme value in the right tail is $S = 7$. Then the two tailed P-value is $P_{p=.6}(S \leq 3) + P_{p=.6}(S \geq 7) = .055 + .382 = .437$. (As mentioned above, this method is equivalent to taking the two tails as equally distant from the midrange.) With the minimum likelihood method, the points which contribute a probability not exceeding $P_{p=.6}(S = 3) = .043$ are $S \leq 3$ and $S \geq 9$, giving a two tailed P-value of $P_{p=.6}(S \leq 3) + P_{p=.6}(S \geq 9) = .055 + .046 = .101$. Notice that if the observed value of S had been in the upper tail, the method of placing an equal number of extreme values in each tail would have given a two tailed P-value smaller than twice the one tailed P-value since the null distribution is skewed to the left in this example.

Neither of these latter two procedures is well known. The first one is applicable only to discrete null distributions which have a finite domain of positive probability. It is meaningless for distributions which permit even a countably infinite number of values, as e.g., the Poisson distribution. Further, this procedure can lead to absurdities if the null distribution is heavily skewed. For example, in the binomial case with H_0: $p = .1$ suppose that $S = 7$ is observed when $n = 10$. The one tailed P-value is then $P_{p=.1}(S \geq 7) = .000$. When an equal number of extreme values are placed in the lower tail, the P-value is $P_{p=.1}(S \leq 3) + P_{p=.1}(S \geq 7) = .987$. Even though $S = 7$ strongly contradicts H_0, a P-value of .987 would lead to the conclusion that the data support the null hypothesis. Another disadvantage of this method is that it can lead to two tailed P-values which are greater than one.

The minimum likelihood method can also lead to absurdities, especially when the distribution is U-shaped, J-shaped, or simply not unimodal. For example, suppose a test criterion X has the null dis-

tribution in Table 2. If $X = 3$ is observed, the P-value by minimum likelihood is $P(X = 0, 1, 3 \text{ or } 10) = .028$. It is difficult to justify the exclusion of $X = 2$ simply on the basis that $X = 2$ is more likely to occur than $X = 3$. Why should $X = 2$ not be considered "more extreme" than $X = 3$?

Table 2

x	0	1	2	3	4	5	6	7	8	9	10
$P(X = x)$	.000	.010	.033	.012	.111	.201	.251	.215	.121	.040	.006

Because of these difficulties, we cannot recommend either of these last two procedures for general use. The other methods described give reasonable results in most cases; the notable exception is when a central value is observed. The practice of doubling the one tailed P-value is perhaps the most popular, but that may be more the result of habit than a thoughtful consideration of the merits. It provides an arbitrary but quite satisfactory result in symmetric distributions. However, if a single procedure were to be recommended as appropriate for two sided tests based on any distribution and any outcome, we prefer reporting the one tailed P-value and the direction of the observed departure from the null hypothesis. The primary basis for this recommendation is that the P-value then retains its clear interpretation, which seems an essential property when it is to be used as input for a practical decision. Further, when the one tailed P-value is small, the sample outcome is extreme in a particular direction and a one sided conclusion may be desirable in view of this observed result. On the other hand, if the P-value is moderate to large, any conclusion about the null hypothesis will probably be unchanged even if the P-value is increased.

The recommendation for reporting the one tailed P-value even with a two sided test can be further reinforced if we consider test procedures which allow a greater variety of conclusions to be reached when a decision is actually to be made. The usual procedure in the two sided situation with a simple null hypothesis permits one to decide only between two possible conclusions, e.g., $\theta = \theta_0$ and $\theta \neq \theta_0$ for some parameter θ. Consider the following four sets of decisions, each involving three possible conclusions about θ.

		(1)	(2)		(3)
S_1:	Decide	$\theta = \theta_0$,	$\theta < \theta_0$,	or	$\theta > \theta_0$
S_2:	Decide	$\theta = \theta_0$,	$\theta \leq \theta_0$,	or	$\theta > \theta_0$
S_3:	Decide	$\theta = \theta_0$,	$\theta < \theta_0$,	or	$\theta \geq \theta_0$
S_4:	Decide	$\theta = \theta_0$,	$\theta \leq \theta_0$,	or	$\theta \geq \theta_0$

Suppose that the test criterion is X and that the values of X are ordered according to how extreme they are in each direction relative to the outcome expected under the null hypothesis H_0: $\theta = \theta_0$. Assume without loss of generality that this ordering is direct rather than inverse in the sense that $F(x;\theta_0) > F(x;\theta_1)$ for $\theta_0 < \theta_1$. Then a logical decision function would be to draw one of the conclusions in column (2) when $X \leq s_l$, column (3) when $X \geq s_u$, and column (1) for $s_l < X < s_u$, for some $s_l < s_u$. Further, the following inequalities hold:

$$P(X \leq s_l \mid \theta = \theta_0) > P(X \leq s_l \mid \theta > \theta_0),$$

$$P(X \geq s_u \mid \theta = \theta_0) > P(X \geq s_u \mid \theta < \theta_0).$$

Suppose that s_l and s_u are chosen such that the lower and upper tail probabilities under H_0 are exactly α_1 and α_2, respectively. Then for the usual two conclusion procedure, the probability of an erroneous rejection of H_0 is $\alpha_1 + \alpha_2$, the two tailed level. The probabilities of erroneously rejecting H_0 are summarized in Table 3 for the decision sets S_1, S_2, S_3 and S_4.

Table 3

Probabilities of Erroneous Conclusions*

Decision Function		True Situation		
Decision Set	Observed Conclusion	$\theta < \theta_0$	$\theta = \theta_0$	$\theta > \theta_0$
S_1	$X \leq s_l \quad \theta < \theta_0$		α_1	$< \alpha_1$
	$X \geq s_u \quad \theta > \theta_0$	$< \alpha_2$	α_2	
	Total Probability	$< \alpha_2$	$\alpha_1 + \alpha_2$	$< \alpha_1$
S_2	$X \leq s_l \quad \theta \leq \theta_0$			$< \alpha_1$
	$X \geq s_u \quad \theta > \theta_0$	$< \alpha_2$	α_2	
	Total Probability	$< \alpha_2$	α_2	$< \alpha_1$
S_3	$X \leq s_l \quad \theta < \theta_0$		α_1	$< \alpha_1$
	$X \geq s_u \quad \theta \geq \theta_0$	$< \alpha_2$		
	Total Probability	$< \alpha_2$	α_1	$< \alpha_1$
S_4	$X \leq s \quad \theta \leq \theta_0$			$< \alpha_1$
	$X \geq s_u \quad \theta \geq \theta_0$	$< \alpha_2$		
	Total Probability	$< \alpha_2$		$< \alpha_1$

* The table entries left blank are those situations where the conclusion does not reject H_0 erroneously. If "accept H_0" is interpreted as no conclusion, an error can be made only by rejecting H_0 and then this table summarizes probabilities for all possible types of erroneous conclusions.

The table shows that no matter what the true situation, the probability that the decision function leads to an erroneous rejection using S_1 is at most $\alpha_1 + \alpha_2$, the two tailed level, while the same probability is at most the larger of α_1 and α_2, the larger one tailed level, using either S_2, S_3 or S_4. S_1 permits a more refined conclusion than S_2 when $X \leq s_l$ is observed, but at the expense of increasing the bound on the probability of erroneous rejection to $\alpha_1 + \alpha_2$. If we are really just as happy to conclude that $\theta \leq \theta_0$ as $\theta < \theta_0$, S_2 would perhaps be more reasonable

than S_1. Similar comments apply to S_1 versus S_3 and S_4.

Thus, unless it is clear that rejecting H_0: $\theta = \theta_0$ can lead only to the conclusion $\theta \neq \theta_0$, or to one of the conclusions $\theta < \theta_0$ and $\theta > \theta_0$, reporting the conclusion at the appropriate one tailed level is more descriptive of the true probability of erroneous rejection, even in a two tailed situation. From this point of view, a one tailed P-value is also more descriptive even in a two sided test. This further suggests the desirability of reporting a one tailed P-value so that when a definite conclusion rather than a P-value is required, the choice of the two tailed procedure which best fits the purposes and problem at hand is left to the ultimate decision-maker.

REFERENCES

[1] Berkson, J.: "Tests of Significance Considered as Evidence," *Journal of the American Statistical Association*, 37 (1942), 325–335.

[2] Berkson, J.: "Experience with Tests of Significance: A Reply to Professor R. A. Fisher," *Journal of the American Statistical Association*, 38 (1943), 242–246.

[3] Birnbaum, A.: "On the Foundations of Statistical Inference," *Journal of the American Statistical Association*, 57 (1962), 269–306.

[4] Efron, B.: "Does an Observed Sequence of Numbers Follow a Simple Rule? (Another Look at Bode's Law)," *Journal of the American Statistical Association*, 66 (1971), 552–559. Comments and Rejoinder, *Ibid.*, 559–568.

[5] Good, I. J.: "Significance Tests in Parallel and in Series," *Journal of the American Statistical Association*, 53 (1958), 799–813.

[6] Good, I. J.: "A Subjective Evaluation of Bode's Law and an 'Objective' Test for Approximate Numerical Rationality," *Journal of the American Statistical Association*, 64 (1969), 23–49.

[7] Jeffreys, H.: *Theory of Probability*, 3rd ed., Oxford: Oxford University Press, 1961.

[8] Kraft, C. H. and C. Van Eeden: *A Nonparametric Introduction to Statistics*, New York: Macmillan, 1968.

[9] Lancaster, H. O.: "Statistical Control of Counting Experiments," *Biometrika*, 39 (1952), 419–422.

[10] Lindley, D. V.: "A Statistical Paradox," *Biometrika*, 44 (1957), 187–192.

[11] Siegel, S.: *Nonparametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill, 1956.

[12] Wonnacott, T. H. and R. J. Wonnacott: *Introductory Statistics for Business and Economics*, New York: Wiley, 1972.

A Supplementary List of Publications of S. S. Wilks

CHURCHILL EISENHART*

T. W. Anderson's memoir, "Samuel Stanley Wilks, 1906-1964" in the February 1965 issue of *The Annals of Mathematical Statistics* (Vol. 36, No. 1, pp. 1–23) is followed immediately (pp. 24–27) by a list of "The Publications of S. S. Wilks", also prepared by Anderson, and arranged in three categories: BOOKS, numbered <1> to <5>; ARTICLES, numbered [1] to [48]; and SOME OTHER WRITINGS, numbered {1} to {12}, consisting of seven book reviews, a book chapter, a mimeographed lecture, a summary of a paper Wilks presented at a Cowles Commission conference, Wilks' contribution to the discussion of a paper on the meaning of probability, and an Educational Testing Service pamphlet. Anderson tells me that he did not attempt to achieve completeness of the last category.

A literature search carried out by the present writer in connection with preparation of an article on Wilks for publication in a forthcoming volume of the *Dictionary of Scientific Biography* (Charles Scribner's Sons, Publishers, 1970–) brought to light thirty-one additional "other writings" but no additional books or articles. It should be noted, however, that Wilks's book with Irvin Guttman that was "to appear" has not only appeared but reached a 2nd edition:

> Guttman, Irwin, and Wilks, S. S., *Introductory Engineering Statistics*, New York: John Wiley and Sons, Inc., 1965; 2nd edition, with J. S. Hunter as co-author, 1971.

The additional "other writings" that the search uncovered are listed below, and numbered (1) to (28) for convenient identification. It will be noticed that the list contains abstracts of Wilks's presentations at meetings of the American Mathematical Society and the Institute of Mathematical Statistics, many corresponding to subsequently published papers, identified in each case by Anderson's "article" number in [ ]. At first I was inclined to omit abstracts that corresponded to published papers, but comparison of the abstracts and the papers revealed that the abstract often contained a clearer statement of the practical usefulness and uses of the results presented than did the paper itself. I decided, therefore, to include them all.

A note of warning to those who may wish to follow the influence of the works of *S*amuel *S*tanley Wilks through the various volumes of the Science Citation Index: His younger brother, *S*yrrel *S*ingleton Wilks, a professor of physiology and an expert on aviation medicine, has the very same initials, and their publications are often lumped together under "S. S. Wilks."

* Applied Mathematics Div., Inst. of Basic Standards, Nat. Bureau of Standards, Washington, DC 20234. (Contribution of the National Bureau of Standards, not subject to copyright.)