SOME VALIDITY CRITERIA FOR STATISTICAL INFERENCES

By Robert J. Buehler

Iowa State University

1. Introduction. This paper is concerned with the ways in which existing statistical theories specify the degree of uncertainty of an inference. For the sake of graphic presentation, problems of inference are described in terms of a game between two players—one who makes the inferences and another who questions their validity. Such a model suggests a number of criteria of validity which depend entirely on classical probability calculations. I anticipate that arguments may be advanced for *not* regarding statistical inference as a game; it is hoped that the cogency of such arguments will not prevent the present model from providing some new insight into the problems here considered.

Many, though not all, problems of inference lead to assertions of the type, "The probability that A is true is equal to α," or, "$P(A) = \alpha$." One may ask whether the person making this assertion should be willing to bet that A is true, risking an amount α to win $1 - \alpha$, and should be equally willing to bet that A is false, risking $1 - \alpha$ to win α, against an opponent who has exactly the same information as he and who is allowed to choose either side of the wager. The affirmative answer will not be defended here, but its consequences will be examined.

The game viewpoint is related to, but not identical with, the ideas of von Mises, who has advanced as a postulate "the impossibility of a gambling system" in his definition of probability (see for example [19], p. 15). It has generally been recognized that modern theories of inference, which avoid the assumption of prior distributions of the parameters, should not have the same interpretation as the classical Bayes-Laplace theory based on prior distributions. The present paper attempts to show the sense in which one pays for weakening the classical assumptions by losing the von Mises postulate for the inferences "$P(A) = \alpha$."

Sections 2 to 5 are devoted to the theory of confidence intervals; in Sections 6 to 9 the ideas are generalized to include other statistical problems. The reader is warned not to expect to find any new problems solved in this paper, for at the present stage of development the theory gives at best new ways of looking at existing solutions.

2. A model for studying interval estimation. In the spirit of the introductory section the problem of interval estimation is here studied in terms of a game between two players. The players have equal knowledge about fixed conditions K (for "known") of a random experiment, for example, knowledge that n values are observed from a normal population having unit variance. Unknown conditions of the experiment, for example the value of the population mean μ, are conveniently referred to as the "state of nature" U (for "unknown"). The

Received October 21, 1958; revised June 19, 1959.

first player, Peter, has the familiar task of setting confidence intervals. It is required that he formulate a rule R which determines the interval as a function of the observations. Then on the basis of the observations he makes a *probability assertion*

$$\text{“}P(A) = \alpha\text{”}.$$

The quotation marks are used to identify an expression as Peter's assertion and to warn that it may not have validity in a direct probability or frequency sense, for that restriction is not imposed. The assertions may have validity as "confidence probabilities" or "fiducial probabilities," these being special cases. In the unit-normal example one would commonly take α to be 0.95 and A to be

$$\bar{x} - 1.96/\sqrt{n} \leqq \mu \leqq \bar{x} + 1.96/\sqrt{n}.$$

In order that the second player, Paul, have information equal to Peter's, it is required that he have knowledge of Peter's rule R as well as of the experimental conditions and observations. Paul adopts a strategy S based on R and on the experimental conditions, and consisting in the specification of two subsets C^+ and C^- of the observation space such that

$$\text{for observations in} \begin{cases} C^+ \\ C^- \end{cases} \text{Paul bets that}$$

$$A \text{ is} \begin{cases} \text{true,} \\ \text{false,} \end{cases} \quad \text{risking} \begin{cases} \alpha \\ 1 - \alpha \end{cases} \quad \text{to win} \begin{cases} 1 - \alpha. \\ \alpha. \end{cases}$$

It is not required that a bet must always be made; thus C^+ and C^- need not be exhaustive. To determine the winner of each bet, we postulate the existence of a referee who knows the true state of nature.

2.1. *The criterion of weak exactness.* If the model is adequately specified, one should in principle be able to calculate the expected gain to Paul. For any fixed experimental conditions K the expected gain would be a function of (i) the state of nature U, (ii) Peter's rule R, and (iii) Paul's strategy S. Different criteria for the sensibility of Peter's rule might be put forward in terms of this expected gain. For example, I propose the following. Suppose Paul's strategy is to bet consistently that A is false, regardless of the observations. Then if Paul's expected gain is zero for all U, Peter's rule R will be defined to be *weakly exact.*[1]

Weak exactness is (at least from the point of view of some theories of probability) essentially equivalent to writing $P(A) = \alpha$ without quotation marks, the definition above being preferred on the grounds that it is less subject to misinterpretation. It will be noted that weak exactness is a propery possessed by Neyman's confidence intervals (see for example [17]), but not necessarily by the fiducial counterparts, as is known.

2.2. *Relevant and semirelevant subsets.* In this paper the ultimate calculation of the expected gain is actually made only once (in Section 4.2). The emphasis lies

[1] Strongly exact is defined in Section 6.1.

more in studying Paul's initial search for a strategy having a guaranteed winning percentage. If we call $P(A \mid C) - \alpha$ the *bias* of C, then Paul's problem is to find subsets C whose bias has the same sign for all U. These will be called *semi-relevant subsets induced by R*. If moreover the bias is bounded away from zero, they will be called *relevant*. That is, if $\epsilon > 0$ is independent of U, then C is called

$$\text{semirelevant}\begin{cases}\text{if} & P(A \mid C) > \alpha \\ \text{or if} & P(A \mid C) < \alpha\end{cases} \qquad \text{for all } U,$$

$$\text{relevant}\begin{cases}\text{if} & P(A \mid C) \geqq \alpha + \epsilon \\ \text{or if} & P(A \mid C) \leqq \alpha - \epsilon\end{cases} \qquad \text{for all } U.$$

The phrase "induced by R" is crucial; the defined properties are necessarily relative to the rule R, which enters the defining equations through A. The definitions are inspired largely by writings of Fisher, of which the following quotations from [11] are typical:

p. 32. "···no such subset can be recognized."

p. 33. "···inability to discriminate any of the different subaggregates having different limiting frequency ratios."

p. 57. "···every subset to which it belongs, and which is characterized by a different fraction must be unrecognizable."

More recently ([12], p. 23) Fisher uses the words "relevant" and "irrelevant." For example:

"The subset of throws made on a Tuesday is easily recognizable, it has, however the same probability as the whole set and is therefore irrelevant."

It would seem that any subset C of the observation space might be called "recognizable" in Fisher's sense since it is determined by known observations. The word "relevant" has been introduced here in a sense intended to be close to Fisher's, but there seems to be at least a small difference arising from the dependence on the rule R. If a need for distinct terms should arise, "induced relevant subset" might be substituted for "relevant subset" as used here.

In typical interval estimation problems the bias of most subsets C will not have the same sign for all U. In particular if C contains only a single point of the observation space, then $P(A \mid C)$ ordinarily will be either zero or unity, depending on U. This corresponds to the statement, sometimes seen in textbooks, that the probability that the true value of the parameter lies within a confidence interval is either zero or unity *after* the interval has been constructed.

2.3. *Relevance of unions and complements.* It will be convenient later to refer to the following elementary results.

LEMMA 1. *If subsets C_1 and C_2 are disjoint, (semi)relevant, and positively [negatively] biased, and if the union $C_1 + C_2$ has nonzero probability for all U, then the union is (semi)relevant and positively [negatively] biased.*

We may remark that for nondisjoint subsets neither the union nor the intersection need have special properties of relevance deriving from the components.

LEMMA 2. *Let C' denote the complement of C. If* (i) $P(A) = \alpha$ *(essentially* *weak exactness*); (ii) $P(C)$ *and* $P(C')$ *are nonzero for all* U; (iii) $P(C) > b > 0$ (b *is a constant independent of* U); *and* (iv) C *is relevant; then* C' *is relevant with bias opposite to* C.

LEMMA 3. *If in Lemma* 2 (iii) *is not assumed, then* C' *is semirelevant with bias opposite to* C.

LEMMA 4. *If in Lemma* 2 (iii) *is not assumed and if* (iv) *is replaced by "*C *is semirelevant," then* C' *is semirelevant with bias opposite to* C.

3. Examples of relevant subsets in interval estimation problems. We now give an assortment of six examples of relevant subsets induced by systems of confidence intervals.

3.1. *An example of intervals based on an insufficient statistic.* This first example, although relatively simple, illustrates a number of interesting points. Let a sample of size $n = 2$ be drawn from a normal population having unknown mean μ and unit variance. A confidence interval based on the first observation x_1, but ignoring the second x_2, corresponds to the weakly exact probability assertion

$$\text{``}P(x_1 - 1.96 \leqq \mu \leqq x_1 + 1.96) = 0.95\text{''}.$$

Paul might logically begin his search for relevant subsets by comparing this questionable assertion with the standard one based on the sufficient statistic $\bar{x} = \frac{1}{2}(x_1 + x_2)$. The comparison shows that when $x_1 = x_2$ the intervals are correctly placed but unduly long, and when $|x_1 - x_2|$ is large they are "badly placed." A clue is thus furnished which suggests conditioning on subsets defined in terms of the statistic $\delta = x_1 - x_2$. If C denotes any such subset and A denotes $|x_1 - \mu| \leqq 1.96$, then the conditional probability is

$$P(A \mid C) = P(AC)/P(C) = \iint_{AC} f\,dx\,dy \Big/ \iint_{C} f\,dx\,dy$$

where $f\,dx\,dy$ is the density $(1/2\pi) \exp\{-\frac{1}{2}(x_1 - \mu)^2 - \frac{1}{2}(x_2 - \mu)^2\}\,dx\,dy$. On setting $y_1 = x_1 - \mu$, $y_2 = x_2 - \mu$, it is seen that both numerator and denominator are independent of μ. Thus $P(A \mid C)$ is independent of μ, and all subsets C defined in terms of δ will be relevant save for exceptional cases for which $P(A \mid C) = 0.95$. In particular the sets (of zero probability) for which $\delta = \delta_0$ can be seen by simple calculation to have negative, zero, positive bias respectively for $|\delta_0| >, =, < 1.593$. For $\delta_0 = 0$, $P(A \mid C)$ achieves its maximum value of 0.997 (= standard normal area within $\pm 1.96\sqrt{2}$); as $|\delta_0|$ tends to infinity, $P(A \mid C)$ tends to zero. Consider now the three subsets

$$C_1\colon\ 0 < \delta < 1; \qquad C_2\colon\ -1 < \delta < 0; \qquad C_2'\colon\ \text{complement of } C_2.$$

C_1 may be regarded as being made up of a continuum of positively biased relevant subsets of zero probability, and by a continuous generalization of Lemma 1 it follows that C_1 has positive bias. By the same reasoning C_2 is positively biased; and by Lemma 2, C_2' is negatively biased. One interesting observation is that C_1 is a subset of C_2'; thus a positively biased subset may sometimes be contained in a negatively biased one. Furthermore it is possible for a particular observation to belong simultaneously to two different subsets having opposite bias. Thus arises the basic question (which will not be resolved in this paper) of the *appropriate* subset to which any particular observation should be referred if it is not to be referred to the universal set of all observations.

3.2. *An example involving shortest average length.* Our second example is inspired by Cox [4], who treats the *testing* situation and the criterion of maximum power, whereas we treat the *estimation* situation and the criterion of minimum expected length. Suppose two populations are known to have standard deviations $\sigma_1 = 1$ and $\sigma_2 = 2$. A random choice between the two populations is made and one normal random variate is observed. If Peter knows which population was sampled he may treat each population separately in the usual way, asserting

$$\text{“}P(x - 1.96\sigma_i \leqq \mu \leqq x + 1.96\sigma_i) = 0.95\text{”}$$

in which both x and σ_i represent observed values. On the other hand, if he wished to minimize the average length of the intervals, he would do best to increase the error rate for the second population to about seven per cent and compensate by decreasing the error rate for the first population to about three per cent, thus maintaining a five per cent average. The first solution may be called a "conditional" one (in the sense of conditional probability) inasmuch as $P(A) = \alpha$ is valid in the frequency sense, with the relative frequency α prevailing separately in the two subsets defined by the value of σ. The second solution is "unconditional" in that $\mathrm{P}(A) = \alpha$ is valid in the frequency sense only when related to the sequence of all observations, ignoring the observed value of σ. Both solutions are weakly exact. The two subsets defined by the value of σ are conspicuously relevant for the unconditional solution; thus the criteria of shortest average length and nonexistence of relevant subsets cannot both be met. I myself would prefer the conditional solution here.

3.3. *An example of relevant subsets defined by an ancillary statistic.* The following example uses some distributional results given by Fisher ([11], pp. 163–165), to illustrate how relevant subsets may be defined by means of an *ancillary* statistic. Consider a sample of size n from the bivariate density

$$e^{-\theta x - y/\theta}\, dx\, dy, \qquad 0 < x, y, \theta < \infty,$$

and define statistics T (the maximum likelihood estimator of θ) and V (an ancillary statistic, according to Fisher) by

$$X = \sum x, \qquad Y = \sum y, \qquad T^2 = Y/X, \qquad V^2 = XY.$$

Then T and V have the joint density

$$\frac{V^{2n-1}}{[(n-1)!]^2} \exp\left\{-V\left(\frac{T}{\theta} + \frac{\theta}{T}\right)\right\} \frac{dT}{T}\, dV,$$

and the marginal and conditional densities of T are respectively

$$K\left(\frac{T}{\theta} + \frac{\theta}{T}\right)^{-2n} \frac{dT}{T} \quad \text{and} \quad K(V) \exp\left\{-V\left(\frac{T}{\theta} + \frac{\theta}{T}\right)\right\} \frac{dT}{T},$$

where K and $K(V)$ are independent of T. The two distributions may be used to give respectively unconditional and conditional systems of confidence intervals for θ. To use the former, put $\beta = T/\theta$ and determine constants β_1 and β_2 such that

$$\int_{\beta_1}^{\beta_2} \beta^{2n-1}(1 + \beta^2)^{-2n}\, d\beta = \alpha \int_0^{\infty} \beta^{2n-1}(1 + \beta^2)^{-2n}\, d\beta.$$

This leads to the assertion

$$\text{“}P(A) = P(\beta_1 \leqq T/\theta \leqq \beta_2) = P(T/\beta_2 \leqq \theta \leqq T/\beta_1) = \alpha\text{”}.$$

But conditionally upon the value of the ancillary V the probability is

$$\begin{aligned} P(A \mid V) &= \int_{\beta_1\theta}^{\beta_2\theta} \exp\left\{-V\left(\frac{T}{\theta} + \frac{\theta}{T}\right)\right\} \frac{dT}{T} \Big/ \int_0^{\infty} \exp\left\{-V\left(\frac{T}{\theta} + \frac{\theta}{T}\right)\right\} \frac{dT}{T} \\ &= \int_{\beta_1}^{\beta_2} \exp\left\{-V\left(\beta + \frac{1}{\beta}\right)\right\} \frac{d\beta}{\beta} \Big/ \int_0^{\infty} \exp\left\{-V\left(\beta + \frac{1}{\beta}\right)\right\} \frac{d\beta}{\beta} \end{aligned}$$

The last expression depends on the ancillary V but is independent of the parameter θ; it will equal α only for a particular intermediate value of V, and for all other values relevant subsets will be defined.

3.4. *Another example involving an ancillary statistic.* Turning to Fisher ([9], and [11], p. 134) for another example involving an ancillary statistic, we let x and y have a circular normal distribution with unit variance and with mean on a circle of known radius R so that the density is

$$(1/2\pi) \exp\{-\tfrac{1}{2}(x - R\cos\theta)^2 - \tfrac{1}{2}(y - R\sin\theta)^2\}\, dx\, dy$$

where θ is to be estimated. If a single observation (x, y) is expressed in polar coordinates by

$$\begin{aligned} a^2 &= x^2 + y^2 \qquad & x &= a\cos\hat{\theta} \\ \hat{\theta} &= \tan^{-1} y/x \qquad & y &= a\sin\hat{\theta} \end{aligned}$$

then $\hat{\theta}$ is the maximum likelihood estimator of θ, and a is ancillary. One easily obtains the joint density

$$(a/2\pi) \exp\{-\tfrac{1}{2}(a^2 + R^2 - 2aR\cos(\hat{\theta} - \theta))\}\, da\, d\hat{\theta}.$$

An unconditional system of confidence intervals is given by

$$\text{“}P(\hat{\theta} - \gamma \leqq \theta \leqq \hat{\theta} + \gamma) = \alpha\text{”}$$

where α is the fraction of the unit circular normal distribution inside a wedge of angle 2γ, the maximum density being centered in the wedge at a distance R from the vertex. To find the probability conditioned on the ancillary, we find the marginal distribution of a to be expressible in terms of the Bessel function I_0 by

$$\begin{aligned} &\frac{a\, da}{2\pi} \exp\{-\tfrac{1}{2}(a^2 + R^2)\} \int_0^{2\pi} \exp\{aR\cos(\hat{\theta} - \theta)\}\, d\hat{\theta} \\ &\qquad = a \exp\{-\tfrac{1}{2}(a^2 + R^2)\} I_0(aR)\, da. \end{aligned}$$

Thus the conditional density of $\hat{\theta}$ given a is

$$\frac{\exp\{aR\cos(\hat{\theta}-\theta)\}}{2\pi I_0(aR)}\,d\hat{\theta},$$

and the conditional probability for a given value of a is

$$P(\hat{\theta}-\gamma \leqq \theta \leqq \hat{\theta}+\gamma \mid a) = \frac{1}{2\pi I_0(aR)}\int_{-\gamma}^{\gamma} e^{aR\cos\phi}\,d\phi$$

As in the previous example, the conditional probability depends on the ancillary and is independent of the parameter. Here it approaches unity as a tends to infinity and approaches γ/π as a tends to zero. Thus subsets constructed from observations lying near the origin will be negatively biased; the variability of $\hat{\theta}$ for these considered separately is larger than the variability of $\hat{\theta}$ for all observations collectively.

3.5. *An example involving the Behrens-Fisher problem.* Behrens' hypothesis states that the means μ_1, μ_2 of two normal populations differ by $\delta(=0$, usually), no assumption of the equality of variances σ_1^2, σ_2^2 being made. We consider the test devised by Welch [21] tabulated by Aspin [1] and appearing as Table 11 in the *Biometrika* Tables of Pearson and Hartley [18]. This test rejects the null hypothesis H_0: $\mu_1-\mu_2=\delta$ when $|\bar{x}_1-\bar{x}_2-\delta| > v(s_1^2/n_1+s_2^2/n_2)^{1/2}$ where n, $\bar{x}$, s^2 denote size, mean, and variance of the samples and v (the analog of Student's t) is a tabulated function of n_1, n_2, s_1^2, s_2^2, and the significance level. A calculation of Fisher [10] shows that for fixed s_2/s_1 the conditional probability P (reject $\mid H_0$, s_2/s_1) can be expressed as a unique function of the ratio σ_2^2/σ_1^2; and that when $n_1=n_2=7$, $s_1=s_2$, and the significance level is 0.1, then P (reject $\mid H_0$, $s_1=s_2$) achieves a *minimum* value[2] of 0.108 when $\sigma_2^2/\sigma_1^2=1$. Now if the tabulated test is translated into a system of confidence intervals, one obtains the probability assertion

$$\text{“}P(A) = P\{|(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)| \leqq v(s_1^2/n_1+s_2^2/n_2)^{1/2}\} = \alpha\text{”}$$

for which Fisher's calculation shows that when $n_1=n_2=7$, $\alpha=0.9$,

$$P(A \mid C) \leqq \alpha-\epsilon \quad \text{for all} \quad \mu_1,\mu_2,\sigma_1^2,\sigma_2^2$$

where $\epsilon=0.008$ and C is the subset $s_1=s_2$.

Thus the tabulated solution induces a negatively biased relevant subset. It is noteworthy that in this example $P(A \mid C)$ depends on the ratio σ_2^2/σ_1^2 whereas in all four preceding examples $P(A \mid C)$ was independent of the parameters. One might well wish to distinguish between these two types of relevance, although separate names have not been advanced here.

In [10] Fisher seems to imply that subsets defined by fixed values of s_2/s_1 are uniquely appropriate reference sets for inferences about Behrens' hypothesis; but like Bartlett [2] and Welch [22], I do not find his reasons to be compelling. Fisher does show that the subset $s_1=s_2$ is not relevant (in the present technical

[2] Fisher gives a graph but no table. The value 0.108 is obtained from Welch [22].

sense) for Behrens' solution, but he does not consider the possible relevance of other subsets. Wallace [20] has since shown that no such relevant subsets exist.

3.6. *An example in which an individual point of the sample space is a relevant subset.* As a rule the property of being relevant belongs to a set of possible observations and not to individual points of the observation space. But in rare cases individual points may be relevant, as can be illustrated by a familiar component-of-variance problem. Let

$$y_{ij} = \mu + a_i + e_{ij} \quad (i = 1, \cdots, r; \; j = 1, \cdots, s)$$

in which μ is an unknown parameter, a_i and e_{ij} are normal and independently distributed with zero means and variances σ_a^2 and σ^2. Using the dot notation to indicate an average over the missing subscript, the sums of squares

$$Q_1 = \sum\sum (y_{ij} - y_{i\cdot})^2 \quad \text{and} \quad Q_2 = \sum\sum (y_{i\cdot} - y_{\cdot\cdot})^2$$

are known to have independent distributions given in terms of chi-square by

$$\sigma^2\chi^2_{r(s-1)} \quad \text{and} \quad (\sigma^2 + s\sigma_a^2)\chi^2_{r-1}\,.$$

Thus the quantity

$$\frac{(\sigma^2 + s\sigma_a^2)\, f_2\, Q_1}{\sigma^2 f_1\, Q_2} = \frac{\chi^2_{f_1}/f_1}{\chi^2_{f_2}/f_2}$$

has the F distribution with $f_1 = r(s - 1)$ and $f_2 = r - 1$ degrees of freedom. If lower and upper percentage points F_1 and F_2 are chosen so that

$$P(F_1 \leqq F \leqq F_2) = \alpha$$

then on rearranging the inequalities one finds that the probability is α that

$$\frac{1}{s}\left\{\frac{f_1\, Q_2}{f_2\, Q_1} F_1 - 1\right\} \leqq \frac{\sigma_a^2}{\sigma^2} \leqq \frac{1}{s}\left\{\frac{f_1\, Q_2}{f_2\, Q_1} F_2 - 1\right\}.$$

If the ratio of mean squares is larger than the upper percentage point,

$$\frac{Q_1/f_1}{Q_2/f_2} > F_2$$

as always can happen with at least some small probability, then the confidence interval includes only negative values. The probability of covering the true ratio σ_a^2/σ^2 conditionally on any such observation is zero; thus such individual points constitute negatively biased relevant subsets.

Now let C denote the collection of all such points, that is, all observations for which the last inequality is satisfied. Then C is relevant and negatively biased. Is increased confidence thereby justified for observations in the complementary set C'? That C' is *semirelevant* follows from Lemma 3. Using $P(A \mid C) = 0$ one obtains

$$P(A \mid C') = \alpha/P(C') = \alpha + \alpha P(C)/P(C')$$

so that the bias is $\alpha P(C)/P(C')$. Since $P(C)$ tends to zero as σ^2/σ_a^2 tends to zero, there is no positive lower bound for the bias. Thus the complementary set C' is not relevant. Loosely summarizing these results: within C one's confidence is necessarily zero; within C', α is the greatest lower bound for the confidence over the possible states of nature.

4. Examples of semirelevant subsets. From the following examples it will be seen that semirelevant subsets are more easily found than relevant ones and that the requirement that none should exist is a very severe restriction indeed.

4.1. *A nonparametric assertion about the median.* Let $g(t)$ be any continuous density which is positive for $-\infty < t < \infty$ and which has median equal to zero. If one observation x is taken from the density $g(x - \theta)$ (θ = median), then the assertion

$$\text{“}P(A) = P(x < \theta < \infty) = \tfrac{1}{2}\text{”}$$

is weakly exact. Consider the subset C defined by $x > 0$. The conditional probability $P(A \mid C) = P(x < \theta \mid x > 0) = P(0 < x < \theta)$ is zero for negative θ while for positive θ one has

$$P(0 < x < \theta) = \int_0^\theta g(x - \theta)\, dx = \int_{-\theta}^0 g(t)\, dt$$

$$= \tfrac{1}{2} - \int_{-\infty}^{-\theta} g(t)\, dt < \tfrac{1}{2}$$

for all $0 < \theta < \infty$. Thus C is semirelevant, negatively biased. The last integral, which gives the bias as a function of θ, is seen to tend to zero as θ tends to infinity; thus C is semirelevant only—not relevant.

4.2. *Student's t.* Let x be normally distributed with unknown mean μ $(-\infty < \mu < \infty)$ and unknown variance σ^2 $(0 < \sigma^2 < \infty)$. If $\bar{x}$ and s^2 denote the sample mean and variance, then the conventional confidence or fiducial interval estimate of μ corresponds to the assertion

$$\text{“}P(\bar{x} - ks \leqq \mu \leqq \bar{x} + ks) = \alpha\text{”}$$

where $k = t_{\alpha,n-1}/\sqrt{n}$ and $t_{\alpha,n-1}$ is the appropriate percentage point of Student's t. We shall first show that for any $a > 0$, the subset $s < a$ is semirelevant, negatively biased. Stated in another way, for any μ, σ^2, "long" intervals cover the true value more frequently than "short" intervals, and this fact holds for any critical length which is used to distinguish "long" intervals from "short" ones.

Denote by A the event $|\bar{x} - \mu| \leqq ks$, by C and C' the subset $s < a$ and its complement $s \geqq a$, by $\sigma^{-1}f(s/\sigma)\, ds$ and $g(\bar{x} - \mu)\, dx$ the independent densities of s and $\bar{x}$. Then

$$P(C) = \int_0^a \sigma^{-1}f(s/\sigma)\, ds = \lambda(a/\sigma), \qquad \text{say,}$$

and

$$P(AC) = \int_0^a \sigma^{-1}f(s/\sigma) \int_{\mu-ks}^{\mu+ks} g(\bar{x} - \mu)\, d\bar{x}\, ds.$$

If we put

$$G(s) = \int_{\mu-ks}^{\mu+ks} g(\bar{x} - \mu)\, d\bar{x} = \int_{-ks}^{ks} g(z)\, dz,$$

then by a mean value theorem

$$P(AC) = \int_0^a \sigma^{-1}f(s/\sigma)G(s)\, ds = G(s_0) \int_0^a \sigma^{-1}f(s/\sigma)\, ds = G(s_0)\lambda(a/\sigma)$$

where $0 < s_0 < a$. This gives

$$P(A \mid C) = P(AC)/P(C) = G(s_0).$$

By a similar argument

$$P(A \mid C') = G(s_0')$$

where $a < s_0' < \infty$. But G is an increasing function so that $s_0 < a < s_0'$ implies $G(s_0) < G(s_0')$. The values of s_0 and s_0' depend on a and σ, but for any particular a, σ, weak exactness implies that C and C' must have opposite bias. Thus for all a, σ, μ,

$$G(s_0) = P(A \mid C) < \alpha < P(A \mid C') = G(s_0')$$

so that C and C' are semirelevant.

To illustrate the situation more clearly, we consider the simple case $n = 2$, $\alpha = \frac{1}{2}$, for which $s^2 = \frac{1}{2}(x_1 - x_2)^2$ and the assertion is that the mean has an even chance of lying between the two observations:

$$\text{“}P(A) = P(x_{\min} \leqq \mu \leqq x_{\max}) = \tfrac{1}{2}\text{”}.$$

For the subset C': $s \geqq a$, the conditional probability is

$$P(A \mid C') = \tfrac{1}{2} + \tfrac{1}{2}\phi(a/\sigma)$$

where $\phi(a/\sigma)$ is the standard normal probability between $-a/\sigma$ and $+a/\sigma$. Thus the bias, $\frac{1}{2}\phi(a/\sigma)$, is always positive, increasing from 0 to $\frac{1}{2}$ as a increases from 0 to ∞. If Paul adopts the strategy: bet even odds that A is true when $s \geqq a$ and bet even odds that A is false when $s < a$, then the probability that Paul wins is

$$P = \tfrac{1}{2} + \phi(1 - \phi), \qquad \phi = \phi(a/\sigma),$$

which lies in the range $1/2 < P \leqq 3/4$, and Paul's expected gain on these bets of 1/2 unit is

$$G = \phi(1 - \phi)$$

which always lies in the range $0 < G \leqq 1/4$. The maxima of P and G are attained when $\phi = 1/2$, that is, when

$$a = 0.67\sigma, \quad \text{approximately.}$$

Thus the optimum value of a is proportional to σ; and for dimensional reasons this is true generally, the proportionality factor depending on n and α.

In an idealized model Paul will have no prior information concerning σ and thus will have no basis for a choice of a in the range from zero to infinity, but in practically any applied problem some prior knowledge of σ will be available. It is noteworthy that Fisher has specifically stated that fiducial inference is valid only in the absence of prior knowledge ([11], p. 51). In contrast, Neyman appears not to have taken a stand on the applicability of confidence interval theory in the presence of prior information. I believe that the above calculations tend to justify Fisher's restriction. Now the assumption that a prior distribution is known and the assumption that *nothing* is known are two boundaries of a vast intermediate area in which there is *partial* prior information. We have seen how Paul can find a crude stategy in this middle area against Peter's conventional solution. But what one would really like to know is how Peter's rule might be altered to use partial prior knowledge. That is a much more subtle and difficult question.

5. An example of nonexistence of relevant subsets. Let $g(t)$ be a continuous density which is nonzero for all t, and let $f(x, \theta) = g(x - \theta)$ so that θ is a location parameter. From a sample of one value of x, θ may be estimated by the assertion

$$\text{“}P(A) = P(-t_1 \leqq x - \theta \leqq t_2) = P(x - t_2 \leqq \theta \leqq x + t_1) = \alpha\text{”}$$

where t_1 and t_2 are chosen so that

$$\int_{-t_1}^{t_2} g(t)\, dt = \alpha.$$

We wish to show that neither of the inequalities

$$(1) \qquad P(A \mid C) \geqq \alpha + \epsilon \quad \text{or} \quad P(A \mid C) \leqq \alpha - \epsilon \qquad (\text{for all } \theta)$$

can hold for any Lebesgue measurable set C of values of x having finite or infinite, but not zero, Lebesgue measure (one is prevented from treating sets of measure zero by the nonuniqueness of the definition of conditional probability; see for example [13], p. 12). Let $a(\theta)$, $b(\theta)$, $A(R)$, $B(R)$ be defined by

$$a(\theta) = P(C), \qquad A(R) = \int_{-R}^{R} a(\theta)\, d\theta,$$

$$b(\theta) = P(AC), \qquad B(R) = \int_{-R}^{R} b(\theta)\, d\theta.$$

Then $P(A \mid C) = b(\theta)/a(\theta)$; and substituting in (1), multiplying by $a(\theta)$, integrating over $-R \leqq \theta \leqq R$, and dividing by $A(R)$ gives

$$\frac{B(R)}{A(R)} \geqq \alpha + \epsilon \quad \text{or} \quad \frac{B(R)}{A(R)} \leqq \alpha - \epsilon \quad \text{(for all } R).$$

It is shown in the appendix that subject to weak conditions on $g(t)$, $B(R)/A(R) \to \alpha$ as $R \to \infty$, and thus the desired result is established by contradiction. Some sweeping generalizations of this result have been found by Wallace [20].

6. Generalization of the model. It may first be noted that the theory of *tolerance* as well as confidence intervals may be treated simply by taking A to be the proposition that at least a certain proportion of the population lies between stated limits.

A useful generalization consists in representing the outcome of the random experiment by a pair of random variables (x, y) of which x is known and y is unknown to the players, both being known to the referee. The theory of *prediction intervals* is then treated by taking x to be "past" observations and y to be "future" observations to be predicted (it is immaterial that the "future" observations y have already been observed by the referee so long as they are unknown to the two players). The proposition A is then a statement depending on x concerning the future observations; a notable distinction is that A does not concern the unknown state of nature U.

In *Bayesian interval estimation* (Cramér [6], p. 508, Neyman [17], p. 162) y plays the role of the unknown parameter having a known prior distribution. Bayesian problems generally have the distinguishing feature that the state of nature is assumed known so that U does not appear. As in prediction interval theory, the proposition A gives a relation between x and y, the known and unknown observations.

A further possible generalization which might have some interest, although it seems not to be required for existing theories, consists in allowing the confidence coefficient α to depend on the known observations: $\alpha = \alpha(x)$. That is, the rule $\boldsymbol{R}$ for probability assertions may specify not only how the proposition A is to depend on x but also how the asserted probability level is to depend on x. The definition of relevant subsets then requires some repair; the logical extension consists in considering only subsets C of values of x for which $\alpha(x)$ takes a fixed value. An example of variable α is given in Section 9.

6.1. *The criterion of strong exactness.* It has been noted that the expected gain to Paul is a function of the unknown state of nature U, Peter's rule R, and Paul's strategy S; thus it may be denoted by $G(U, R, S)$.

Definition. A rule R_0 will be called *strongly exact* if

$$G(U, R_0, S) = 0 \quad \text{identically for all } U, S.$$

In other words: Whatever the true state of nature and whatever strategy Paul may use, the expected gain to Paul is zero. It is essentially equivalent to write $P(A \mid C) = \alpha$ for all U, C.

It appears to be impossible to satisfy the very stringent condition of strong exactness except in rather special cases, e.g., in Bayesian estimation where the condition "all U" is in fact no requirement at all since U is not variable. Thus strong exactness is not so much a practical requirement as a goal toward which one might strive even though it cannot actually be reached.

7. Remarks on Bayesian interval estimation. In a typical Bayesian situation an experiment consists in obtaining a random value of the "parameter" from a known distribution $w(\theta)$ and subsequently observing values $x_1, \cdots, x_n$ from a distribution $f(x; \theta)$ which is known but for the value of θ. Thus in the model of Section 6, x represents $x_1, \cdots, x_n$ and y represents θ. The conditional distribution of θ given $x_1, \cdots, x_n$ is

$$h(\theta \mid x_1, \cdots, x_n) = \frac{w(\theta)\{f(x_1; \theta) \cdots f(x_r; \theta)\}}{\int w(\theta)\{f(x_1; \theta) \cdots f(x_n; \theta)\}\, d\theta}$$

A Bayesian estimate of θ is given by the assertion

$$\text{“}P(k_1 \leqq \theta \leqq k_2) = \alpha\text{”}$$

where k_1 and k_2 are any two numbers depending on the observations by

$$\int_{k_1}^{k_2} h(\theta \mid x_1, \cdots, x_n)\, d\theta = \alpha$$

Apart from the arbitrary allocation of the probability $1 - \alpha$ between the two tails of the distribution, k_1 and k_2 are unique functions of the observations. It may further be noted that the interval (k_1, k_2) based on any particular observation $(x_1, \cdots, x_n)$ has its own validity without reference to intervals which might be defined for other observations; this is in contrast to confidence interval theory in which any particular interval is meaningful only when referred to a system of intervals defined for all possible samples. A closely connected fact is that the above Bayesian solution gives a rule for assertions which is strongly exact.

Two variations of the above are found in the literature. Cramér ([6], p. 508) replaces $h(\theta \mid x_1, \cdots, x_n)$ by

$$h(\theta \mid \hat{\theta}) = \frac{w(\theta) g(\hat{\theta}; \theta)}{\int w(\theta) g(\hat{\theta}; \theta)\, d\theta}$$

where $\hat{\theta}$ is an estimator of θ and g is the density of $\hat{\theta}$. Three remarks can be made: (i) If $\hat{\theta}$ is sufficient for θ and if the allocation of the probability $1 - \alpha$ to the two tails is fixed, then this solution will not differ from the preceding one. (ii) If $\hat{\theta}$ is not sufficient and if the sample $(x_1, \cdots, x_n)$ is known to the players, then relevant subsets will exist and the solution is weakly exact but not strongly exact. (iii) If $\hat{\theta}$ is not sufficient, but the players have knowledge only of $\hat{\theta}$ and not of $(x_1, \cdots . x_n)$, then the last solution is strongly exact.

A second variation is given by Neyman's "modernized Bayes' estimating intervals" ([17], pp. 165–181) in which the expected length of the intervals is shortened by requiring that the asserted probability α refer only to the relative frequency in the sequence of all experiments, not separately to sequences of fixed $(x_1, \cdots, x_n)$. Thus in the "modernized Bayes' solution" we have a clear-cut example of a rule that is weakly but not strongly exact. This situation is quite similar to the Cox example of Section 3.2.

8. A familiar prediction example. To show how a prediction problem fits into the general scheme and to point out some analogies between prediction and confidence intervals, we consider a familiar prediction example. Let x have a continuous density of unspecified form, let x_1 and x_2 represent two known observations, and let x_3 represent an unknown or future observation. Then since all six of the permutations of $x_1 < x_2 < x_3$ are equally likely, the probability assertion

$$\text{“}P(x_{\min} < x_3 < x_{\max}) = \tfrac{1}{3}\text{”}$$

is weakly exact. Some years ago this particular example figured in a dispute between Fisher [7], [8] and Jeffreys [15], [16]. Without claiming to understand the subtleties of the dispute we note that Fisher [7] points out that "for any particular population the probability will generally be larger when the first two observations are far apart than when they are near together." From Fisher's remark it follows that just as in the Stud nt's t example of Section 4.2, long intervals are valid more often than short ones, and the subsets $x_{\max} - x_{\min} \leqq a$ or $\geqq a$ are semirelevant for any $a > 0$. Fisher goes on to say that "the fallacy of Jeffreys' argument consists just in assuming that the probability shall be 1/3, *independently of the distance apart of the first two observations*." In the present terminology we would say that Jeffreys is accused of treating an assertion as if it were strongly exact when in fact it is only weakly exact.

9. An example of intersecting relevant subsets. In this section some rather artificial examples are constructed primarily to illustrate intersecting relevant subsets. The examples differ from those preceding in that the proposition A is made independent of the observations and the "confidence level" α is allowed to be random. The examples raise the question of whether interesecting relevant subsets might exist elsewhere, for example, in confidence interval situations.

Let an experiment consist in drawing one ball from twelve contained in an urn. Let A, B, D denote attributes, and let A', B', D' denote the respective negations. We suppose that Peter and Paul have knowledge of the contents of the urn as shown in Table 1, and that the referee draws one ball and announces whether B or B' is observed ($x = B$ or B' = observation known to players). It is not revealed whether A or A' is observed ($y = A$ or A' = unknown observation). Peter's probability assertion concerns the probability that the ball drawn has attribute A. The assertion "$P(A) = 5/12$" is weakly exact. It is objectionable on the grounds that the (only) subsets B and B' are relevant.

A strongly exact assertion clearly can be made by allowing a variable "confidence level":

$$\text{“}P(A) = \begin{Bmatrix} 4/9 & \text{if} & x = B \\ 1/3 & \text{if} & x = B' \end{Bmatrix}\text{”}$$

The complexion of the problem is changed if the players are given further information. Suppose the twelve balls are also classified in categories D and D' and that the values in Table 2 are also known to the players. The referee announces one of four possible results: $x = BD, BD', B'D$, or $B'D'$. It will be seen that no probability assertion of the form "$P(A) = \alpha(x)$" is strongly exact inasmuch as the values of $P(A \mid BD)$, etc., are dependent on an unknown variable state of nature, for the two *marginal* 2×2 tables fail to specify a unique $2 \times 2 \times 2$ table which would describe the contents completely. Table 3 shows that there are six possible states of nature which may be obtained by putting $a = 2, 3$ (= number of ABD) and $a' = 1, 2, 3$ (= number of $A'BD$).

Table 1

	A	A'	
B	4	5	9
B'	1	2	3
	5	7	12

Table 2

	A	A'	
D	3	3	6
D'	2	4	6
	5	7	12

Table 3

	A			A'		
	B	B'		B	B'	
D	a	$3 - a$	3	a'	$3 - a'$	3
D'	$4 - a$	$a - 2$	2	$5 - a'$	$a' - 1$	4
	4	1	5	5	2	7

Intersecting relevant subsets are obtained from the weakly exact probability assertion, "$P(A) = 5/12$," for which B and D are relevant, positively biased and B' and D' are relevant and negatively biased. BD and $B'D'$ are intersections of similarly biased subsets; BD' and $B'D$ are intersections of dissimilarly biased subsets. It is interesting to note that BD itself is *not* relevant, for with $a = 2$, $a' = 3$ one has $P(A \mid BD) = 2/5 < 5/12$. Thus Paul has a positive expectation if he bets A is true when B is observed, or if he bets A is true when D is observed; but he may have a negative expected gain if he bets A is true only when *both* are observed!

Is there a uniquely appropriate reference set for Peter's assertions in the last example? Use of the universal set ignores information about B and D. Use of BD, BD', $B'D$, and $B'D'$ seems inappropriate because the relative frequency of A is not known within these subsets. It would seem about equally as appropriate to use B and B' as to use D and D'. Thus there appear to be no uniquely appropriate subsets.

10. Summary and conclusions. Statistical inferences having the form, "The probability that A is true is equal to α" can be studied within the framework of a game between two players, one who makes such inferences (or *probability assertions*) and an opponent who questions their validity. The model suggests a number of criteria of validity of such inferences; four criteria which have been defined and illustrated by examples are: (i) weak exactness, (ii) strong exactness, (iii) nonexistence of relevant subsets, and (iv) nonexistence of semirelevant subsets. Definitions of these concepts, too lengthy to be repeated here, are given in Sections 2.1, 6.1, 2.2, and 2.2, respectively. Some general observations are:

(i) Weak exactness is a criterion suggested by a familiar requirement of Neyman's in the theory of confidence intervals; in the general model of Section 6 the definition is extended to apply to more general problems.

(ii) Strong exactness is a much more severe requirement which can be satisfied in classical Bayesian estimation situations, but which appears to be unreasonably demanding in most non-Bayesian problems. Those who eschew the prior distributions of the classical theory pay for weakening the classical assumptions by losing the property of strong exactness of the inferences. To mistake weak exactness for strong exactness is to attribute to an inference a more desirable property than it actually possesses. The logical fallacy is neatly stated by Sir Macklin [14]:

"Then I shall demonstrate to you,
According to the rules of Whately,
That what is true of all, is true
Of each, considered separately."

(iii) The criterion of nonexistence of relevant subsets is largely inspired by some recent work of Fisher. Various examples of relevant subsets have been given in order to provide a better understanding of their nature. Nonexistence is established only for one simple case; for much more general results the reader is referred to Wallace [20].

(iv) From the examples of Section 4 it is seen that nonexistence of semirelevant subsets is a very severe requirement indeed. One may conjecture that fiducial intervals do not induce *relevant* subsets, but from the example of Student's t it is seen that the same conjecture is not true for *semirelevant* subsets.

It is to be hoped that eventually there will be found some generally accepted notion of an "appropriate reference set" for inferences. Some readers may find that the examples of Sections 3.2, 3.3, and 3.4 indicate that the universal set is not always as appropriate as some suitably chosen subset. On this subject, Fisher ([11], p. 110) states:

> "If, therefore, any portion of the data were to allow of the recognition of such a subset, to which the predicand belongs, a different probability would be asserted using the smallest such subset recognizable."

Perhaps Fisher's idea should be formulated mathematically in terms of minimal subsets on which probabilities are known independently of the unknown state of nature. That uniqueness of appropriate reference sets might be a problem is indicated by the example of Section 9, in which two different subdivisions of the observation space give different reference sets which are about equal in merit.

The appropriate reference set has been a subject of controversy in testing situations (i.e., tests of significance and tests of hypothesis); contingency tables and regression problems (Fisher [11], pp. 82–88) are old examples. Some recent examples can be found in Cox [5] and in Cohen [3]. It is to be noted that the present development has been based largely on problems of interval *estimation*. The usual translation of criteria to *testing* situations is of course possible in many cases. Thus certain testing situations have been treated implicitly in this work; but perhaps the reader will find that whatever force the arguments may have for estimation problems is diminished in the translation to testing problems.

11. Acknowledgement. I wish to thank Prof. Wallace for giving me a draft of his paper prior to publication.

APPENDIX

Proof of the result of Section 5

To establish the result of Section 5 it will be assumed that the cumulative distribution $G(t) = \int_{-\infty}^{t} g(t)\,dt$ satisfies the mild restrictions

$$\int_{-\infty}^{0} G(t)\,dt < \infty \quad \text{and} \quad \int_{0}^{\infty} (1 - G(t))\,dt < \infty .$$

Denote by $\phi_C(x)$ the set characteristic function of the set C. Then

$$a(\theta) = P(C) = \int_{-\infty}^{\infty} \phi_C(x) g(x - \theta)\,dx$$

$$b(\theta) = P(AC) = \int_{-t_1+\theta}^{t_2+\theta} \phi_C(x) g(x - \theta)\,dx$$

Further define $\mu(R)$ by

$$\mu(R) = \int_{-R}^{R} \phi_C(x)\,dx$$

Then if $\mu(\infty)$ is finite, a straightforward calculation shows that $A(R) \rightarrow \mu(\infty)$ and $B(R) \rightarrow \alpha\mu(\infty)$ as $R \rightarrow \infty$ which establishes the desired contradiction.

If $\mu(R) \rightarrow \infty$, we may compare $A(R)$ with

$$\begin{aligned} A'(R) &= \int_{\theta=-\infty}^{\infty} \int_{x=-R}^{R} \phi_C(x) g(x - \theta)\,dx\,d\theta \\ &= \int_{x=-R}^{R} \phi_C(x) \int_{\theta=-\infty}^{\infty} g(x - \theta)\,d\theta\,dx \\ &= \int_{-R}^{R} \phi_C(x)\,dx = \mu(R). \end{aligned}$$

The difference $A(R) - A'(R)$ is made up of four integrals described (ignoring signs) by: x (or θ) in the range $-R$ to R, and θ (or x) in the range $\pm R$ to $\pm\infty$.

All four are bounded by virtue of the assumption on G. A typical calculation follows:

$$\int_{x=-R}^{R}\int_{\theta=R}^{\infty}\phi_C(x)g(x-\theta)\,d\theta\,dx \leqq \int_{x=-R}^{R}\int_{\theta=R}^{\infty}g(x-\theta)\,d\theta\,dx$$

$$= \int_{x=-R}^{R}\int_{t=-\infty}^{x-R}g(t)\,dt\,dx$$

$$= \int_{-R}^{R}G(x-R)\,dx$$

$$= \int_{-2R}^{0}G(x')\,dx'$$

Thus $A(R) = \mu(R) + K(R)$ where $| K(R) |$ is bounded as R tends to infinity. The expression for $B(R)$ may be treated as follows:

$$B(R) = \int_{\theta=-R}^{R}\int_{t=-t_1}^{t_2}\phi_C(t+\theta)g(t)\,dt\,d\theta$$

$$= \int_{t=-t_1}^{t_2}\int_{\theta=-R}^{R}\phi_C(t+\theta)g(t)\,d\theta\,dt$$

$$= \int_{t=-t_1}^{t_2}\int_{x=t-R}^{t+R}\phi_C(x)g(t)\,dx\,dt$$

$$= K'(R) + \int_{t=-t_1}^{t_2}\int_{x=-R}^{R}\phi_C(x)g(t)\,dx\,dt$$

$$= K'(R) + \alpha\mu(R)$$

where $K'(R)$ is the sum of two positive and two negative terms, each term being an integral over a region whose area is independent of R. Thus $| K'(R) |$ is bounded. Consequently the ratio

$$\frac{B(R)}{A(R)} = \frac{\alpha\mu(R) + K'(R)}{\mu(R) + K(R)}$$

tends to α and R tends to infinity, and the contradiction is established.

REFERENCES

[1] A. A. ASPIN, "Tables for use in comparisons whose accuracy involves two variances, separately estimated," *Biometrika*, Vol. 36 (1949), pp. 290–293.

[2] M. S. BARTLETT, "Comment on Sir Ronald Fisher's paper: 'On a test of significance in Pearson's Biometrika Tables (No. 11),' " *J. Roy. Stat. Soc., Ser. B*, Vol. 18 (1956), pp. 295–296.

[3] LEONARD COHEN, "On mixed single sample experiments," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 947–971.

[4] D. R. COX, "Some problems connected with statistical inference," *Ann. Math. Stat.*, Vol. 29 (1958), pp. 357–372.

[5] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc., Ser. B*, Vol. 20 (1958), pp. 215–242.

[6] Harald Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[7] R. A. Fisher, "The concepts of inverse probability and fiducial probability referring to unknown parameters," *Proc. Roy. Soc. London, Ser. A*, Vol. 139 (1933), pp. 343–348.

[8] R. A. Fisher, "Probability, likelihood, and the quantity of information in the logic of uncertain inference," *Proc. Roy. Soc. London, Ser. A*, Vol. 146 (1934), pp. 1–8.

[9] Sir Ronald Fisher, "Statistical methods and scientific induction," *J. Roy. Stat. Soc., Ser. B*, Vol. 17 (1955), pp. 69–78.

[10] Sir Ronald Fisher, "On a test of significance in Pearson's Biometrika Tables (No. 11)," *J. Roy. Stat. Soc., Ser. B*, Vol. 18 (1956), pp. 56–60.

[11] Sir Ronald A. Fisher, *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh, 1956.

[12] Sir Ronald A. Fisher, "Mathematical probability in the natural sciences," *Technometrics*, Vol. 1 (1959), pp. 21–29.

[13] D. A. S. Fraser, *Nonparametric Methods in Statistics*, Wiley, New York, 1957.

[14] W. S. Gilbert, "Sir Macklin" in *The Bab Ballads*, Macmillan, London, 1898.

[15] H. Jeffreys, "On the theory of least squares," *Proc. Roy. Soc. London, Ser. A*, Vol. 138 (1932), pp. 48–55.

[16] H. Jeffreys, "Probability, statistics, and the theory of errors," *Proc. Roy. Soc. London, Ser. A*, Vol. 140 (1933), pp. 523–535.

[17] Jerzy Neyman, *Lectures and Conferences on Mathematical Statistics and Probability*, U. S. Department of Agriculture, Washington, D. C., 1952.

[18] E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians, Vol. I*, Cambridge University Press, 1954.

[19] Richard von Mises, *Probability, Statistics and Truth*, Allen and Unwin, London, 1957.

[20] David L. Wallace, "Conditional confidence level properties," *Ann. Math. Stat.*, Vol. 30 (1959), pp. 864–876.

[21] B. L. Welch, "The generalization of 'Student's' problem when several different population variances are involved," *Biometrika*, Vol. 34 (1947), pp. 28–35.

[22] B. L. Welch, "Note on some criticisms made by Sir Ronald Fisher," *J. Roy. Stat. Soc., Ser. B*, Vol. 18 (1956), pp. 297–302.