COMMENTARIES

Commentaries are informative essays dealing with viewpoints of statistical practice, statistical education, and other topics considered to be of general interest to the broad readership of *The American Statistician*. Commentaries are similar in spirit to Letters to the Editor, but they involve longer discussions of background, issues, and perspectives. All commentaries will be refereed for their merit and compatibility with these criteria.

Nonprobabilistic Statistical Inference: A Set-Theoretic Approach

HENRY ROUANET, JEAN-MARC BERNARD, and BRUNO LECOUTRE*

The familiar sampling procedures of statistical inference can be recast within a purely set-theoretic (ST) framework, without resorting to probabilistic prerequisites. This article is an introduction to the ST approach of statistical inference, with emphasis on its attractiveness for teaching. The main points treated are unsophisticated ST significance testing and ST inference for a relative frequency (proportion).

KEY WORDS: Teaching of statistics; Probability.

1. THE PARADIGM OF SET-THEORETIC INFERENCE

It is so rooted a habit to make probability an indispensable ingredient of statistical inference that suggesting an approach to statistical inference without probability prerequisites will presumably be found rather paradoxical. Yet consider the following paradigm:

1. The data consist of one group of observations.
2. In addition, there is a known population of observations available that can serve as a reference for the data.
3. The question is raised whether the data can be considered as more or less "typical"—intuitively speaking—of the population with regard to some particular aspect, such as the mean of a numerical character of interest or the relative frequency of some attribute.

Clearly, this is a common situation. As an example, take a committee of n members, appointed from an assembly of N persons. Suppose we ask whether the committee is "typical" of the assembly—for example, with respect to the mean age or the sex ratio—or on the contrary "differs significantly" (again, intuitively speaking) from the assembly. In such examples, no probabilistic considerations are involved. In particular, the group of observations being examined is in no sense a random sample of the reference population. In many instances, it is not even a subset of it. For example, the group of observations might consist of the scores of n gifted children, and the reference population, of the scores of N normal children.

Naturally, if we want to avoid extraneous probabilistic considerations, the notions of "typical" and "significant difference" will have to be conceptualized in some novel, nonprobabilistic way. It turns out that this is perfectly feasible with a purely set-theoretic (ST) framework. The name *ST paradigm* will hence be used to refer to the preceding paradigm. An ST inferential procedure will be similar to a conventional one, except that no notion of randomness is involved. Indeed, one can recast all familiar sampling procedures of statistical inference within an ST framework. At a more advanced level, the viewpoint of ST inference will also directly apply to permutation test theory, especially rank test theory. In fact, permutation procedures, like elementary sampling procedures, can easily be dissociated from probabilistic considerations, in this specific case, randomization assumptions. (We intend to investigate the relations between ST inference and permutation tests more thoroughly in a future paper.)

What about the novelty of the ST approach? Nonprobabilistic formulations can certainly be found in standard textbooks, such as "95 percent of calculated confidence intervals will cover the parameter's true value" (a typical ST formulation, as we shall see). Nonetheless, such sentences appear isolated in textbooks, which almost universally stress the necessity of a probabilistic framework. One notable exception known to us is Faverge's (1956) textbook (a classic among French-speaking psychologists). This is why we feel that ST inference surely deserves a clear-cut, explicit presentation.

Here we will present a few examples of the ST approach to statistical inference, with emphasis on its use in an introductory statistical course. We first describe an example of unsophisticated ST inference (Sec. 2), which will lead to some general comments (Sec. 3). Then combinatorial ST inference is illustrated by ST significance testing and confidence methods for a relative frequency (Sec. 4). The ex-

*Henry Rouanet, Jean-Marc Bernard, and Bruno Lecoutre are researchers at the Groupe Mathématiques et Psychologie, Université René Descartes, Sciences Humaines–Sorbonne, 12 rue Cujas, 75005 Paris, France. This article is an outgrowth of a paper presented at the International Conference on Teaching Statistics held in Sheffield, England, in 1982. The authors are especially grateful to Alan R. Hoffer, who first encouraged the writing of this article, and to Vincent Duquenne, Marie-Claude Bert, David Hand, Colin Taylor, and P. O. White for their helpful comments and suggestions.

tension to infinite sample spaces is then outlined (Sec. 5). After a short discussion, practical teaching considerations are presented (Sec. 6).

2. UNSOPHISTICATED ST INFERENCE: AN ST SIGNIFICANCE TEST

Unsophisticated ST inference readily follows from the ST paradigm. Suppose that we want to compare a group of n numerical observations to a reference population of size N with respect to the mean. We may proceed along the following steps.

1. Define an ST sample of size n from the population (or in brief, a sample) as an n-element subset of the set constituted by the population. The set X of all $\binom{N}{n}$ samples will be called the ST sample space.

2. Consider the mapping $M: X \to R$ that associates with each sample x its mean $M(x)$ (M is the mean statistic). Let m denote the mean of the group of observations being examined. Whenever the mean $M(x)$ of a sample x is such that $M(x) \geq m$, the sample will be said to satisfy the property on X denoted by $(M \geq m)$.

3. Let $\bar{p} = \text{Prop}(M \geq m)$ be the proportion of the samples of X that satisfy the property $(M \geq m)$. This proportion can be taken as a directional index of departure of the group of observations from the reference population, with respect to the mean. For any given $\alpha \in [0, \frac{1}{2}]$ (a specified significance level), if $\bar{p} \leq \alpha$, we will say that in an ST sense, the departure of the group of observations from the reference population, with respect to the mean, is upwardly significant at the level α. The proportion $\bar{p}$ is clearly the greatest specified level rendering the departure upwardly significant and will be called the *observed upper level* of the test.

Example. Let a group of $n = 3$ observations with mean $m = 39$ be compared to the following numerical population of size $N = 9$:

$$(31; 31; 34; 34; 37; 37; 37; 40; 43).$$

There are $\binom{9}{3} = 84$ samples (i.e., subsets) of size 3; their means generate the ST sampling distribution of the statistic M, as shown in Figure 1.

By inspection, it is found that out of the 84 samples, 8 satisfy the property $(M \geq 39)$, hence $\bar{p} = 8/84$. For any $\alpha \geq 8/84$, the departure of the group of observations—indeed any group of observations with mean $m = 39$—from the reference population, with respect to the mean, is upwardly significant at the level α.

Downwardly significant departures and observed lower-level $\underline{p} = \text{Prop}(M \leq m)$, or again two-sided significant departures, will be defined along the same lines.

The foregoing construction exemplifies ST significance testing (directional or absolute). Clearly the construction applies to any numerical statistic. To provide an answer to the ST paradigm, we consider all ST samples of the reference population having the same number of observations as the data. We take as an index of departure from the reference population the proportion of samples that are more extreme than the data, with respect to this statistic, either in a directional or in an absolute way.

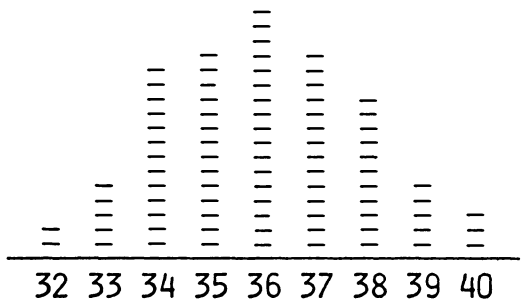

Figure 1. ST-Sampling Distribution of the Statistic M.

Fundamental Significance Property. Given an ST significance test, the procedures it defines can be applied to every sample of X, viewed as a particular set of observations. For any specified level α, the test will separate out those samples whose departure is significant at the level α. The definition of the ST significance test implies that the proportion of such samples is at most α. This will be called the *fundamental ST significance property*. Of course, the qualification "at most α" rather than "equal to α" is due to the discreteness of the distribution.

3. FIRST COMMENTS ABOUT ST INFERENCE

3.1 ST Inference and Conventional Probabilistic Inference; the Conversion Property

The link between the foregoing ST notions and those of conventional inference is apparent: for example, an ST sample is simply the familiar unordered sample without replacement from a (finite) population, except that for each property of interest, there is no probability attached to it and we consider instead the proportion of samples that satisfy this property. Of course, the conventional probabilistic inference can be "recovered" from ST inference if we include appropriate assumptions. Suppose that the group of observations being examined is one of the $\binom{N}{n}$ samples of some unknown parent population of size (N) and that prior to the experiment, all $\binom{N}{n}$ samples have been assigned equal probabilities, either by assumption or by design (random sampling). Now let H_0 be the following hypothesis: "the parent distribution coincides with the reference distribution." It is easily seen that $\text{Prop}(M \geq m)$ becomes the familiar probability under H_0 that the test statistic M (now a random variable on X) is greater than or equal to the observed mean m. This conversion property—from proportions to probabilities—allows conventional probabilistic inference to be firmly attached to ST inference, since under appropriate randomness assumptions, ST procedures will produce probabilistic procedures.

3.2 ST Inference and Descriptive Statistics; the Notion of α Typicality

On the other side, ST significance testing appears to be the direct extension, for $n \geq 1$, of the familiar descriptive procedure that consists of evaluating the degree of typicality of a given subject, vis-à-vis a population of numerical scores, by the proportion of scores exceeding the score of this subject. In this procedure, the population is simply used as a

reference, and no probabilistic judgment is implied. The notion of typicality of a group of observations, vis-à-vis a reference population, with respect to a statistic of interest, which we took as an intuitive starting point of the discussion, can thus be conceptualized in a natural way as the converse to ST significance (whether directional or absolute). A group of observations whose departure is not α significant will be said to be α typical of the reference population, with respect to the statistic of interest—whether upwardly, downwardly, or absolutely—and the observed level can be taken as a degree of typicality of the group of observations. For instance, a committee elected within an assembly may be found to be 10% typical of the assembly with respect to the age, for the two-sided ST test of the mean discussed previously. From the fundamental significance property, for any ST significance test, the proportion of samples of a population that are α typical is at least $1 - \alpha$.

As a conclusion, although technically ST inference will produce procedures akin to those of conventional probabilistic inference, conceptually it is a straightforward extension of descriptive statistics, to which it obviously reduces for $n = 1$. The proportion and typicality terminology introduced here are in harmony with those of "typical values" and quantiles of empirical distributions.

4. COMBINATORIAL ST INFERENCE: ST SIGNIFICANCE TESTING AND CONFIDENCE METHODS FOR A RELATIVE FREQUENCY

With categorized data, combinatorial techniques can be used, leading to explicit formulas. As an illustration, we describe ST significance testing and confidence methods for a relative frequency (proportion).

4.1 ST Significance Testing: Comparing an Observed Relative Frequency to a Reference Value ϕ_0 for a Population of Size N

For a data set of size n, let $f = k/n$ be the observed relative frequency of a character of interest—that is, k out of the n observations possess this character. Suppose that in a reference population of size N, K individuals possess this same character. If we call $\phi_0 = K/N$ a *reference value* (for f), the significance test here will amount to "comparing the observed relative frequency f to the reference value ϕ_0" (for a population of size N).

Let F: $X \to [0, 1]$ be the mapping that associates with each sample x the corresponding relative frequency $F(x)$. Using combinatorial (not probabilistic) reasoning, one finds that the number of samples for which the relative frequency is equal to k/n is $\binom{K}{k} \times \binom{N-K}{n-k}$. Hence the observed upper level

$$\bar{p} = \text{Prop}(F \geq k/n) = \sum_{k'=k}^{n} \binom{K}{k'} \binom{N-K}{n-k'} \Big/ \binom{N}{n}$$

Numerical Example. $n = 5, f = 4/5 = .80, \phi_0 = .30$, and $N = 20$. There are $\binom{20}{5} = 15{,}504$ samples and, among them, $\binom{6}{4} \times \binom{14}{1} + \binom{6}{5} \times \binom{14}{0} = 216$, for which the value of the relative frequency is equal to or greater than $4/5$. Hence $\bar{p} = 216/15{,}504 = .0139$. For any specified $\alpha \geq .0139$, the observed relative frequency, $f = 4/5$, is significantly higher, in an ST sense, than the reference value $\phi_0 = .30$ (for a population of size $N = 20$).

Similarly, the observed lower level is found to be $\underline{p} = .9996$; hence the observed two-sided level $p = 2 \min(\bar{p}, \underline{p}) = 2 \times .0139 = .0279$. Thus for any $\alpha \geq .0279$, $f = 4/5$ differs significantly (absolutely speaking) from $\phi_0 = .30$ (for $N = 20$); and conversely, for any $\alpha < .0279$, f is α typical of ϕ_0, and so forth. All numerical results, of course, are those of the classical hypergeometric test, where probabilities are replaced by proportions.

4.2 ST Confidence Limits for a Relative Frequency: The Notion of α Compatibility

A given observed relative frequency $f = k/n$ can be compared by the preceding tests to every one of the $N + 1$ values belonging to the parameter set $\Phi = \{0, 1/N, \ldots, (N-1)/N, 1\}$. Thus for the upper test, we define $f \bar{S}_\alpha \phi \Leftrightarrow \text{Prop}^\phi(F \geq f) \leq \alpha$, the superscript ϕ being a reminder that the calculated proportions of samples do indeed depend on ϕ. The notation $f \bar{S}_\alpha \phi$ will be read "f is significantly higher than ϕ at the level α." Conversely, we define $f N\bar{S}_\alpha \phi \Leftrightarrow \text{Prop}^\phi(F \geq f) > \alpha$.

Let us now envisage all pairs $(f, \phi) \in F \times \Phi$, where $F = \{0, 1/n, \ldots, (n-1)/n, 1\}$. The property $f \bar{S}_\alpha \phi$ defines a binary relation on the cartesian product $F \times \Phi$, which will be called the *upper significance relation* at the level α. Now for a given f, this relation will discriminate between those ϕ values for which $f \bar{S}_\alpha \phi$ holds and those for which the converse relation $f N\bar{S}_\alpha \phi$ holds.

It can easily be shown that the set $\{\phi | f N\bar{S}_\alpha \phi\}$ lies above the set $\{\phi | f \bar{S}_\alpha \phi\}$. The smallest element of the former set will be called the *ST lower confidence limit* for ϕ given f at the level α (or with confidence $1 - \alpha$). Denoting this limit as $\underline{l}_\alpha(f)$, we thus have by definition:

$$\underline{l}_\alpha(f) = \min_{\phi \in \Phi} \{\phi | f N\bar{S}_\alpha \phi\}.$$

The following equivalences readily follow:

$$f \bar{S}_\alpha \phi \Leftrightarrow \underline{l}_\alpha(f) > \phi$$

and

$$f N\bar{S}_\alpha \phi \Leftrightarrow \underline{l}_\alpha(f) \leq \phi.$$

The set $\{\phi | f N\bar{S}_\alpha \phi\} = \{\phi | \underline{l}_\alpha(f) \leq \phi\}$ will be called the *ST lower confidence region* (for ϕ given f) at the level α (or with confidence $1 - \alpha$).

Numerical Example. $n = 5$, $N = 20$, and $\alpha = .05$. Here we have $F = \{0, 1/5, \ldots, 4/5, 1\}$ and $\Phi = \{0, 1/20, \ldots, 19/20, 1\}$. The upper significance relation at the level $\alpha = .05$ will be constructed by determining, for each of the 21 ϕ values, the set of f values that are upper significant at this level. In this way we get the upper left part of the diagram of Figure 2, thus constructed "vertically." Then reading the diagram "horizontally," we easily determine, for each of the six values of f, the corresponding lower confidence limits at the level .05, namely $(0, 1/20, 2/20, 5/20, 8/20, 12/20)$.

The notion of lower significance relation, leading to those

of upper confidence region and upper limit, will be defined similarly. For the numerical example, they appear in the lower right part of the diagram of Figure 2. Finally, a balanced confidence interval at the (two-sided) level α will be defined as the set of ϕ values for which the observed relative frequency f is both upper and lower typical of ϕ at the level $\alpha/2$. All of these notions are pictured on Figure 2 for the two-sided level $\alpha = .10$.

Again all numerical results coincide with conventional ones. In Figure 2, the reader will have recognized the well-known hypergeometric confidence belts expressed in ST terms.

The Language of Compatibility. As is apparent from the preceding discussion, ST confidence relations are just typicality relations read in the reverse order, that is, from f to ϕ. As a specific name to this notion, we suggest that of compatibility. For example, the value ϕ will be said to be lower α compatible with f iff f is upper α typical of ϕ, that is, if $fN\bar{S}_\alpha\phi$. Special notation would also be useful. We suggest the notation $\underline{C}_\alpha$, with

$$\phi\underline{C}_\alpha f \Leftrightarrow fN\bar{S}_\alpha\phi \Leftrightarrow \phi \geq \underline{l}_\alpha(f)$$

(read "ϕ is lower compatible with f at level α iff etc.").

Remark. When for a pair (f, ϕ), $\text{Prop}^\phi(F \geq f) = 0$, ϕ can be said to be strictly lower compatible with f. In this case, the relation $f\bar{S}_\alpha\phi$ is trivially satisfied for any $\alpha \geq 0$ (in Fig. 2, it is represented by small stars). Thus α incompatibility can be regarded as an extension, for $\alpha > 0$, of the strict incompatibility notion.

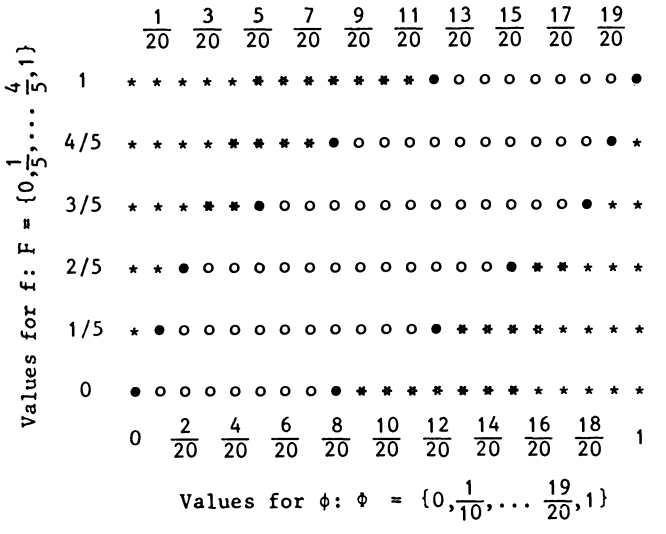

Figure 2. Diagram of Significance and Confidence (or compatibility) Relations at the (balanced) Two-Sided Level $\alpha = .10$, for $n = 5$ and $N = 20$. Significance (*) is read vertically, compatibility (○) is read horizontally.

Confidence Limit Statistics and Fundamental Confidence Property. The value $\underline{l}_\alpha(f)$ can effectively be calculated from the observed relative frequency f, regardless of the value of ϕ. Therefore, the mapping from X to Φ, which associates with each sample x the limit $\underline{l}_\alpha(F(x))$, is a statistic that will be called the *lower confidence limit statistic* and denoted (as a function of the F statistic) by $\underline{l}_\alpha(F)$. The upper confidence limit $\bar{l}_\alpha(F)$ will be defined similarly.

The fundamental significance property (Sec. 2), when applied to the upper test for a relative frequency, reads: for any $\phi \in \Phi$, $\text{Prop}^\phi(F\bar{S}_\alpha\phi) \leq \alpha$. Now owing to the equivalence $(F\bar{S}_\alpha\phi) \Leftrightarrow (\underline{l}_\alpha(F) > \phi)$, this property can be stated in terms of the lower limit statistic, yielding $\text{Prop}^\phi(\underline{l}_\alpha(F) > \phi) \leq \alpha$, or $\text{Prop}^\phi(\underline{l}_\alpha(F) \leq \phi) \geq 1 - \alpha$. Under this last form, the property will be called the *fundamental confidence property* (for the lower confidence limit). Similar properties will be stated for upper and absolute confidence. For instance, for a balanced confidence interval at the level α, the fundamental confidence property will read:

$$\text{For any } \phi \in \Phi, \text{Prop}^\phi([\underline{l}_{\alpha/2}(F), \bar{l}_{\alpha/2}(F)] \ni \phi) \leq 1 - \alpha.$$

Or written out fully: For any $\phi \in \Phi$, the proportion of confidence intervals at the level α that contain ϕ is at least $1 - \alpha$ (again, "at least $1 - \alpha$" stands instead of "equal to $1 - \alpha$" due to the discreteness of the distribution).

5. INFINITE SAMPLE SPACE AND MEASURE-THEORETIC INFERENCE

It should not be inferred from the preceding sections that ST inference is confined to a finite sample space. Indeed, ST inference can be extended to cover all procedures involving an infinite sample space. This will be done by extending the basic notions and making appropriate use of classical convergence theorems about distributions, employed as purely mathematical tools.

As a first example, take the comparison of an observed relative frequency $f = k/n$ with a reference value $\phi_0 \in [0, 1]$, when no population size N is specified. We may consider a sequence of populations indexed by ν, with an ordered pair of integers (K_ν, N_ν), such that as ν tends toward infinity, $K_\nu \to +\infty$, $N_\nu \to +\infty$, and $\phi_\nu = K_\nu/N_\nu \to \phi_0$. For each ν, the hypergeometric ST procedure described earlier applies, yielding $\text{Prop}^{\phi_\nu}(F \geq k/n)$. When $\nu \to +\infty$, $\text{Prop}^{\phi_\nu}(F \geq k/n)$ converges to the familiar binomial expression

$$\sum_{k'=k}^{n} \binom{n}{k'} \phi_0^{k'} (1 - \phi_0)^{n-k'}.$$

This expression can be taken as defining the proportion $\text{Prop}^{\phi_0}(F \geq k/n)$ when the population size N is not specified. Intuitively, the procedure will be thought of as an inference for an arbitrarily large population. We may continue to call $\text{Prop}^{\phi_0}(F \geq k/n)$ a proportion of samples, even though the total number of samples is not finite in the limit.

In the general case, the basic notion will be that of a sample from an ST distribution, as opposed to a sample from a finite population of the elementary case. An ST distribution will be conceptualized as a measure space (U, Π), where Π is a positive measure of unit total mass over the measurable space U. Then an ST sample of size n from the ST distribution (U, Π) will be defined as an element of

the product measure space $(U^n, \Pi^{(n)})$, and a property of the sample space will be characterized by a measurable subset of U^n. All of these notions are measure-theoretic in character, thus ST inference in the general case might appropriately be called *measure-theoretic inference*. Formally, measure-theoretic inference will be equivalent to conventional probabilistic inference, since a sample from an ST distribution is equivalent to the familiar ordered sample from a probability distribution. On the other hand, the formulations will be close to those of the elementary case if we carry over the language of proportions and speak of the $\Pi^{(n)}$ measure of a property of the sample space as the "proportion of samples" that verify this property.

As an example, let us rephrase in ST terms the familiar property of the Student ratio $T = (M - \mu)/(S/\sqrt{n})$: Among the n samples of a normal distribution of mean μ, the proportion of those for which the Student ratio exceeds t_α is equal to α, where t_α denotes the upper α percentage point of the Student distribution; that is, for any $\mu \in R$, $\text{Prop}^{\mu}(T > t_\alpha) = \alpha$. Using such a formulation, we may compare, in ST terms, a group of observations with a normal distribution of a given mean.

6. DISCUSSION AND TEACHING CONSIDERATIONS

Every procedure of statistical inference, such as a significance test, has two aspects:

1. an *algorithm* (properly speaking) that tells us how to perform the procedure, for example, the calculation of a test statistic from the data and the checking of its value against a standard distribution, and
2. a *probabilistic framework* that tells us how to justify and interpret the procedure.

Those two aspects involve two different types of logic: computing an observed significance level is one thing; assessing, for example, the probability of an erroneous rejection conclusion is another thing. ST inference makes the point that the algorithm can be completely dissociated from the probabilistic framework.

In our opinion, there are numerous situations in which the ST framework makes perfect sense by itself. Resuming an example mentioned earlier, let us assume that from an assembly of $N = 20$ members, 14 of them women, a committee of 5 members is appointed, of which only 1 is a woman. Such a committee is liable to be disqualified on the ground that it is not typical with respect to the sex ratio; the fact that it is not a random sample is irrelevant. In addition, it is a notorious fact that in actual practice, inference procedures are commonly used even when probabilistic assumptions are not seriously founded. By offering the possibility of a nonprobabilistic interpretation, ST inference in a sense simply preaches what makes sense of actual practice.

We now turn to the specific teaching aspects of ST inference. The difficulties of interpreting probabilistic statements in statistical inference are well known, if only because the specification of the space to which such statements pertain is too easily omitted or forgotten. The teaching of ST inference allows the student to concentrate on learning algorithms without being prematurely concerned by the difficulties of a probabilistic phraseology. Statements in terms of proportions compel one to spell out the relevant sets of objects. Conceivably, the teaching of ST inference can be undertaken immediately after descriptive statistics, without probability prerequisites.

For more than three years now, we, along with other colleagues at our university, have gradually introduced ST inference in courses and seminars for audiences of various backgrounds. In what follows, we briefly describe the place ST inference has come to occupy in the three-year undergraduate curriculum for psychology students (with which the senior author has been involved for several years).

1. The first-year course deals exclusively with descriptive statistics. Standard continuous distributions (such as the normal distribution) are introduced at this stage, not as probability distributions, but as conceptual extensions of observed frequency distributions. With such distributions, proportion formulations are used in a natural way, preparing the student for their use in ST inferences (e.g., see Lecoutre and Lecoutre 1979).
2. The second-year course is divided into two parts of equal importance. The first part is an introductory probability course of the usual type, stressing the use of probability for evaluating the uncertainty of unknown events. The second part is an introduction to statistical inference. Unsophisticated ST inference is first developed, along the lines of Section 2 of this article. The enumerating and counting operations are conveniently carried out by computer programs, enabling the experimental investigation of various ST sampling distributions. ST inference on relative frequencies follows. Significance testing and confidence limits are presented and discussed at length on examples of the kind discussed in Section 4, again resorting to computers for the computations. Finally, probabilistic inference is introduced, using the conversion property to transform ST formulations into probabilistic ones (Sec. 3.1). At this point, probabilistic inference appears as a synthesis of probability and ST inference.
3. The third-year course is mainly concerned with normal inference techniques (comparisons between means, etc.). Here again, algorithms and ST formulations are presented first, followed by probabilistic inference. Experience has definitely shown that carrying over the proportion formulations to infinite sample spaces comes quite naturally to students.

Remarks

1. With the ST approach, the asymmetry between parameter and observations is apparent, since there is no obvious proportion of populations to match the proportion of samples. As a consequence, the standard mistake of interpreting observed significance levels or confidence limits in terms of inverse probabilities is avoided. (On the other hand, it is worth mentioning that at an advanced stage, the Bayesian approach appears to be more clearly understood.)
2. The crucial step of the second year is based on a very limited set of theoretic notions: subset of a set, Cartesian product, binary relation, and mapping from one set to an-

other. All of these notions, which are taught at high school level in our country, are taken up again in the first-year course.

Yet this does not mean that teaching ST inference is a trivial matter. Progression in teaching has to be cautious and slow. Much care has to be taken with the notations used. For example, we tried several alternatives to denote significance and related relationships. Instead of $f\bar{S}_\alpha\phi$ and $f\underline{S}_\alpha\phi$, we tried $fS'_\alpha\phi$ and $fS''_\alpha\phi$, as well as $f\gg_\alpha\phi$ and $f\ll_\alpha\phi$. The notations $f\bar{S}_\alpha\phi$ and $f\underline{S}_\alpha\phi$ were found to be the best ones, perhaps because they can be read as shorthand and require little mathematical sophistication.

Again, the use of colored transparencies has proven efficient. When figures like Figure 2 are drawn with green and red dots, the shape of confidence belts emerges at first sight and provides insights into the influence of α level, sample size, and so forth.

In conclusion, our teaching experience with ST inference is now firmly established. Practical examples and exercises have been tried out, and we are now engaged in preparing extensive material for publication.

[*Received January 1984. Revised August 1985.*]

REFERENCES

Faverge, J.-M. (1956), *Méthodes Statistiques en Psychologie Appliquée*, Paris: Presses Universitaires de France.

Lecoutre, M.-P., and Lecoutre, B. (1979), *Enseignement Programmé sur l'utilisation d'une Table de la Distribution Normale*, Paris: Editions C.D.U.-S.E.D.E.S.

A History of the Development of Craig's Theorem

MICHAEL F. DRISCOLL and WILLIAM R. GUNDBERG, JR.*

Craig's theorem on the independence of quadratic forms in normal variates is traced from its first form, for iid standard normal variates, to the form for variates following an arbitrary nonsingular joint normal distribution. This article gives the main thrust of the development and makes recommendations on coverage of the theorem in courses and textbooks. The history of Craig's theorem is not a happy one. The authors of the earlier articles in the literature tended to make errors of a linear-algebraic nature. Authors of more recently published textbooks have given incorrect or misleadingly incomplete coverage of Craig's theorem and its proof.

KEY WORDS: Quadratic forms; Independence; Normal variates.

1. INTRODUCTION

This article gives a history of Craig's theorem (Craig 1943) on the independence of quadratic forms in a normal vector. The theorem states that

$$\text{For } x \sim N(\mu, V),\ x'Ax \text{ and } x'Bx \text{ are stochastically independent iff } AVB = 0. \quad (1)$$

As usual, $x \sim N(\mu, V)$ denotes that x is a random vector following a fixed multivariate normal distribution with mean vector μ and covariance matrix V; V is assumed to be nonsingular and thus positive definite (the singular case is beyond the scope of our treatment). The matrices A and B are taken to be real and symmetric (special subcases, such as A or B nonnegative definite, are omitted from our presentation).

The sufficiency part of (1) is a central tool in the theory and application of linear models. The proof that $AVB = 0$ is in fact a sufficient condition for independence is straightforward. All that needs to be done is to show that when $AVB = 0$, the joint moment-generating function of $x'Ax$ and $x'Bx$ (e.g., see Searle 1971, chap. 5, Lemma 10) is the product of their marginal moment-generating functions. Since the proof of sufficiency is just a direct application of the factorization criterion, we concentrate on the proof of necessity.

The necessity part of Craig's theorem is of little importance in applied statistics, but it is important to theoreticians precisely because it establishes $AVB = 0$ as a characterization of independence. The proof of necessity is difficult—disturbingly so, in view of the simplicity of the statement of the result (1). As we shall see, the difficulty of the proof has been a source of error for many authors, past and present.

A correct proof for the general case of necessity requires the use of results from the theory of functions of complex variables (and depending on the approach taken, also from algebraic field theory). Consequently, the proof is inaccessible to many statisticians. Unfortunately, there is some evidence that a more accessible proof does not exist.

Our work is motivated by a desire for mathematical completeness. It should remove some of the persistent mystery that accompanies Craig's theorem and provide a passkey for those who wish to understand its underpinnings.

In the next section we outline the growth of Craig's theorem from the central $N(0, I)$ and $N(0, V)$ cases to the threshold of the noncentral $N(\mu, I)$ and $N(\mu, V)$ cases. Section 3

*Michael F. Driscoll is Associate Professor, Department of Mathematics, Arizona State University, Tempe, AZ 85287. William R. Gundberg, Jr., is Student Actuary and Plan Administrator, Arizona Qualified Plan Services, Inc., Phoenix, AZ 85012. The authors thank S. R. Searle for his prompt and generous reply to their inquiries about his 1971 proof and for suggestions on revision of a draft of this article, and R. V. Hogg for his comments about A. T. Craig's work on proving (1).