A DISCIPLINE FOR THE AVOIDANCE OF UNNECESSARY ASSUMPTIONS [1])

LEWIS H. ROBERTS
New York

Introduction

Although unnecessary assumptions are something we all try to avoid, advice on how to do so is much harder to come by than admonition. The most widely quoted dictum on the subject, often referred to by writers on philosophy as "Ockham's razor" and attributed generally to William of Ockham, states "Entia non sunt multiplicanda praeter necessitatem". (Entities are not to be multiplied without necessity.) As pointed out in reference [1], however, the authenticity of this attribution is questionable.

The same reference mentions Newton's essentially similar statement in his Principia Mathematica of 1726. Hume [3] is credited by Tribus [2c] with pointing out in 1740 that the problem of statistical inference is to find an assignment of probabilities that "uses the available information and leaves the mind unbiased with respect to what is not known." The difficulty is that often our data are incomplete and we do not know how to create an intelligible interpretation without filling in some gaps. Assumptions, like sin, are much more easily condemned than avoided.

In the author's opinion, important results have been achieved in recent years toward solving the problem of how best to utilize data that might heretofore have been regarded as inadequate. The approach taken and the relevance of this work to certain actuarial problems will now be discussed.

Bias and Prejudice

One type of unnecessary assumption lies in the supposition that a given estimator is unbiased when in fact it has a bias. We need not discuss this aspect of our subject at length here since what we

[1]) Originally presented at the seminar on Mathematical Theory of Risk and allied topics, auspices of the Committee on Mathematical Theory of Risk, Casualty Actuarial Society, November 16, 1966.

might consider the scalar case of the general problem is well covered in textbooks and papers on sampling theory. Suffice it to say that an estimator is said to be biased if its expected value differs by an *incalculable* degree from the quantity being estimated. Such differences can arise either through faulty procedures of data collection or through use of biased mathematical formulas. It should be realized that biased formulas and procedures are not necessarily improper when their variance, when added to the bias, is sufficiently small as to yield a mean square error lower than the variance of an alternative, unbiased estimator.

As an example of bias due to sampling procedure, suppose we sample a population in a non-random, haphazard manner so that probabilities of selection vary in an unknown way. There is no method by which to calculate the difference between the expected value of the mean of such a sample and the mean of the population. Hence, the sample mean is a biased estimator. On the other hand, if probabilities of selection are known, appropriate weighting will provide an unbiased estimator. An example of bias due to choice of mathematical formula is the use of ratio-estimates, as where the ration of y to x obtained by sampling is multiplied by a known population total of x to estimate the population total of y. The combined bias and standard error of a ratio estimate is often less, however, than the standard error of the best alternative unbiased estimate. An estimator is not considered to be biased if there is any way of removing the bias. Thus, the sum of the means of random samples of x and of y is considered to be an unbiased estimator of the expected value of x if we know the expected value of y. This is because we can subtract the latter quantity leaving $\bar{x} + \bar{y} - Ey$, the expected value of which is clearly Ex.

Our concern here is not primarily with point estimations but with complete statistical distributions. We shall consider any distribution function characterized by parameters *or form* not directly derived from the data as" prejudiced". This seems an apt characterization since different analysts may derive different functions from a given set of data if they go beyond the data in their specifications. These differences can inferentially be imputed to differing personal prejudices (perhaps unconscious) in favor of one function over another.

It is shown by Jaynes and Tribus that the assignment of the p_i for which S is at a maximum (K being an arbitrary constant) is

$$p_i = \exp.\ [-a_0 - a_1 g_1(x_i) - a_2 g_2(x_i) - \ldots .] \tag{3}$$

in which the a's are Lagrangian multipliers satisfying the requirements of $\bar{g}_r(x)$ and

$$a_0 = \ln \sum_i \exp.\ [\sum_r a_r g_r(x_i)\ ] \tag{4}$$

while

$$\bar{g}_r(x) = -\partial a_0/\partial a_r \qquad = \text{mean of } g_r(x)$$

$$\text{Var.}\ [g_r(x)\ ] = \partial^2 a_0/\partial a_r^2 \qquad = \text{variance of } g_r(x)$$

and

$$S = K a_0 + K \sum_r a_r\, g_r(x)$$

Specific Derived Distributions

Known Data	*Distribution with Maximum Entropy*
Range	Uniform
$\sum_{i=1}^{m} p_i = 1$	$p_i = \exp.\ (-a_0) = \frac{1}{m}$
Mean*	Exponential
$\sum_{i=0}^{\infty} p_i x_i = \bar{x}$	$p_i = \exp.\ (-a_0 - a_1 x_i)$
Mean and variance*	Truncated Gaussian
$\sum_{0}^{\infty} p_i x_i = \bar{x}$	$p_i = \exp.\ (-a_0 - a_1 x_i - a_2 x_i^2)$
$\sum_{0}^{\infty} p_i x_i^2 = x^2$	

* $\sum_{i=0}^{\infty} p_{ii} = 1$

Success in these areas suggested that valid applications might be found in the area of statistical inference [2d, 2e]. Shannon's measure, which he called the "entropy" or "uncertainty" of a distribution, is defined by:

$$S = -K\Sigma p_i \ln p_i \tag{1}$$

where p_i is the probability associated with the i'th discrete possibility and the summation is taken over all possibilities having non-zero probability. K is an arbitrary scaling factor. "Ln" refers to natural logarithms although inclusion of a scaling factor would permit use of logarithms to any base.

An amusing sidelight on the naming of this measure is related by Tribus [2f]:

> When Shannon discovered this function he was faced with the need to name it, for it occurred quite often in the theory of communication he was developing. He considered naming it "information" but felt that this word had unfortunate popular interpretations that would interfere with his intended uses of it in the new theory. He was inclined towards naming it "uncertainty" and discussed the matter with the late John Von Neumann. Von Neumann suggested that the function ought to be called "entropy" since it was already in use in some treatises on statistical thermodynamics . . . Von Neumann, Shannon reports, suggested that there were two good reasons for calling the function "entropy". "It is already in use under that name," he is reported to have said, "and besides, it will give you a great edge in debates because nobody really knows what entropy is anyway." Shannon called the function "entropy" and used it as a measure of "uncertainty," interchanging the two words in his writings without discrimination.

Shannon showed that this measure is unique in satisfying the following criteria:

(a) It should depend only upon the probability distribution, i.e., S is a function of $p_1, p_2 \ldots p_n$.
(b) If all of the p_i are equal, then $p_i = 1/n$ and S is a monotonically increasing function of n.
(c) The measure should be consistent in the sense that if we con-

sider events A and B in the context of a state of knowledge X, then we should have

$$S(AB|X) = S(A|BX) + S(B|X)$$

That is, the entropy ascribed to A and B jointly in the context of X equals the entropy that would be ascribed to A in the context of B and X plus the entropy that would be ascribed to B alone in the context of X. This parallels the law of compound probabilities.

Formal Results

Defining the minimally prejudiced distribution function as that for which S is at a maximum, let us look at the derivations of some familiar distributions. These problems will be characterized by the information available and the solution derived by maximizing S. We assume that nothing whatever is known about each distribution beyond what is stated. In practice there might be additional, non-quantitative data that would preclude use of the functions derived here in certain cases. Derivation of the minimally prejudiced distribution subject to common qualitative constraints would be an important extension of presently known results.

In a wide variety of problems, available information may be in the form of averages such as the mean first power, mean square, mean cube, etc. of the variate x. The following results would apply to means of any single-valued continuous functions, for example trigonometric or logarithmic functions, as well as to the usually reported integral power functions. We can denote these various means as

$$\bar{g}_r(x) = \Sigma p_i g_r(x_i) \qquad (2)$$

where $r = 1, 2, 3 \ldots m$ for m different functions of x and $\Sigma p_i = 1$.

The measure just presented enables us to compare statements about a distribution in such a way that we can select that one among all satisfying the given data which, by virtue of maximum entropy, best complies with Ockham's dictum in the sense of asserting the least information. As noted by Tribus, "By using this principle, the observer reduces his subjectivity to the minimum possible value." In problems where this procedure inevitably leads different analysts to the same result, the author considers that subjectivity,

or prejudice, has been reduced to zero. The only challenge that might be made to this claim would seem to rest upon the degree of subjectivity entailed in adopting the principle of maximum entropy as a criterion in the first place. Whether the case for adoption of this principle is so overwhelming as to remove all possibility of subjectivity on that point (so that its rejection is outright error) will not be argued here. It does seem clear, however, that as between persons who adopt the principle as a convention, there is no room for personal prejudice. This alone is a strong recommendation for any convention not demonstrably in error.

We now make certain observations concerning Shannon's measure:

1. If the logarithm is taken to base 2 (rather than to the base e) S is equal to the expected number of questions in a taxonomic game, such as Twenty Questions, that would be needed to remove all doubt. [2b]
2. In general, S is a measure of the "flatness" of a distribution, hence of the relative equality with which probabilities are assigned. This is consistent with the intuitive notion that event A should not be assumed, without reason, to be more likely than event B. (It seems obvious that consistent results cannot be expected if probabilities are assigned whimsically.)
3. The measure is differentiable, hence can be maximized by classical methods (i.e., without resort to linear programming or other iterative procedures) to yield minimally prejudiced functions as extremals.
4. The fact that the measure employs a summation of probabilities, rather than an integral, apparently precludes its use in problems that require continuous distributions. Yet, the class of phenomena involving only a finite number of particles and the emission or absorption of discrete quanta of energy may be sufficiently broad as severely to limit, if not to rule out, the occurrence of physical events for which continuous distributions are strictly appropriate. Physical considerations aside, the digitalization of measurements converts data representing even theoretically continuous distributions into discrete form. This author does not see it as a flaw, therefore, that the measure of entropy has not been defined for continuous distributions.

While we presumably exercise no conscious favoritism for one one type of distribution function over another and we test all plausible choices impartially, we are necessarily limited to those functions with which we are familiar and which we can handle mathematically. The phenomena we study are not necessarily so constrained. In some problems, however, we are fortunate in that the data include information that a process is involved which can produce only a particular kind of distribution, so there is no possibility of prejudice.

The Logical Inconsistency of Prejudice

Let us suppose that data X imply conclusions C_X. Let us suppose, further, that we do not quite know how to interpret X and cannot draw any conclusion unless we assume Y also to be true. Then we draw the conclusion C_{XY} and tender it as C_X. That this is clearly a false coin is seen when someone else similarly finds it necessary to make an assumption, say Z, and tenders C_{XZ} as C_X. More embarrassing, we ourselves may at a later date find assumption W to be more agreeable than Y so we now find ourselves with a different conclusion, C_{XW}, from the same data. Alternatively, we may telescope the process and offer two or more conclusions simultaneously, at the same time admitting their dubious nature by revealing the alternative assumptions we found ourselves obliged to adopt but between which we are at a loss to choose.

The thesis of this paper is that there is a way out of this dilemma in an important class of problems.

Entropy

By way of wielding Ockham's razor, we might devise some measure whereby different functions could be compared as to number of "entia". Of all functions consistent with the data we might select the one, or ones, requiring the fewest "entia", i.e., the least information, as being minimally prejudiced. The author joins others, cited in the references hereto, in proposing a measure employed by Shannon [4] in the development of information theory and subsequently adopted by Jaynes [5], Tribus [2] and others in re-derivations of the theorems of statistical mechanics and thermodynamics.

Mean and mean logarithm*	Gamma
$\sum_0^\infty p_i x_i = \bar{x}$	$p_i = \exp.\ (-a_0 - a_1 x_i - a_2 x_i^2)$
$\sum_0^\infty p_i \ln x_i = \overline{\ln x}$	$= x_i^{-a_1} \exp.\ (-a_0 - a_1 x_i)$
Mean logarithm and mean logarithm of complement where $0 \leqslant x \leqslant 1$	Beta distribution $p_i = \exp.[-a_0 - a_1 \ln x_i - a_2 \ln(1-x_i)]$ $= x_i^{-a_1} (1-x_i)^{-a_2} e^{-a_0}$

From theory and the foregoing examples it can correctly be inferred that for every distribution there is at least one specification as to the data which must be known for that distribution to be the minimally prejudiced distribution. Also, there is a unique minimally prejudiced distribution for each specification of known data. In general, for $f(x)$ to be the minimally prejudiced distribution, the known data must be the expected value of the natural logarithm of $f(x)$. For example, what data must be known in order that $f(x) = \sin x$ where $0 < x < \pi/2$? Evidently we shall have $p_i = \exp.\ (-a_0 - a_1 \ln \sin x) = \sin x$ if a_0 is set equal to zero and $a_1 = 1$.

An Apparent Paradox

An apparent paradox can arise in the fitting of distributions of the generalized exponential type, $p_i = \exp.\ (a_0 + a_1 x_i + a_2 x_i^2 + \ldots)$, which more or less typify the system of maximum entropy, when actual distributions are better fitted by some other curve. At such a time we are inclined to ask what is so good about a system that does not give the best fit. The point to remember here is that if we have the distribution function, or if we have a summary of it in the form of grouped data, there is no particular reason to prefer the generalized exponential over any other curve. Equation (3) applies strictly only when our data are limited to the expected values of $g_1(x)$, $g_2(x)$ etc. If we have more information we should use it. Theoretically, of course, by calculating the mean values of a sufficient number of functions of x we can approximate any arbitrary distribution as closely as we please.

The discipline advanced here does not tell us what function best fits a more or less completely specified distribution. It does tell

* See page 380.

us, however, what data to summarize in order that a given kind of distribution function shall be best characterized by that data. For example, if a class of distributions are found to be of the log-normal type, the data we should be collecting are the mean and variance of log x. Similarly, if the distributions for a certain kind of variable are typified by a Gamma distribution, then we should compile mean values of x and log x, and so on. Such knowledge is economical since necessary data can often be summarized in the course of ordinary processing of cases without the necessity of compiling a great many separate distributions.

It is obviously advantageous, by judicious selection of the function of x to be averaged, to reduce the number of statistics that must be compiled.

Of more importance, in the author's opinion, is that for any given data the criterion of maximum entropy leads to what he believes to be a mathematically optimum compliance with the principles attributed at the outset of this paper to Ockham and Hume for the avoidance of prejudice and unwarranted assumptions.

Entropy as a Measure of Homogeneity

Let a classification plan subdivide a population of risks into n classes such that for any particular layer of loss the probability of occurrence of a loss during a specified time interval is p_i for the i'th class. Then for that layer of loss the entropy of this classification scheme is as defined in Eq. (1). As between two classification plans applied to the same population of risks, the plan for which S is smaller contains the more information (less entropy). As between two populations classified according to the same plan, S is greater for the more homogeneous population. This measure is of interest in comparison with the coefficient of variation, proposed by Bailey [6] as a measure of homogeneity. It is not clear how much advantage, beyond consistency with the general theory advanced here, entropy offers over Bailey's measure.

Applications to Composite and Convoluted Distributions

We define a composite distribution as the result of mixing two or more dissimilar distributions. It is obvious that for the mixture all of the functions x, x^2, x^3 etc. will have as their expected values

the weighted averages of the distributions brought together. This enables us to describe the composite distribution without further analysis in terms of Equation (3). It does not, however, assure that the distribution so determined will provide a good fit to the data unless the functions being averaged are appropriate to describe each of the separate distributions.

We define an n-fold identically convoluted distribution as the distribution of the sum or mean of n values selected independently from the same (infinite) parent population. The parameters of such a distribution are shown by Kendall [7] to vary as follows:

Parameter	*Parent Population*	*Convolution*
Mean	$\bar{x}$	Sum $n\bar{x}$, mean $\bar{x}$
Relative Variance	$V^2 = \sigma^2/\bar{x}^2$	V^2/n
Skewness	$\beta_1 = \frac{[E(x-\bar{x})^3]^2}{[E(x-\bar{x})^3]^2}$	β_1/n
Kurtosis	$\beta_2 = \frac{[E(x-\bar{x})^2]^2}{E(x-\bar{x})^4}$	$\frac{\beta_2 - 3}{n} + 3$

Parameter values shown for the convolution can be used to compute Ex^2, Ex^3, Ex^4, etc. and similarly substituted in Eq. (3). Of course, if the parent distribution function is known explicitly its convolutions can be calculated by standard methods [8].

Comparison with Other Schools of Statistical Inference

The method of minimum prejudice, or maximum entropy, is distinguished from the Neyman-Pearson school of statistical inference in that whereas the latter school sets up hypotheses and judges their plausibility in terms of the probability of occurrence of an observed event given the truth of a hypothesis, the former method goes straight from the data to the answer without any testing whatsoever. No testing is theoretically even possible if the method of maximum entropy has been strictly followed, since all available data will have gone into the calculation and no further information is obtainable, in principle, by testing or otherwise.

As a practical matter, the two approaches apply under different circumstances. If the only available data are several different kinds

of means, the distribution with maximum entropy is asserted to be the appropriate distribution *on these data*. As more data, such as a histogram, are acquired an entirely different curve may be indicated from what was derived from limited data. In principle it should be possible to derive a maximum entropy distribution from any arbitrary data. Very little is known, however, as to just how to go about incorporating data other than averages. This should be a fruitful field for study. Fully developed, it ought to obviate the need for Chi-square and other tests in a great many cases. In the meantime, however, it is entirely possible to conceive of using a Chi-square test, for example, upon receipt of more data, to confirm or revise any earlier choice of curve based upon maximum entropy. It might also be used where a generalized exponential function has been fitted to given data on the basis of selected parameters computed from more detailed data such as a histogram. The necessity for such a mixing of methods is less than satisfying.

That the need for testing can be eliminated may come as a surprise to persons, such as the author, trained under the Neyman-Pearson influence. Yet it is readily apparent that a solution derived strictly according to Bayes' theorem requires no testing. Application of this theorem does, however, require knowledge of prior probabilities. It is only in the attempt to "fudge" an answer in the absence of such knowledge that we find ourselves obliged to resort to confidence tests and the like. The method of maximum entropy, as a logical outgrowth and extension of Bayes' theorem, provides a solution to this dilemma in a wide class of cases.

Actuarial Implications

An obvious actuarial implication arises in the calculation of deductibles under conditions of inadequate data. Given only the mean of a non-negative variable, we know the exponential distribution is the minimally prejudiced estimate of the distribution. Sometimes we may have more information, such as that $f(0) = 0$. This implies that $\ln x$ has a finite mean *. Hence we might let $f(x) = \exp.\ (-a_0 - a_1 x - a_2 \ln x) = (x) \exp.\ (-a_0 - a_1 x)$ if $a_2 = 1$.

*) This implication holds without qualification only for discrete distributions which are the only distributions for which entropy has been defined here.

Whether such a solution is valid is one of the questions to be studied. (If we knew the mean value of $\ln x$, this equation would be minimally prejudiced — but is it minimally prejudiced when only the existence, not the value, of $E(\ln x)$ is known? How do we know the exponent of $\ln x$ should be unity? Does the arbitrary selection of this value for the exponent betray a prejudice?)

It appears that in many important practical cases involving constraints of a form inexpressible as averages, it is not feasible to maximize the entropy through use of the calculus of variations to find extremals. Correct answers in such instances may be calculable only through iterative procedures. [9]

In collective risk theory it seems unlikely that we shall ever have satisfactorily specified distributions of the claims arising from heterogeneous portfolios. It may be that Eq. (3) provides our best estimate of such distributions for practical purposes.

Finally, in such imponderables as the probability distribution of the error in existing rates — which must be estimated if credibility is to be calculated using Gauss's theorem on minimum variance, complete specification of distributions is apparently out of the question. In this and many other cases we must settle for a good deal less information. It seems clear that in such instances, as in others, we are well advised to use such information as we have with a minimum of prejudice and unsupported assumptions.

REFERENCES

[1] The authenticity of such attribution is questionable, as observed by C. Kenneth BRAMPTON, editor of the volume, *The De Imperatum et Pontificum Potestate of William of Ockham*, (University Press, Oxford, 1927) who states in a note on page 80. "By a curious fate Ockham is in many quarters known solely for his 'razor', which Mr. W. M. Thorburn ably proves (Mind, no 107, July 1918) to be an invention of a later age, occurring first in the works of Condillac less than two centuries ago, and introduced into England by Sir William Hamilton in 1852. But Ockham's meaning is clear enough, that if there is no 'humanity' existing apart from the individuals which collectively form it, it is gratuitous to postulate its objective existence (Log i, cap. lxvi): 'frustra fit per plura quod potest fieri per pauciora' (Sent. ii, Dist 15, o). These words as Mr. Thorburn points out, are actually quoted by Sir Isaac Newton in his third edition of his Principia Mathematica of 1726 (De Mundi Systemate, lib. iii, p. 387). This is Regula i, and continues, 'Natura enim simplex est et rerum causis superfluis non luxuriat'· but the

garbled version in the form 'entia non sunt multiplicanda praeter necessitatem' was invented by John Ponce of Cork in 1639 and took its present shape for the first time in the Logica Vetus et Nova of John Clauberg of Groningen in 1654. Even in his philosophy there is much that is untrue in the name, weapon, and formula bestowed upon Ockham by posterity."

The Encyclopaedia Brittanica, however, says that "The famous dictum, 'pluralites non est ponenda sine necessitae' (multiplicity ought not to be posited without necessity) has become known as 'Ockham's razor' though it had already been stressed by other Scholastics," without commenting upon the variation in wording nor challenging the attribution to Ockham. In the following paragraph it says ". . . Ockham did not make much of the philosophical arguments of earlier theologians, and applied to theology his famous 'razor' . . ."

This author relinquishes the task of any further research into the authenticity of 'Ockham's razor' to qualified medievalists

[2] TRIBUS, Myron

(a) "The Probability Foundations of Thermodynamics", Myron TRIBUS and Robert B. EVANS, *Applied Mechanics Review*, Vol. 16, No. 10, October 1963.

(b) "Why Thermodynamics Is a Logical Consequence of Information Theory", Myron TRIBUS, Paul T. SHANNON and Robert B. EVANS, *A.I Ch.E. Journal*, March 1966.

(c) "Information Theory as the Basis for Thermostatics and Thermodynamics", Myron TRIBUS, *Journal of Applied Mechanics*, March 1961.

(d) "The Maximum Entropy Estimate in Reliability" in *Recent Developments in Information and Decision Processes* Macmillan Co. 1962.

(e) "The Use of Entropy in Hypothesis Testing", Myron TRIBUS, Robert EVANS and Cary CRELLIN, paper presented at *the Tenth National Symposium on Reliability and Quality Control*, January 7-9, 1964.

(f) "Information Theory and Thermodynamics", *Boelter Anniversary Volume*, McGraw Hill Book Co. 1963.

[3] HUME, David, "A Treatise of Human Nature", 1740. A more pertinent reference, in this author's opinion, is provided in Volume 4 of Hume's *Philosophical Works*, Edition of 1777. This edition was "corrected by the author for the press, a short time before his death, and which he desired might be regarded as containing his philosophical principles", according to the "Advertisement" prefacing Volume 1 of the 1854 reprint, published by Little, Brown and Co. of Boston and by Adam and Charles Black of Edinburgh, of the 1777 edition. Most to the point, perhaps, is Hume's rhetorical question (page 35) "All these suppositions are consistent and conceivable. Why then should we give the preference to one, which is no more consistent or conceivable than the rest?" In what follows he argues that past experience is our only guide where no *a priori* connection can be demonstrated between cause and effect. This author agrees that Hume's discussion of inductive principles is consistent with Tribus's formulation but thinks it may be reading too much into Hume's rather prolix text to find there so clear a statement of the problem as given by Tribus.

[4] SHANNON, C. E., "A Mathematical Theory of Communication", *Bell System Technical Journal*, Vol. 27, 379, 623. 1948.
[5] JAYNES, E. T., "Information Theory and Statistical Mechanics" *Phys. Rev.* 106, p. 620 and 108, p. 171 (1957); *AMR* 11 (1958), Rev. 2293. Other references to Jaynes are given in [2].
[6] BAILEY, R. A., "Any Room Left for Skimming the Cream", *P.C.A.S. XLVII*, 1960.
[7] KENDALL, Maurice, *The Advanced Theory of Statistics*, Vol. 1, p. 302, Charles Griffin & Sons, Ltd. 1948 *).
[8] FELLER, William, *An Introduction to Probability Theory and Its Applications*, Vol. 1, p. 250, John Wiley & Sons, 1950.
[9] Besides linear programming, possible directions such calculations might take are suggested in *Nonlinear Mathematics* by Thomas L. SAATY and Joseph BROM. McGraw Hill Book Co. 1964.

*) The expected values M_1 and M_4 for sample means, as given there for sampling from a finite population of N cases, can be reduced to β_1/n and $(\beta_2-3)/n+3$ on taking the limit as $N \to \infty$