Inference for a Bernoulli Process (A Bayesian View)
Author(s): D. V. Lindley and L. D. Phillips
Source: *The American Statistician*, Vol. 30, No. 3 (Aug., 1976), pp. 112-119
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: https://www.jstor.org/stable/2683855
Accessed: 11-05-2019 08:57 UTC

29. Rosenblatt, M. (1969). "Conditional probability density and regression estimators," *Multivariate Analysis, 2,* 25–31.
30. Rosenblatt, M. (1970). "Density estimates and Markov sequences," *Nonparametric Techniques in Statistical Inferences,* ed., M. L. Puri, Cambridge University Press, Cambridge, 199–213.
31. Rosenblatt, M. (1971). "Curve estimates," *Ann. Math. Statist., 42,* 1815–1842.
32. Royston, E. (1956). "Studies in the history of probability and statistics. III. A note on the history of the graphical presentation of data," *Biometrika, 43,* 241–7.
33. Schwartz, S. C. (1967). "Estimation of probability density by an orthogonal series," *Ann. Math. Statist., 38,* 1261–65.
34. Specht, D. F. (1971). "Series estimation of a probability density function," *Technometrics, 13,* 409–24.
35. Stanat, D. F. (1966). "Nonsupervised pattern recognition through the decomposition of probability functions," *Technical Report, University of Michigan Sensory Intelligence Laboratory.*
36. Tarter, M. E.; Holcomb, R. L.; and Kronmal, R. A. (1967). "A description of new computer methods for estimating the population density," *Proceedings, Association for Computing Machinery, 20,* Thompson Book Co., Washington, D. C., 511–19.
37. Tarter, M. E. and Kowalski, C. (1972). "A new test for and class of transformations to normality," *Technometrics, 14,* 735–43.
38. Tarter, M. E. and Kronmal, R. A. (1968). "Estimation of the cumulative by Fourier series methods and application to the insertion problem," *Proceedings, Association for Computing Machinery, 23,* Thompson Book Co., Washington, D. C., 491–97.
39. Tarter, M. E. and Kronmal, R. A. (1970). "On multivariate density estimates based on orthogonal expansions," *Ann. Math. Statist., 41,* 718–22.
40. Tarter, M. E. and Raman, S. (1971). "A systematic approach to graphical methods in biometry," *Proceedings, Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume Four,* (ed., L. LeCam, J. Neyman and E. L. Scott), University of California Press, 192–221.
41. Tarter, M. E. and Silvers, A. (1975). "Implementation and Applications of Bivariate Gaussian mixture decomposition," *J. Amer. Statist. Assoc., 70,* 47–55.
42. Tarter, M. E. (1975). "A method for maximum likelihood parameter estimation based upon a reduced data file," *Proceedings, Conference on Computer Graphics, Pattern Recognition and Data Structure, May 14–16,* Institute of Electrical and Electronics Engineers.
43. Tocher, K. D. (1961). *The Art of Simulation.* The English Universities Press LTD, London.
44. Van Ryzin, J. (1966). "Bayes risk consistency of classification procedures using density estimation," *Sankhyā Series A, 28,* 261–70.
45. Watson, G. S. (1964). "Smooth regression analysis," *Sankhyā Series A., 26,* 359–72.
46. Watson, G. S. and Leadbetter, M. R. (1965). "On the estimation of the probability density I," *Ann. Math. Statist., 34,* 480–91.
47. Watson, G. S. (1969). "Density estimation by orthogonal series," *Ann. Math. Statist., 40,* 1496–98.
48. Wegman, E. J. (1969). "A note on estimating a unimodal density," *Ann. Math. Statist., 40,* 1661–67.
49. Wegman, E. J. (1972). "Nonparametric probability density estimation: I. A Summary of available methods," *Technometrics,* Vol. 14, No. 3, 513–546.
50. Wegman, E. J. (1972). "Nonparametric probability density estimation: II. A comparison of density estimation methods," *Journal of Statistical Computation and Simulation,* Vol. 1, 225–245.
51. Whittle, P. (1958). "On smoothing of probability density function," *J. Roy. Stat. Soc., 20,* 334–43.

Inference For a Bernoulli Process (a Bayesian View)*

D. V. LINDLEY**AND L. D. PHILLIPS***

It would have been good to have entitled this paper "Elementary Statistics from an Advanced Standpoint", but we could not hope to approach the high standard set by Felix Klein (1932) in his magnificent work on mathematics, and to use such a title would invite comparison. Nevertheless, that is what this paper is about: we propose to look at a very simple statistical situation, in which an event keeps occurring or not under stable conditions, from a viewpoint which, if not advanced, is at least only to be found in articles. We feel this approach has several advantages over that used in the elementary texts, and we hope that the reader will feel so too when he has read and considered the argument.

An alternative way of regarding the paper is to think of it as providing an introduction to Bayesian statistics. Experience suggests that there is a lot of misunderstanding in the statistical community about Bayesian ideas, and our hope is that, by discussing their use in a familiar context, we shall help people to comprehend them better, and, hopefully, feel them superior to other approaches to statistics.

1. *A Scientific Problem*

You are invited to consider a situation which commonly arises in statistics. A scientist, having performed an experiment (or a sociologist, having conducted a survey), visits a statistician to ask for his help in analyzing the results. We will play the role of the scientist and you, that of the statistician. We are going to tell you about an experiment that one of us performed, and invite you to consider your reactions.

* This paper is the basis of three closed-circuit television programs entitled "Introduction to Bayesian Statistics" (programs #1–3) by the authors. Details are available from Brunel University Television Service, Kingston Lane, Uxbridge, Middlesex, England.

** Dept. of Statistics & Computer Science, University College London, Gower St., London WC1E 6BT, England.

*** Brunel Univ., Kingston Lane, Uxbridge, Middlesex, England.

It is a very simple experiment, but not too unlike many that are carried out and subsequently analyzed.

I took a box of ordinary metal drawing pins (thumb tacks) such as might be purchased at any stationers (drug store) and selected one of the pins. Having inspected it to make sure it was not obviously defective, I tossed it, allowing it to fall onto a table covered with a cloth, and observed whether it fell with the point uppermost (abbreviated to U) or with the point resting on the table, downwards (abbreviated to D). In tossing the pin, I tried to give it a fair amount of spin so that its subsequent position on the table was unpredictable. I did this a few times in order to gain some practice in tossing and to make sure that I could do it under reasonably reproducible conditions. Having assured myself of this, I then tossed it 12 times in all, observing that on 9 of these it fell point uppermost, U, and on the remaining 3 point down, D. I noticed that the pin did not appear to suffer any damage during the tossing and that the external conditions remained stable throughout the experiment.

The scientific problem, upon which I need your advice, is that of assessing the chance that the pin will fall uppermost on a further, thirteenth, similar toss. You may think this to be somewhat trivial, but Karl Pearson (1920), one of the founders of modern statistics, described the following as one of the more important problems in applied statistics. On n_1 occasions an event has been observed to occur r_1 times: what is the chance that on n_2 further occasions it will occur r_2 times? We have $n_1 = 12$, $r_1 = 9$, $n_2 = r_2 = 1$. The problem arises with repeated trials, each of which can give one of two results (up, down; heads, tails; male, female; success, failure; dead, alive).

Before we try to answer the question, do you want to know anything more about the experiment beyond what you have already been told? Statisticians, in their capacity as consultants, properly quiz their scientific clients about their experiments or surveys. Do you want to quiz me about my experiment, simple though it is?

I have repeatedly given a lecture, upon which this article is based, and have stopped to ask that question of the audience. Let me report on the typical responses that are produced. They fall into three groups. First, I am often asked about the pin, how big was it, where was its centre of gravity, etc. Questions of this sort I decline to answer on the grounds that to provide the answer would mean performing an additional experiment. Maybe I should have carried out those measurements, but I did not. These questions refer to another experiment; you are being asked about this experiment. In more realistic and complicated situations additional experiments may be costly and difficult, so that the scientist has to proceed on the basis of what he has.

The second question is a simple request for the order of the 9 U's and 3 D's. This was retained, and was

UUUDUDUUUUUD.

We will see later that, in a sense, this gives no information: technically, 9 and 3 are jointly sufficient. But people presumably ask for the order because they want to see if there is anything strange about it—for example, did the 3 D's occur at the end? As it is, the sequence seems reasonably 'mixed-up' and no suspicions are aroused.

The last point is an important one, but is not raised as often as it might, perhaps because people are polite. It is simply: did I cheat? For example, did I suppress some of the results because I didn't like them; and if so, why? The answer is "No". As explained, I did have a few practice results at the beginning, but here the falls were not recorded and my interest was centered on my tossing ability rather than on the experimental outcomes. Such questions are important because bias, unconscious or not, can easily enter into a scientific experiment. I am not aware of any such bias in my performance of, or reporting on, this experiment.

There is one other question that is sometimes raised by a statistician, but never by someone unfamiliar with modern statistical ideas, and which we will return to later. Aside from this, these cover the main points that are mentioned by a typical, live audience.

Now for an astonishing fact.

On the information so far provided about the pin-tossing experiment, it would be impossible for a non-Bayesian statistician to perform *any* of the standard statistical procedures (with one doubtful exception). The performance would require either more information from me, or the introduction of an assumption that seems to many to be unwarranted, and to all of doubtful justification. He could not perform a significance test, construct an unbiased estimate (with its standard error) nor find a confidence interval. He could not begin to act like a statistician. And yet, let us repeat, such information is rarely requested by typical, intelligent non-statisticians (and even by many statisticians, who make the above-mentioned assumption without realizing they have done so). To see what this additional information is let us consider a significance test of the hypothesis that the pin is unbiased: that is, is equally likely to come U as D. Incidentally, this hypothesis and the resulting significance test is one of the first ever to be performed, by Arbuthnot in 1710. He was concerned with human births, not pins, and with the events male and female. He rejected the hypothesis that male and female births occur equally often and concluded that ". . . this inequality of Males and Females is not the Effect of chance but Divine Providence". The argument goes as follows. Obtaining 9 U's out of 12 suggests that the chance of its falling uppermost exceeds $\frac{1}{2}$. The results that would even more strongly support this suggestion are

10,2; 11,1 and 12,0;

so that, on the null hypothesis, the chance of the observed result, or more extreme, is

$$\left\{\binom{12}{3}+\binom{12}{2}+\binom{12}{1}+\binom{12}{0}\right\}\left(\frac{1}{2}\right)^{12}.$$

This easily gives 299/4096, or about 7½%. In usual statistical parlance, the result is not significant; at least if the favoured 5% level is employed. In other words, the observed result, or more extreme results, could reasonably be expected to occur by chance if the pin was equally likely to fall in either position.

But stay. Who said anything about 10 U's and 2 D's being another possibility? A method of experimentation, sometimes used and often ascribed to J. B. S. Haldane (1945), is to toss until a prescribed number, here 3, of D's is observed. Under this method of tossing the result (10,2) is not a possibility. The more extreme values are

10,3; 11,3; 12,3; and so on.

The probability of all these results, including that obtained, is easily found to be 134/4096 or about 3¼%. Now the result *is* significant at 5%, and deserves a single asterisk in many scientific journals. Indeed the probability is under one-half the value obtained by the previous argument in which the total number of tosses, 12, was supposed fixed, rather than the number of D's.

In other words, the significance (in the technical sense) to be associated with the hypothesis of equal chances depends heavily on what *other* results could have been achieved besides the 9 U's and 3 D's reported. Thus was (10,2) or (10,3) an alternative possibility? And yet it is rare for anyone to ask a scientist for this information.

In fact, in the little experiment with the drawing pin I continued tossing until my wife said "Coffee's ready". Exactly how a significance test is to be performed in these circumstances is unclear to me.

Another way of describing the information required for a significance test is to say that we need to specify the *stopping rule*: the rule that caused the experiment to terminate. In reality the rule depended on my wife's coffee-making, a habit that seems to have little to do with drawing pins. Haldane's rule is to stop when $D = 3$: the usual rule is stop when 12 tosses have been performed. Yet another way is to specify the *sample space*, a technical phrase for a listing of all the values that could have been obtained in the experiment. We had two such partial lists above. The usual statistical significance test requires the sample space, or alternatively, the stopping rule to be specified. Many people's intuition says this specification is irrelevant. Their argument might more formally be expressed by saying that the evidence is of 12 honestly reported tosses, 9 of which were U; 3, D. Furthermore, these were in a particular order, that reported above. Of what relevance are things that might have happened, but did not?

The same need for the sample space underlies the unbiased estimate of the chance of the pin falling point uppermost. With 12 fixed, it is 9/12; in Haldane's experiment it is 8/11. Confidence limits similarly demand consideration of the sample space. Indeed, so does every statistical technique, with the possible exception of maximum likelihood. Even this weakly requires it in the calculation of the associated standard error.

2. *The Bayesian Analysis*

What happens in the Bayesian argument?

We shall see that it uses nothing more than the observed sequence of tosses. At least, nothing more *from the experiment*. It does use something else, but something that arises from outside the experiment. Let us now give you the Bayesian argument and try to persuade you that the requirement of this additional knowledge is reasonable.

We described to you, with some care, the conditions under which the experiment was performed; in particular, how we tried to ensure that the conditions of toss were reasonably held constant. The effect of this is to make a result obtained on any one toss equivalent to that on any other—and similarly results on pairs or triplets of tosses. This idea is made precise by assuming that the sequence of tosses has a property called *exchangeability*. This means that the probability for any sequence of r U's and s D's is the same as that for any other sequence containing the same numbers of U's and D's. In other words, position and order are irrelevant. (This refers to any length of sequence and not just to 12.) The concept of exchangeable sequences is due to de Finetti (1937). The notion seems very acceptable in the case of the pin-tossing experiment.

Notice that familiar sequences of independent trials with constant probability, θ, of success, so-called Bernoulli trials, are exchangeable: for any sequence of r successes and s failures has probability $\theta^r(1 - \theta)^s$ irrespective of the order of the successes and failures. But there are exchangeable sequences that are not of this familiar type. Some of you may remember d'Alembert's argument that if he tossed a coin twice, there were three possible results: 2 heads, 1 head or no head. Thence he argued each had probability ⅓. Presumably for d'Alembert:

$$p(HH) = \tfrac{1}{3},\ p(HT) = \tfrac{1}{6} = p(TH),\ p(TT) = \tfrac{1}{3}\,.$$

These are trivially exchangeable but the probabilities are not of the form $\theta^r(1 - \theta)^s$ for any θ. Consequently if the sequences generated by pin-tossing are assumed exchangeable, *less* is being assumed than is usual.

De Finetti then asked what exchangeable sequences looked like. He proved a most remarkable theorem. If we denote by $p(r,s)$ the probability associated with a sequence of r U's and s D's—the order being immaterial by exchangeability—then he proved* that

$$p(r,s) = \int_0^1 \theta^r(1 - \theta)^s p(\theta)\, d\theta$$

for some $p(\theta) \geq 0$ with $\int_0^1 p(\theta)\, d\theta = 1$. Notice that $p(\theta)$ has all the properties of a continuous probability distribution over the unit interval: namely, it is non-negative and integrates to one. (The Stieltje extension brings in the discrete distributions as well.) So the theorem says that exchangeable sequences are essentially *mixtures* of Bernoulli sequences (independent, constant probability of success) giving the $\theta^r(1 - \theta)^s$ term; the mixture being by a distribution over the value of θ. In particular, if the distribution is discrete (we really do need that Stieltje form!) on a single θ-value, we get just $\theta^r(1 - \theta)^s$, the special case used in the familiar statistical arguments.

In d'Alembert's example, $r + s = 2$ and his probabilities can be obtained by taking $p(\theta) = 1$ for all θ. Of course, for coins, d'Alembert's result would be regarded by most people as eccentric; but for pins, as we shall see below, it is not too unreasonable.

De Finetti proved a second result which is of importance. Write $r + s = n$, the total number of tosses. He proved that for exchangeable sequences $\lim_{n\to\infty} r/n$, θ say, exists with probability one and has the distribution $p(\theta)$. In words, the proportion of U's in exchangeable sequences tends to a limit as the sequence gets indefinitely long, and this limit has the distribution described by $p(\theta)$. The special case of Bernoulli trials, where the distribution is concentrated on a single value, shows that the law of large numbers is a special case of de Finetti's theorem.

It is tangential to our main argument, but it helps to notice that θ is not a probability in the Bayesian meaning of probability. Probability is a relation between you and the external world, expressing your opinion of some aspect of that world: here, your opinion about drawing pins. On the other hand, θ is a property of that world, a property of that pin. We may refer to θ as the propensity (of the pin to fall uppermost).

We can now solve the original problem of assessing the chance that the same pin will fall uppermost if tossed in a similar manner for a thirteenth time. For U on the thirteenth toss will mean that there will be 10 U's and 3 D's in total, so

$$p(U_{13}|9,3) = p(10,3)/p(9,3)$$

and both numerator and denominator are available by de Finetti's formula. To see that this gives the same result as Bayes theorem with prior $p(\theta)$ and likelihood $\theta^r(1 - \theta)^s$, we need only write out the numerator in full, obtaining for the complete expression,

$$p(U_{13}|9,3) = \int_0^1 \theta\{\theta^9(1 - \theta)^3 p(\theta)\, d\theta/p(9,3)\},$$

and recognize that the term in braces is $p(\theta|9,3)$ the posterior distribution when $p(\theta)$ is the prior distribution. Hence

$$p(U_{13}|9,3) = \int_0^1 \theta p(\theta|9,3)\, d\theta.$$

The point is that exchangeability, itself weaker than a Bernoulli assumption, produces $p(\theta)$ and demonstrates the soundness of the subsequent Bayesian manipulations. There is no 'assumption of a prior' as many critics of the approach claim.

Thus, from a single assumption of exchangeability the Bayesian argument follows. This is one of the most beautiful and important results in modern statistics. Beautiful, because it is so general and yet so simple. Important, because exchangeable sequences arise so often in practice. If these are, and we are sure there will be, readers who find $p(\theta)$ distasteful, remember it is only as distasteful as exchangeability; and is that unreasonable?

3. *Interpreting $p(\theta)$, and its Effect on $p(\theta|r,s)$*

Let us have a closer look at $p(\theta)$ and, to facilitate the study, suppose

$$p(\theta) \propto \theta^{a-1}(1 - \theta)^{b-1}$$

for some a, $b > 0$. This is a Beta-distribution with parameters a and b. We assume the reader is familiar with these distributions. Details will be found in many texts; for example, Phillips (1973). The point about this assumption is that

$$p(\theta|r,s) \propto \theta^r(1 - \theta)^s p(\theta) \propto \theta^{r+a-1}(1 - \theta)^{s+b-1}.$$

In other words, whatever the result of the experiment, the distribution remains of the Beta form. All that happens is that a increases to $a + r$ and b to $b + s$. The probability can now be inserted in the general result to give

$$p(U_{n+1}|r,s) = \frac{a + r}{a + b + n}.$$

We now have an explicit answer to our problem. Assuming that the experiment generated exchangeable sequences, and supposing that our opinion of the propensity of the pin to fall uppermost is described by a Beta distribution with values a and b, the chance of the thirteenth toss resulting in U is $(9 + a)/(12 + a + b)$. In particular, the stopping rule or sample space is irrelevant: it does not matter whether 12 was fixed, a Haldane experiment used or we relied on my wife's coffee-making habits. 12 honestly reported tosses gave 9 U's and 3 D's.

But still there is the problem of a and b: and in general about $p(\theta)$. The orthodox statistical argu-

* The mathematically knowledgeable will protest, quite rightly, that this is incorrect. The integral here ought to be Stieltje, not Riemann, and we should have $\int \ldots dP(\theta)$ with $P(\theta)$ a distribution function. But not everyone is familiar with the general form; so, accurately, de Finetti proved a result which is nearly that stated.

ments have been criticized for bringing in extraneous material, namely the sample space. Now the Bayesian includes $p(\theta)$.

It is generally true, that whereas the usual procedures add one ingredient, the stopping rule, the Bayesian argument uses another, the distribution, $p(\theta)$. That is the key difference between the two approaches: sample space or distribution. But notice a difference—*the former is imposed by the statistician, whereas the latter is a result of mild assumptions; assumptions, moreover already made by the statistician*. Remember, Bernoulli trials are a special case of exchangeable trials.

What does $p(\theta)$ mean? θ is the long-run frequency of the event, U in our case. It is a property of the particular drawing pin that was used. $p(\theta)$ is therefore a probability statement about the long-run frequency of it falling point uppermost. To understand this more thoroughly, contrast the original experiment with another which is identical to it except that an ordinary coin replaces the pin. Now, for coins we think θ is very near $\frac{1}{2}$. We would be astonished if an ordinary coin had $\theta = \frac{1}{4}$, or even $\frac{5}{12}$. But with pins, both these values are quite reasonable for us. In other words our attitude to long-run frequencies is quite different with pins from coins.

Now this is interesting; the Bayesian argument distinguishes between pins and coins because $p(\theta)$ changes. The orthodox argument does not. The result r/n for an unbiased estimate is used for coins or pins, or deaths, or social class, or whatever. But a Bayesian cannot proceed until he thinks about whether it is a coin or a pin that he is tossing, or whether, with Arbuthnot, he is studying sex. (And sex is surely different from pins!) So Bayesian statistics is, in this respect, much more practical than orthodox statistics.

For us, a reasonable distribution* for θ in the case of a pin is $6\theta(1 - \theta)$: namely, Beta with $a = b = 2$. Our opinions about U and D are symmetrical and we think values near $\frac{1}{2}$ are more likely than those near zero or one. Bayes originally took $a = b = 1$. Hence the chance of the thirteenth toss being U is for us, $\frac{11}{16}$, for Bayes, $\frac{10}{14}$: 0.69 or 0.71 to two decimal places, a relative difference of about 3%, not very much.

But with coins we would take a distribution with $a = b = 50$, at least. Again this is symmetrical about $\frac{1}{2}$, but is now sharply peaked around $\frac{1}{2}$, so that values even a little way from $\frac{1}{2}$ are, for us, pretty unlikely. (The standard deviation is about 0.05.) Hence if 12 tosses of a *coin* had given 9 heads and 3 tails, we would assess the chance of heads on the next toss as $\frac{59}{112}$, about 0.53. This is still close to $\frac{1}{2}$ and a lot different from 0.69 obtained with a pin.

Now isn't that conclusion reasonable? That our statements about pins should differ from those about coins? Wouldn't you be inclined to stay with the value $\frac{1}{2}$ with a coin, but, if you knew as little about drawing pins as we did, be easily influenced by the result of even a few tosses? Consequently, not only does $p(\theta)$ arise from a natural assumption of exchangeability, but its presence makes good, practical sense. Indeed the orthodox statistical arguments are suspect precisely because they do not include any reference to it; pins and coins are alike to sample-space statisticians.

Notice that as the number of tosses increases the probability for the next toss is almost r/n whatever be a or b. This asymptotic agreement with the usual results does not always obtain. In some aspects of significance tests the Bayesian and orthodox arguments can differ substantially, one value going to one, the other to zero: see, for example, Lindley (1957), or Phillips (1973, Chapter 14).

The concept of exchangeability finds many applications beside Bernoulli sequences. Consider, for example, a one-way analysis of variance with observation x_{ij} normally distributed with means θ_i. It is often reasonable to suppose the means exchangeable. This leads to quite different estimates for θ_i from the usual $x_{i.}$. Indeed, the full benefit of the Bayesian results only appear when dealing with several parameters, and not, as here, with just one. See, for example, Lindley and Smith (1972).

We now consider a practical application of these results. We could have included an illustration from industry but our little problem has the merit that it involves little specialist (for example, industrial) knowledge and is easily performed in class. Bayesian examples are always more difficult to present than orthodox ones because they involve extraneous (prior) knowledge about θ whereas this is irrelevant in the usual approach where θ can remain a Greek letter: a Bayesian has to consider its meaning.

4. *An Example*

A curious property of a small flower native to South Africa provides an interesting example. Some years ago one of us discovered that he was unable to detect the scent of freesias, yet his wife reported that the flowers are very, very fragrant. A florist who was told of this said that most people can detect the smell of freesias, but not everyone. Further enquiry revealed that the ability to smell freesias has a genetic basis. So let us consider the question, what proportion, θ, of people can smell freesias?

We will use the theory developed in previous sections to make inference about θ, whose true value is unknown to us. In line with the arguments used there we suppose there is a group of people, say the white population of Britain, that we judge exchangeable with respect to the ability to smell freesias. In the language of Section 2, θ is the propensity of people in that group to have the ability. Then we will take a sample of these and, from observations on the members of the sample, make inferences about θ. We have seen that the Bayesian argument requires an assess-

* More detailed discussion of the practical determination is given below, section 4.

ment of $p(\theta)$, and since this aspect is not present in the more frequently used types of statistics, some time is spent in discussing it.

First, it is necessary to assess $p(\theta)$, a probability density function that represents our judgment of the proportion of people who can smell freesias. As a starting point, suppose we consider that any value of θ is as likely as any other value. That gives a uniform distribution over possible values of θ, which, as we noted previously, can be represented by a Beta distribution with $a = b = 1$. But wait! We've ignored two solid pieces of information: one of us cannot smell freesias, but his wife can. Now our distribution might have $a = b = 2$, both having increased by one due to the two different observations. Actually, we have a little more information. The florist says most people can smell freesias; that seems reasonable for if only a few people could smell them they probably would not sell, and yet we have seen them on many flower stalls in London. So, we would like more of our opinion to the right of 0.5 than to the left of 0.5. A Beta with $a = 3$ and $b = 2$, seems to provide a more satisfactory representation of our judgment.

No assessment of $p(\theta)$ should be considered final until a few consistency checks are made. For example, the median of this distribution, which can be found by referring to tables of the cumulative distribution (Pearson & Johnson, 1968), is 0.61, so that with $a = 3$, $b = 2$ we should be just as happy to bet that the true value of θ lies above 0.61 as below 0.61. If we prefer to bet on one interval rather than the other, then we should reconsider our original assessment. With a little practice, it becomes fairly easy to find a distribution whose median defines two equally good bets.

In fact, we are not indifferent between those bets. It seems to us that the interval from 0.61 to 1.0 is the better bet, and also that the distribution is too spread out—there's still too much density in the vicinities of 0 and 1.

At this point we could try larger values of a and b, for as they increase, the distribution becomes more peaked. It is often easier to choose the appropriate distribution by referring to graphs of the Beta distribution, drawn for different values of a and b. These can be found in most textbooks of Bayesian statistics, for example, Phillips (1973).

However, there are alternative methods which leave to a computer the work of searching for a distribution that represents our judgment. Computer programs that do this have been developed at the University of Iowa by a team led by Melvin Novick (Novick, 1971), and at Harvard University by Robert Schlaifer and his associates (Schlaifer, 1971).

Both these methods consist of computer interrogation about certain aspects of the distribution followed by computer calculations which show to the user what these mean about other features—just as selection of a and b above implies statements about the median. With these facilities, the user is easily able to understand the implications of what he is saying and arrive at a sensible expression of his opinions.

For the freesia example we judge the median to be 0.7, the upper and lower quartiles to be 0.8 and 0.57 respectively. For a Beta prior the last two values imply a median of 0.69, in good agreement with our selected value. The corresponding values of a and b are about 5.0 and 2.5 respectively, so we use these in the subsequent analysis. It is shown in Figure 1.

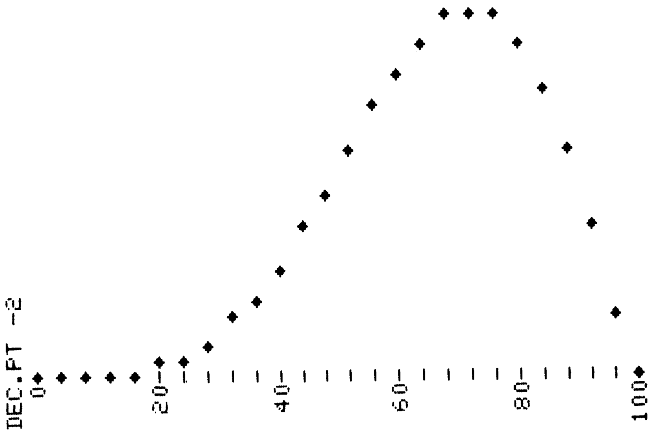

Figure 1: Beta density function with $a = 5$, $b = 2.5$

The next stage in our experiment is to gather some data. We asked 43 people who said their sense of smell was normal to sniff a bunch of freesias. We found that 36 people said the scent is very fragrant, and 7 people told us either they could not smell them or they found the scent to be very faint. That is all the information we need: neither the actual sequence of successes and failures nor the stopping rule is relevant to our inference. We assume only that the people in this experiment are exchangeable, with respect to ability to smell freesias, with the people whose propensity we are trying to evaluate. This is necessary in order that the results of this investigation will generalize to the population at large. We feel that the University students who participated in the experiment are not special in their ability to smell freesias.

Now we are ready to find $p(\theta|r,s)$, the distribution of θ with the data taken into account along with the collateral information which was summarized by $p(\theta)$. Recall that

$$p(\theta|r,s) \propto \theta^{r+a-1}(1 - \theta)^{s+b-1}.$$

We are still dealing with a Beta distribution, but with new parameters $a + r$ and $b + s$. In our experiment, a and b are the original parameters, whose values we assessed at 5 and 2.5.

Next, r and s represent the numbers of people who either can or cannot smell freesias; their values are $r = 36$ and $s = 7$. This gives new parameter values of 41 and 9.5. What does this new distribution of θ look like? By giving Schlaifer's computer program the parameter values, we can obtain a printout of the new density function; it is shown in Figure 2.

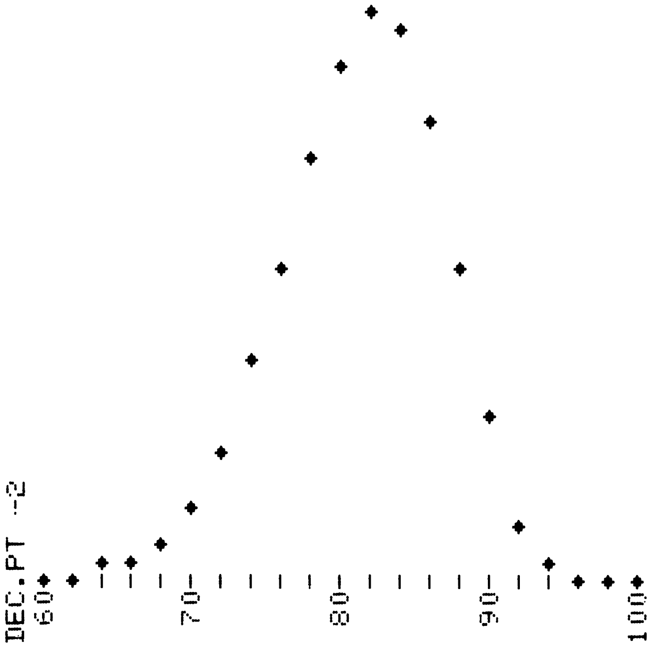

Figure 2: Beta density function with $a = 41$, $b = 9.5$

Now let's compare the two distributions we have obtained. The first one, in Figure 1, is the *prior distribution* because it is based on information without regard for the data of the experiment, or prior to including the experimental results, while the second distribution, in Figure 2, is the *posterior distribution*. Most of the prior distribution can be found between values of θ of .2 and 1.0, but the posterior distribution is concentrated mainly over values of θ that extend from 0.65 to 0.95. The data have caused the distribution to become more peaked; the posterior standard deviation is less than the prior standard deviation. Our judgment about the true value of θ has been sharpened as a result of the experiment.

At this point we might consider reporting some summary of $p(\theta|r,s)$. For example, we might give our best guess about the true value of θ. One possibility is the mean; it can be calculated from the parameters of the prior and the experimental results as

$$\frac{a + r}{(a + r) + (b + s)}.$$

This is the expression we gave for the drawing pin coming uppermost on the next toss. For our case here, the mean is

$$\frac{41}{41 + 9.5} = 0.81.$$

We might also report the mode if we wanted to give the value of θ we thought to be most likely. This is

$$\frac{a + r - 1}{(a + r - 1) + (b + s - 1)} = 0.82,$$

hardly different from the mean. The standard deviation might be instructive as well. This is

$$\sqrt{\frac{(a + r)(b + s)}{(a + r + b + s)^2(a + r + b + s + 1)}} = .0545.$$

Another useful statistic is a credible interval, a range of values of θ encompassing a specified proportion of the distribution. Highest density regions are usually reported, and extensive tables for most distributions have recently been computed (Isaacs, Christ, Jackson and Novick, 1974). The 95% highest density region for the case here extends from $\theta = 0.70$ to $\theta = 0.91$. We might, then, report this in the following way:

$$p(0.70 \leq \theta \leq 0.91) = .95.$$

That brief expression, which tells us that there is a 95% chance that the true value of θ lies between 0.70 and 0.91, seems to us far more informative and easier to understand than a confidence interval.

Perhaps you are interested in knowing whether θ is larger or smaller than some specific value, θ^*. It is a simple matter to find $p(\theta > \theta^*)$ by computing the area under $p(\theta|r,s)$ to the right of θ^*. For example, if $\theta^* = 0.5$, then

$$p(\theta > 0.5) = \int_{0.5}^{1.0} p(\theta|r,s)\, d\theta = .9999983.$$

We can be almost dead sure that more than half the population can smell freesias. The practical implication of this result is rather like that of the significance test of $\theta = \frac{1}{2}$ discussed above. We can say with considerable confidence that $\theta > \frac{1}{2}$. It is possible to develop Bayesian forms of significance tests but usually we feel that calculations of the type exemplified here are more appropriate.

	Standard approach	Bayesian approach
Assumption regarding experiment	Events independent, given a probability	Events form exchangeable sequences
Interpretation of probability	Relative frequency; applies only to repeated events	Degree of belief; applies both to unique and to sequences of events
Statistical inferences	Based on sampling distribution; sample space or stopping rule must be specified	Based on posterior distribution; prior distribution must be assessed
Estimates of parameters	Requires theory of estimation	Descriptive statistics of the posterior distribution
Intuitive judgment	Used in setting significance levels, in choice of procedure and in other ways.	Formally incorporated in the prior distribution

5. *Summary*

In the above table, we summarize the major differences between the Bayesian approach and standard statistical practice.

REFERENCES

Arbuthnot, J. (1710): An Argument for Divine Providence, taken from the constant Regularity observed in the Births of both Sexes, *Philosophical Transactions,* **27,** 186–190.

de Finetti, B. (1937): La prévision: ses lois logiques, ses sources subjectives, *Annales de l'Institut Henri Poincaré,* **7,** 1–68 (Reprinted in English translation as "Foresight: Its logical laws, its subjective sources", in *Studies in Subjective Probability* (eds. H. E. Kyburg, Jr., H. E. Smokler), New York: Wiley, (1964)).

Haldane, J. B. S. (1945): On a method of estimating frequencies. *Biometrika,* **33,** 222–225.

Isaacs, G. L., Christ, D. E., Novick, M. R. and Jackson, P. H. (1974): *Tables for Bayesian Statisticians,* The University of Iowa. London: Dawsons.

Klein, F. (1932): *Elementary mathematics from an advanced standpoint.* New York: MacMillan. (Translation from 3rd German edition.)

Lindley, D. V. (1957): A statistical paradox. *Biometrika,* **44,** 187–192.

Lindley, D. V. and Smith, A. F. M. (1972): Bayes estimates for the linear model. *J. Roy. Statist. Soc. B,* **34,** 1–41.

Novick, M. R. (1971): Bayesian Computer-Assisted Data Analysis, *ACT Technical Bulletin No. 3,* The American College Testing Program.

Pearson, E. S. and Johnson, N. L. (1968): *Tables of the Incomplete Beta-Function.* Biometrika Trust, (Second edition).

Pearson, K. (1920): The fundamental problem of practical statistics. *Biometrika,* **13,** 1–16.

Phillips, L. D. (1973): *Bayesian statistics for social scientists,* London: Nelson; New York: Crowell, 1974.

Savage, L. J. (1962): *The Foundations of Statistical Inference,* London: Methuen.

Schlaifer, R. (1971): *Computer Programs for Elementary Decision Analysis,* Boston: Harvard University, Graduate School of Business Administration.

The Use of Proxy Variables When One or Two Independent Variables are Measured with Error

BURT S. BARNOW*

Social science researchers are often confronted with situations where one or more independent variables in a regression model are available only with measurement error. These fallible but unbiased measures of variables are generally referred to as *proxy variables*. When only proxy variables are available for a subset of the independent variables one must choose between the strategies of including the set of proxy variables in the regression or omitting them. In two recent papers McCallum (1972) and Wickens (1972) demonstrate that if one is interested in estimating the regression coefficient of a particular variable and one or more additional independent variables are available only with measurement error, an asymptotically less biased estimate of the coefficient of interest is always obtained by including the proxy variables rather than omitting them.[1] In this paper it is shown that in the case of two independent variables, if the variable of interest is measured with error, asymptotically less biased estimates can sometimes be obtained by omitting the fallible proxy variable.

The underlying structural model for the analysis is

$$Y_i = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \epsilon_i \qquad (i = 1, 2, \ldots, n), \qquad (1)$$

where it is assumed that there is no intercept. The disturbance term ϵ is assumed to have $E(\epsilon) = 0$ and finite variance. At several points it will be assumed that X_1^* and/or X_2^* are not observable but that we do observe fallible measures of these variables. Denoting these fallible proxy variables as X_1 and X_2, respectively, we may write

$$X_{1i} = X_{1i}^* + u_i, \qquad (2)$$

$$X_{2i} = X_{2i}^* + v_i, \qquad (3)$$

where u and v are error terms with zero means and finite variances that are assumed to be independent of ϵ_i, Y_i, X_{1i}^*, X_{2i}^*, and one another. It is further assumed that the random variables X_{1i}^*, X_{2i}^*, ϵ_i, u_i, and v_i are independent over i. The assumption of error independence implies that var (X_1) = var (X_1^*) + var (u) and var (X_2) = var (X_2^*) + var (v). Denote the ratios var (X_1^*)/var (X_1) as P and var (X_2^*)/var (X_2) as Q and observe that $0 \leq P \leq 1$ and $0 \leq Q \leq 1$. The variances and covariances for the other variables can then easily be obtained. To simplify the notation, define σ_1^2 = var (X_1^*), σ_2^2 = var (X_2^*), and σ_{12} = cov (X_1^*, X_2^*). The population correlation coefficient between X_1^* and X_2^* is defined as $\rho = \sigma_{12}/\sigma_1\sigma_2$.

It is assumed that the primary goal is to estimate β_2 with as little expected bias as possible under various assumptions concerning the availability of X_1^* and

* University of Pittsburgh, ASPER, U.S. Dept. of Labor, Office of the Secretary, Washington, DC 20210. The author is grateful to James Kenkel, Asatoshi Maeshiro and an anonymous referee for comments on an earlier draft of this paper.

[1] If mean square error rather than minimizing asymptotic bias is used as the procedural criterion, it is demonstrated in [1] that the use of proxy variables is not always the dominant strategy.