DOCUMENT RESUME

ED 362 559 TM 020 606

AUTHOR Neel, John H.
TITLE Induced Probabilities.
PUB DATE Apr 93
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Research; Elementary Secondary Education; Equations (Mathematics); *Mathematical Models; Monte Carlo Methods; *Probability; *Research Methodology; *Sample Size
IDENTIFIERS *Induced Probabilities; Power (Statistics); *Variance (Statistical)

ABSTRACT

Induced probabilities have been largely ignored by educational researchers. Simply stated, if a new or random variable is defined in terms of a first random variable, then induced probability is the probability or density of the new random variable that can be found by summation or integration over the appropriate domains of the original random variable. The technique is often simple and can lead to useful results. Among these are a mode of teaching and learning, the development of new techniques, and the study of existing methods. The technique is sometimes applicable where Monte Carlo techniques would otherwise be used. Induced probabilities offer the advantage of exact rather than approximate effort and require less computer time. The approach is described and four examples of its use are presented to: (1) solve a simple probability problem; (2) define a new technique; and (3) find the sample size for a variance used in a power study. Four tables and one figure illustrate these analyses. (Contains 4 references.) (SLD)

Induced probabilities

John H. Neel

Department of Educational Foundations
Georgia State University
University Plaza
Atlanta, Georgia 30303

Phone (404) 651-2582

Running head: Induced Probabilities

Induced Probabilities

Abstract

Induced probabilities have been largely ignored by educational researchers. The technique is often simple and can lead to useful results. Among these are 1) a mode of teaching and learning, 2) the development of new techniques, and 3) the study of existing methods. The technique is sometimes applicable where Monte Carlo techniques would otherwise be used. When used instead of monte carlo techniques, induced probabilities offer the advantage of exact rather than approximate effort and further require less computer time.

The technique is described and four examples of its use are presented: 1) to solve a simple probability problem, 2) to define a new technique, and 3) to find the sample size for a variance used in a power study.

Keywords: induced probabilities, probabilities, monte carlo techniques,

Induced Probabilities

Induced probability, IP, was discussed by Paul L. Meyer (1965, P72). Simply stated, if a new or second random variable is defined in terms of a first random variable, then induced probability is the probability or density of the new random variable which can be found by summation or integration over the appropriate domains of the original random variable. The technique is often simple and can lead to useful results. Researchers have tended to ignore the possibilities of induced probabilities. A search of Current Index to Statistics from 1975 to 1989, ERIC from January, 1964 to March, 1990, Psychlit from 1974 to March 1990, and of 10 statistics textbooks published between 1970 and 1992 revealed no mention of induced probabilities. Yet examples given here will demonstrate that the technique has great applicability. Among such applications are:

1) a mode of teaching and learning,

2) the development of new techniques, and

3) the study of existing methods.

As a mode of instruction and learning, IP has the advantage of making some probability problems clear since it identifies a common pattern which appears in probability problems. An example of the derivation of a new technique will be presented here. In the study of existing methods, induced

probabilities may have particular utility as a replacement for Monte Carlo techniques. Monte Carlo techniques yield approximate answers at best while the technique of induced probabilities, when appropriate, yields exact answers. IP has the additional advantage that it does not require repeated sampling and hence may not require as much computer time as a monte carlo technique.

I. Definition

I assume set theory notation and set theory definitions of random variables are familiar to the reader. Recall that a random variable is a function whose domain is a sample space and whose range is a proper or improper subset of the real numbers. With this definition it is possible to consider the random variable itself as the sample space. Thus we can have random variables whose domains are random variables. It is this definition of random variables as functions defined on other random variables which leads us to induced probabilities.

Random variables are also the domains of probability functions, probability density functions, and the cumulants of these functions. The reader should be cautious and remain clear as to whether a particular random variable is being discussed as:

1. the range of a function defined on a sample space,

2. the domain of a second random variable,

3. the range of second random variable, or

4. the domain of a probability function or cumulative density function.

The following discussion of induced probabilities follows Meyer (1965, Pp 70-73). Any errors or difficulties in the discussion should be attributed to my modifications rather than to Meyer's work.

Definition, equivalent events:

Let $B \subset R_g$. Define $A \subset R_f$ as:

$A = \{x \mid x \in R_X , f(x) \in B \}$,

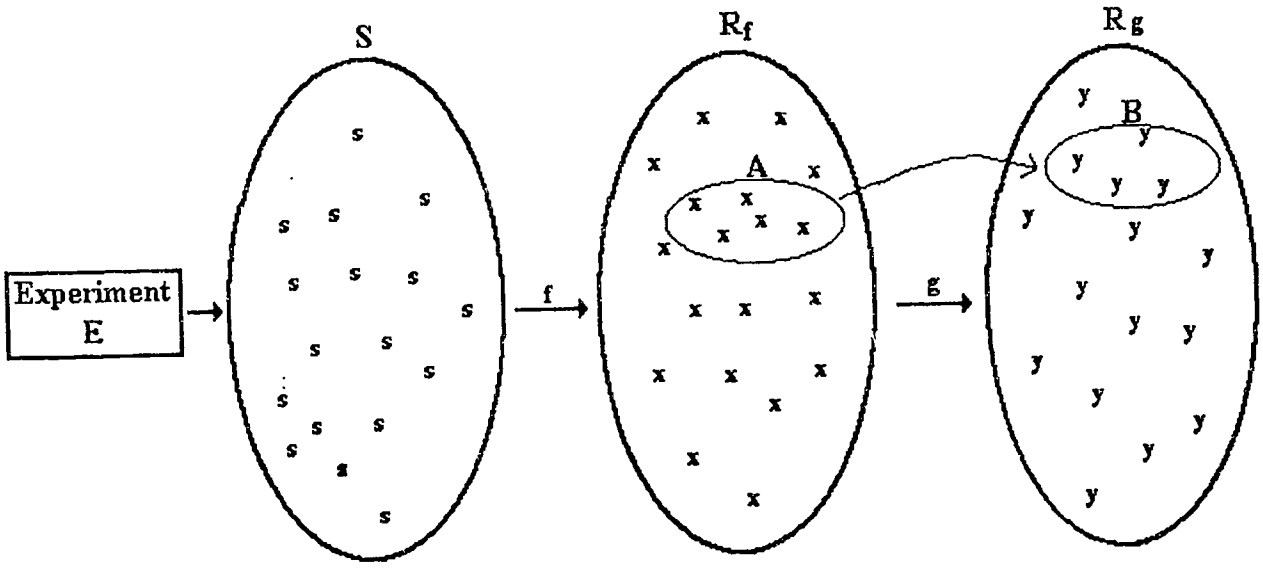
then A and B are equivalent events.

In words, A and B are equivalent events if A is the set of elements in R_f which map into B in R_g. In set theory terminology, B is the image (Halmos, 1960 P31) of A under the function g.

Let E be an experiment and let S be a sample space associated with E. Let X be a random variable defined on S, $f(s) = x$. Suppose that $y = g(x)$ is a real valued function of x. Then $Y = h(s)$ is a random variable since for every $s \in S$, a value of Y is determined since $y = g(x) = g[f(s)] = h(s)$.

Figure 1

Illustration of Induced Probability



Schematically we have Figure 1. In Figure 1 E is an experiment resulting in the possible outcomes in the sample space S. A function f defined on S determines a random variable X. S is the domain of f R_f is the range of f. A second function, g, defined on R_f determines a second random variable, Y, whose domain is R_f and whose range is R_g. A and B are equivalent sets in R_f and R_g fitting the following definition. It is important to note that A and B are events in different sample spaces. A is an event in the random variable X while B is an event in the random variable Y.

Definition, induced probability:

> Let X be a random variable defined on the sample space S. Let R_X be the range of X. Let h be a real-valued function and consider the random variable $Y = g(X)$ with range R_Y. For any event $B \subset R_Y$, we define the induced probability of B, P(B), as:
>
> $$P(B) = P(\{x \mid x \in R_X , g(x) \in B\})$$
>
> or
>
> P(B) = P(A), where B and A are equivalent.

In words, the probability of B is defined as the probability of the equivalent set A (Meyer,1965). It is important to note that the probability of

the event B may be found by finding the probability of the event A. Events A and B can come from different sample spaces as long as they are equivalent. This means that the probability of either event may be determined by determining the probability of the other event.

Figure 1 illustrates the use of IP. Here E, f, g, X, Y, R_x, and R_Y are defined as before. A and B are equivalent sets in R_f and R_g, respectively. P_X and P_Y are the probability functions of X and Y, respectively. $P_Y(B)$ may be found by finding $P_Y(B)$ directly in R_Y or by finding the probability of B's equivalent set A, $P_x(A)$, since these two probabilities are equal.

Induced probabilities become useful when we find the task of determining the probability of A easier than determining the probability of B. This can be the case when the random variables X and Y have a one-to-one relationship and is likely to be the case when the function from X to Y, h(X), is a many-to-one function. It is only necessary to sum or integrate over the appropriate set in X or Y to find the probability of the other set. Students often solve simple probability problems this way.

Applications

I present two applications of induced probability in finite sample spaces and two examples of IP in infinite sample spaces. The first finite example and the first infinite example use only simple induced probabilities with 1-to-1 correspondence between the new random variable and the original random variable. The second finite example illustrates an induction where the function is many-to-one.

Application 1 Finite Sample Space

Consider the following example: A sample space, S, consists of the results when three coins are tossed. The random variable X is defined as the number of heads in the toss. Table 1 illustrates this situation with the probabilities calculated assuming a fair coin.

Table 1

S	x	P(X = x)
HHH	3	1/8
HHT HTH THH	2	3/8
HTT THT TTH	1	3/8
TTT	0	1/8

Now consider a second random variable, Y, defined as 1 if $X < 2$ and 2 if $x \geq 2$. Table 2 illustrates this situation. The probability distribution of Y is easily found by using the method of induced probabilities. The probability that $Y = 2$ is the probability that $X = 3$ or $X = 2$. Similarly the probability that

Table 2

S	x	P(X = x)	y	P(Y = y)
HHH	3	1/8		
HHT HTH	2	3/8	2	1/2
THH HTT				
THT TTH	1	3/8	1	1/2
TTT	0	1/8		

$Y = 1$ is the probability that $X = 1$ or $X = 0$. In this common situation Ip provides a framework for considering the probability problem.

Application 2 Finite Sample Space

The discrepance was defined by Neel (1970) as a measure of interjudge reliability. Consider a case where k judges each give a rating ranging from 1 to M. Each rating has a possible range of values from 1 to M. Label these ratings as x_1 x_2 ... x_k from the respective k judges. Define a function of these ratings, the spread, S, as:

$$S(x_1 \ x_2 \ \ldots \ x_k) = \Sigma (|x_i - x_j|),$$

where the summation is taken over all possible pairs of the k ratings.

Define a second function D, the discrepance, on S as

$$D(S) = S/\max(s).$$

That is, the discrepance of a set of ratings is the spread of that set of ratings divided by the maximum possible spread.

Now consider all possible all possible sets of ratings by the judges. Since there are k judges each producing 4 possible ratings. Simple counting rules give us M^k possible sets of ratings by the judges. We define $N = M^k$. If the judges are viewed as rating at random, then each of the sets of ratings by the judges is equally likely and each thus has probability 1/N. Since there is a

one-to-one relationship between the spread and the discrepance, the probabilities of the discrepance are the same as the corresponding probabilities of the spread. This is illustrated by Table 3 which is for the situation of 3 judges using a rating scale ranging from 1 to 4. In Table 3 the first column, labeled J, represents all possible ratings by the three judges in lexicographical order. The second column represents the spread calculated from the 3 ratings to the left on the same line. Columns 3 and 4 have been sorted by the size of the spread to make the induced probabilities easy to see. To the right of these longer columns is a column of spreads and their induced probabilities. Notice, for example that the probability that the spread is equal to 0 is 4/64 since there are four possible ratings which lead to a spread of zero. The 2 rightmost columns give the discrepance values and their probabilities. Since the discrepance values are simply the spread values divide by their maximum, 4, the discrepance probabilities are identical to those of the spreads by a second induced probability.

This listing of probabilities is probability distribution for the discrepance under the assumption of random ratings. The probability distribution can be used to calculate the mean and variance of the discrepance which can be used in other procedures such as hypothesis testing. Induced probability has thus provided a means of testing hypotheses regarding random judging.

Table 3
Defining the Discrepance through Induced Probability

Unsorted J	s	Sorted J	s
111	0	111	0
112	2	222	0
113	4	333	0
114	6	444	0
121	2	112	2
122	2	121	2
123	4	122	2
124	6	211	2
131	4	212	2
132	4	221	2
133	4	223	2
134	6	232	2
141	6	233	2
142	6	322	2
143	6	323	2
144	6	332	2
211	2	334	2
212	2	343	2
213	4	344	2
214	6	433	2
221	2	434	2
222	0	443	2
223	2	113	4
224	4	123	4
231	4	131	4
232	2	132	4
233	2	133	4
234	4	213	4
241	6	224	4
242	4	231	4
243	4	234	4
244	4	242	4
311	4	243	4
312	4	244	4
313	4	311	4
314	6	312	4
321	4	313	4
322	2	321	4
323	2	324	4
324	4	331	4
331	4	342	4
332	2	422	4
333	0	423	4
334	2	424	4
341	6	432	4
342	4	442	4
343	2	114	6
344	2	124	6
411	6	134	6
412	6	141	6
413	6	142	6
414	6	143	6
421	6	144	6
422	4	214	6
423	4	241	6
424	4	314	6
431	6	341	6
432	4	411	6
433	2	412	6
434	2	413	6
441	6	414	6
442	4	421	6
443	2	431	6
444	0	441	6

s Spread	P(S = s)	d Discrepance	p(D = d)
0	4/64	.00	4/64
2	18/64	.33	18/64
4	24/64	.67	24/64
6	18/64	1.00	18/64

Application 3 Infinite Sample Space

An interesting question is how good an estimate of the error variance is needed for the estimation of a sample size in a t-test. Given random sampling, the goodness of the estimator depends on the sample size. That is, the distribution of error variance depends on the sample size used to estimate the error variance. One way to answer this question might be to conduct a Monte Carlo study on the error variance for selected sample sizes and see how the distribution of power varies as the sample size varies. Induced probabilities offer an alternative method for studying the effect of sample size on the power estimation. Let us assume that a large effect size (Cohen, 1977, P63) is being investigated and that σ_e^2 is estimated from a sample of size 30. We can assume σ_e^2 of 100 and an effect of 8 without loss of generality. Power is then a random variable which is then defined on the random variable s_e^2. Probabilities for power thus become the induced probabilities induced on power by the probabilities of s_e^2. Under random sampling and normality assumptions s_e^2 has a chi-square distribution and since the relationship between s_e^2 and power is one-to-one, the induced probabilities for power are the same as the chi-square probabilities for s_e^2. Thus to find the cumulative power points it is only necessary to find the cumulative s_e^2 points and calculate the power based on these s_e^2 points. s_e^2 points may be found by solving the well known formula for chi-square:

$$\chi^2 = ((df)\, s^2) / \sigma^2 ,$$

where df = the degrees of freedom for s_e^2 or χ^2

χ^2 = a cumulative point of chi-square.

Solving for s^2 yields:

$$s^2 = \sigma^2 \chi^2 / (n-1).$$

Substituting the correct cumulant of chi-square in this formula gives the same cumulative point of the distribution of s_e^2. From a given value of s_e^2 one power value may be calculated. This value is then the second function whose probabilities have been induced by the distribution of s_e^2. Table 4 contains cumulative probability points of .025, .05 .25, .50, .75, .95, and .975 for s_e^2 and the corresponding power. Cases of df of 5, and 30 are listed. The cumulative probability points are found from the chi-square distribution which is the distribution for a sample variance. Note that the error variance terms cumulate up the page while the power terms cumulate down the page. This is due to the inverse relationship between error variance and power.

If table 4 were expanded to include a wide range of df for s_e^2, it would be possible to select a criteria for power then to make recommendations based on the table. For example, we might select the criteria that power be found to within .05 of the correct power with probability .95. From Table 4 we see that the probability is .95 that power will be within (.42, >.99) for df = 5 and will be within (.62, >.97) for df = 30. Clearly neither of these df's is large enough to meet the criteria. More extensive tables would answer the question of what df is needed to meet the selected criteria..

Table 4

Cumulative Distributions of s_e^2 and Power for

df = 5, 30, with α = .05, and for Large Effect Size.

df for s_e^2 = 5

Cumulative Probability for s_e^2	s_e^2	power	Cumulative Probability for power
.975	256.	.42	.025
.95	222.	.48	.05
.75	133.	.79	.25
.50	87.0	.83	.50
.25	53.6	.97	.75
.05	22.8	> .99	.95
.025	16.6	> .99	.975

df for s_e^2 = 30

Cumulative Probability for s_e^2	s_e^2	power	Cumulative Probability for power
.975	156.	.62	.025
.95	146.	.66	.05
.75	116.	.76	.25
.50	97.6	.82	.50
.25	81.7	.88	.75
.05	61.7	> .95	.95
.025	56.0	> .97	.975

Discussion

The three examples show that induced probabilities have uses. First, it has been my experience that students find it easier to work simple probability problems when presented with the framework of induced probabilities. If for not other reason this seems to be that the procedure is then a named technique. Named techniques are probably easier to recall than unnamed techniques and thus easier applied when solving problems. Second, induce probabilities have been shown to have application in developing a statistical technique. This is only an outgrowth of their use in solving problems. Third, induced probabilities have been shown to be an alternative to monte carlo investigations with the added advantage that the induced probability technique gives an exact answer. Induced probabilities are thus a useful technique for learning, developing, and studying statistical techniques.

References

Cohen, J. (1977). Statistical Power Analysis. New York: Academic Press.

Halmos, P.R. (1960). Naive Set Theory. Princeton: D. Van Nostrand

Meyer, P.L. (1965). Introductory Probability and Statistical Applications. Reading: Addison Wesley.

Neel, J.H. (1970). "The Discrepance, a Measure of Interjudge Reliability", paper presented at the 1970 AERA convention.