THE EFFECTS OF GRID SIZE AND APPROXIMATION TECHNIQUES ON THE SOLUTIONS OF MARKOV DECISION PROBLEMS

Roy Mendelssohn
Southwest Fisheries Center
National Marine Fisheries Service, NOAA
Honolulu, Hawaii 96812

August 1978

DRAFT FOR COMMENT

ABSTRACT

Computational results are reported for the effect grid choice has on a solution to a Markov decision problem. The method used to discretize and reduce a continuous state problem can have significant effects on both the optimal policy calculated and on the estimated value of that policy. Tail probabilities of the ergodic distribution of an optimal policy appear to be quite sensitive to grid choice. Several a posterior measures of the error due to approximating the original problem on a reduced grid are compared with the actual error found.

I. In applying Markov decision processes (MDP's) to real life situations, an initial decision the modeler must face is the choice of grid over which to solve the problem. This arises because the problem is often described first as a stochastic model over a continuous state space; or else, the initial problem may be too large to solve.

The modeler must select an appropriate grid, "collapse" the original problem into the smaller problem by some method, solve the smaller problem, and extend both the optimal policy and the optimal value function back to the original problem. In this paper, several different techniques for choosing a grid are explored. Collapsing an MDP into a smaller grid is mathematically equivalent to aggregating or approximating MDP's as discussed in [8, 9, 10]. Section II explores how well different techniques perform for a model that arises in salmon management. For each technique, several different measures of performance are explored. Section III compares the results of section II with bounds that have been proposed for aggregated MDP's [3, 8, 9]. These bounds are checked both for how close they are to the true value of the performance measure, and also for how well they reflect the superiority of one particular method of aggregating and disaggregating over another. The reader is assumed to be familiar with both the notation and terminology in [3, 8, 9, 11].

II. Salmon spawn upriver in Alaska, and then the young salmon swim downriver to the ocean. Depending on the river, in a few years

the salmon return upriver to spawn again. The salmon are harvested by various types of fishermen as they run upriver. The released salmon are called spawners, the returning salmons recruits.

Mathews [2] has postulated the following spawner-recruit model for salmon in the Naknek River in Alaska: let x_t be the number of recruits that come upriver in period t, y_t the number of spawners released at the end of period t, and:

$$x_{t+1} = e^{d} 6.727\, y_t \exp\{-0.859\, y_t\}$$

where d is a random variable distributed as N(0, 0.1444). For computation the problem is discretized on the following grid (in units of 10^6 fish):

$$X = \{0,\ 0.125,\ 0.25,\ 0.375,\ 0.5,\ 0.625,\ 0.75,\ 0.875,\ 1,\ 1.175,$$
$$1.25,\ 1.375,\ 1.5,\ 1.625,\ 1.75,\ 1.875,\ 2,\ 2.5,\ 3,\ 3.5,\ 4,$$
$$4.5,\ 5,\ 5.5,\ 6,\ 6.5,\ 7,\ 7.5,\ 8,\ 8.5,\ 9\}$$

as follows:

State space = X

Decision space $Y(x) = \{y \mid 0 \leq y \leq x;\ y \in X,\ x \in X\}$

Transition probabilities: $\Pr\{x_{t+1} \leq \varepsilon \mid y_t\} = \Pr\{d \leq \ln(\varepsilon) - a \mid y_t\}$

where $a = \ln(6.727) + \ln(y_t) - 0.859\, y_t$

Given any $y \in X$, the probability of going to $x \in X$ is calculated as the difference of the cdf given y evaluated at x and the next smallest value in X.

This 31 point grid is considered for the purposes of this paper as the "original" problem. A discount factor of $\alpha = 0.97$ is used throughout, and the problem is to:

$$\text{maximize} \sum_{t=1}^{\infty} \alpha^{t-1}(x_t - y_t)$$

The optimal value function, optimal policy, ergodic distribution and gain rate for the original problem are given in Table 2.1. It is known a priori (see [4]) that a base stock policy is optimal. The true optimal policy has a base stock size of 750,000 fish, a mean annual harvest of 1,856,382 fish, and only a 3% chance of long run extinction, given the model.

The original grid was partitioned into 12 subsets as follows:

Subset No.	States in subset
1	0
2	0.125, 0.25, 0.375, 0.5
3	0.625, 0.75, 0.875, 1
4	1.125, 1.25, 1.375, 1.5
5	1.625, 1.75, 1.875, 2
6	2.5, 3
7	3.5, 4
8	4.5, 5
9	5.5, 6
10	6.5, 7
11	7.5, 8
12	8.5, 9

For the first trials, a reduced grid of

$$X_1 = \{0, 0.5, 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9\}$$

was selected. $Y(x)$ is defined as $\{y | 0 \leq y \leq x;\ y \in X_1,\ x \in X_1\}$ and the transition probabilities are calculated as:

$$\Pr\{x_{t+1} = x \in X_1 | y_t\} = \sum_{\substack{s \in \\ \text{subset} \\ \text{including } x}} \Pr\left\{x_{t+1} = s \in X | y_t\right\}$$

This follows [1, 9].

Two different one-period return functions were tried: the first uses the point estimate x-y. The second uses the average return from all states in a subset for which y is a feasible decision. For example, the return from choosing zero from state 0.5 is $^5/_{16}$. This second one-period return function was tried because it is the one-period return that results from aggregating the LP tableau for the MDP using fixed weight row and column aggregation (see Zipkin [11]) in order to arrive at the same reduced problem. For the second trials, a grid of

$$X_2 = \{0, 0.25, 0.75, 1.25, 1.75, 2.5, 3.5, 4.5, 5.5,$$
$$6.5, 7.5, 8.5\}$$

was used. This is the midpoint approximation to the partition. Transition probabilities and one-period returns are calculated as before.

Solutions to the aggregate problems were extended to the larger problem in a variety of ways. The optimal policy, since it is known to be of base stock form, was extended by the following rule: let x_i be the state used in the grid from the i^{th} partition, then:

$$y = \begin{cases} x, \text{ if } A(x_i) = x_i \\ \min \left(A(x_i), x \right) \text{ otherwise} \end{cases}$$

where $A(\cdot)$ is an optimal policy function.

Extensions of the optimal value function were calculated two ways. Method 1 involves a constant extension across a subset [1, 9], that is:

$$f(x) = f(x_i) \quad \text{for } x \in \text{subset } i$$

Method 2 consists of performing one iteration of successive approximations on the extended value function of method 1.

Five measures of comparison are used between the true solution and an approximate solution. Let $f(x)$ be the optimal value function of the original problem, and $e\tilde{}_{f(x)}$ the extension of the optimal value function for a reduced problem. Then measure one [9] is:

$$\sup_{x \in X} \left| f(x) - e\tilde{}_{f(x)} \right|$$

The second measure is $\left| \sum_x f(x) - \sum_x e\tilde{}_{f(x)} \right|$. $\sum_x f(x)$ is the value of the

Measures 3, 4, and 5 arise from the relationship between the dual variables of a linear programming solution to an MDP, and the discounted normalized fraction of years that we observed state x and take action y (Sobel [7]). If the sequence v_i, $i = 1, \ldots,$ number of states, is such that $v_i / \sum_i v_i$ equals the initial probability of being in state i, then:

$$\frac{(1-\alpha)\ \bar{u}_y^x}{\sum_i v_i} = \text{discounted normalized fraction of years that state x is observed and action y is taken.}$$

where $\bar{u}_y^x$ is an optimal dual variable in the LP. Measure 3 is:

$$\left| \frac{0.03}{31} \times \sum_x f(x) - \frac{0.03}{12} \times \sum_x \tilde{f}(x) \right| .$$

If the initial probability distribution is uniform, then

$$\frac{(1-\alpha)}{\text{Number of states}} \times \sum_{x \in X} f(x) = (1-\alpha) \sum_{x \in X} \sum_{y \in Y(x)} \bar{u}_y^x\, G(x, y)$$

which is the sum over all states and actions of the discounted fraction of years that (x, y) is observed times the return during those periods. In a sense, it is like a "discounted mean" harvest.

Measure 4 is the same as measure 3, except now the optimal policy from the smaller model is evaluated on the larger grid.

Measure 5 compares the cumulative discounted fraction of years that the Markov chain arising from an optimal policy is less than or equal to a given value on the different grids.

The results presented here are for the discounted fraction of years that a state-action pair are taken, starting from an initial uniform distribution. It should be mentioned, however, that similar qualitative results as what follows have also been found for the following cases:

(1) Discounted fraction of years starting from a non-uniform distribution

(2) Mean and tail behavior of the stationary distribution of the Markov chain that arises from following an optimal policy.

The particular results are chosen because they allow for consistent comparison of aggregation results from dynamic programming and from linear programming.

The eight trials are as follows:

Trial No.	State space	One-period return function	Method of extension
1	X_1	Point return	Constant over subset
2	X_1	Average return	Constant over subset
3	X_1	Point return	Successive approximation step
4	X_1	Average return	Successive approximation step
5	X_2	Point return	Constant over subset
6	X_2	Average return	Constant over subset
7	X_2	Point return	Successive approximation step
8	X_2	Average return	Successive approximation step

The results are summarized in Table 2.2 and in Figure 2.1. Average return from a given action over the partitions consistently performs better than the point return, for just about all grids and for most of the measures of comparison.

The midpoint of the partition as the selected state appears to give a better result than using endpoints. For this example it is necessary to be cautious in drawing conclusions, because the midpoint grid X_2 has the true base stock size in it, while the endpoint grid X_1 does not.

One iteration of successive approximation on the extended optimal value function produced only a small increase in accuracy in terms of measures 1 or 2. However, one iteration in all eight cases found the true optimal policy. Thus, though our estimate of the value function has not improved greatly, an optimal policy has been found. This one-step procedure is similar to that of [5], where the value of the equivalent deterministic model is calculated first. The computational equivalent of one iteration of successive approximation must be performed in calculating the bounds discussed in section III.

The one disturbing feature of this analysis is the comparisons of measure 5. In practice, in managing a fishery, both the probability of extinction and the probability of zero catch are of some concern, and often sensitivity analysis will be done describing the tradeoff between average per period catch and percentage of time there is no catch. Both grids X_1 and X_2, evaluated at their best policies (in the case of X_2 the true optimal policy) nearly triple discounted probability of extinction and the discounted fraction of years when there is no allowable catch as compared to the true value calculated on X (from 3.9% to 9.3%). Further, the estimated "discounted mean" harvest is off by amounts ranging from 12,000 fish to 1.6 million fish. This suggests that particularly when post-optimality analysis is being done, the grid should be expanded step by step until some stability is achieved in the key areas of interest.

One final trial was done. The absorbing state (zero) was dropped from the grid. The results are summarized in Table 2.3. The

value of the LP solution is 1,709.6359. The optimal policy is base stock with stock size 0.75. The "discounted mean" harvest is 1.6028. It is worth mentioning that when this problem is solved using successive approximations it took 320 iterations to converge at a tolerance of 0.0001. However, when methods of extrapolating forward to put bounds on the true value function are used [6], the same procedure converges in five iterations at the 0.0001 level. Computations with the absorbing state can be modified similarly in order to reduce the computational effort necessary.

III. In the previous section, a comparison is made of several ways of choosing a grid and aggregating a larger problem into a smaller problem. In that instance, the "true" original solution is known. However, usually grids and aggregation procedures are applied to problems where the "true" solution is never found. Several authors [3, 8, 9, 10] have developed bounds to compare the approximate solution with the unknown true solution. The reader is assumed to be familiar with the terminology and notation of these papers, and is referred to those articles for further details.

In this section, these bounds are calculated for several of the trial runs of the previous section. The different bounds are compared both as to how "tight" they are, and also as to how well they reflect the fact that one reduced problem gives a "better" solution than another. This certainly would be a desirable feature of any bound.

Whitt [9] derives both a priori and a posterior bounds of the form of measure 1:

$$\underset{\substack{x \in X \\ y \in Y(x)}}{\text{maximum}} \left| h\left(x, y, e_{\tilde{f}(x)}\right) - \tilde{h}\left(x, y, \tilde{f}(x)\right) \right|$$

where $$h\left(x, y, e_{\tilde{f}(x)}\right) = G(x, y) + \alpha \sum_j P^y_{xj} e_{\tilde{f}(j)}$$

and $\tilde{h}\left(x, y, \tilde{f}(x)\right)$ is similarly defined for the reduction of the full model into the smaller model. The a posterior bounds are calculated for trials 1, 2, 5, and 6, and the results are summarized in Table 3.1.

Table 3.1

Trial No.	A posterior bound	True value
1	379.5298	6.5643
2	347.57	2.3800
5	353.3910	6.4973
6	314.996	1.022

These bounds are very loose. For all of the trials, the bounds are much greater than the true optimal value function. The bounds do reflect, however, the superior performance of the grid X_2 and the average return function over the partition.

Bounds for measure 2 are given by [3, 8, 11]. These are summarized in Table 3.2. These bounds are calculated for the extended value function.

Table 3.2

Trial No.	Unimproved bound	Improved bound	Actual value of the absolute difference
1	$772.499 \leq z^* \leq 2{,}137.013$	$772.499 \leq z^* \leq 2{,}137.013$	164.2652
2	$689.79 \leq z^* \leq 2{,}978.533$	$689.79 \leq z^* \leq 2{,}314.5908$	60.4832
5	$646.1337 \leq z^* \leq 2{,}639.7954$	$646.1337 \leq z^* \leq 2{,}278.2525$	179.6471
6	$707.908 \leq z^* \leq 2{,}738.113$	$707.908 \leq z^* \leq 2{,}194.9325$	12.0681

The upper bounds for the partitions, p_k in the notation of [11], are given below.

Partition	Variables aggregated into choosing this action from each state in the partition										
	0	0.5 (0.25)	1 (0.35)	1.5 (1.25)	2 (1.75)	3 (2.5)	4 (3.5)	5 (4.5)	6 (5.5)	7 (6.5)	8 (7.5)	9 (8.5)
1	0	--	--	--	--	--	--	--	--	--	--	--
2	0	10	--	--	--	--	--	--	--	--	--	--
3	0	10	25	--	--	--	--	--	--	--	--	--
4	0	0	125	125	--	--	--	--	--	--	--	--
5	0	0	500	500	125	--	--	--	--	--	--	--
6	0	0	500	500	125	50	--	--	--	--	--	--
7	0	0	250	250	125	50	25	--	--	--	--	--
8	0	0	250	250	125	50	25	10	--	--	--	--
9	0	0	250	250	125	50	25	10	--	--	--	--
10	0	0	250	250	125	50	25	10	--	--	--	--
11	0	0	250	250	125	50	25	10	--	--	--	--
12	0	0	250	250	125	50	25	10	--	--	--	--

These bounds are reasonably tight, however they do not reflect which method of aggregation actually performs better. These bounds are tighter than the previous one for two reasons. Firstly, the bound in [9] essentially uses $\alpha/1-\alpha$ as the value for p_k, assuming there are 31 individual partitions. In this case, $\alpha/1-\alpha = 32.333$ which is 50% of any of the true values. The bounds in [8] calculated above allow for greater precision in the choice of p_k.

Secondly, the first bound includes any difference, whether positive or negative in sign. The second bound has a zero contribution to the error term from any column for which the extended vector is dual feasible in the LP solution to an MDP. This eliminates the entry that causes the larger absolute deviations in the first bound.

IV. Discussion

Several methods of approximating an MDP have been explored, and a posteriori bounds have been tested. The results seem to indicate:

1) When aggregating, it is best to use a state at a midpoint of the partition as the representative state from each subset of the grid.

2) The return function should be aggregated as the average return from each state in the partition for which the given action is feasible. Whenever feasible, one iteration of successive approximation should be done on the extended value function.

3) The actual error, by several criteria, is quite small. However, several key features of the long run probabilistic behavior of an optimal policy are sensitive to the grid size.

4) A posterior bounds can often be very loose, and may not reflect in fact how well one aggregate problem approximates the true problem as compared with a second aggregate problem.

5) The most important factor influencing the a posterior bounds is the value used to "blow up" the error terms by some appropriate amount. Little seems to be known of systematic or better ways to choose the values of these terms.

6) Convergence of successive approximations for the real models considered here is greatly speeded up by extrapolating an upper and lower bound for the true optimal value function, and using the largest of these differences as the convergence criterion. For one example, at a 0.0001 tolerance level, convergence was achieved in 5 iterations using this method as compared to 320 iterations required otherwise.

These results are not definitive, as they deal with one model over a very limited choice of grids. However, in most instances, the results can be explained in terms of the problem; this makes it reasonable to assume that these results are valid beyond this problem.

LITERATURE CITED

[1] D. P. Bertsekas, Dynamic Programming and Stochastic Control, Academic Press, New York (1976) - Chapter 6, pp. 225-295.

[2] Stephen B. Mathews, The Economic Consequences of Forecasting Sockeye Salmon Runs to Bristol Bay, Alaska: A Computer Simulation Study of the Potential Benefits to a Salmon Canning Industry From Accurate Forecasts of the Runs, Ph.D. Dissertation, Univ. of Washington (1967), Seattle.

[3] R. Mendelssohn, "Improved Bounds for Aggregated Linear Programs," S.W.F.C. Admin. Report, 16H, NMFS, NOAA (1978), 19 p.

[4] R. Mendelssohn and M. J. Sobel, "Capital Accumulation and the Optimization of Renewable Resource Models," Submitted to Journal of Economic Theory (1977), 38 p.

[5] J. Norman and D. White, "A Method for Approximate Solutions to Stochastic Dynamic Programming Problems Using Expectations," Operations Research 16, 296-306 (1968).

[6] E. Porteus, "Some Bounds for Discounted Sequential Decision Processes," Management Sci. 18, 7-11 (1971).

[7] M. J. Sobel, "Dual Variables of Discounted MDP's," Paper Presented at Joint National TIMS/ORSA Meeting, New York City, May 1-3 (1978).

[8] M. J. Sobel and P. Zipkin, Unpublished Manuscript.

[9] W. Whitt, "Approximations of Dynamic Programs, I," To appear in Mathematics of Operations Research (1978).

[10] W. Whitt, "Approximations of Dynamic Programs, II," To appear in Mathematics of Operations Research (1978).

[11] P. Zipkin, "Bounds on the Effect of Aggregating Variables in Linear Programs," Research Paper No. 211, Graduate School of Business, Columbia University (1977), 29 p.

Table 2.1.--Solution to the original problem.

State	Optimal value function	Discounted fraction of years that x_t is no greater than the given value
0	0	0.0323
0.125	59.7464	0.0333
0.25	60.3404	0.0342
0.375	60.7503	0.0353
0.5	61.0846	0.0364
0.625	61.3370	0.0376
0.75	61.5213	0.0391
0.875	61.6463	0.0416
1	61.7713	0.0460
1.125	61.8963	0.0536
1.25	62.0213	0.0660
1.375	62.1463	0.1078
1.5	62.2713	0.1389
1.625	62.3963	0.1760
1.75	62.5213	0.2183
1.875	62.6463	0.2645
2	62.7713	0.4620
2.5	63.2713	0.6398
3	63.7713	0.7718
3.5	64.2713	0.8594
4	64.7713	0.9140
4.5	65.2713	0.9472
5	65.7713	0.9668
5.5	66.2713	0.9786
6	66.7713	0.9858
6.5	67.2713	0.9903
7	67.7713	0.9933
7.5	68.2713	0.9954
8	68.7713	0.9971
8.5	69.2713	0.9991
9	69.7713	1.0
	1,918.167158	

Table 2.1.--Continued.

Optimal policy: Base stock given by y = minimum (x, 0.75)

"Discounted mean" harvest: 1.856382

Value of LP solution: 1,918.167158

Table 2.2a

Trial No.	Optimal policy	$\Sigma\tilde{f}(x)\left(\Sigma e_{\tilde{f}(x)}\right)$	Measures of "error" [1] Supremum norm on value function	[2] Deviation of LP solution	[3] Deviation of "discount mean" harvest	[4] Subopt. policy in full model
1	y = min(x, 1)	772.4990 (2,082.4323)	6.5643	164.2652	0.0750	0.0498
2	y = min(x, 1)	689.7904 (1,857.6839)	2.3800	60.4832	0.1318	0.498
3	y = min(x, 0.75)	(2,075.0572)	5.2263	156.8900	0.1518	0
4	y = min(x, 0.75)	(1,855.6727)	2.4056	62.4944	0.0605	0
5	y = min(x, 0.75)	646.1337 (1,738.5200)	6.4973	179.6471	1.6153	0
6	y = min(x, 0.75)	707.9084 (1,906.0990)	1.022	12.0681	0.0865	0
7	y = min(x, 0.75)	(1,742.8149)	6.1141	175.3523	0.1697	0
8	y = min(x, 0.75)	(1,905.8321)	0.4174	12.3351	0.0119	0

Table 2.2b

Discounted fraction of years that x_t is no greater than given value.

	Base stock size = 0.75 Midpoint of partition used	Base stock size = 1 Endpoint of partition used
x	Fraction of years	Fraction of years
0	0.0833	0.0833
0.25	0.0858	
0.5		0.0858
0.75	0.0934	
1		0.0909
1.25	0.1804	
1.5		0.1531
1.75	0.4852	
2		0.4187
2.5	0.7795	
3		0.7238
3.5	0.9151	
4		0.8858
4.5	0.9658	
5		0.9529
5.5	0.9844	
6		0.9792
6.5	0.9922	
7		0.9903
7.5	0.9965	
8		0.9957
8.5	1	
9		1

Table 2.3.--Optimal value function when zero is removed from the grid.

State	f(x)
0.125	53.1691
0.25	53.7283
0.375	54.0825
0.5	54.3591
0.625	54.5684
0.75	54.7241
0.875	54.8491
1	54.9741
1.125	55.0991
1.25	55.2241
1.375	55.3491
1.5	55.4741
1.625	55.5991
1.75	55.7241
1.875	55.8491
2	55.9941
2.5	56.0991
3	56.5991
3.5	57.0991
4	57.5991
4.5	58.0991
5	58.5991
5.5	59.0991
6	59.5991
6.5	60.0991
7	60.5991
7.5	61.0991
8	61.5991
8.5	62.0991
9	62.5991
Value of LP solution =	1,709.6359
Opt policy =	y = min (x, 0.75)

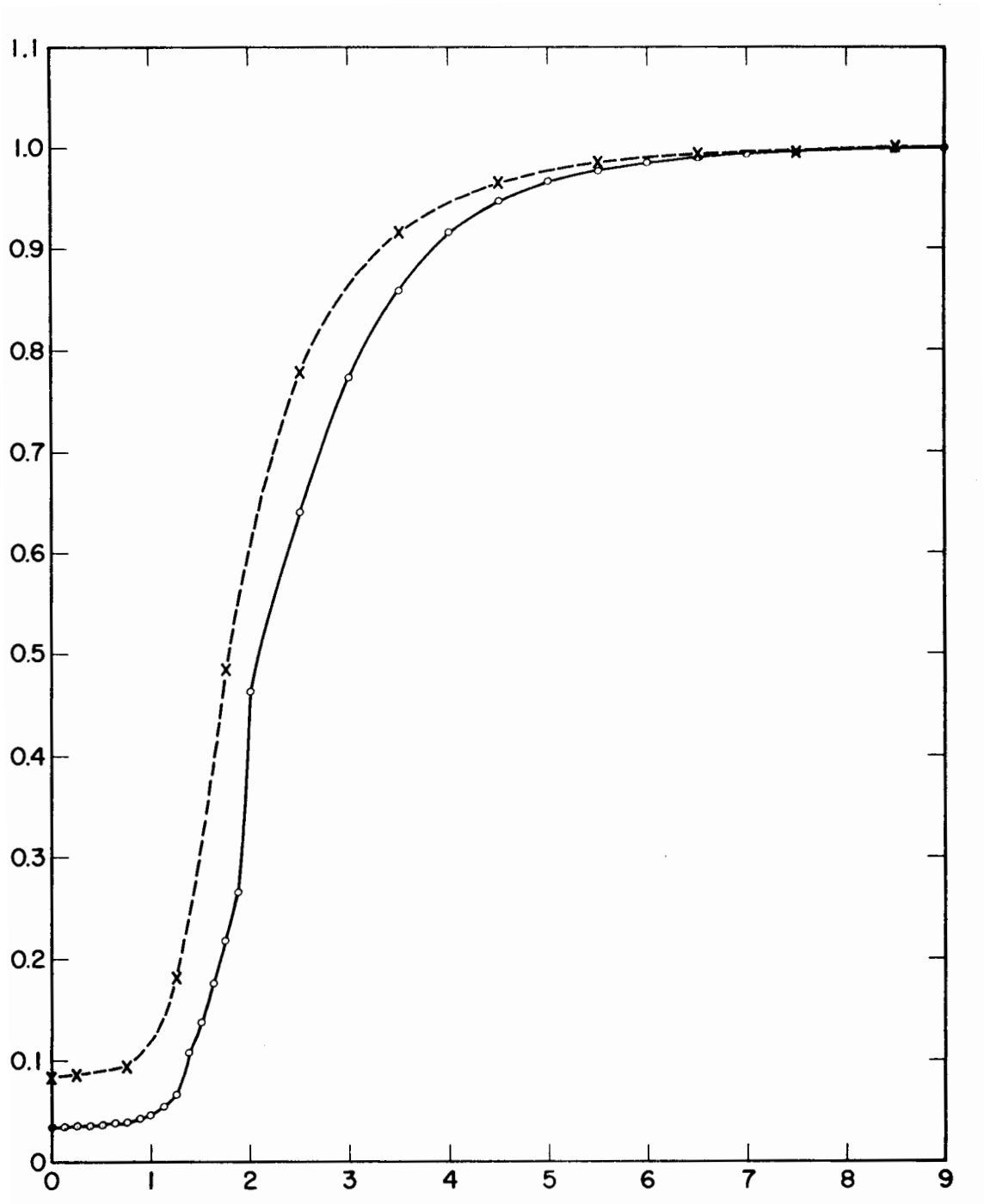

Figure 2.1.--Graph of distributions from Table 2.2b.