A preliminary model for lot sizing in semiconductor manufacturing

Tali F. Carmon Steven Nahmias

1. Introduction

In this paper, we develop a preliminary mathematical model for determining lot sizes for a single product which has several quality levels. The quality level of each unit is determined ("binned") by testing the product after it is produced. Binning based on performance after production is common in several industries, but the application which motivated the model developed here is semiconductor manufacture. A typical example is a microprocessor that operates at one of several different speeds. Specifically, consider the 80486 microprocessor developed by Intel Corporation of Santa Clara, California. (The 486 chip is currently the processor of choice for IBM compatible personal computers). At the current time, the 486 operates at three speeds: 25, 33 and 50 MHz. The production process is exactly the same in each case. After the chip is produced, it is tested and binned based on its performance.

The issue that we address in this paper is to determine the value of the lot size (x) so that we are guaranteed with specified probabilities that the demands for the various grades of product are satisfied. We also assume, as is often done for problems of this type, that we may substitute higher quality products for ones of lower quality, but not vice versa. While several related studies have addressed similar issues in more general settings, our approach will yield a solution which is easy to compute and easy to implement.

2. Literature review

Uncertainty of supply is a concept that has been studied by many researchers. Karlin [1] considered supply uncertainty in the context of one-stage models with uncertainty. Karlin [1] considered the standard newsboy type single-period stochastic inventory model but generalized this model to the case where the amount received is a random variable whose distribution depends on the size of the order. Giffler [2] formulated a model in which the uncertainty of the supply is a consequence of a production process in which items might be defective. Giffler [2] assumed that the number of defects is a random variable whose distribution depends on the number of items produced. The goal of the analysis is to find an optimal "reject allowance", which is an amount over and above the normal lot size to compensate for defectives in the lot. Levitan [3] considered the same model but provided a more rigorous analysis, and a different computing algorithm.

Because of the increased emphasis on manufacturing competitiveness and, more specifically, on quality in manufacturing, there has been considerable interest in extensions of the types of models considered by Giffler [2] and Levitan [3]. We will not provide a complete review here, but refer the interested reader to the recent related papers by Lee and Yano [4] and Henig and Gershak [5] and references contained therein.

Two studies which are much closer to the spirit of ours are due to Bitran and Dasu [6] and Bitran and Leong [7]. In the former study, the authors formulate a multi-period, multi-product production planning problem in which each product is binned according to a single attribute after production. Determining the optimal lot size for production of each product is referred to as the "morning" problem. Once production is completed, the yields in each quality category are realized. The next problem, referred to as the "evening" problem, is to allocate the existing stocks in an optimal fashion. The assumption is that there are known demands in each quality category. Both optimal and heuristic solution techniques are-explored. The latter reference explores essentially the same problem, but assumes that customer demand is satisfied from inventory $\alpha\%$ of the time, where α is a given constant between zero and one representing a type 1 service level (see [8] for a discussion of service levels). The model developed is also multi-product and multi-period and results in a large chance constrained formulation of the problem. Heuristic solution methods are recommended.

3. Assumptions and notation

We consider a single product, which after production, is classified into one of n quality levels according to a single attribute. The levels will be designated $1, 2, \ldots, n$. We assume that the quality levels are ordered in descending order. That is, quality level 1 is the best level, then level 2 and so forth. Furthermore, we will assume that customers requesting product of quality level 1 will not accept product of lower quality as a substitute, but customers demanding a lower quality product will accept a higher quality product as a substitute.

Suppose that x is the total size of a production run. We assume that the numbers of items which are binned into the n quality levels follows a multinomial distribution. That is there are probabilities $p_1, p_2, \ldots, p_n$, representing the likelihood that a unit of production is binned into each category. Since this is a multinomial distribution we require the condition that $p_i \geqslant 0$ for each i and $\Sigma p_i = 1$. It is possible that one or more of the items produced in a batch may be defective. In that case we could consider quality level n representing defective items and assume it has zero demand and zero service level. Define the random variables y_i as the number of units of quality level i produced. Then according to our assumptions, y_i is a binomial random variable with parameters x (# of trials) and p_i (probability of success). (It is well known that marginals of the multinomial have the binomial distribution.) Suppose that the demand for product of quality level i is a known constant d_i. Finally suppose that there are given constants $\alpha_1, \alpha_2, \ldots, \alpha_n$ between zero and 1 representing the required level of type 1 service for each quality level. Type 1 service means that we would like to satisfy *all* the demand for quality level i with probability α_1. The problem is to find the minimum value of x to accomplish this goal.

4. Analysis

The service level requirements and the assumption that higher quality levels may be substituted for lower quality levels gives rise to the following nested set of inequalities. For quality level of grade 1,

$$P\{y_1 \geqslant d_1\} \geqslant \alpha_1.$$

This says that the number of units of quality level 1 produced should meet all the demand for quality level 1 with probability at least α_1. Consider the demand for quality level 2. Allowing for the possibility that $y_1 > d_1$, part of the demand for quality level 2 may be satisfied by the overage for level 1 production. This gives rise to the chance constraint:

$$P\{y_1 + y_2 \geqslant d_1 + d_2\} \geqslant \alpha_2.$$

Note that these two constraints *taken together* guarantee the demand for quality levels 1 and 2 are both met with the desired service level. The general constraint is

$$P\{y_1 + y_2 + \cdots + y_i \geqslant d_1 + d_2 + \cdots + d_i\} \geqslant \alpha_i \quad \text{for } 1 \leqslant i \leqslant n.$$

The goal of the analysis is to find the minimum value of x so that these n constraints are simultaneously satisfied. We should point out that our model is considerably simpler than that considered by Bitran and Leong [7] who also assumed a type 1 service level criterion, multiple periods and multiple products. The advantage of our approach is that we allow for different service levels. We also obtain an explicit algebraic solution.

Define the random variable W_i as the partial sum of the random variables $y_1, \ldots, y_i$. That is,

$$W_i = y_1 + y_2 + \cdots + y_i.$$

It is easy to see that W_i also has the bionomial distribution with parameters x and $r_i = p_1 + p_2 + \cdots + p_i$. For large x it is well known that the binomial distribution can be closely approximated by the normal distribution (see [9] for example). The approximate values for the mean and variance are

$$\mu_i = x\left(\sum_{j=1}^{i} p_j\right) = x r_i,$$

$$\sigma_i^2 = x\left(\sum_{j=1}^{i} p_j\right)\left(\sum_{j=i+1}^{n} p_j\right) = x r_i(1 - r_i).$$

For convenience of notation, define $D_i = d_1 + d_2 + \cdots + d_i$. Then the i chance constraints can be written in the shorthand form:

$$P\{W_i \geqslant D_i\} \geqslant \alpha_i \quad \text{for } 1 \leqslant i \leqslant n.$$

Approximately each W_i by a normal random variable with mean μ_i and standard deviation σ_i means that the chance constraints become

$$P\left\{Z \geqslant \frac{D_i - x r_i}{\sqrt{x r_i(1 - r_i)}}\right\} \geqslant \alpha_i \quad \text{for } 1 \leqslant i \leqslant n,$$

where Z has the standard normal distribution. Because Z is continuous, we are guaranteed that there is a solution if we set the probability on the left equal to α_i. Doing so results in the equation

$$\frac{D_i - x r_i}{\sqrt{x r_i(1 - r_i)}} = \Phi^{-1}(\alpha_i),$$

where Φ^{-1} is the inverse of the complementary cumulative standard normal distribution function. This is equivalent to

$$D_i - x r_i = \sqrt{x r_i(1 - r_i)}\, \Phi^{-1}(\alpha_i).$$

The goal is to solve for x. Since each value of i between 1 and n defines a different equation and, hence, a different value of x, we will henceforth append x with a subscript. That is, the solution to this equation will be denoted x_i. This is easily seen to be a quadratic equation in x_i. Squaring both sides and rearranging terms so that the equation is in the standard form of $ax^2 + bx + c$, we have:

$$x_i^2[r_i^2] + x_i[-2r_i D_i - z_i^2 r_i(1 - r_i)] + D_i^2 = 0,$$

where we have defined $z_i = \Phi^{-1}(\alpha_i)$ for convenience.

The solution to the quadratic equation is

$$x_i = \frac{1}{2r_i}\{2D_i + z_i^2(1 - r_i) \pm z_i\sqrt{z_i^2(1 - r_i)^2 + 4D_i(1 - r_i)}\}.$$

Because z_i is the *complementary cumulative* distribution function, it follows that z_i is negative for service levels more than 50%. The solution, x_i, is the larger of the two roots. The value of the lot size x must satisfy *all* n chance constraints. Hence it follows that the required lot size, x, is given by

$$x = \max(x_1, x_2, \ldots, x_n).$$

5. Computations

To better understand the interrelationships among the various parameters in this model we have computed the optimal values of x_i for $n = 2$ and $n = 3$ for several values of the system parameters. The results for $n = 2$ appear in Table 1 and the results for $n = 3$ appear in Table 2. In Table 1 we compute the value of x_1 for nine values of the demands (d_1, d_2) and nine values of the probabilities associated with the two quality levels (p_1, p_2). We consider only the case where the service level for both quality levels is 90%. The numbers in both Tables 1 and 2 will just be scaled upward proportionally if higher service levels are considered. Since $d_1 + d_2 = 100$ in all cases, $x_2 = 100$ in each case in Table 1. Similarly, we assume $d_1 + d_2 + d_3 = 100$ in Table 2, resulting in $x_3 = 100$ in each case in Table 2. For this reason, Table 1 only contains values of x_1 and Table 2 only contains values of (x_1, x_2).

Consider first Table 1. Since $x_2 = 100$ in all cases what is interesting is which cases give $x_1 > 100$. Some of the results are different from what one might expect. For example, for the case $(p_1, p_2) = (0.5, 0.5)$ and $(d_1, d_2) = (50, 50)$ one might think that the values of x_1 and x_2 would be close since all parameter values are the same. However, Table 1 shows that in this case $x_1 = 203$. The reason that the value of x_1 is so much larger than x_2 is a consequence of the one way substitution. Excess amounts of product 1 can be substituted for product 2, but not vice-versa. Hence, we need to produce a much larger lot to satisfy the demand for *both* quality levels that we would need just considering quality level 2 by itself. Only when the demand for quality level 2 is substantially higher than that for quality level 1 or the yield for quality level 2 is substantially lower than that for quality level 1, is $x_1 < x_2$ (in which case $x = 100$). It is also interesting to note from Table 1 that for very low yield rates for quality level 1, the lot size required to satisfy a demand of 100 units can be as much as 25 times higher!

Table 2 shows results for the case of 3 quality levels. Again, we fixed the sum $d_1 + d_2 + d_3 = 100$, so that $x_3 = 100$ in each case. The values of x_1 and x_2 are shown under each parameter setting. The results are consistent with those observed for $n = 2$. When the yield rate for quality level 1 is low, the lot size must be increased substantially over the demand to be sure that the demand for level 1 product is satisfied. When the yield rates for quality level 1 are high, it is more common that the values of either x_2 or x_3 exceed x_1. Higher or lower service levels for quality level i result in proportionally higher or lower values of x_i.

If the yield rates, demands, and service levels for the various quality levels are approximately the same, it is clear that the value of x_1 will determine the optimal lot size, x, as we saw in the $n = 2$ case in Table 1. When one or more of these parameter values differ markedly, all values of x_i need to be computed to find the optimal lot size that guarantees that all quality levels are met with the desired levels of confidence.

6. Conclusions and extensions

We have presented a preliminary model of a production system in which a single product is binned into one of several quality grades after production.

Table 1
Values of x_1 for various parameter settings ($n = 2$) ($x_2 = 100$ and $\alpha_1 = \alpha_2 = 0.90$ in each case)

	(d_1, d_2)								
(P_1, P_2)	(10, 90)	(20, 80)	(30, 70)	(40, 60)	(50, 50)	(60, 40)	(70, 30)	(80, 20)	(90, 10)
(0.1,0.9)	307	587	867	1147	1427	1707	1987	2267	2547
(0.2, 0.8)	142	272	402	532	662	792	922	1052	1182
(0.3, 0.7)	87	167	247	327	407	487	567	647	727
(0.4, 0.6)	60	115	170	225	280	335	390	445	500
(0.5, 0.5)	43	83	123	163	203	243	283	323	363
(0.6, 0.4)	32	62	92	122	152	182	212	242	272
(0.7, 0.3)	24	47	70	93	116	139	162	185	208
(0.8, 0.2)	18	36	53	71	88	106	123	141	158
(0.9, 0.1)	14	27	40	54	67	80	94	107	120

Table 2

Values of (x_1, x_2) for various parameter settings $(n = 3)$ $(x_3 = 100$ and $\alpha_1 = \alpha_2 = \alpha_3 = 0.90$ in each case)

	(d_1, d_2, d_3)									
(p_1, p_2, p_3)	(10, 10, 80)	(30, 10, 60)	(50, 10, 40)	(70, 10, 20)	(10, 30, 60)	(10, 50, 40)	(10, 70, 20)	(20, 20, 60)	(20, 60, 20)	(60, 20, 20)
(0.1, 0.1, 0.8)	307	867	1427	1987	307	307	307	587	587	1707
	272	532	792	1052	532	792	1052	532	1052	1052
(0.1, 0.3, 0.6)	307	867	1427	1987	307	307	307	587	587	1707
	115	224	334	445	225	335	445	225	445	445
(0.1, 0.5, 0.4)	307	867	1427	1987	307	307	307	587	587	1707
	62	122	182	242	122	182	242	122	242	242
(0.1, 0.7, 0.2)	307	867	1427	1987	307	307	307	587	587	1707
	36	71	106	140	71	105	141	71	140	140
(0.3, 0.1, 0.6)	87	24	407	567	327	87	87	167	167	487
	115	225	335	445	335	335	445	225	445	445
(0.5, 0.1, 0.4)	43	123	203	283	43	43	43	83	83	243
	62	122	182	242	122	182	242	122	242	242
(0.7, 0.1, 0.2)	24	70	138	184	24	24	24	47	47	138
	36	71	123	158	71	106	141	71	141	141
(0.4, 0.4, 0.2)	60	170	280	390	60	60	60	115	115	335
	36	71	106	141	71	106	141	71	141	141

The model is only preliminary as it fails to consider several aspects of the real problem. One is that in most environments in which binning takes place there are capacity restrictions on the optimal lot size. While adding an upper bound to x would be a trivial extension in this case (one would simply produce the min (x, b) if b is the production capacity), the capacity issue is more complex than this. Most production processes involve multiple steps with differing capacities and yield rates at each step.

Another limitation is that our model is a static one. We considered a single production decision and a fixed and known demand pattern. In practice, log-sizing decisions are made on a continual basis. A multiperiod model would be much more realistic. Another feature of the real problem not considered here is demand uncertainty. However, even though our model lacks many features that would make it a realistic model of semiconductors manufacture, it can be useful for providing a "rough cut" first approximation for determining the size of the production run needed to satisfy demand for multiple quality levels.

Acknowledgement

Dr. Nahmias's contribution to this study was supported by the Competitive Manufacturing Institute at Santa Clara University. The authors are grateful to Professor Robert Leachman of Berkeley and Frank Barone of AMD Corporation for helpful discussions.

References

[1] Karlin, S., 1958. One stage models with uncertainty, in: Arrow, Karlin and Scarf (Eds), Studies in the Mathematical Theory of Inventory and Production. Stanford University Press, Stanford, CA.

[2] Giffler, B., 1960. Determining an optimal reject allowance. Naval Research Logistics Quarterly, 7: 201–206.

[3] Levitan, R.E., 1960. The optimal reject allowance problem. Mgmt. Sci., 6: 172–186.

[4] Lee, H.L. and Yano, C., 1988. A production control in multistage systems with variable yield losses. Oper. Res., 36: 269–278.

[5] Henig, M. and Gerchak, Y., 1990. The structure of periodic review policies in the presence of random yield. Oper. Res., 38: 634–643.

[6] Bitran, G.R. and Dasu, S., 1989. Ordering policies in an environment of stochastic yields and substitutable demands. Working Paper No. 3090-89-MS, MIT Sloan School.

[7] Bitran, G.R. and Leong, T., 1989. Deterministic approximations to co-production problems with service constraints. Working Paper No. 3071-89-MS, MIT Sloan School.

[8] Nahmias, S., 1989. Production and Operations Analysis. Irwin, Homewood, IL.

[9] Watson, C.J., et. al., 1990. Statistics for Management and Economics, 4th Ed., Allyn & Bacon, Boston, MA.