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 a\% of the time, where a 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, ..., 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, ..., 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 \geq 0 for each i and \sum 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, ..., \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 \alpha_i. 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,

\[ P\{y_1 \geq d_1\} \geq \alpha_1. \]

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 \alpha_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:

\[ P\{y_1 + y_2 \geq d_1 + d_2\} \geq \alpha_2. \]
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

\[ P\{y_1 + y_2 + \cdots + y_i \geq d_1 + d_2 + \cdots + d_i \} \geq \alpha_i \]

for \( 1 \leq i \leq n \).

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 binomial 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:

\[ P\{ W_i \geq D_i \} \geq \alpha_i \quad \text{for} \quad 1 \leq i \leq n. \]

Approximately each \( W_i \) by a normal random variable with mean \( \mu_i \) and standard deviation \( \sigma_i \) means that the chance constraints become

\[ P\left\{ Z \geq \frac{D_i - xr_i}{\sqrt{xr_i(1 - r_i)}} \right\} \geq \alpha_i \quad \text{for} \quad 1 \leq i \leq n, \]

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 \( \alpha_i \). Doing so results in the equation

\[ \frac{D_i - xr_i}{\sqrt{xr_i(1 - r_i)}} = \Phi^{-1}(\alpha_i), \]

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

\[ D_i - xr_i = \sqrt{xr_i(1 - r_i)} \Phi^{-1}(\alpha_i). \]

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:

\[ x_i^2 \left[ r_i^2 \right] + x_i \left[ -2r_iD_i - z_i^2r_i(1 - r_i) \right] + D_i^2 = 0, \]

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

The solution to the quadratic equation is

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

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 \( x_i \) 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 \(x_1 = 203.0\). 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 \(x_1 = x_2 = 0.90\) in each case)

<table>
<thead>
<tr>
<th>((P_1, P_2))</th>
<th>((10, 90))</th>
<th>((20, 80))</th>
<th>((30, 70))</th>
<th>((40, 60))</th>
<th>((50, 50))</th>
<th>((60, 40))</th>
<th>((70, 30))</th>
<th>((80, 20))</th>
<th>((90, 10))</th>
</tr>
</thead>
<tbody>
<tr>
<td>((0.1, 0.9))</td>
<td>307</td>
<td>587</td>
<td>207</td>
<td>1147</td>
<td>1427</td>
<td>1707</td>
<td>1987</td>
<td>2267</td>
<td>2547</td>
</tr>
<tr>
<td>((0.2, 0.8))</td>
<td>142</td>
<td>272</td>
<td>402</td>
<td>532</td>
<td>662</td>
<td>792</td>
<td>922</td>
<td>1052</td>
<td>1182</td>
</tr>
<tr>
<td>((0.3, 0.7))</td>
<td>87</td>
<td>167</td>
<td>247</td>
<td>327</td>
<td>407</td>
<td>487</td>
<td>567</td>
<td>647</td>
<td>727</td>
</tr>
<tr>
<td>((0.4, 0.6))</td>
<td>60</td>
<td>115</td>
<td>170</td>
<td>225</td>
<td>280</td>
<td>335</td>
<td>390</td>
<td>445</td>
<td>500</td>
</tr>
<tr>
<td>((0.5, 0.5))</td>
<td>43</td>
<td>83</td>
<td>123</td>
<td>163</td>
<td>203</td>
<td>243</td>
<td>283</td>
<td>323</td>
<td>363</td>
</tr>
<tr>
<td>((0.6, 0.4))</td>
<td>32</td>
<td>62</td>
<td>92</td>
<td>122</td>
<td>152</td>
<td>182</td>
<td>212</td>
<td>242</td>
<td>272</td>
</tr>
<tr>
<td>((0.7, 0.3))</td>
<td>24</td>
<td>47</td>
<td>70</td>
<td>93</td>
<td>116</td>
<td>139</td>
<td>162</td>
<td>185</td>
<td>208</td>
</tr>
<tr>
<td>((0.8, 0.2))</td>
<td>18</td>
<td>36</td>
<td>53</td>
<td>71</td>
<td>88</td>
<td>106</td>
<td>123</td>
<td>144</td>
<td>158</td>
</tr>
<tr>
<td>((0.9, 0.1))</td>
<td>14</td>
<td>27</td>
<td>40</td>
<td>54</td>
<td>67</td>
<td>80</td>
<td>94</td>
<td>107</td>
<td>120</td>
</tr>
</tbody>
</table>
Table 2
Values of $\{x_t > x^*\}$ for various parameter settings ($n = 3$) ($x^*_3 = 100$ and $x_1 = x_2 = x_3 = 0.90$ in each case)

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