Quality Improvement at a Semiconductor Equipment Manufacturing Facility Through Error Re-categorization and Proper Inventory Management

by

Elyud Ismail

B.Sc., Mechanical Engineering, Massachusetts Institute of Technology (2015)

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Master of Engineering in Advanced Manufacturing and Design

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2016

Author . . . .
Department of Mechanical Engineering
August 12, 2016

Certified by . . . .
Dr. Stanley Gershwin
Senior Research Scientist
Thesis Supervisor

Accepted by . . . .
Rohan Abeyaratne
Quentin Berg Professor of Mechanics
Chair, Committee for Graduate Student

Quality Improvement at a Semiconductor Equipment Manufacturing Facility Through Error Re-categorization and Proper Inventory Management

by

Elyud Ismail

Submitted to the Department of Mechanical Engineering
on August 12, 2016, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Advanced Manufacturing and Design

Abstract

This thesis addresses the improvement of first pass yield on the assembly floor of Varian Semiconductor Equipment and Associates (Varian) - a low volume complex semiconductor capital equipment manufacturing facility. First pass yield refers to the proportion of fully built modules that pass testing without the need for additional rework. The first pass yield (FPY) project began in 2011 and showed steady improvement for its first three years. But over the following two years, the primary yield metric has stayed level. The goal of the work presented here is to analyze the reasons for the leveling and propose novel ways of improving the metric once more. Two major quality improvement recommendations are presented in this thesis. The first is a new way of categorizing error reports that will lead to targeted corrective action to reduce the recurrence of pre-identified failure modes. A multi-step methodology is presented on how to determine an efficient corrective action, along with a case study based on data acquired from Varian. The second quality improvement recommendation involves reducing the number of part and sub-assembly shortages on the assembly floor that were shown to be correlated with poor first pass yield. A new management system for a subset of the total inventory at Varian is presented, along with a discussion on how to execute the recommendation.

Thesis Supervisor: Dr. Stanley Gershwin
Title: Senior Research Scientist

Acknowledgments

First and foremost, I would like to thank God for his unending grace, and the joy He has given me in attending MIT for the last 5 years. I thank Him for the strength and courage He has given me during the tough times and for the humility He has given me during times of success. Praise be to His name forever.

I would like to say a big thank you to my entire family for their support, encouragement and prayers as I went about completing this degree. A special thanks goes to my parents for being worthy role models in my pursuit of success.

I would like to thank the faculty and staff of the Master of Engineering in Manufacturing and Design for the effort and dedication they put in to making the program a success.

I express my heartfelt gratitude to everybody at Varian Semiconductor Equipment and Associates for their kindness and hospitality during my work their. I would like to thank my supervisor, Dan Martin, for his guidance and constant availability to discuss any issues. I would also like to thanks all the managers I had the previlege to work with and learn from during the project - Tom Faulkner, Tim Wood, Peter Justice, Ron Dognazzi, James Firth, Phil Hobbs, Howie Amaral, Mike Rathe, Terry Greel, Randy Fontaine, Paula Perry, Dave Adkins, Adam Mahoney and everyone else unnamed here.

I also would like to thank my advisor, Dr. Stanley Gershwin for his guidance in the completion of the internship project and the writing of this thesis. His way of thinking and approaching problems was immensely helpful to perform effectively work at Varian.

I would like to thank the Singapore University of Technology and Design for sponsoring my stay at MIT the past year.

I would like to thank my teammates, Sean Daigle and Shaswat Anand, for being intelligent and creative and making our work quite fun. I wish them well in their future endeavors.

Lastly I would like to thank all my friends and my roommates for their constant

support and encouragement during the past year.

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

One of the hallmarks of lean manufacturing, as championed by Toyota in the beginning of the 21st century, is known as *kaizen* - Japanese term for "continuous improvement". It is the relentless pursuit of optimization of a process such that workers and managers can never be satisfied with defects of any kind. The kaizen way of thinking one of the things that made the Toyota Production System (TPS) the classic case study for operational efficiency and superb yield.

This thesis documents an effort to optimize the continuous quality improvement project at Varian Semiconductor Equipment and Associates (Varian), a subsidiary of Applied Materials (AMAT). Located in Gloucester, MA, Varian is a manufacturer of semiconductor capital equipment that primarily does ion doping of wafers. It also services customers for repairs and replacement parts. The primary metric that is the focus of the continuous improvement project is First Pass Yield (FPY), which refers to the proportion of modules that pass testing without the need for additional rework. While the project yielded excellent results during the first 3 years since its inception in 2011 when Varian was acquired by AMAT - over the next couple of years the metric has been level at around 82%.

While Varian does not have a specific goal as to what the metric should read, the kaizen principle that they are trying to adhere to makes it impossible for them to be satisfied anything less than perfect yield. That is why every error report (referred to as "Quality Notification (QN)" at Varian) is investigated by a team of manufacturing managers to determine the root cause so that they can attempt to implement corrective measure to that cause. Despite the meticulous review of every QN, the FPY metric hasn't improved ever since it leveled out. That indicates a presence of an inefficiency somewhere in the facility that is currently not being addressed by Varian. This thesis aims to identify any source of inefficiency within the facility and provide recommendations to get rid of those inefficiencies so that the FPY metric can start improving again.

1.1 Thesis Project Execution

The work at Varian, which is partially documented in this thesis, was conducted by three MIT graduate students pursuing their Masters of Engineering in Advanced Manufacturing and Design degree. The students, referred to as "the MIT team" in this thesis, spent between February and May of 2016 working part time at Varian and spent between May and August working full time. The full body of work is distributed among the separate theses of the three students, one of which is the author of this thesis. Since the project was a collaborative effort, most of the work documented here will be described as being done by the MIT team. Any work that was the result of the insight of a particular team member will be cited to that person. Below are the names of the MIT team members, as well as citations to their individual theses that can be found in the bibliography section of this thesis.

1. The author - Elyud Ismail

2. Sean Daigle [5]

3. Shaswat Anand [6]

1.1.1 Summary of Contributions to Project

The effort of identifying inefficiencies in the current FPY improvement project led the MIT team into several avenues of analysis that are documented in the separate theses of the team members. While the introduction and problem statement as well as the description of the FPY program (chapters 1-3) will be similar across the three theses, the projects that were undertaken to eliminate those inefficiencies are written separately and will be unique to each thesis.

This thesis, in chapter 4, will emphasize the shortcomings of the current way of assigning QNs into categories that do not effectively lead the FPY team to implement targeted corrective measures to recurring failure mechanisms. It will propose a new way of categorizing the QNs that give the FPY team a better chance of addressing failure modes that prove to be problematic to the FPY metric, and present case studies based on an error re-categorization effort undertaken by the MIT team at Varian.

Daigle's thesis [5] will dedicate a chapter to discussing a bottleneck in one area of sub-assembly production that causes shortages of sub-assemblies that lead to quality problems. Those quality problems result in a QN being written and the FPY metric suffering as a result. His thesis will go in depth with what it would take to remove that bottleneck - both operationally and with added personnel.

Anand's thesis [6], will dedicate a chapter to discussing some of the short comings of the data collection system at Varian. It will go in depth about how bad data collection leads to reduced quality and provides specific recommendations on what type of data would need to be collected on the various processes that could lead to tangible improvement in FPY.

Chapter 5 of this thesis and one chapter of all Daigle's and Anand's theses will share the same theme of how part and sub-assembly shortages affect quality of the fully assembled machine, what the MIT team discovered as major errors in inventory management and their recommendations to mitigate them. Separate portions of this work will be presented in the three theses. Chapter 5 of this thesis, as well as chapter 4 of Daigle's thesis [5], will analyze the existing methodology of inventory management and present an alternative way of managing the inventory that would result in less part and sub-assembly shortages. Chapter 4 of Anand's thesis [6] will provide a detailed cost-benefit analysis of the alternative inventory management scheme developed by the MIT team.

1.1.2 Thesis Organization

Chapter 2 of this thesis will introduce Varian in a more detailed manner, as well as provide an in-depth discussion on the FPY project and the various metrics involved in tracking the state of quality at Varian. Chapter 3 will provide a diagnosis of FPY stagnation through statistical tests of various hypotheses that the MIT team developed. Chapter 4, as discussed above, will describe the short comings of the existing QN categorization system and recommend a specific methodology to make better use of the gathered QN data. Chapter 5 will discuss in depth how the existing inventory management system is flawed for certain categories of parts and the recommendation by the MIT team as to the proper way of managing that inventory. Conclusions and future work will be given in chapter 6.

Chapter 2

Background Information and Problem Description

2.1 Background Information and Problem Introduction

2.1.1 Background Information

This thesis project has been carried out at the manufacturing facility of Varian Semiconductor Equipments, a subsidiary of Applied Materials, Inc. (Nasdaq: AMAT), located in Gloucester, MA. Applied Materials, Inc. is the global leader in providing innovative equipment, services and software to enable the manufacture of advanced semiconductor, flat panel display and solar photo-voltaic products. Applied Materials purchased Varian Semiconductor and their ion implantation equipment manufacturing facility in Gloucester, MA in 2011. The Varian division of Applied Materials produces a variety of product lines all involved with ion implantation. The primary customers of Applied include all the major semiconductor chip producing companies.

Ion implantation is the most common process of doping semiconductors in the manufacturing of semiconductors. This process of doping a silicon wafer involves presenting the wafer to a focused and filtered ion beam. The beam begins as an ionized gas and is focused through a beam line of magnets that filters the gas to only the desired ions by the time the beam hits the wafer. This beam line equipment is complex and bulky, and as a result the equipment to refine this beam is manufactured in a series of modules. A typical complete machine would be made of around 7-8 modules. Each module is tested separately, and from time to time an entire machine would be assembled in a clean room within the facility for a complete test before shipping. Since not all machines undergo complete test at the Varian facility, it is imperative that each module is working properly before it ships out, lest it fail at a customer site.

A list of the modules commonly built at Varian, listed according to order the of gas flow through completed machine, is given below. Refer to cited work ([7], [8]) for a detailed description of each module.

1 Gas Box

2 90 Module

3 Beam Line

4 55/70 Module

5 UES Module

6 Buffer Module

7 Facilities Module

2.1.2 Layout of Assembly Floor

All assembly at Varian is done by hand tools. There are two major areas in which assembly takes place at Varian. The first is "the Supermarket". The supermarket is where piece parts are put together to form sub-assemblies. These sub-assemblies either go into one of the modules listed above, or they are made available for replenishing a broken part at the site of a customer's manufacturing facility, or they go into one of the many part banks located around the world. Whenever there is a demand for a sub-assembly, a shop order is opened for that sub-assembly. The shop order is opened by SAP, several days in advance of when that sub-assembly is needed unless the sub-assembly is one of few sub-assemblies that are managed on the Gold Square system. The Gold Square system forms a buffer between the supermarket and the demand so that whoever needs a sub-assembly can go a pick it up from the buffer instead of wait for it to be built. Sub-assemblies that are placed on the Gold Square system are typically high-use sub-assemblies that are critical for the build of a module. Maintaining the Gold Squares and making sure there are adequate number of sub-assemblies within each Gold Square are part of the inventory management scheme at Varian.

The second area of assembly is known as the "Flow line". This is the area in which modules are built from the sub-assemblies out of the supermarket as well as some piece parts. Each module has its own flow line, assembly procedure and test protocol. As mentioned before, every module is tested individually before being shipped to the customer. Once in a while, all the modules are integrated into a full machine. The full machine is then tested in a clean room that replicates the conditions of the customer site. The results of this test make up the backbone of the quality improvement project that is the focus of this thesis.

2.1.3 Brief Problem Description

After Applied Materials acquired Varian Semiconductors in 2011, the "First Pass Yield (FPY)" program was instituted as part of a continuous improvement effort. The program is targeted at reducing the number of quality defects per module and thereby minimizing the rework caused as a result of these quality defects. At the heart of all these is the reduction of cost to build a module as well as to provide products of superior quality to its customers. FPY is the percentage of modules manufactured without registering any manufacturing-related defects on the first test. The scope of the project deals with defects arising on the shop floor which are attributable to workmanship errors. This does not include errors arising out of a defective part from a supplier nor because of any inherent design issues.

This quality project carried out by Varian resulted in significant reduction in number of defects per module across all the modules in the first few years. The FPY metric increased from around 55% across modules in the fiscal year 2011 to about 80% in 2013, but has not shown meaningful improvement since then. The primary goals of this thesis project was to ascertain reasons for this lack of improvement, to critique their FPY program and to present Varian with a methodology that improves yield in the future.

2.2 First Pass Yield Program

First Pass Yield (FPY) is one of the most important metrics for the manufacturing division at Applied Materials, MA. It is the percentage of the modules that pass the final testing without any quality issue. FPY is calculated on a monthly basis. While there are FPY values associated with various types of quality issues, the scope of this thesis project is limited to those quality issues that arise from workmanship related errors. Any quality issue found out during the process of build or testing of a module is logged in as a Quality Notification(QN) in the ERP(SAP) system. QNs are failure reports written by workers on the assembly line during build or test. A detailed description of QNs is presented in section 2.3.

The recording of QNs makes up the building block of the FPY program and the calculation of the FPY metric, which will be discussed in section 2.4.

2.3 Quality Notifications

Manufacturing defects at Applied Materials are recorded into SAP using what are known as Quality Notifications (QNs). QNs are typically written when a defect is detected during the build sequence or during module testing. Workers are encouraged to report any defects they come across and put as much detail as necessary to give an accurate representation of the problem. In SAP, there is a field to classify the QN into predetermined categories, or "buckets," that aid manufacturing engineers that investigate the QN later on what to focus on. A discussion of the major buckets will be given in section 2.5.

A QN can be assigned against three major possible sources of error, also known as "cause codes". These are:

1 Manufacturing: This cause code is for QNs that arise as a result of workmanship related errors. These could range from a mishandled part to a wrong electrical connection or an improperly sealed vacuum chamber.

2 Supplier: This cause code is for QNs that arise as a result of material that are either not working when they arrive at the facility, or are not built to the required specifications by the supplier. These parts would typically be returned to the supplier for rework.

3 Design: This cause code is for QNs that arise as a result of poor design decisions that lead to parts being damaged or connections not being mated appropriately etc. These QNs would typically be investigated by a design engineer in the facility.

The scope of this thesis project was limited to QNs that were assigned the **Manufacturing** cause code. This was primarily due to relatively easy access to manufacturing engineers that would be available to answer questions as well as assist in executing possible corrective actions based on recommendations that would be made during the project.

All QNs are read and investigated by manufacturing and quality engineers at the facility. These engineers would go to the person who wrote a QN, request more details about what failed and where the problem might have originated. Quality engineers take note of parts that have been repeatedly failed and get to the bottom of what might be wrong with that particular part. Manufacturing engineers would look for process improvements to to avoid future failures of the kind in each QN.

Module Type	Nos. of Modules Built	Nos. of Modules without QNs	Module FPY %
A	x	u	$100 \times \frac{u}{x}$
B	y	v	$100 \times \frac{v}{y}$
C	z	w	$100 \times \frac{w}{z}$

Table 2.1: FPY Sample Calculation.

2.4 Calculating FPY

The FPY for any particular module for a time period is defined as the ratio of the number of modules built without any quality notifications (QNs) written against it to the total number of modules built in that period - typically one month.

Extending this definition, we calculate the FPY for the manufacturing unit as the weighted average of the individual modules' FPYs. Refer to Table 2.1 and Equation 2.1 for a symbolic representation of an FPY calculation.

$$FPY = \frac{(\frac{u}{x} \times x) + (\frac{v}{y} \times y) + (\frac{w}{z} \times z)}{x + y + z} = \frac{u + v + w}{x + y + z} \tag{2.1}$$

2.4.1 A related metric: QNs per Module

Along with First Pass Yield, another metric that the FPY team looks at and keeps track of is QNs per Module. This metric shows in more detail which modules are accumulating the most defects and hence require special attention. This information is masked in the FPY metric because the FPY only tracks the binary output of pass/fail. It only takes one QN on a module to generate a hit against the FPY of that module. The FPY metric doesn't distinguish between a module that had 1 QN logged against it and another module that had 10 QNs logged against it. It weights all modules the same way regardless of how many QNs might have been recorded against each module. By looking at QNs per Module, the team gets a more precise representation of the state of manufacturing quality at the facility.

QNs per Module is calculated according to equation 2.2

Location	Total Build	Passed	No. Defects	%FPY	Average QNs/Module
55/70 Mod Assy/Test	9	8	1	89	0.11
90 Mod Assy/Test	9	6	8	67	0.89
Gas Bod Mod Assy/Test	2	2	0	100	0.00
MC Term Assy/Test	2	2	0	100	0.00
MC BL Assy/Test	2	2	0	100	0.00
UES Mod Assy/Test	20	6	25	30	1.25
Final Assembly/Shipping	10	10	0	100	0.00
Buffer	10	10	0	100	0.00

Table 2.2: FPY for the month of December 2015.

Location	Total Build	Passed	No. Defects	%FPY	Average QNs/Module
55/70 Mod Assy/Test	14	12	3	86	0.21
90 Mod Assy/Test	14	6	12	43	0.86
Gas Bod Mod Assy/Test	14	14	0	100	0.00
MC Term Assy/Test	6	4	2	67	0.33
MC BL Assy/Test	6	6	0	100	0.00
UES Mod Assy/Test	20	6	25	30	1.25
Final Assembly/Shipping	4	4	0	100	0.00
Buffer	20	20	0	100	0.00

Table 2.3: FPY for the month of January 2016.

$$\text{QNs per Module} = \frac{\text{Total number of QNs}}{\text{Total number of modules}} \tag{2.2}$$

2.4.2 FPY Sample Data: December 2015 and January 2016

Two months of representative data on FPY as it would be presented to the FPY team are given in Table 2.2 and Table 2.3. For the given month, they tables show the locations, or the modules, how many of those modules where built, how many passed without any QNs, how many QNs were written in total against those modules that did not pass, the FPY percentage for that module and the average number of QNs for that modules.

2.5 The Bucketing Approach

As mentioned above, workers writing QNs have the ability to assign their QNs under predetermined categories or buckets. These buckets were established during the inception of the FPY program and were meant to capture the vast majority of QN types that were being recorded. The four major buckets a QN typically gets designated to are listed below:

1 Connections: This bucket is made up of QNs having to do with issues related to mechanical connections such as water fittings.

2 Harnessing: This bucket is made up of QNs having to do with electrical connections, both standard and fiber optic. These wires usually appear in large bundles or harnesses with some carrying as many as 20 individual wires.

3 Vacuum: This bucket is made up of QNs having to do with issues related to vacuum leakage. QNs in this bucket would typically contain failures associated with broken vacuum seals, damaged o-rings or debris located on mating surfaces that introduce leaks.

4 Parts: This bucket is made up of QNs having to do with broken or malfunctioning parts

Each bucket is assigned to a manufacturing lead who is tasked with investigating every QN under his/her bucket. These leads document their findings and brief the rest of the team in a weekly meeting. At this meeting, the leads present the results of their investigations, a discussion ensues about possible fix projects and the end result of the discussion is conveyed to the people on the assembly floor who originally wrote the QNs. The results might lead to a determination to contact vendors, or make design revisions or procedural changes.

2.5.1 Shortcomings of the Approach

The bucketing approach was in use for the past 5 years and yielded positive results in the first 2 years as a lot of systemic issues were solved across the different buckets. FPY improved dramatically in those years and QNs per module went drastically down across all modules.

But once all the "low hanging fruits" were removed, the FPY team has struggled to register meaningful improvement for the past 3 years. The FPY metric has consistently hovered around the low 80% mark. The rigidity of the existing bucket system did not allow the engineers to identify common failure mechanisms that might transcend across part numbers and different buckets. The system also restricts workers who are writing QNs from classifying the QN to an appropriate category. The MIT team at Applied Materials was given the freedom to think outside of the prescribed boundaries and think of novel ways to approach the quality problems. As a result the MIT Team was able to come up with a way of categorizing, or "re-bucketing", the QNs to alert the FPY team to possible common failure modes as opposed to strictly QN types. A more detailed description of the new QN re-bucketing approach is presented in chapter 4.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Analysis of the Lack of Improvement of the FPY

This chapter will cover the reasons why the FPY metric has not improved over the last previous years. It will analyze the metric itself for its appropriateness, delving into the various elements that go into its formulation, as discussed in the previous chapter. It will outline some ways in which the MIT team thought the metric was not sufficiently informative, and make recommendations on how to make the it carry more meaning. This chapter will also delve into several hypotheses regarding process-related reasons for the stagnation of the metric and the analyses that were done do validate those hypotheses.

3.1 Detailed presentation of FPY Stagnation

A brief overview of the FPY metric stagnation has been given in the introduction. Below, actual figures are presented that show the sharp increase in FPY in the first 2-3 years since the commencement of the project and the subsequent stagnation in the years since.

Figure 3-1 shows the overall trend in the FPY metric starting from fiscal year (FY) 2011 until the first quarter of FY 2016. It can be observed from Figure 3-1 that the FPY metric increased from 55% in 2011 to 66% in 2012 then to 76% in 2013 and

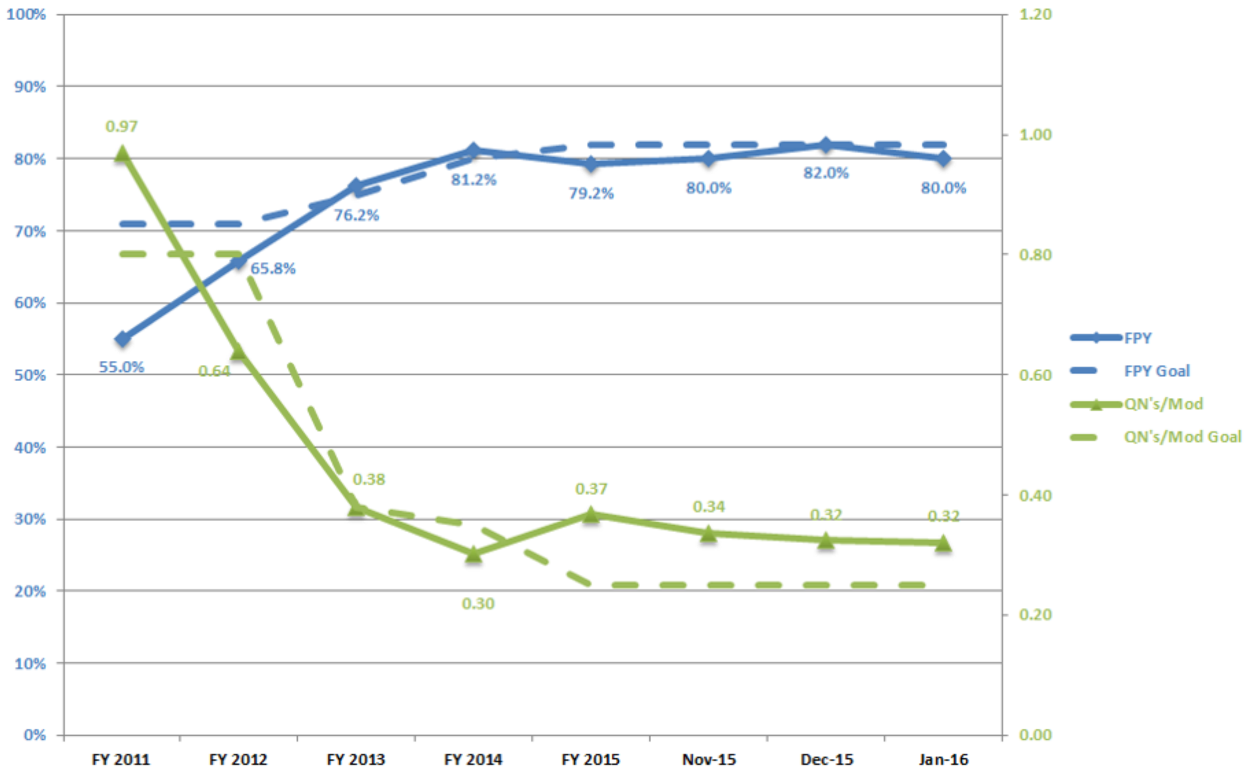
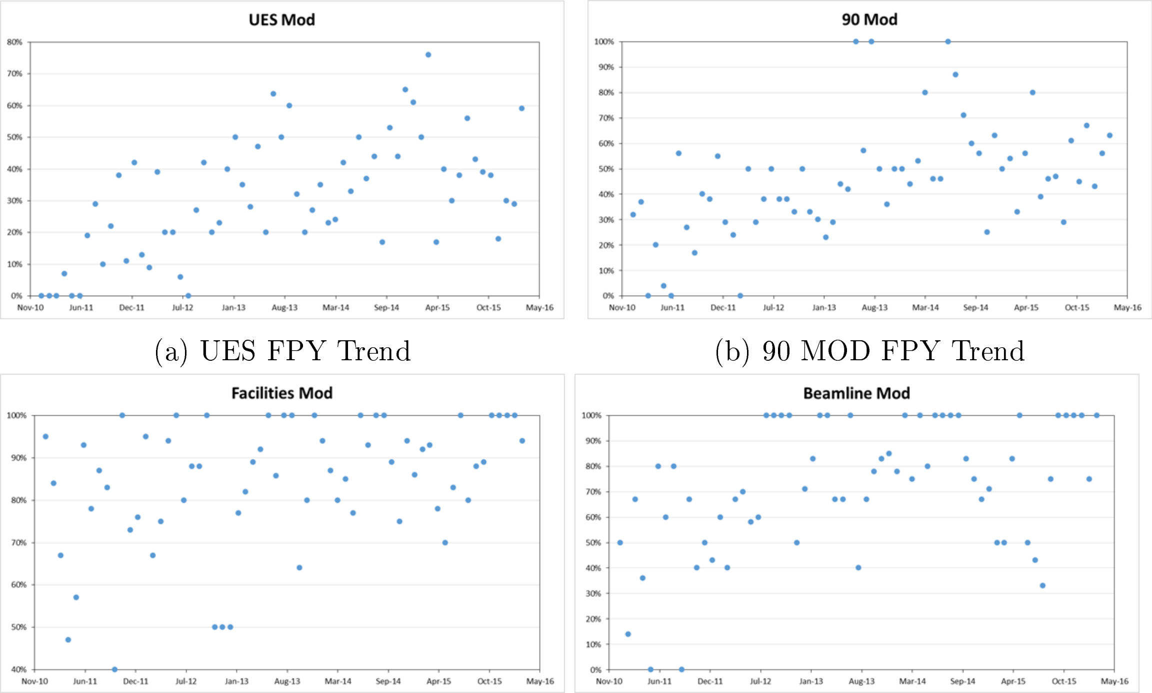


Figure 3-1: FPY Trend over past 5 years.

finally to 81% in 2014. But since 2014, the metric has been stuck around the 81% mark.

Similarly, the QNs per module metric has shown the same trend as the FPY, only in the reverse, during the same time period. It can be seen from Figure 3-1 that back in 2011 when the project was started, there were roughly 1 QNs per module on average. That value went down to about 0.3 in 2014 and has hovered around there ever since.

While looking at the overall trend is useful to get a big picture perspective of the quality improvements at Applied Materials, more information can be gained by looking at module wise trends in FPY and in QNs per module over the same time period. Figure 3-2 shows scatter plots of FPY values for a few modules during the duration of the FPY project.



(a) UES FPY Trend

(b) 90 MOD FPY Trend

(c) Facilities FPY Trend

(d) Beam Line FPY Trend

Figure 3-2: FPY Trends for Various Modules

3.2 Is FPY Metric Appropriate?

First Pass Yield is a widely used metric to get an understanding of the quality of a manufacturing facility. It gives a good snapshot of how optimized the processes are. It is general enough to report to executives of a company to give them a sense of the big picture yield of their manufacturing plant. Having said that, from a continuous improvement point of view, the FPY metric as used by Applied Materials has several flaws and doesn't paint a detailed enough picture of what is happening on the assembly floor. These flaws are discussed in the following subsections.

3.2.1 Weighting of Modules in FPY Formulation

The current formulation of the FPY metric was given above in Equation 2.1 and Table 2.1. It can be observed from that formulation that it is simply the total number of passed modules during a given month normalized by the total number of modules that was built in that month. Such a lumped parameter treats all modules the same way, each with equal opportunity to improve the metric. But what the metric does not take into account is the fact that not all modules are able to contribute the same way towards yield improvement. That is because some modules are more complex, thus have more opportunities for error than other modules.

If the aim of tracking FPY is for continuous quality improvement, it is essential that the FPY team identify modules that, if their own FPY was to be improved, it would make a large impact on the overall metric. Therefore any metric, in order for it to be meaningful, needs to take into account the number of error opportunities that exist in each module.

3.3 QNs per Opportunity is a better metric

Normalizing by number of opportunities is neither easy nor intuitive to do on the FPY metric. It is much more meaningful to normalize the QN per Module metric. That is because the relationship between the number of QNs written against a module during a given period of time can be compared against the total number of possible errors that could have been had during that same period. If, for example, a month was to be considered the time unit of interest, the appropriate normalized formulation of QNs per Module for a given month would be given as QNs/Opportunity as shown in Equation 3.1, instead of the one given in Equation 2.2:

$$\text{QNs/Opportunity} = \frac{\text{Number of QNs during month}}{\text{No. of Modules built} \times \text{Total Error Opportunities per Module}} \tag{3.1}$$

As long as QNs/Mod is greater than 1.0, any improvement yields small marginal change on FPY

The monthly QNs per Module metric can be understood as coming from a certain probability distribution that is characterized by a mean value. This mean value would be analogous to the QNs per Module metric. The QN probability distribution of a module would change if the state of quality in the assembly flow line of that module was to change. Therefore in a month where the quality is high, the distribution would be narrower and skew to the left whereas if the quality was poor, the distribution would be wider and skew to the right.

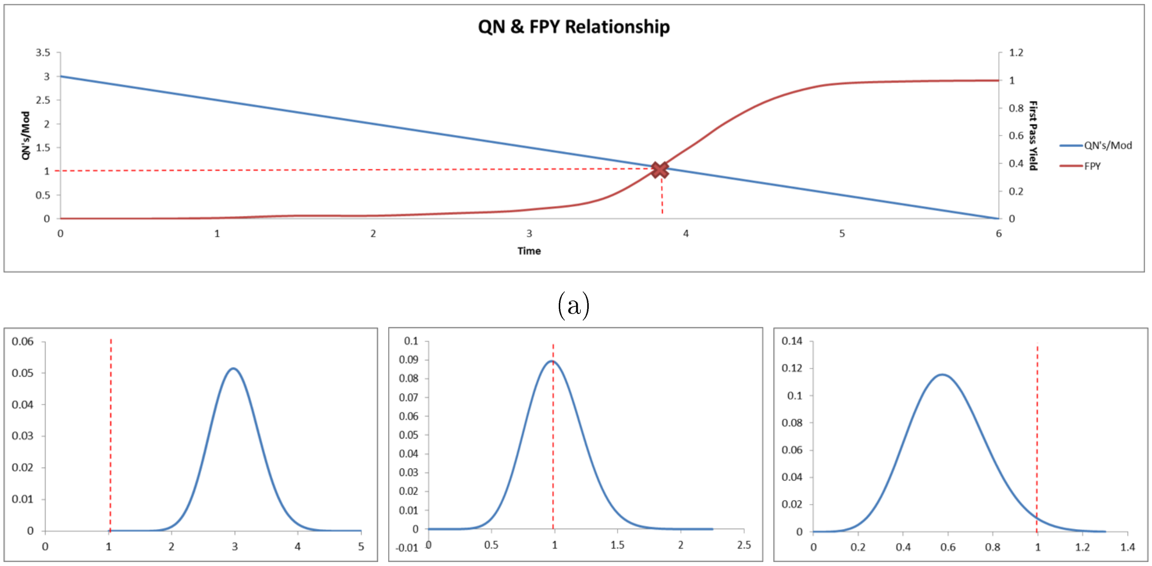
Taking a closer look at the relationship between the FPY and QNs per Module metrics, an interesting hypothesis can be made. It is explained mathematically in section 3.3.1. As has been discussed in previous sections, all it takes for a module to take a hit against FPY is one QN. In other words, no matter how many QNs may be logged against a module, it is the first one that deems it failed.

In the aggregate going from month to month, as long as the QNs per Module for a given module, or the mean of the QN distribution, is kept under 1.0, that module would enjoy a high FPY. Additionally, if the mean is kept under 1.0, a small improvement in QNs per Module would result in a significant increase in FPY. On the flip side, a module that has a QNs per Module greater than 1.0 would suffer from relatively low FPY. Similar to the preceding case, as long as the QNs per Module for a module is greater than 1.0, a small improvement in QNs per Module will only result in minute increase in FPY.

A visual illustration of the hypothesis given above is presented in Figure 3-3.

To test the hypothesis presented above, a chart of QNs per Module vs FPY can be made for every module. Then the resulting curve can be analyzed as to its behavior when QNs per Module is greater than 1.0 and when QNs per Module is less than 1.0. The hypothesis implies that the rate of change of the curve would be high in the region where QNs per Module is less than 1.0 and low in the region where QNs per Module is greater than 1.0.

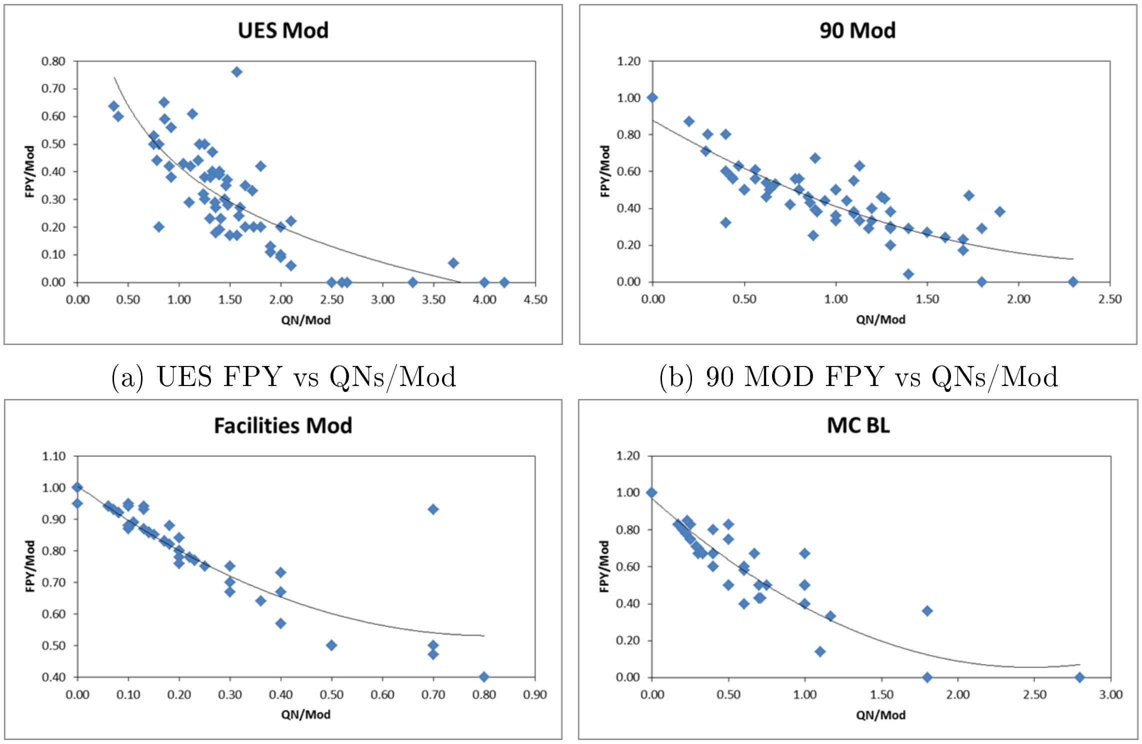
A few graphs of the kind introduced above are given in Figure 3-4. Each dot represents a month between FY2011 and FY2016 for which FPY and QNs per Module metrics exist. The line is the best regression of the plotted points. It can be seen that towards the right side of the graphs, small improvements in QNs/mod do not make a significant change in the FPY. But towards the left side of the graphs, small changes in QN/Mod result in greater improvement in FPY. That inflection point



(a)

(b)

Figure 3-3: Symbolic Illustration of QN distribution and relatinship bn FPY and QN/Mod. (a) shows a downward trending line that signifies improvement in QNs/-Mod, as well as an upward trending curve showing improvement in FPY. (b) shows probability distributions of QNs/Mod from month to month for a given year. When the mean of the QN/Mod distribution is much greater than 1.0, the FPY curve increases very slowly. As the distributions in (b) shift leftward to where the mean of the distribution approaches 1.0, the FPY curve in (a) gets steeper and steeper. Once the mean of the distributions gets below 1.0, the FPY curve in (a) will pass an inflection point after which it increases much quicker than before the curve gets to the inflection point.



(a) UES FPY vs QNs/Mod

(b) 90 MOD FPY vs QNs/Mod

(c) Facilities FPY vs QNs/Mod

(d) Beam Line FPY vs QNs/Mod

Figure 3-4: Module-wise FPY vs QN/MOD

occurs somewhere in the region of 1.0 QNs/Mod, which confirms the hypothesis given above.

3.3.1 Mathematical model to relate FPY with QN/Mod

Looking at the trends in Figure 3-4 a mathematical model can be developed to relate the FPY metric to the QNs per Module metric. In order to construct the model, a few variables and symbols would need to be defined as follows.

N = Total Opportunities for Failure per Module

Q = Probability of Failure per Opportunity

n = number of failures in a given time, a random variable

$\bar{n}$ = expected (Average) number of failures - analogous to QNs per Module metric

The mathematical model can be formulated as follows:

$$\bar{\mathrm{n}} = \mathrm{E}(\mathrm{n}) = Q \times N \tag{3.2}$$

$$\text{Probability of having no failures} = \mathrm{FPY}_{\mathrm{mod}} = (1 - Q)^N \tag{3.3}$$

Equation 3.2 can be rewritten as:

$$Q = \frac{\bar{\mathrm{n}}}{N} \tag{3.4}$$

Equations 3.3 and 3.4 can be combined as:

$$\mathrm{FPY}_{\mathrm{mod}} = \left(1 - \frac{\bar{\mathrm{n}}}{N}\right)^N \tag{3.5}$$

A uniform convergence of the right side of Equation 3.5, when N gets to be very large, results in the exponential function given in Equation 3.6:

$$e^{-x} = \lim_{k \to \infty} \left(1 - \frac{x}{k}\right)^k \tag{3.6}$$

By combining equation 3.5 and equation 3.6 an expression for $\mathrm{FPY}_{\mathrm{Mod}}$ can be written as follows:

$$\mathrm{FPY}_{\mathrm{mod}} = e^{-\bar{\mathrm{n}}} \tag{3.7}$$

Equation 3.7 makes the assumption that N is very large, which is a reasonable assumption since a module is a very complex piece of equipment with thousands of opportunities for error within each one.

From equation 3.7 and equation 3.2 , the final mathematical model is:

$$\mathrm{FPY}_{\mathrm{mod}} = e^{-Q \times N} \tag{3.8}$$

Equation 3.7 explains why the shapes of the fit curves in Figure 3-4 look exponential. An attempt was made to fit the FPY and QNs per Module data presented in Figure 3-4 but it was unsuccessful because the actual number of error opportunities in each module had not been determined. Nevertheless the fact that the trends in the figure is validated by the model presented in Equation 3.7 is encouraging.

3.4 Are standard operating procedures to blame?

One of the questions that was raised early by the MIT team was regarding how well standard operating procedures (SOPs) were written and followed. Especially given that the all manufacturing processes that take place at Varian Semiconductors are manual assembly operations, special care must be taken when compiling SOPs. It is imperative to make sure each assembly step is written clearly and that the sequence of operations is intuitive enough so as to ensure a low probability of making an error. Once the procedures are written well, the workers on the assembly line need to follow them carefully.

This section will describe several hypothesis that were made by the MIT team regarding the impact of SOPs on FPY at Varian specifically. These hypotheses range from the impact on FPY of a change in sequence of operations, to a possible correlation between the experience levels of the workers on a particular module to the FPY trend of that module. These hypotheses and associated tests are presented below.

3.4.1 Does sequence of operations have a significant impact on FPY?

One of the early tests that was conducted during the analysis of FPY stagnation as it relates to procedures was to validate if a change in sequence of operations is correlated to FPY in any way. In the summer of 2014, a different MIT co-op team had undertaken a project to redesign the assembly flow line for the UES module. It was known as the "critical path project". The outcome of that project was a more streamlined flow line that optimized for cycle time. The UES line was broken up into a handful of sub-module areas which would then be integrated into the full module. The new design led to a complete overhaul of the sequence of operations, the arrival of materials was more synchronized with the progress of the assembly process and the number of workers required to build the sub-modules and the complete module was determined.

In order to carry out the analysis of the impact on FPY of this change of sequence of operations, QN data was gathered from a period of 6 months prior to the start of the critical path project as well as from a period of 6 months proceeding the conclusion and full implementation of the project. A 2 tail Analysis of Variance was conducted on the two sets of data to see if there was a statistically significant difference between them.

Analysis of the impact the critical path project on UES line had on FPY

Figure 3-5 shows an interval plot of the two data sets described above. The p-value, which is a measure of the likelihood that two data sets come from the same probability distribution, of the statistical analysis was 0.374. Tyipically if the p-value for a statistical analysis is greater than 0.05, it is concluded that the data sets come from the same distribution, and any difference between them is a result of random chance. This implies that there will not be a statistically significant difference between the means of the two data sets. In this case, with a p-value of 0.374, it can be concluded that the difference in the means of the two data sets is **not** statistically significant. But it can be observed from the box plot in Figure 3-5 that the variation drops visibly from the earlier data set to the more recent data set. In fact, the standard deviation of the pre-critical path project data set is 19.08 while the standard deviation of the post-critical path project data set is 13.23. That is a reduction of close to 30.67%. Therefore it can be concluded that the critical path project did not affect the average instance of a QN being written against the UES module, but significantly reduced the variation of QN generation from month to month.

This phenomenon can be explained by taking into consideration the standardization of procedures that was put in place as a result of the critical path project. The procedures were now more precise in what they directed the worker to do. They laid out where a worker ought to be next, how many workers are needed to perform a particular operation etc. Kits would arrive close to when they are needed and do not

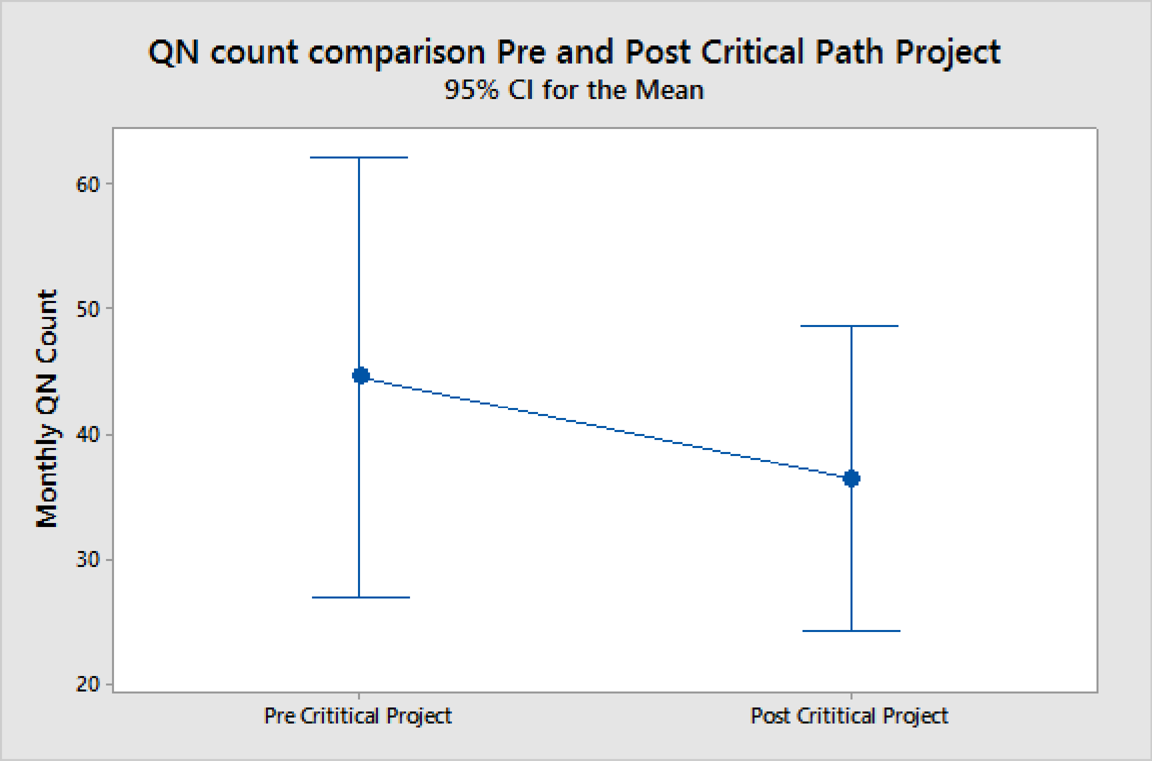


Figure 3-5: Effect of Critical Path Project on MFG Quality

have to stay exposed to possible damage, as they were before. All in all, the set of operations were more repeatable than they were prior to the critical path project. What did not change significantly was the operations themselves. Therefore the number of opportunities that exist to make errors stayed the same - hence the noticed lack of difference in the means of the two data sets. However the fact that the build process is more repeatable explains why the variation in errors would drop the way it did. Workers are not more likely to commit significantly more errors from one month to another, nor would it the case anymore that one set of workers would commit errors that another set would not.

In conclusion, based on the statistical analyses presented above, the sequence of operations does not have a significant impact on the FPY metric.

3.4.2 How much is FPY affected by the experience level of workers?

In any manufacturing facility, worker experience is positively correlated with higher quality of work. Experienced workers are more knowledgeable about the operations they perform on a daily basis. They do not need constant supervision, nor does it take them a long time to perform their operations. The MIT team decided to ascertain this widely understood notion in the context of the assembly operations that take place on the manufacturing floor of Varian. The reason the team wanted to investigate this is because of the labor hiring policy of Varian. Not all workers on the various assembly stations are permanent workers. About 15-25% of all workers on the assembly floor are hired as contractors. The contractors are working at Varian for no more than 18 months at a time. After the 18 months have passed, the contractors are let go, and a new group of contractors are hired. Due to labor regulatory laws, in order for the same contractor to be hired back to Varian, he/she should not have worked at Varian for at least 3 months after being let go the last time. Therefore at any given time, there could be distinct mix of new workers and experienced workers who are working on the same sub-module or module.

Testing if experience level is a factor

Given the labor turnover scenario outlined above, the MIT team made the hypothesis that modules that were built by a group of workers who logged a high ratio of hours by inexperienced workers to total number of hours logged would tend to have more QNs written against them. This ratio will be referred to as the **inexperience index**. The formulation of the inexperience index is given in Equation 3.9 . The smaller inexperience index, the less errors would be expected and hence less QNs.

$$\text{Inexperience Index} = \frac{\text{Total hours logged against module by inexperienced workers}}{\text{Total hours logged against module}} \tag{3.9}$$

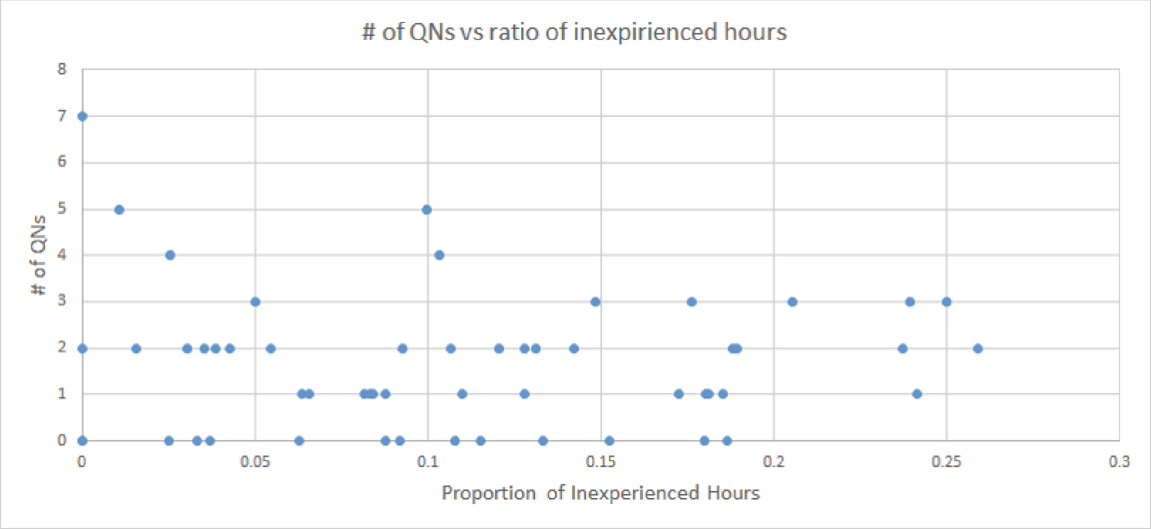


Figure 3-6: QNs against a module vs inexperience index for that module

To simplify the testing of the hypothesis, only data from the UES module was considered. QN data was filtered to the UES module, and a time period of 1 year. To calculate the inexperience index, the team first obtained a labor finance sheet containing each module that was built during the time period of interest. The sheet contains the names of the workers that worked on each module and the number of hours logged by each worker on each module. Once a list of names of workers was prepared, the team went to the lead supervisor of the UES line with that list to get his opinion on who he considered to be experienced and who he deemed inexperienced. Then, for each module built, the hours logged by the workers who were deemed inexperienced by the supervisor were tallied and divided by the total hours logged by all workers during that time period. Finally, for each module, the number of QNs logged was compared against the ratio presented above. Figure 3-6 shows a scatter plot of the comparison.

It is clear from 3-6 that there is no obvious correlation between the number of QNs on a module to the experience composition of workers that built it. The plot shows 7 QNs logged against a module that was built exclusively by experienced workers while showing only 2 QNs logged against a module for which more than a quarter of the hours were from inexperience workers. This could be explained in a number of possible ways. The first is that there truly is no correlation between experience level and quality of output, which is doubtful. The second, explanation is that the sheer number of possible factors that could result in a QN makes it very difficult to find a correlation with any single factor - especially if that factor is not very significant. It is possible that while there might be some correlation between experience level and QN generation, it is not strong enough to be detected in the presence of other stronger factors. Another possibility is that when the MIT team asked the managers on the assembly floor for their opinions on who is experienced and who is not, the managers might not have given accurate information, or shown bias towards certain workers whom they might have seen more favorably.

3.5 What Effect Do Part Shortages Have on Manufacturing Quality?

In the early phase of the project, the MIT team talked to workers on the module assembly areas to get their thoughts on what might be a significant contributor to errors being made. The workers responded by listing several possible causes for errors, but the one they stressed as very problematic was critical shortages. They described their frustration when a part that was supposed to have been brought to the module lay down does not arrive in time, and force the workers to build an assembly out of the prescribed sequence outlined in the procedures. Having to do that, they claimed, would lead to possible error being committed. In addition, when the part that was missing eventually arrives, it may cause workers to do a bit of disassembly before they are able to insert the part. The process of disassembling and re-assembling parts in this way introduces more opportunities for errors to be made and for quality to be compromised.

Shortages are costly to Varian, not just due to the monetary burden caused by the quality issues that might arise from shortages, but also because workers log labor hours against time spent tracking down and addressing shortages. Therefore it was deemed necessary by the MIT team that closer attention ought to be paid to part shortages.

To confirm the series of hypothesis laid out above, the MIT team performed an analysis to find a quantitative correlation between shortages of parts that go into tools, to manufacturing related quality issues on those tools. The methodology and results of the analysis is presented in the following subsections.

3.5.1 Correlating part shortages with manufacturing related quality problems

In order to perform the analysis, two sets of data needed to be gathered and compiled. The first is QN data from tools produced in a one year period. The other is part shortage data for all parts that are required for those tools. The one year period was from May 2015 - May 2016.

The QN data was gathered via a search on the SAP database where QNs are stored. The part shortage data on the other hand was gathered from a database that contained a list of "cross-docked" parts. A part that is cross-docked is routed directly to the assembly floor, bypassing the stock location. A part is only cross-docked if there was a shortage of that part reported by workers on the assembly floor. Therefore it represented the most comprehensive data set for piece part shortages the MIT team was able to acquire.

The cross-dock data was analyzed in Excel to isolate the number of shortage occurrences on each part that was later rolled up into the tools under consideration. A shortage occurrence is defined as when a worker needed a part and it was not available. It is independent of how many of the part was needed, just that the worker was unable to get it when he/she went looking for it. It is when a part shortage occurrence happens that workers are forced to work out of sequence as described above, therefore it served as the appropriate variable to correlate with QN data.

The two data sets introduced above were put on a scatter plot to perform a visual inspection of a possible correlation. The scatter plot is presented in Figure 3-7.

In Figure 3-7, each red dot represents a tool that was built during the one year

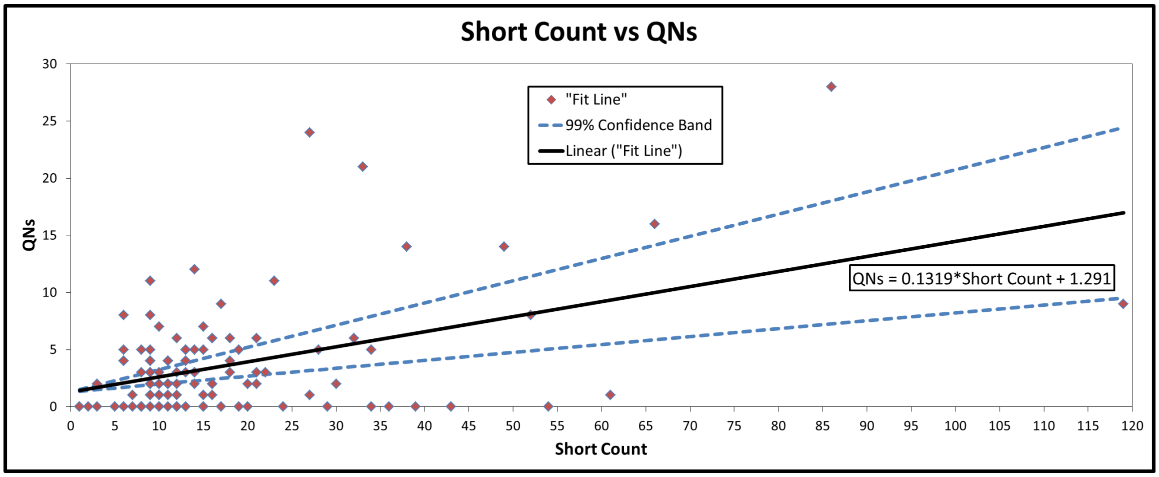


Figure 3-7: Scatter Plot of Part Shortages vs QNs for one year

period under consideration. For each tool built, the y-axis represents the total number of manufacturing related QNs written against that tool. The x-axis represents the total number of shortage occurrences of parts that went into that tool. A 99% confidence band on the slope of a linear regression that was performed showed a rising trend, even at the lowest range of the band. The linear regression equation is given in Equation 3.10. Furthermore the MIT team decided to perform a statistical test see if whatever correlation that is present in Figure 3-7 is statistically significant.

$$\text{QNs} = 0.1319 \times \text{Short Count} + 1.291 \tag{3.10}$$

For a statistical test, the tools were put into three categories based on the number of part shortage occurrences on them. The first category included tools which had a high number of shortage occurences - greater than 18 shortages. The second category included tools that include an intermediate number of shortage occurences - between 11 and 18. The last category included tools that had a low number of shortage occurences - less than 11. All categories had roughly around 30 tools in them. Thereafter, a mean difference test was conducted. The mean in this case is of the number of QNs against the tools in each category. The null hypothesis for the test was that the means were equal. If the null hypothesis is rejected by the test to a reasonable confidence level, then a statistically significant correlation between QNs

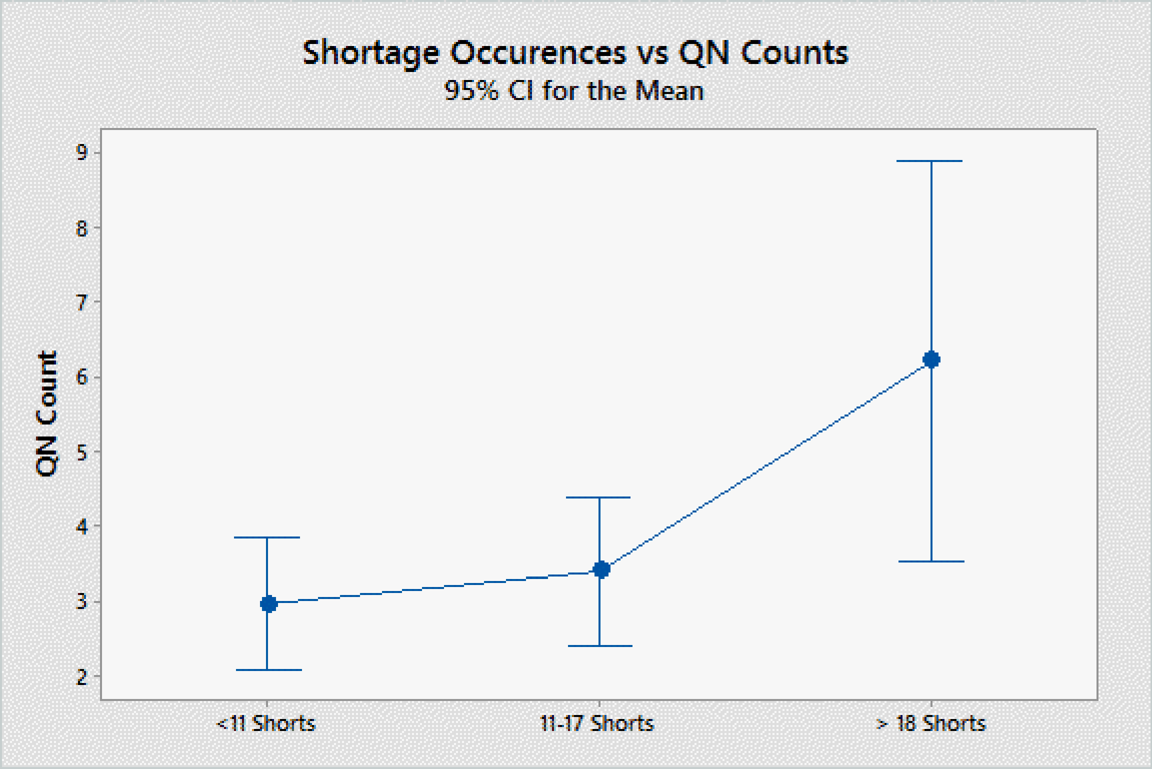


Figure 3-8: Interval Plot of Part Shortages vs QNs for one year

and shortages can be established.

Figure 3-8 shows a interval plot of the three categories. It can be seen visually that the first category looks markedly distinct from the last two categories. The statistical test revealed the three categories came from different distributions to a p-value of 0.074. That represents a 92.7% confidence that the null hypothesis can be rejected, and hence make the conclusion that the three categories are distinct from each other.

3.6 Mathematical correlation between shortage occurences and the FPY metric

Based on the relationship between shortage occurrences and QNs given in Equation 3.10 in section 3.5, as well as the relationship between QNs and the FPY given in Equation 3.8 in section 3.3.1, a mathematical relationship can be made between shortage occurrences and the FPY. This relationship will inform the FPY team what impact a reduction in shortages will have on the FPY metric that they monitor. This relationship is derived given below.

Let SO be the shortage occurrences observed over some period of time, and let Q be the number of QNs observed during the same time. Then Equation 3.10 can be re-written as:

$$Q = 0.1319 \times \mathrm{SO} + 1.291 \tag{3.11}$$

Inserting Q from Equation 3.11 into Equation 3.8, the expression below arises.

$$\mathrm{FPY} = e^{-(0.1319 \times \mathrm{SO} + 1.291) \times N} \tag{3.12}$$

If the number of shortage occurrences changes such that the new number of shortage occurrences is a times the old (where $a > 0$), Equation 3.13 gives the resulting change in FPY. For a detailed derivation of this equation, refer to Appendix A.

$$\mathrm{FPY}_{\mathrm{new}} = \left(\mathrm{FPY}_{\mathrm{old}} \times e^{-\left(\frac{1.291}{a} - 1.291\right) \times N}\right)^{a} \tag{3.13}$$

3.7 Conclusion

Now that a statistically significant correlation between shortages and quality issues has been established and that the mathematical link between shortages and FPY has been demonstrated, the MIT team proceeded to investigate the cause of these part shortages and took steps to address those causes. A detailed explanation of that investigation and its outcomes are presented in Chapter 5.

Chapter 4

Error re-categorization Project

At the end of chapter 2, a discussion was given regarding the short-comings of the current approach used by Varian to improve their FPY. It was stated that the buckets were to broad for the quality and manufacturing engineers to identify proper trends in failure modes. That prompted the MIT team to re-evaluate the existing buckets and problem solving approach and devise a new strategy that would likely lead to better improvement of FPY. This chapter will discuss this effort and present a new approach of identifying recurring failure modes and a way to tackle those failure modes.

4.1 Motivation

As introduced above, the reason the MIT team believed re-evaluating the existing improvement strategy is because the buckets that Varian uses to categorize QNs are not really being helpful in coming up with comprehensive preventative measures for common failure mechanisms on the assembly floor. That is because the 4 buckets are, in essence, just categories based on the types of QNs that fall under them. For example, the harnessing bucket contains all QNs that are logged due to a failure in a harness. The same applies for all the other buckets. The categories do not inform the engineers of the failure mechanism of the QNs within them, just what type of failure it is or what type of component failed. This way, it is difficult to identify recurring failure mechanisms that might exist within the same bucket or across buckets. If such trends are not being detected, then an effective mitigation strategy will be difficult to formulate and it would be more likely that the trend will continue.

In addition, the existing buckets focus a lot on repair once an error was committed, but not so much on preventing similar errors from being committed. This way of approaching yield improvement will lead to frustration as root causes of failures are not adequately being found out and addressed. As long as the root cause remains, any action as a result of a QN is bound to not have lasting impact.

Another problem that the MIT team noticed in attending the weekly meeting among the members of the FPY team is the tendency of the bucket leaders to dismiss QNs as being the result of a lack of "attention to detail". While it is often easy to arrive at this conclusion and not pursue it any further, such attitude also masks the real underlying problem that leads to the error difficult to discern. For example, a common QN would involve a worker connecting a cable to the wrong port. Since each cable and port are labeled, a manager looking into this QN would readily attribute it to a lack of attention to detail. If the investigation doesn't go further, then the real root cause of the error would not be found. The root cause could be that the labels are not legible and need to be a bigger font. Or perhaps the ports were so close to each other that the worker mistakenly plugged into the wrong port. Perhaps the numbering sequence is not intuitive. All these are potential root causes that could lead to the failure mode of a mis-connected cable. While it is true that the error would probably not have happened if the worker was paying closer attention to where he/she was plugging the cable, as long as the underlying root causes are not identified and solved, this error is likely to be committed again by a different worker.

Part of Varian's existing strategy to deal with repeat errors is to wait until a specific part fails repeatedly. While it is important to do this, as it could lead the quality engineers to look into what is wrong with that particular part, this approach on its own does not help to identify failure mechanisms that could be affecting multiple parts at once.

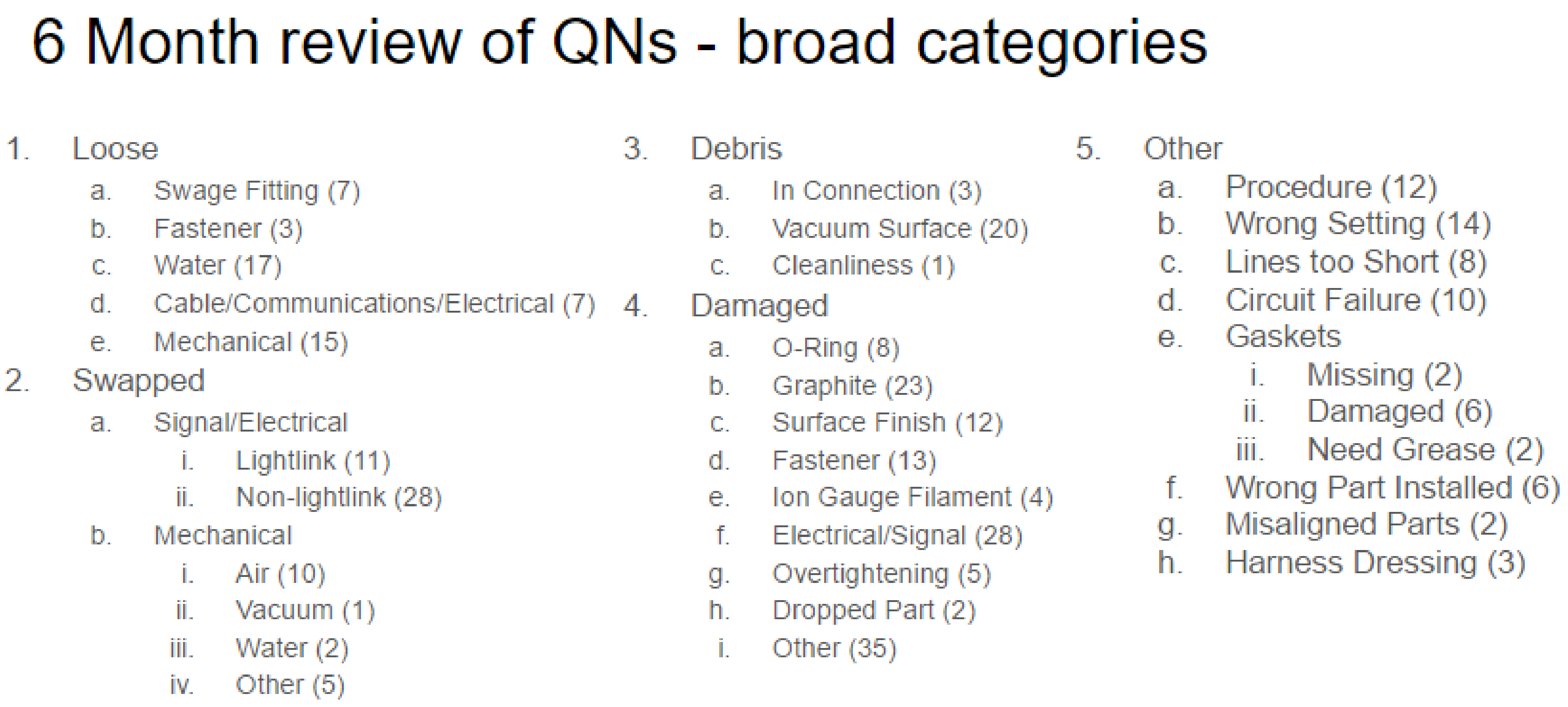
4.2 Proposed Alternative - Failure Mode Categories

Based on the analysis given above of the existing problem solving approach, the MIT team developed a new way of categorizing the QNs - by failure mode. QNs would no longer fall into 4 broad buckets based on their type, but into categories that describe the particular failure mechanism outlined in the QN. The process of developing a category should be undertaken with the mindset of identifying specific corrective measures that would prevent all the QNs in that category from happening again. For example, the QN of the mis-connected cable above would not be assigned to a "Harnessing" bucket, but would fall into a category of "Cable connected to wrong port". The mis-connection is the failure mechanism that lead to the QN being written. Any QN of a mis-connected cable would fall under this category. Forming the category this way would better enable quality engineers to devise a problem solving strategy that would prevent cases mis-connected cables from coming up again.

4.2.1 Case-study - categorizing 6 months of QNs by failure mode

In order to validate the alternative categorization method proposed above, the MIT team made an effort to create failure-mode categories by analyzing all manufacturing related QNs from a 6 month period between January and June of 2016. Each QN's description was searched for key-words that helped the team come up with very broad categories, from which they proceeded to sub-categorize further until all QNs within a category had the same failure mode. Figure 4-1 shows the initial stages of the categorization process. These categories are not necessarily failure mode categories, but rather a step along the process of forming failure mode categories. Figure 4-2 shows a sample of specific failure mode categories based on the general categories shown in Figure 4-1.

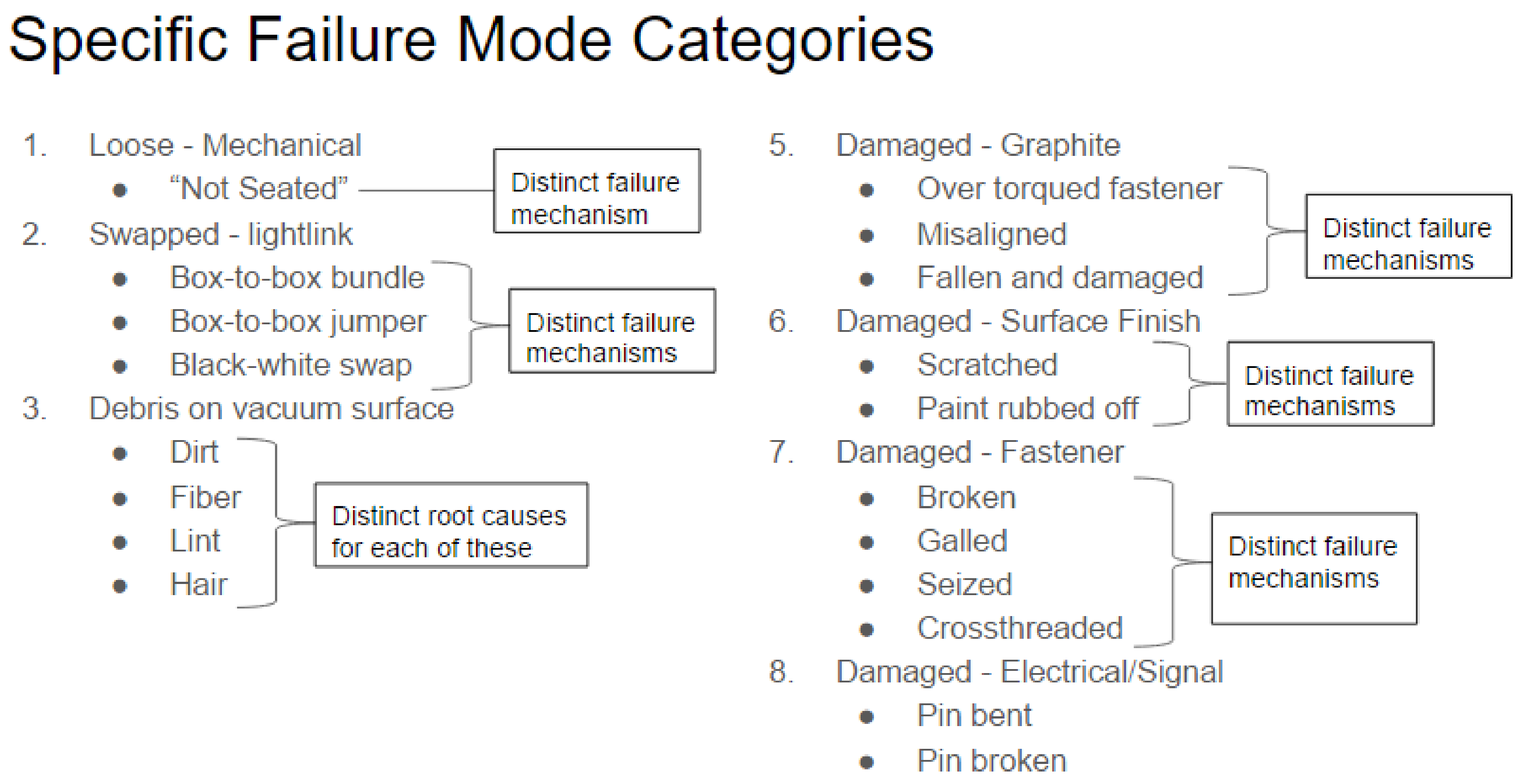
It can be seen from Figure 4-2 that each of the failure mode categories are formed such that corrective action is unique to each of them. The solution to a graphite piece being damaged due to an over-torqued fastener is different to the solution to a



6 Month review of QNs - broad categories

1. Loose
   a. Swage Fitting (7)
   b. Fastener (3)
   c. Water (17)
   d. Cable/Communications/Electrical (7)
   e. Mechanical (15)
2. Swapped
   a. Signal/Electrical
      i. Lightlink (11)
      ii. Non-lightlink (28)
   b. Mechanical
      i. Air (10)
      ii. Vacuum (1)
      iii. Water (2)
      iv. Other (5)
3. Debris
   a. In Connection (3)
   b. Vacuum Surface (20)
   c. Cleanliness (1)
4. Damaged
   a. O-Ring (8)
   b. Graphite (23)
   c. Surface Finish (12)
   d. Fastener (13)
   e. Ion Gauge Filament (4)
   f. Electrical/Signal (28)
   g. Overtightening (5)
   h. Dropped Part (2)
   i. Other (35)
5. Other
   a. Procedure (12)
   b. Wrong Setting (14)
   c. Lines too Short (8)
   d. Circuit Failure (10)
   e. Gaskets
      i. Missing (2)
      ii. Damaged (6)
      iii. Need Grease (2)
   f. Wrong Part Installed (6)
   g. Misaligned Parts (2)
   h. Harness Dressing (3)

Figure 4-1: Broad Categories of 6 months of QNs



Specific Failure Mode Categories

1. Loose - Mechanical
    - “Not Seated”
2. Swapped - lightlink
    - Box-to-box bundle
    - Box-to-box jumper
    - Black-white swap
3. Debris on vacuum surface
    - Dirt
    - Fiber
    - Lint
    - Hair

5. Damaged - Graphite
    - Over torqued fastener
    - Misaligned
    - Fallen and damaged
6. Damaged - Surface Finish
    - Scratched
    - Paint rubbed off
7. Damaged - Fastener
    - Broken
    - Galled
    - Seized
    - Crossthreaded
8. Damaged - Electrical/Signal
    - Pin bent
    - Pin broken

Distinct failure mechanism
Distinct failure mechanisms
Distinct root causes for each of these
Distinct failure mechanisms
Distinct failure mechanisms
Distinct failure mechanisms

Figure 4-2: Failure Mode Categories of 6 months of QNs

graphite piece being damaged as a result of misalignment. The solution to avoiding fiber from damaging vacuum surfaces is different than the solution to avoiding hair from damaging vacuum surfaces. By forming the categories in this way, quality engineers are better equipped to come up with preventative measures to minimize the recurrence of the QNs within those categories.

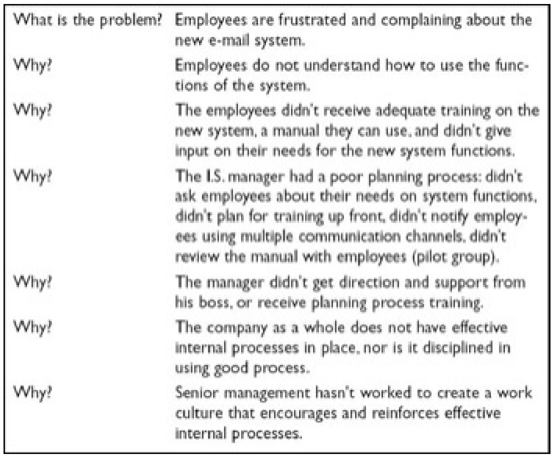
4.3 Using New Categories for Problem Solving

Getting to these failure mechanisms is often tedious but not particularly difficult to do. Where the problem solving process becomes more challenging is in identifying root cause of those failure mechanisms and determining corrective measure to ensure that failure mechanism doesn't reoccur. A good strategy for coming up with corrective action to these failure modes was the 5-why strategy used in the Toyota Production System to identify root causes. The 5-why analysis involved asking the question **why?** repeatedly until the root cause of that problem is determined [2]. Even though the strategy is named 5-whys, it doesn't necessarily mean the analysis stops at the fifth why. It should continue until root cause has been found. That could take more or less than 5 whys to get to. Figure 4-3 shows an example of the 5-why strategy in use to identify the root cause of a problem in an office setting. The same strategy can be used in a manufacturing setting to determine root causes of QNs. Once the root cause is determined, then a corrective measure can be instituted to fix that root cause. Eliminating the root cause of a failure mode will reduce the chances of QNs of that failure mode from resurfacing.

4.3.1 Proposed Problem Solving Methodology for Varian

The MIT team used the 5-why root cause analysis system described above to propose a problem solving methodology for the FPY team that it believes should lead to better long term improvement of the FPY metric. This methodology is outlined below.

1. Once the failure mode categories have been determined, then at each week's



What is the problem?	Employees are frustrated and complaining about the new e-mail system.
Why?	Employees do not understand how to use the functions of the system.
Why?	The employees didn't receive adequate training on the new system, a manual they can use, and didn't give input on their needs for the new system functions.
Why?	The I.S. manager had a poor planning process: didn't ask employees about their needs on system functions, didn't plan for training up front, didn't notify employees using multiple communication channels, didn't review the manual with employees (pilot group).
Why?	The manager didn't get direction and support from his boss, or receive planning process training.
Why?	The company as a whole does not have effective internal processes in place, nor is it disciplined in using good process.
Why?	Senior management hasn't worked to create a work culture that encourages and reinforces effective internal processes.

Figure 4-3: 5-why analysis to identify root cause of office problem [2]

meeting the team should assign the QNs from the previous week to the appropriate category. It is important to resist the temptation to dismiss a QN as a problem of "attention to detail".

2. Keep a running count of all categories.

3. If a category accumulates a lot of QNs, form a small team that will use the 5-why system described above to identify the root-cause of that failure mode.

4. Once the root cause is determined, the small team should brainstorm and propose a corrective measure to address that root cause.

5. Implement the proposed corrective measure

6. Keep the categories fluid. If a QN is written that doesn't belong to any established failure mode category, then form a new category with the appropriate failure mode. On the other hand, if a failure mode category goes a long time without any additional QNs, then remove that category.

Failure Mode	Light link cables out of a harness are plugged to the wrong port on a grid of ports
Why?	Written label on cable was not enough to correctly identify the desired port
Why?	There is no other stimuli, besides visual, to assist worker to find and plug into the correct port
Why?	Light link cables come bundled in pairs but a pair of cables do not necessarily go to adjacent ports
Why?	**The harness manufacturer does not pay close enough attention to the relative location of the ports the cables will be plugged to**

Table 4.1: 5-why analysis to identify root cause of fiber optic swapping

4.4 Case studies - Possible Projects Based on Failure Mode Categories

Based on the methodology given above, the MIT team decided to analyze the failure-mode categories they identified to propose potential corrective actions to those failure modes. A presentation of 2 possible projects is given in the subsections below.

4.4.1 Case study 1 - Grouping Fiber-optic Cables

Failure Mode

The first failure mode the team looked at is the "Swapped - lightlink, box-to-box bundle" category shown in Figure 4-2. This failure mode is manifested in workers plugging a light link (fiber optic) cable out of a bundle (harness) to the wrong port on a grid of ports.

Root Cause Analysis

Table 4.1 shows a root cause analysis based on the 5-why method described above.

Light link cables out of a big harness usually come in pairs of two, but it is not necessarily the case that the pair of two go to adjacent ports. Although both the cables and the ports are labeled, the labels are often small and difficult to read unless from close range. The lack of an alternate stimuli to guide the worker to the correct port would increase the chance that he/she will mistake a wrong port for the right one.

Proposed Project

The root cause analysis calls for harness manufacturers to pay closer attention to where adjacent cables are plugged. A closer look at the grid of ports these cables are plugged into shows that the grid is made up of 4 rows of ports with 4 ports per row. The MIT team observed that if the cables were to come in groups of 4 that all plug into adjacent ports, instead of 2, then that would simplify the job of the worker immensely. Now all the worker needs to get right is the first connection out of each of the groups of 4. Then, since each group of 4 contains cables that plug into adjacent ports, the worker can easily plug the remaining cables by touch. As long as the grouping is done correctly, this grouping strategy should reduce the chances of swapping cables on a grid of ports. The proposed project would call for Varian to make arrangements with its harness manufacturers to have them deliver the specific cables that go on a grid in groups of 4 that each go to adjacent ports on the grid.

4.4.2 Case study 2 - Procedure to Confirm Tight Connections

Failure Mode

The second failure mode the team looked at is the "Loose - water" category shown in Figure 4-1. This failure mode is manifested in workers leaving a water connection loose or only finger tight.

Root Cause Analysis

Table 4.2 shows a root cause analysis based on the 5-why method described above.

Connections are often found not tightened to specification because of the way workers generally make these connections. The general practice is to hand tighten all the connections in a batch, then to go back and use a torque wrench to tighten to specification. When there are a lot of connections that are being made, workers

Failure Mode	Water connections are left not tightened to specification
Why?	Workers forget to tighten with torque wrench after initial positioning of connection
Why?	General practice is to finger tighten connections before tightening to specification
Why?	**The assembly procedure is not explicit enough in calling out connections that need further tightening**

Table 4.2: 5-why analysis to identify root cause of Loose Water connections

sometimes forget that some of the connections are only hand tight. The assembly procedure is not written with this phenomena in mind, which led to the close to 20 cases of loose connections in a period of 6 months.

Proposed Project

As is evident from the root cause analysis that the appropriate corrective action for this failure mode is to put a reminder in the procedure to remind the worker to check all connections for their tightness, and perhaps to have them apply the torque wrench to each connection to confirm they are all tightened according to specification.

4.5 Conclusion

This chapter described a novel way of implementing corrective actions to QNs by categorizing them by failure mode, as opposed to by type as is done at the time. A methodology to form those categories, as well as a criterion for judging the effectiveness of a category was given. Once appropriate categories have been formed, a strategy to track the categories and implement corrective measures for large categories was proposed. The methodology was partially implemented by the MIT team - failure mode categories were established, root cause analysis for a couple of those categories was performed and recommendations were made for corrective actions to those root causes. It is the belief of the MIT team that following the recommendations in this chapter will lead to better identification of failure trends and lead the FPY team to take better targeted corrective action to mitigate those failure trends more effectively.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Critical Shorts Project

The previous chapter gave a detailed explanation for various ways to account for why the FPY metric at Varian Semiconductors has stagnated over the prior 3 years. At the end of chapter 3, a discussion was given on how part shortages on the assembly line had a significant impact on the state of quality at Varian. This chapter will discuss the efforts of the MIT team to investigate the root cause of these part shortages and make recommendations to the company on how to mitigate the problem. A cost analysis of the recommended solution that was performed is also presented in this chapter along with a discussion on why the MIT team believes it is in the best interest of the company to adopt the recommendations.

It is clear that a part shortage problem is analogous to an inventory management problem. Various reasons could explain why a part is not where it is supposed to be when it is supposed to be there. Some reasons include the vendors not delivering on time, parts arriving but getting damaged during processing etc. These reasons do not necessarily point to a systemic problem that is responsible for a high number of shortages. But a possible systemic problem in the inventory management scheme would be if not enough parts are being ordered and stocked to mitigate possible variation in demand. This chapter will focus on the MIT team's investigation of the inventory management system at Varian to see if there are potential systemic causes for shortages that need to be addressed.

5.1 Decision making on what inventory to focus on

Varian Semiconductors holds close to $90 million in inventory. The whole inventory is categorized by the way the parts are procured. Those categories are given below. There are more procurement types but these are the major ones.

- PO - Purchase Order: Parts that are ordered by buyers on an individual basis. These parts have variable lead times.
- VO - Varian Order: Parts that are Varian designed that ship from machine shops or contract manufacturers.
- KB: Large assemblies that are ordered based on a 2-bin Kanban system. These parts have more or less a fixed lead time based on agreements with vendors and suppliers.
- KC: Small piece parts that, like KB parts, are ordered based on a 2-bin Kanban. Like the KB parts, these parts arrive based on an agreed upon schedule. An explanation of KC procurement is given in the next subsection.

The above list is of procurement categories for piece parts that go into assemblies and then into modules. Another type of inventory at Varian are the Gold Squares. These are inventories of a subset of finished assemblies out of the supermarket area that eventually go into modules. The Gold Squares are located in an area behind the supermarket for easy tracking and maintenance. The assemblies that are placed on a Gold Square space are high-need assemblies that are mostly common among multiple modules. The way the number of squares is determined for each assembly affects whether those assemblies will be available in the Gold Square space when workers form the flowline come to pick them up. In an ideal state, the Gold Squares are either always full, or at least not empty while on track to being full within the assembly lead time of those assemblies. But that has not been happening at Varian. The Gold Squares for a lot of assemblies are frequently empty. Unfortunately, the shortage occurrences of Gold Square assemblies has never been tracked in a systematic way that would have led the MIT team to make a correlation with quality issues as was done with piece part shortages overall. Due to lack of archived data, the MIT team decided they would start tracking every Gold Square assembly shortage. After two weeks of tracking, it was determined that Varian was on track to have close to 500 Gold Square shortages in a year. Therefore in addition to investigating shortage occurrences on piece parts, the MIT team still also decided to look into the way the number of squares is determined if there are any improvement opportunities there.

As far as the piece parts are concerned, since it was not feasible to look into all the procurement types during the time spend at Varian, the MIT team had to determine which procurement type was worth investigating. That decision was made based on the return on investment an improvement would have on the policy associated with each procurement type. To make this decision, the number of shorts of parts on each procurement category was compared to the total number of parts that fall under that procurement category. The procurement category on which the largest percentage of shorts occur would be selected for analysis of its inventory management policy.

Figure 5-1 shows a bar graph of the shorted occurences on various procurement categories as percentages of the total number of parts on those categories.

As can be seen from Figure 5-1, compared to the total parts, the shortage on KC procurement type is significant - close to 50%. Clearly KC is where the biggest improvement opportunity lies, and hence is where the MIT team went into for analysis.

5.1.1 KC part inventory management - 2-bin Kanban

The parts in the KC procurement category are ordered according to a 2-bin Kanban system. The 2-bin Kanban system is a stock ordering policy in which parts are consumed from 2 appropriately sized bins. When the first bin is fully consumed, an order is made to replenish it. The second bin is consumed while the first is being replenished. The bins must be sized so that each can last through the replenishment time without a stock out.

Varian has contractual agreements with its various suppliers and vendors to replenish KC bins at a fixed commitment time. For the majority of KC parts, that

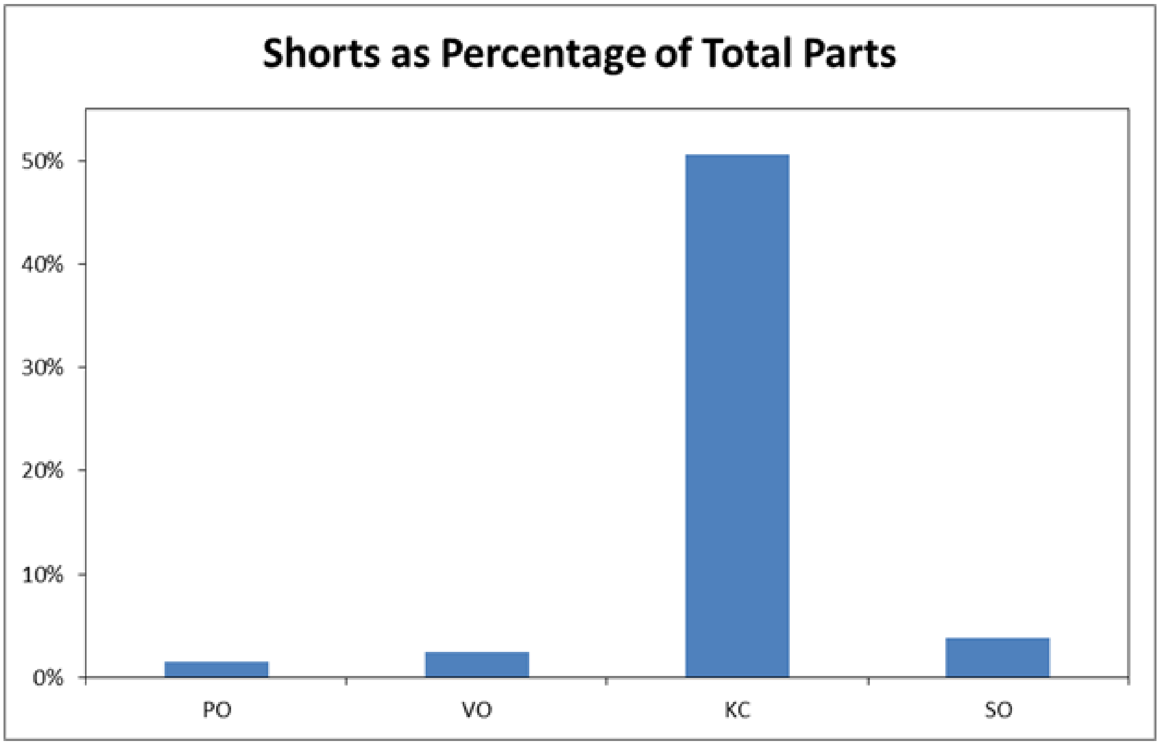


Figure 5-1: Comparison of Shortage Percentages Across Different Procurement Types

replenishment time is usually 2-5 days. A small proportion of the KC parts have varying replenishment policies, such as longer replenishment times, minimum order quantities, or incremental ordering. All these special cases were accounted for in the analysis presented later in this chapter.

5.2 Demand Forecast Discussion

KC bin sizing is driven according to a forecast of demand for the following 3 months. That forecast includes three distinct demand streams. They are presented below

- Production: This demand stream is for parts that consumed for production tools at Varian.
- Applied Global Services (AGS): This demand stream is for parts used to re-stock parts banks around the world to service customers. These parts are not consumed by production tools at Varian. AGS is a separate business unit under Applied Materials that handles sales.

- Emergency Orders: This stream is for parts or assemblies that need to go to customers in an emergency situation, such as when a part on the customer site fails. The forecast for this stream is not very reliable but the forecasting team do their best to get ahead of emergency orders.

Varian's inventory policy is what is referred to as "co-mingled inventory". This means that there is one pool of inventory for all three demand streams to dip into whenever there is a need. As such, the KC bins are sized according to a pooled forecast that takes into account the demand from all those streams. There are advantages and disadvantages to having a co-mingled inventory. The major advantage of co-mingled inventory is customer satisfaction. AGS can more quickly serve its customers by simply grabbing the parts it needs from the assembly floor with less worry of a stock out. By the same token, production is also able to get parts from the bin that might have been meant for an AGS demand, which reduces the chances of a stock out of parts for tools being built. The major disadvantage of the co-mingled inventory is that it gives license to AGS to come and grab as many parts as it needs to satisfy its order, hence potentially wiping out a bin and not leave any for production. AGS usually comes near the start or the end of a month and cuts transactions for large quantities of orders all at once. So while the bin might be padded up for most of the month, there is a chance that nothing is left in the bins, causing shortages on the assembly floor that could potentially result in quality issues.

There are some parts that AGS needs to service its customers and part banks that are never marked as needed for production. Therefore, in essence, Varian is serving as a stock room for some parts that it would never use to make machines. Varian has to pick up the cost of acquiring and holding that inventory. Therefore the co-mingled inventory doesn't only the have disadvantage of possible quality problems, but also real monetary burden of servicing the customers of a different business group. As will be discussed later on in this chapter, the MIT team considered the possibility of segregating the AGS demand stream from the production demand stream in their inventory analysis.

Even though, as mentioned above, that forecasts are being made for the upcoming 3 months, it is updated monthly. Therefore the 3 month forecast varies from one month to the next, and as such KC bins should be updated accordingly. But the current method of calculating bins, which will be described in the next section, is only updated quarterly. This means, for two-thirds of a quarter, the KC bins are sized for the wrong forecast. This could be problematic if there is a big change in the forecast from the first month of the quarter to the next two. A more detailed description of the way KC and Gold Square inventory is manged is presented next.

5.3 Current method of managing KC bins and Gold Square inventory

5.3.1 KC bins sizing

Currently, the the KC bins are sized according to Equation 5.1.

$$\text{Pull} = (\text{LT} \times WSF \times \mu_{\text{day}}) + \frac{1}{2} \times \sigma_{\text{day}} \tag{5.1}$$

Where:

- Pull is the size of 1 bin of inventory for a part
- LT is the replenishment lead time for the part
- WSF is the weekly safety factor
- μ_{day} is the average of the daily demand forecast
- σ_{day} is the standard deviation of the daily demand forecast

A pull is supposed to last through the replenishment time of a bin, represented by LT in Equation 5.1. The weekly safety factor (WSF) is meant as sort of a safety stock in case the bin does not last through the typical LT of 5 days - the equivalent of 1 business week. For most parts, the LT is 5 days and the WSF is 2. The $\frac{1}{2} \times \sigma_{day}$

is meant to account for spikes in the daily demand caused as a result of variation in the forecast that comes from the AGS stream. Originally this term was just σ_{day} but due to pressure from upper management to cut inventory, the $\frac{1}{2}$ was inserted into the equation.

Immediately from Equation 5.1, a mathematical error can be spotted. The $\frac{1}{2} \times \sigma_{day}$ is only the standard deviation of the daily demand forecast. But if the intent was to account for demand spikes, then it is necessary to use the standard deviation of demand across the number of days that the pull is being calculated for - which would be the LT × WSF. A corrected version of Equation 5.1 is given in Equation 5.2.

$$\text{Pull} = (\text{LT} \times WSF \times \mu_{\text{day}}) + \left(\frac{1}{2} \times \sigma_{\text{day}} \times \sqrt{\text{LT} \times \text{WSF}}\right) \tag{5.2}$$

There are several things that are wrong with the way the KC bins are sized:

1. Inventory is supposed to be sized to ensure demand variation does not lead to part shortages. The way to do that would be to cover as much area under the demand distribution as possible. For a normally distributed demand it would mean adding multiples of the standard deviation to the mean, instead of having multiples of the mean as is done currently as shown in Equation 5.2. This is especially important if the distribution has a high variance, as multiples of the mean would still not capture enough area under the distribution to be adequately account for the variation.

2. In Equation 5.2, the standard deviation term is multiplied by $\frac{1}{2}$. Adding half of one standard deviation only covers 69% of the area, which means 32% of the time a shortage can be expected. To cover closer to 95% of the area, the standard deviation would have to be multiplied by 1.65, and by 2.33 to cover 99% of the area.

3. KC bins are updated quarterly whereas the demand forecast is updated monthly. Therefore the formula in Equation 5.1 is essentially invalid for 2 out of the 3 months of the quarter. The MIT team, while presenting the bin sizing policy

they developed, insisted on a monthly review of the bins.

In his thesis, Daigle described the inventory variation for the KC parts using Figure 5-2. It describes the steady state behavior of the inventory level at different times. Since there are 2 bins for every part, the maximum amount of stock that could exist for a part is 2 bins worth of parts. The minimum amount of stock for any part is 0 bins worth of parts. Therefore at any given time, the inventory level is somewhere in between these two extremes. As described above, the typical lead time for parts is 5 days, and the typical safety factor is 2. Therefore, each bin contains 2 business weeks worth of parts. In steady state, a bin would be refilled when the other bin is at about half its capacity. As a result, immediately after replenishment there is about a bin and a half worth of parts for each KC part type. Therefore the overall expected inventory level is the average of the inventory level immediately before and after replenishment, which is about 1 bin's worth of parts.

5.3.2 Gold Square sizing

The determination for how many Gold Squares would be allocated for each assembly that is on the Gold Square system is done much in the same way as the KC bins are sized for the piece parts. Equation 5.3 shows the formula used to determine the number of Gold Squares assigned for a given assembly.

$$\text{Number of Gold Squares} = \mu_{week} + \sigma_{week} \tag{5.3}$$

Where:

- μ_{week} is the average of weekly demand of assemblies on Gold Square
- σ_{week} is the standard deviation of weekly demand of assemblies on Gold Square

There is a major weakness in Equation 5.3 that will lead to a lot of shortages. That weakness is the fact that the standard deviation term is being multiplied by 1. One standard deviation only covers 84% of the area under the demand distribution

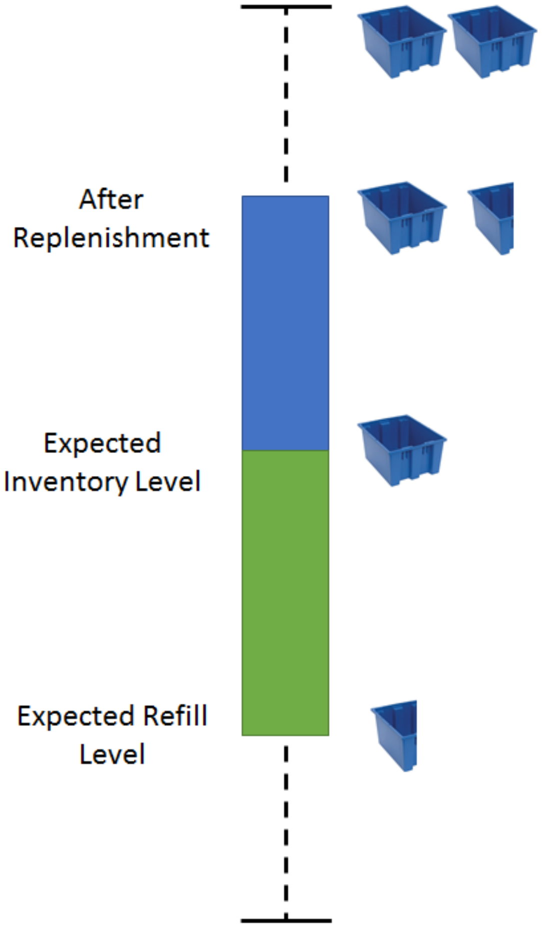


Figure 5-2: Inventory Characteristic of KC parts [5]

curve (assuming it is normally distributed), which implies shortages occurring 16% of the time.

The demand stream for assemblies on the Gold Square system is analyzed in weekly buckets, instead of the daily buckets used in the forecast of the KC piece parts. That is reason the average and standard deviation shown Equation 5.3 are of the weekly figures. When the MIT team sought to recalculate the sizing of the Gold Squares, they used a forecast of the assemblies in daily buckets instead of the weekly buckets used in the current formulation.

A detailed description of how the MIT team analyzed the formulations given in Equations 5.1 and 5.3 as well as how they developed a novel bin and Gold Square sizing formula is presented in the next section.

5.4 MIT proposed method of calculating Bin sizes

Looking at Equations 5.2 and 5.3, it is possible to draw similarities to a continuous review policy stock sizing formula with a few minor changes. A sample continuous review policy stock size formula is given in Equation 5.4 [1].

$$\text{Stock Size} = (\text{LT} \times \mu_{\text{daily demand}}) + \left(\text{z} \times \sigma_{\text{daily demand}} \times \sqrt{\text{LT}}\right) \tag{5.4}$$

Where:

- LT: is the lead time in days
- z: is the z score that covers a desired range of demand

If the LT × WSF term in Equation 5.2 was replaced by just LT and if the z score was assumed to be $\frac{1}{2}$, then it would be identical to Equation 5.4.

Similarly, Equation 5.3 would look a lot like Equation 5.4 if LT was 5 days (equivalent of 1 business week) and the z score was 1.0.

A major assumption that has to be made when using the base stock formula is that the demand forecast distribution is normally distributed. The criticism above of the current way of sizing the KC and Gold Square inventory is based on this assumption. The MIT team proceeded to check the validity of this assumption with the forecast used to size the current bins. The results of that analysis showed that the daily part demand was geometrically distributed, while the weekly demand is distributed according to the negative binomial distribution. Since the replenishment time for a lot of the parts is 5 days or more, the negative binomial distribution will be well approximated by a normal distribution, with parameters mean and standard deviation of the weekly demand. The full analysis of demand characterization is presented in Appendix C.

5.5 Bin sizing execution - comparison of current and proposed method

5.5.1 Methodology

There were two objectives that the MIT team had in mind when perfoming the analysis of bin sizing and Gold Square determination. The first was to introduce a more scientific way of analyzing a demand forecast and arrive at an appropriate formulation for stock sizing. The second was to perform a cost and shortage analysis to determine how much more would need to be invested by Varian to adopt the new recommended policy and what the pay off of the new policy would be when it comes to shortages.

In order to make an appropriate comparison, the MIT team applied the current method of inventory sizing calculation to the most up to date forecast, and used that same forecast to size the inventory according to the recommended policy. For the KC piece parts, the replenishment time was ensured to be the correct one for each part (by referencing what was being used to calculate the current bin sizes). For each assembly on the Gold Square system, the replenishment time was determined as the average time between the assembly being picked by workers on the flow line to when the supermarket is expected to replace it. The replenishment time has several components, described in Equation 5.5.

$$\text{Replenishment Time} = \text{Review Period} + \text{Supermarket Backlog} + \text{Assembly Build Time} \tag{5.5}$$

Where:

- Review Period: is the time between an assembly being taken of the shelf to when an order is cut to the supermarket to replace that assembly and put a new one back on the shelf. Currently the review period is about 30 hours or **1 day**.

- Supermarket Backlog: is the time an assembly stays in queue in the supermarket waiting for its turn to be built. Currently the supermarket backlog is about **4 days**.

- Assembly Build Time: is the time it takes to build the assembly. Currently the average build time is **1 day**.

Therefore, the current replenishment time for assemblies on the Gold Square system is **6 days**. For both the KC bins and the Gold Square assemblies, the replenishment time was used as the parameter LT in the continuous review policy formula given in Equation 5.4.

The mean and standard deviation of the daily demand for both KC parts and Gold Square assemblies were determined and used in Equation 5.4. The value of z was determined based on what service level gave the best projected shortage performance while also being financially feasible. The service level is defined as the percentage of the times a part is available when it is needed to satisfy demand. In the case of Varian Semiconductors, that demand could come from any of the demand streams discussed in a previous subsection. In many industries, service level ranges from about 95% to 99%. For cost calculations discussed in section 5.6, the MIT team picked various values of service level from this range.

In Microsoft Excel, the syntax to determine the appropriate value for z is given in Equation 5.6.

$$z = \text{NORM.S.INV(SL)} \tag{5.6}$$

Where SL is the desired service level.

Once the appropriate values for the parameters of the normal distribution were determined, the MIT team used the Microsoft Excel function for the inverse CDF of the standard normal distribution given in 5.6 and Equation 5.4 to determine the proper bin sizes and number of Gold Squares for each KC part and Gold Square assembly respectively.

5.6 Summary of results and economic recommendations

Given the demand stream scenarios and inventory sizing calculation methodology outlined in previous sections, the MIT team makes the following recommendations with regards to determining the sizes of the KC bins and the number of Gold squares for each eligible sub-assembly. For a detailed explanation of the cost calculations both for KC parts and Gold Square assemblies, refer to Anand's thesis [6].

5.6.1 KC Bin sizing recommendations

The MIT team recommends that all KC bins be sized based on the assumption that the probability density for weekly demand is approximated by a normal distribution. The service level that the KC bins provide should be set to 0.97 (97%). This service level is chosen because it provides good shortage performance at a modest increase in cost.

If Varian was to follow the two recommendations given above, the MIT team estimates that it will lead to an approximated 25% increase in total KC inventory. Considering the total cost of holding inventory, adding in the estimated monetary cost of shortage occurrences, the total cost increase on the Varian unit is approximately $\$320,000$ (19%). Adopting the recommendations above would lead to an estimated 80% reduction in shortage occurrences each year. Anand's thesis [6] presents a detailed cost breakdown of KC part bin sizing under different service level assumptions.

5.6.2 Gold Square sizing recommendations

Similar to the KC bins, it is the recommendation of the MIT team that the number of Gold Square slots for each eligible sub-assembly out of the supermarket be determined with the assumption that the daily demand is geometrically distributed, but that the weekly demand can be approximated by a normal distribution. A service level of 0.99 (99%) should be adopted, as it ensures good shortage performance at the least cost increase. It is essential that the backlog in the supermarket be reduced, then monitored closely in order to have the correct parameters in the negative binomial function.

As mentioned in section 5.5, the existing total lead time to make a sub-assembly in the supermarket is 6 days, of which 4 days the shop orders are waiting in several queues. Using 6 days in the negative binomial distribution would lead to a drastic increase in Gold Square slots which will overburden the supermarket and be costly to hold. Therefore, before the Gold Square numbers are determined, an effort needs to be made to reduce the existing backlog in the supermarket and optimize the production system so that in the long term, the backlog is manageable. That way the Gold Square numbers will be reasonable for the supermarket to keep up with.

It is the belief of the MIT team that if the total lead time was to be reduced to about 3 days, then the resulting Gold Square numbers would lead to a total annual cost increase of approximately $\$130,000$ (20% increase). It would also lead to a shortage reduction of about 74% over the existing scenario. Anand's thesis [6] presents a detailed cost breakdown of Gold Square sizing under different lead time and service level assumptions.

Chapter 6

Conclusion

This thesis analyzed the lack of improvement of FPY at Varian, after steady increase in the first 3 years since the start of the FPY project. Various reasons were hypothesized as to the reasons for the lack of improvement, statistical tests were performed to confirm or disprove those hypothesis. The hypotheses that were confirmed by statistical tests were followed up with projects to further analyze the problems hypothesized and recommend corrective action.

6.1 Error Re-categorization

The first problem that was identified was the way QNs that are written on the assembly floor are categorized as a way of identifying corrective action. The MIT team recognized that the QNs are only categorized by type, and don't provide a clear way of proposing specific projects to prevent the QNs. In response to this problem, the MIT team proposed a new way of categorizing the QNs based on the failure modes of the QNs. A methodology based on lean manufacturing practices was presented that will lead the quality engineers to identify the root cause of the failure modes and propose corrective action. Two case studies were presented in which the MIT team identified failure mechanisms after analyzing QN data for 6 months. Then the MIT team used the methodology they developed to identify root causes for the failure mode and proposed specific corrective action for those failure modes.

6.2 Inventory Management

The second problem that was identified was the way inventory for some part types was being managed. Part shortages was identified as a significant contributor to QN generation. When the MIT team looked into inventory management for KC parts and Gold Square Assemblies, they realized that the formulation to size the KC bins was not done in accordance to the proper frequency distribution of the part demand. In response, the MIT team identified the correct probability distribution for the daily demand and used that distribution to calculate bin sizes for the KC parts and number of squares for the Gold Square assemblies. Cost and predicted shortage analysis was performed on the new recommended inventory sizes and recommendations were made based on what service level will be the best economic choice for Varian. Additional recommendation was made to Varian regarding the Gold Square sizing, which involved reducing the assembly build lead time in the supermarket to replenish Gold Squares faster. The recommended service level for KC bins is 97%, with a predicted shortage reduction of 53%. The recommended service level for Gold Square assemblies 99%, the recommended supermarket lead time is 2 days, and a 75% reduction is predicted if Varian follows these recommendations.

Appendix A

Effect of Shortage Occurrence Reduction on FPY

In chapter 3, section 3.6, an expression was given that relates a change in the number of shortage occurrences of parts to the associated improvement in FPY. The full derivation of that expression is given below.

Being with equation 3.12, reproduced below:

$$\text{FPY} = e^{-(0.1319\times\text{SO}+1.291)\times N} \tag{A.1}$$

Let SO' be the new number of shortage occurrences and let SO be the old number shortage occurrences. Then let a be defined such that:

$$\text{SO'} = a \times \text{SO} \tag{A.2}$$

Let FPY' be the new FPY after shortage reduction and let FPY be the old FPY before shortage reduction. The expression for FPY' is:

$$\text{FPY'} = e^{-(0.1319\times\text{SO'}+1.291)\times N} \tag{A.3}$$

Substitute Equation A.2 into Equation A.3:

$$\text{FPY'} = e^{-(0.1319\times a\times\text{SO}+1.291)\times N} \tag{A.4}$$

Take the natural log of both sides of Equation A.4:

$$\ln(\text{FPY'}) = -(0.1319\times a\times\text{SO}+1.291)\times N \tag{A.5}$$

Divide both sides of Equation A.5 by a:

$$\frac{\ln(\text{FPY'})}{a} = -\left(0.1319\times\text{SO}+\frac{1.291}{a}\right)\times N \tag{A.6}$$

Divide both sides of Equation A.6 by N:

$$\frac{\ln(\text{FPY'})^{\frac{1}{a}}}{N} = -\left(0.1319\times\text{SO}+\frac{1.291}{a}\right) \tag{A.7}$$

Subtract 1.291 from both sides of Equation A.7:

$$\frac{\ln(\text{FPY'})^{\frac{1}{a}}}{N} - 1.291 = -0.1319\times\text{SO}-\frac{1.291}{a}-1.291 \tag{A.8}$$

Multiply both sides of Equation A.8 by N:

$$\ln(\text{FPY'})^{\frac{1}{a}} - 1.291\times N = -0.1319\times\text{SO}\times N-\frac{1.291}{a}\times N-1.291\times N \tag{A.9}$$

Rearrange the right side of Equation A.9:

$$\ln(\text{FPY'})^{\frac{1}{a}} - 1.291\times N = -\left(0.1319\times\text{SO}+1.291\right)\times N-\frac{1.291}{a}\times N \tag{A.10}$$

Add 1.291 × N to both sides of Equation A.10 and rearrange the right side:

$$\ln(\text{FPY'})^{\frac{1}{a}} = -\left(0.1319\times\text{SO}+1.291\right)\times N-\left(\frac{1.291}{a}-1.291\right)\times N \tag{A.11}$$

Take the exponential of both sides of Equation A.11:

$$\text{FPY'}^{\frac{1}{a}} = e^{-(0.1319\times\text{SO}+1.291)\times N} \times e^{-\left(\frac{1.291}{a}-1.291\right)\times N} \tag{A.12}$$

Notice the first expression on the right side of Equation A.12 is equal to FPY as definded in Equation A.1. Substitute FPY for that expression:

$$\text{FPY'}^{\frac{1}{a}} = \text{FPY} \times e^{-\left(\frac{1.291}{a}-1.291\right)\times N} \tag{A.13}$$

Finally, raise both sides of Equation A.13 to the power of a to get the new FPY after shortage reduction.

$$\text{FPY'} = \left(\text{FPY} \times e^{-\left(\frac{1.291}{a}-1.291\right)\times N}\right)^{a} \tag{A.14}$$

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

MATLAB Code For Demand Characterization

In chapter 5, section 5.5, a methodology was presented as to how the MIT team analyzed the daily demand distributions to come up with the right KC bin sizes and Gold Square numbers. In that methodology, the MIT team fit a geometric distribution to the daily demand to determine the parameter p, then used that value of p and the replenishment time to fit a negative binomial distribution to the demand across the replenishment time. The MIT team used MATLAB to find the the parameter p, and the code used is given below.

```
function a = geoDistParameterGenerator(demandFileName,demandSheetNum,demandRange,outputF

% Load the filename, sheet number and range of the demand data
% Make sure to include the column with part numbers (without letters)
filename = demandFileName;
sheet = demandSheetNum;
xlRange = demandRange;

% Read the excel file and save to the rawData matrix
rawData = round(xlsread(filename, sheet, xlRange));
sizeOfData = size(rawData);
```

```
% numberOfParts is the number of rows in rawData
numberOfParts = sizeOfData(1);

% divvy up the first column of rawData as the vector of part_number
% divvy up the second rest of the rawData matrix as the dailyBuckets matrix
part_number =rawData(:,1);
dailyBuckets = rawData(:,2:end);

numberOfDays = sizeOfData(2)-1; % subtract 1 because first column is part nubmers

% numberOfWeeklyBuckets is the number of lead times there are in the range of
% the demand data. Lead time for KC parts is 5 days.
numberOfWeeklyBuckets = (numberOfDays - mod(numberOfDays,leadTime))/leadTime;

% Initialize the weekly buckets, first column is the part number vector
% defined above
weeklyBuckets = zeros(numberOfParts,numberOfWeeklyBuckets+1);
weeklyBuckets(:,1) = part_number;

% initialize an output matrix that will serve as summary for each part.
% first column is the part numbers; second column is the fitted parameter
% of the geometric distribution for the daily demand, third parameter is
% the goodness of fit (R^2) of the negative binomial distribution to the
% weekly demand. We will only send the first two columns of the output
% matrix to excel.
output = zeros(numberOfParts,3);
output(:,1) = part_number;
partNumberCol_strings = cell(numberOfParts,1); % initialize as cell array to export to e

% start indexing through each part
for row=1:numberOfParts
% for row=1:100
    part = weeklyBuckets(row,1);

    % create the weekly buckets by aggregating demand every 5 days
```

```
for x=1:numberOfWeeklyBuckets
    weeklyBuckets(row,x+1) = sum(dailyBuckets(row,(((x-1)*leadTime)+1):(leadTime*x)
end

part_weekly = weeklyBuckets(row,2:end); % demand in weekly buckets
part_daily = dailyBuckets(row,:);       % demand in daily buckets

% create a histogram of the weekly demand to get demand bins and
% their frequencies. Normalize the fequencies to get a probability mass
h = histogram(part_weekly);
h.BinEdges = 0:1+max(h.Data);
x = h.BinEdges(1:length(h.BinEdges)-1)';
y = h.Values';
y_norm = y/sum(y);
close all

% fit a geometric distribution to the daily demand buckets to get the
% parameter to use to create the appropriate negative binomial
% distribution for the weekly demand
phat = mle(part_daily,'distribution','geo');
y_binom = makedist('NegativeBinomial','R',leadTime,'p',phat);
binom = cdf(y_binom,x);

% calculate the goodness of fit (R^2) of the negative binomial distribution
% on the weekly demand buckets created above
square_residuals = (binom-y_norm).^2;
r_square = sqrt(sum(square_residuals));

% store the parameter of the geometric distribution and the goodness of
% fit (R^2) to the second and third column of the output matrix
% respectively.
output(row,2) = phat;
output(row,3) = r_square;

% Store the part number as a string in the partNumberCol_strings
% intitialized above. Modify the if and elseif parameters depending on
```

```
    % the structure of the demand file.
    if strcmp(isKC,'TRUE') % If fitting for KC parts

        if row <=22 % For parts that don't have a letter in front of them
            partNumberCol_strings(row) = cellstr(strcat(num2str(part)));
        elseif row == 23 % For parts that have the letter B in front of them
            partNumberCol_strings(row) = cellstr(strcat('B',num2str(part)));
        elseif row == 24 % For parts that have the letter C in front of them
            partNumberCol_strings(row) = cellstr(strcat('C',num2str(part)));
        else % All other parts should have the letter E in front of them.
             % If not, add more elseif statements
            partNumberCol_strings(row) = cellstr(strcat('E',num2str(part)));
        end

    else % If fitting for Gold Square Assemblies; all assemblies should
         % have the letter E in front of them.
        partNumberCol_strings(row) = cellstr(strcat('E',num2str(part)));
    end

end

% create the phatCol, a column of the parameters for the geometric fit for
% each part, and copy the second column of the output matrix to it.
phatCol = output(:,2);

% store the name of the excel files you're doing the analysis in as
% fileName.
% store the sheet where the matlab output will go to as sheetNum.
fileName = outputFileName;
sheetNum = outputSheetNum;

% store the range of cells where the output will go in the partWriteRange
% and phatWriteRange variables.
partWriteRange = strcat('A2:A',num2str(numberOfParts+1));
phatWriteRange = strcat('B2:B',num2str(numberOfParts+1));
```

```
% write the part numbers and the phat values to the specified file, sheet
% and cell range above.
xlswrite(fileName,partNumberCol_strings,sheetNum,partWriteRange)
xlswrite(fileName,phatCol,sheetNum,phatWriteRange)

a = 'Success'; % Confirmation that the fitting was successfully executed

end
```

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix C

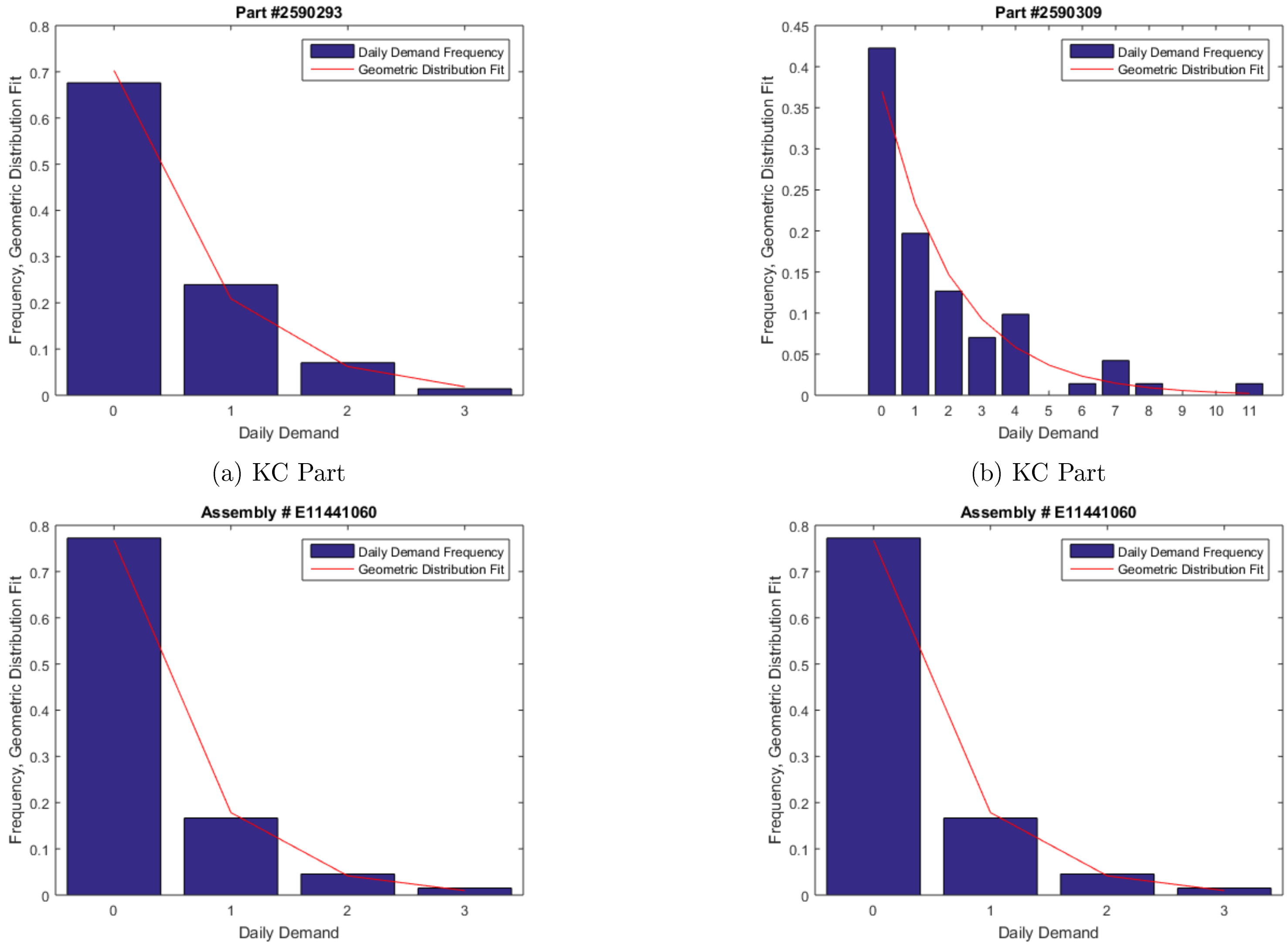
Demand Characterization

C.1 Demand Characterization

To determine the probability distribution of the daily demand for each part, the author created a histogram of the daily demand amounts and normalized them by the total number of entries. A few of these histograms are shown in Figure C-1 in the dark blue bars. The histograms show a distribution that resembles a decaying exponential function. Since demand is a discrete value, the author proceeded to fit a geometric distribution to the normalized histograms. A geometric distribution is the discrete form of the exponential distribution. The geometric distribution appears to be a decent fit to the normalized histograms of the demand for a lot of parts.

Having come to the realization that the daily demand forecasts are not normally distributed, but rather geometrically distributed, the MIT team decided that the current way of calculating the bin and Gold Square sizes is not appropriate. A new way to determine the bin sizes and the number of Gold Squares had to be formulated, given a geometrically distributed demand. Before getting into that formulation, it is important to define an important term - the service level.

The service level is defined as the percentage of the times a part is available when it is needed to satisfy demand. In the case of Varian Semiconductors, that demand could come from any of the demand streams discussed in a previous subsection. In many industries, service level ranges from about 95% to 99%. For cost calculations



(a) KC Part

(b) KC Part

(c) GS Assembly

(d) GS Assembly

Figure C-1: Demand Probability Distributions and Fits for KC Parts and GS Assemblies discussed in section 5.6, the MIT team picked various values of service level from this range.

C.2 The Geometric Probability Distribution

The probability mass function (PMF) of the geometric distribution gives the probability of a certain number of trials occurring exactly before success is achieved, and is characterized by the probability of a success being achieved on any of the trials. In mathematical form, the PMF for the geometric distribution looks like the one given in equation C.1 [3].

$$\text{PMF}_{\text{geometric}} = \text{P}(x) = p(1-p)^{x};\ \text{where x} = 0,\ 1,\ 2,\ 3,\ ... \tag{C.1}$$

The cumulative distribution function (CDF) of the geometric distribution gives the probability of certain number of trials or less occurring before success is achieved. It is also characterized by the probability of a success being achieved on any of the trials. In mathematical form, the CDF for the geometric distribution looks like the one given in equation C.2 [3].

$$\text{CDF}_{\text{geometric}} = \text{P}(X \leq x) = 1-(1-p)^{x+1};\ \text{where x} = 0,\ 1,\ 2,\ 3,\ ... \tag{C.2}$$

In both equations C.1 and C.2, p, the parameter of the geometric distribution, is the probability of success on any trial. p is always a number between 0 and 1 since it is a probability.

Figure C-2 shows a sample plot of the PMF and CDF of a geometric distribution for various values of p.

In the context of demand characterization, the geometric distribution can be understood as follows. In equation C.1, x would correspond to the number of parts needed during a given day and p would correspond to the probability that no part is needed to build a particular machine. A "trial," in this case would be the assembly

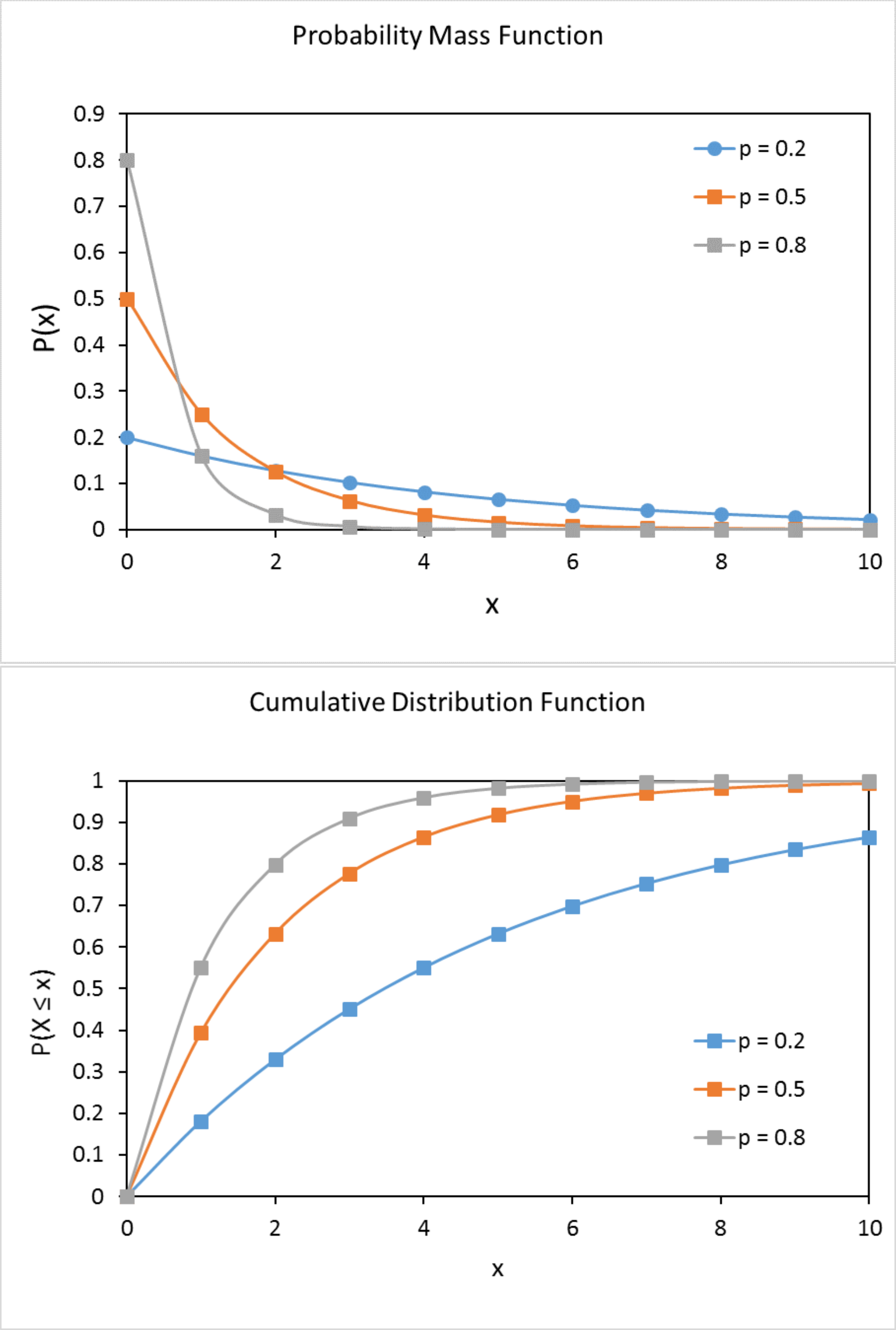


Figure C-2: PDF and CDF for a geometric Distribution process that takes place during 1 day of production. "Success" in this trial would be defined as the event in which 0 counts of a part were needed for assembly during a day; and "failure" would be defined as the event in which 1 part was needed for production. A "success" occurs at the end of the day, therefore x is the count of "failures" during that day. In accordance to the definition of the geometric PMF given above, $P(x)$ is the probability that exactly x parts are consumed for production during a given day (or x failures), where p is the probability that no part was needed.

C.3 Working with a geometrically distributed demand

When attempting to create any stock policy, it is important to make sure there is enough inventory to prevent stock-outs during periods of part replenishment. That is done by characterizing the demand distribution, and covering as much area under the probability density as is financially feasible. Typically the cumulative distribution function (cdf) is used to determine what level of stock is required to satisfy demand a given percentage of the time - based on the service level the business wants to provide.

In the simple case of the replenishment time being 1 day, then the stock is sized to cover just a day's worth of demand. Hence the cdf of the daily demand can be used. But if the replenishment period (RP) is greater than 1 day, then the stock needs to be sized to cover demand variability on each day of the RP. During each day of the RP, demand is being sampled from its distribution. Therefore to size the stock appropriately, those instances where the daily demands are being sampled from a distribution need to be aggregated in some way. For example, if the demand was normally distributed and if each day's demand was being sampled from that distribution during the replenishment period, then the aggregate amount would come from a distribution whose parameters are simple algebraic functions of the parameters of the normal distribution from which each day's demand was being sampled. Equation C.3 illustrates this point further.

$$\begin{aligned}\mu_{\text{aggregate distribution}} &= \text{RP} \times \mu_{\text{daily demand}} \\ \sigma_{\text{aggregate distribution}} &= \sqrt{\text{RP}} \times \sigma_{\text{daily demand}}\end{aligned} \tag{C.3}$$

Where:

- $\mu_{\text{aggregate distribution}}$ is the average of the distribution from which the aggregate of the individually sampled values comes from
- RP is the replenishment period which corresponds to how many days the sampling takes place for
- $\mu_{\text{daily demand}}$ is the average of the daily demand from which sampling takes place during each day of the RP
- $\sigma_{\text{aggregate distribution}}$ is the standard deviation of the distribution from which the aggregate of the individually sampled values comes from
- $\sigma_{\text{daily demand}}$ is the standard deviation of the daily demand from which sampling takes place during each day of the RP

When the underlying daily demand is not normally distributed, but rather geometrically distributed, then the relationship between the distribution from which the daily demand comes from and the distribution from which the aggregate value comes from is not as simple. In fact, Daigle [5] insisted that the aggregate comes from a different distribution entirely. Upon further research, it was found out that the aggregate of several geometrically distributed samples comes from the negative binomial distribution.

The probability distribution function of the negative binomial distribution describes the probability of observing a certain number of "successes" before observing a pre-defined number of "failures." The distribution has 2 parameters - k and p [4]. k is the pre-defined number of "failures" as described above and p is the probability of "success." In the special case that $k = 1$, the negative binomial distribution converges to the geometric distribution with parameter p. In the case of the KC bins and the Gold Squares, k would correspond to the replenishment time for a bin or the production lead time of an assembly. Meanwhile p would be the parameter obtained from fitting both the KC part and the Gold Square assembly daily demand to the geometric distribution [5].

Figure C-3, from Daigle's thesis, shows the PDF and CDF of a negative binomial distribution for various values of k and p.

The negative binomial distribution is a built in function in Microsoft Excel, which accepts as inputs the value of interest,x (the number of "successes" to be observed), and parameters k and p as described above, and returns the output of either the PDF or the CDF. The syntax in Excel looks like Equation C.4.

$$\begin{aligned} \text{PDF} = P(x) = \text{NEGBINOM.DIST}(x, k, p, FALSE) \\ \text{CDF} = P(X \leq x) = \text{NEGBINOM.DIST}(x, k, p, TRUE) \end{aligned} \tag{C.4}$$

The $TRUE$, or $FALSE$ binary in Equation C.4 is to indicate whether the user desires to use the CDF of the distribution or not. $TRUE$ indicates a desire to do so, whereas $FALSE$ indicates the user would like to use the PDF instead.

In sizing the KC bins as well as determining the number of Gold Square, the MIT team used the inverse of the CDF of the negative binomial distribution to determine what number of parts would satisfy demand during the replenishment period according to a given service level. The inverse negative binomial distribution is not a built in Excel function, but an add-in from an external source is easily obtainable. The syntax of the negative binomial function from the add-in is given in Equation C.5.

$$x = \text{NEGBINOM\_INV}(\text{SL}, k, p) \tag{C.5}$$

Where:

- SL is the desired service level

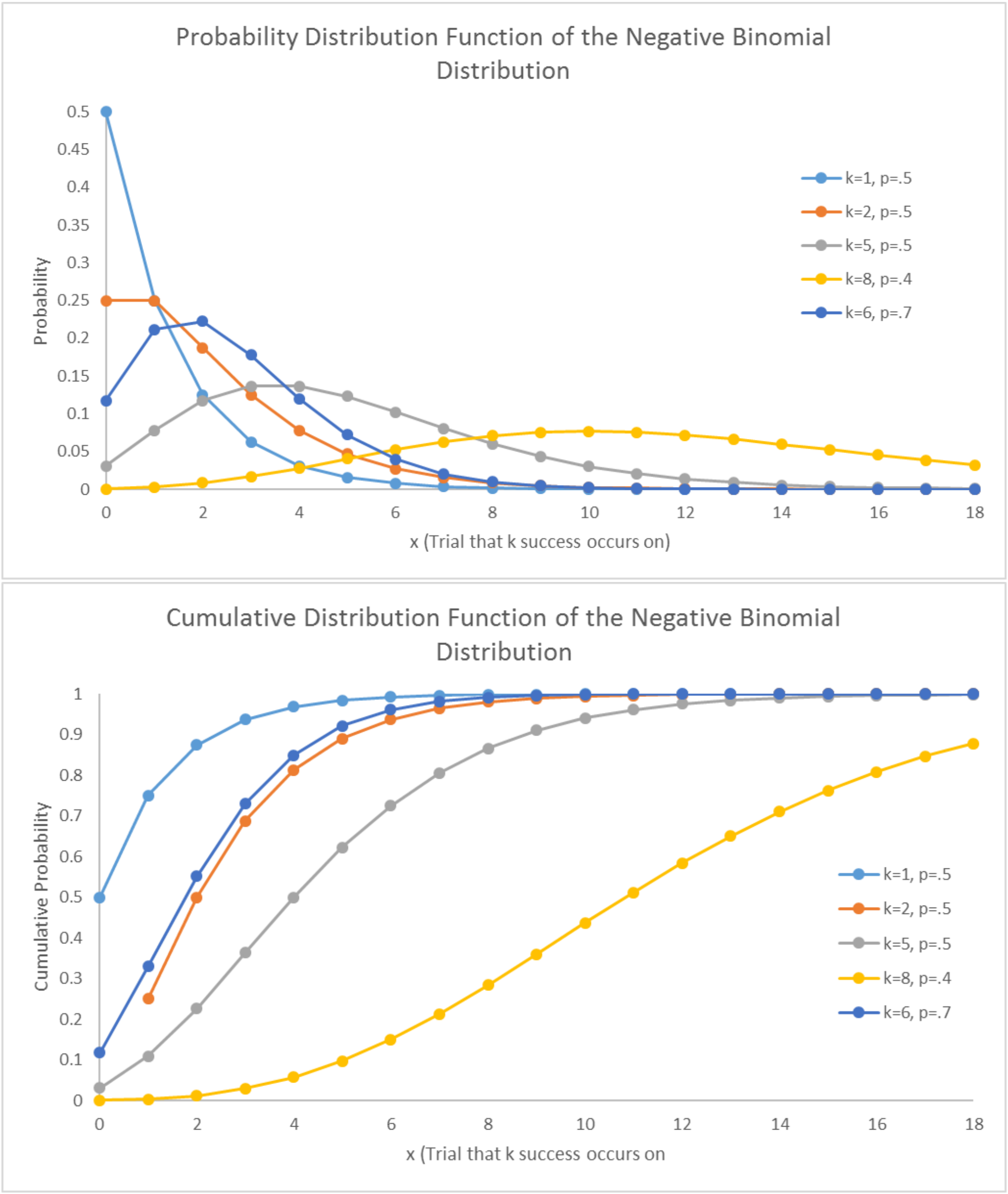


Figure C-3: PDF and CDF for a Negative Binomial Distribution Distribution

While the negative binomial distribution captures the demand distribution over multiple days of RP well, over a typical replenishment time of 5 days, it tends to converge to a normal distribution. The normal distribution is widely known and is ubiquitous in its application. The MIT team believes that Varian is more likely to adopt their inventory recommendations if they were presented in terms of the normal distribution than lesser known distributions such as the negative binomial distribution. Therefore the MIT team made the normality approximation on the weekly demand when recommending bin sizes for KC parts and number of squares for Gold Square assemblies. The methodology and execution of that sizing is presented in section 5.5

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] D. Simchi-Levi, P. Kaminsky, and E. Simchi-Levi, *Designing and Managing the Supply Chain*, 3rd ed. McGraw-Hill, 2007.

[2] J. Liker, The Toyota Way - *14 Managememt Principles From the World's Greaters Manufacturer*. 2 Penn Plaza, Newyork, NY 10121: McGraw-Hill, 2004.

[3] "Geometric Distribution - MATLAB Simulink." [Online]. Available: http://www.mathworks.com/help/stats/geometric-distribution.htmlbtxgmni. [Accessed: 06-Aug-2016].

[4] "The Negative Binomial Distribution." [Online]. Available: http://www.math.uah.edu/stat/bernoulli/NegativeBinomial.html. [Accessed: 11-Aug-2016].

[5] S. Daigle, "Title to be determined; thesis in progress at time of this writing", M.Eng Thesis, Massachusetts Institute of Technology, January, 2017.

[6] S. Anand, "First Pass Yield Analysis and Improvement at a Low Volume, High Mix Semiconductor Equipment Manufacturing Facility," M.Eng Thesis, Massachusetts Institute of Technology, September, 2016.

[7] S. Jain, "Assembly lead time reduction in a semiconductor capital equipment plant through improved material kitting," M.Eng Thesis, Massachusetts Institute of Technology, 2014.

[8] A. S. Bhadauria, "Production lead time reduction in a semiconductor capital equipment manufacturing plant through optimized testing protocols," M.Eng Thesis, Massachusetts Institute of Technology, 2014.