AALTO UNIVERSITY
School of Science

Mat-2.4108 Independent Research Project in Applied Mathematics

Optimizing sampling structures in multilevel testing

Mansikkamäki Ilkka
26.11.2011

Table of Contents

1 Introduction

1.1 Background

Modern corporations face difficulties when trying to develop devices. Small differences in technical qualities, flexibility in industry standards and a variety of products with similar functions force device manufacturers to develop their products so that they can be used as widely as possible. Globally functional products further force the corporations to test their products extensively.

Companies typically test their products based on the technical properties of the product. All the aspects of the user experience need to be taken into account. Furthermore, when testing globally functional products the companies need to perform tests with several products that can be compatible with theirs. This is important because it increases the market opportunities for their products. When testing some specific capability of a product, it is important to have statistical support for testing. Naturally, statistical proof of performance may not be sufficient, for instance, aircraft testing needs to be done thoroughly because errors cannot be there. On the other hand, for products that are used daily by millions of people, such as computer mice, statistical testing of compatibility and performance is sufficient, because the cost of device failure is small.

1.2 Objective and scope

In this study we analyze the compatibility testing process of an electronic device manufacturer. We examine the testing process to determine the minimum effort and cost of proving that products are compatible with as many products as possible. Especially, we need to minimize the number of undetected failed tests and the associated costs. We explore data from earlier tests to discover possible patterns and correlations in the testing data. In what follows, we refer to products and devices, products being the manufactured items object to test and devices meaning the external items we want our product to function with.

The rest of this study is divided into four sections. The second section introduces reliability engineering in more detail. We review the notion of reliability and what it means in the context of this study. We also discuss tools presented in the literature for solving reliability problems. Third section describes the data and applied methods and calculation principles. Fifth section introduces the results. The last section concludes by discussing the study and future implementations for further studies.

2 Reliability engineering and testing

First, we discuss reliability in general and look at the engineering process phases. Second, we look at the affecting factors in reliability based decision making. In the last section, we discuss the different aspects of sampling and choice of sample size.

2.1 Reliability

Automization and mass production as well as hazardous failures of important equipment have necessitated the development of reliability calculations [4]. These processes enabled data gathering for statistical analysis and allowed better analysis of processes.

Reliability is directly connected to testing results and errors in testing. Two kinds of errors exist. Random errors are due to lack of repeatability. For a single test, the result may vary, but with repeated tests, the result becomes more accurate. If the accurate testing results achieved via repeating still differ from the expected value, the reason can be a systematic error. Systematic error causes measurements to deviate from the actual value systematically [1]. Furthermore, testing failure may be caused by several reasons, such as human factors or software elements. Also environmental factors influence the results [3].

Reliability is defined by failure rates, failure being determined as failure within some time. Smith [3] described reliability as "the probability that an item will perform a required function, under stated conditions, for a stated period of time". This gives also the opportunity to calculate the mean time to failure.

Reliability can mean several things. For instance, it may indicate the length of life of a device, the functioning of all the properties of a device or in general the satisfaction of the customer with a device. When determining reliability, we need to remember, though, that there is only one definition of failure, which is, in the words of Smith, "non-conformance to some defined performance criterion" [3]. In our system, reliability is seen as compatibility. Moreover, failure means that our product does not function properly with the tested device.

2.2 Testing program planning

Wheeler and Ganji [1] present a framework that opens the different phases of experimental program, as outlined in Table 1. The first two phases of a testing program define the problem and the experiments. In the next two phases the actual testing is performed and documented. Last phases turn the gathered results into actions and help in decision making.

Table 1 Reliability testing planning phases

Program phases
1. Problem definition
2. Experiment design
3. Experiment construction and development
4. Data gathering
5. Analysis of data
6. Interpreting results and reporting

Our study focuses on the last phases of the testing program, starting from the stage where we have already received the testing results and seek to analyze their importance and inclusiveness. Based on the results we try to give recommendations for the next phases of the process. Figure 1 illustrates the phases in the focus of this study.

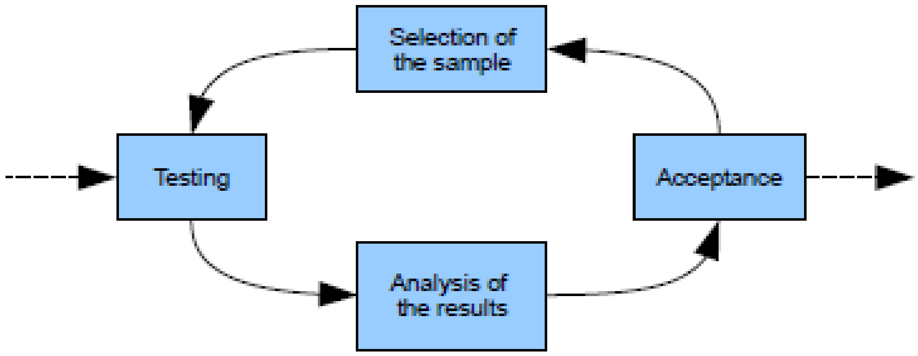


Figure 1 Testing system

Companies should pay close attention to the importance of experiment design. They need to consider all the various aspects of reliability and possibilities for error. In principle, the more complex the design and manufacturing structure, the more prone to error the system is.

2.3 Reliability and costs

Reliability and especially failures are costly. Table 2 presents the typical costs related to quality and failure to conform to the quality [4].

Table 2 Costs related to reliability.

	Testing cost	Specification	Failure cost	Specification
Costs before sales	Prevention costs	Product or service requirements Quality planning Quality assurance Inspection equipment Training Miscellaneous	Internal failure costs	Scrap Rework or rectification Reinspection Downgrading Waste Failure analysis
Costs after sales	Appraisal costs	Verification Quality audits Inspection equipment Vendor rating	External failure costs	Repair and servicing Warranty claims Complaints Returns Liability Loss of goodwill

When determining the desired confidence level and pass rate of products, we need to estimate the costs, especially those costs that might be high in value. In the study case, the highest are the external failure costs. Customer complaints and liability fights may be costly for the company.

There are several different methods to model the reliability in a system. Some methods are more or less conceptual methods that help to understand the process and draw attention to the flaws of the system [2]. Some methods, on the other hand, have been created to better estimate reliability of the system. These methods use mathematics in calculations

Block diagrams are used to present the system. Visualization of the system often helps identify the problems and improvement areas of the system. Block diagrams also work as a way of presenting the mathematical model [3]. Section 3.1 shows a block diagram of the testing system we are analyzing.

When the system involves circular references, simulation may be needed to determine the optimal probability level [2]. Complex mathematical models of the system usually require computer assistance and simulation. In this study, we did not need to simulate the process but used computer assistance and graphical solving of equations.

2.4 Risk analysis

Risk analysis links financial aspects with reliability and gives occurring probabilities for the undesirable events [8]. It is often advisable to over-estimate the effect of realized risk to find a sufficient testing sample, for instance. Assessing the risks includes several phases as listed in Table 3 [7].

Table 3 Phases of risk assessment process.

1. Identify the hazards
2. Identify the barriers for hazards
3. Estimate the capacity of the barriers for hazards
4. Estimate the exposure
5. Estimate the risks

Given that this study focuses on modeling the required sample size for statistical coverage of errors, it is assumed that the first three phases of the risk assessment process have been completed. Furthermore, although it is hard to determine the impact of an incompatible product for a company, estimates of the monetary exposure should exist. Thus this study focuses mainly on estimating the risks.

2.5 Reliability engineering and mathematics

Besides theoretical frameworks and guides about how companies should tackle the issues related to product reliability, reliability engineering also considers mathematical models for calculating the reliability of systems and products.

Most approaches in reliability engineering focus on detecting failures over time as the product ages. Another common testing approach considers cases where we test some measurable quality of a product such as length or strength [2]. Fewer models consider the cases where there are only two outcomes that do not depend on time. Our study focuses on two outcomes; pass and fail. The product either passes or fails the compatibility test. Analyzing this scenario with binomial distribution is a common approach.

3 Methods and data

3.1 Testing process

The process that we want to analyze is shown in Figure 2. All the test cases are divided based on two variables. First, we have two kinds of devices that we test our products with. Second, all the test cases have been divided into three categories. We expect to see a different failure rate for each joint group of device type and each testing category. If the device passes all the tests, the device is determined compatible with the product. Otherwise it is defined as partially compatible or incompatible.

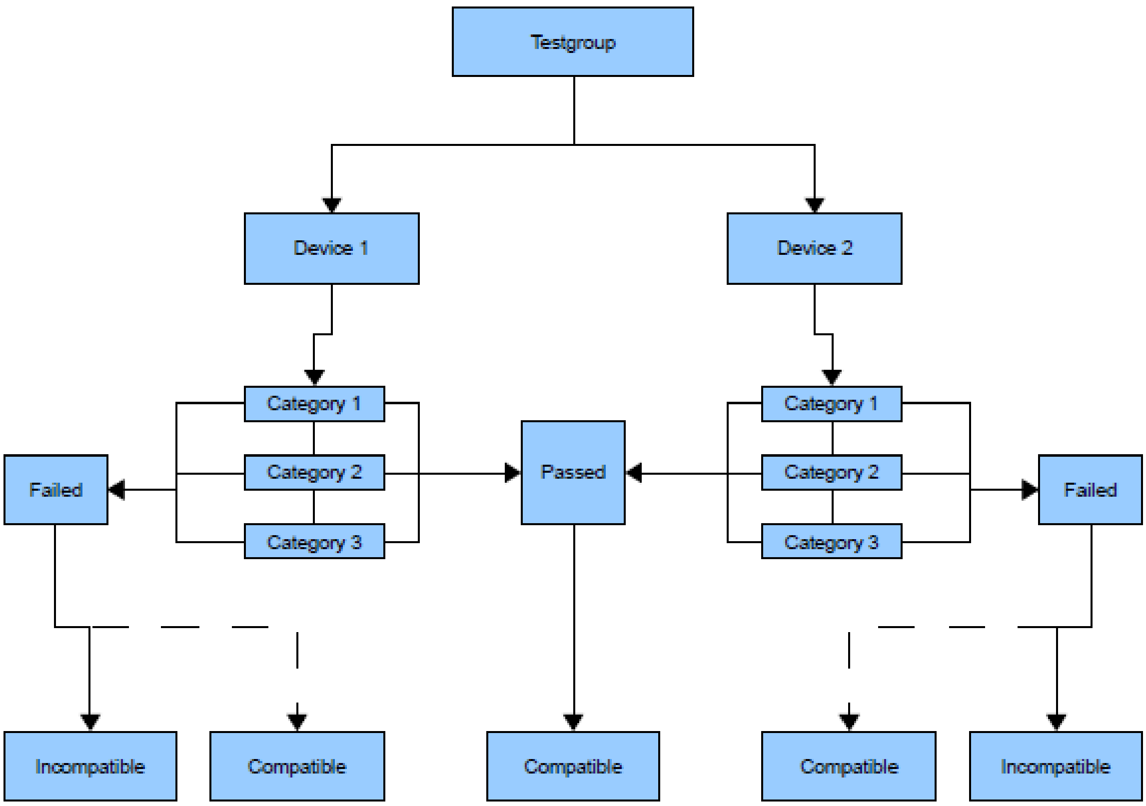


Figure 2 Testing process.

We assume that the number of failed tests that we discover from the tests follows binomial distribution. That is, if a test is repeated n times and each test fails with probability p, then the number of failed tests, X follows binomial distribution $X \sim Bin(n, p)$ [5]. The probability that exactly x tests fail is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, 0 < p < 1, x = 0,1,2, \dots, n. \tag{1}$$

The expected value and variance of a binomial distribution are

$$E(X) = np, \qquad Var(X) = np(1-p).$$

Binomial distribution has the property that a transformation

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

approaches standard normal distribution $N(0,1)$ as the number of tests n increases. The estimated number of tests n after which the approximation is statistically relevant varies between different sources, it is commonly estimated that when $n > 30$, $np > 5$ and $n(1-p) > 5$ the approximation is sufficient [6]. We expect that the failure rates in this case are quite close to zero and thus we need to check the sufficiency of the normal distribution approximation. The probability p defines the failure rate for one group unless other notice.

3.2 Groupwise failure rates

For determining the failure rates for each group, we had test results for a set of similar products. We calculated the average failure rate for each group across the whole test set. The test results included more than 34 000 test cases and nearly 2000 test cases for one group, so the average estimate was quite comprehensive. From the data, we could calculate the failure rates p for each group. To ensure that the differences between the groups would be statistically significant, we performed variance analysis for the failure rates.

3.3 Determining the number of undetected failures

We tested several different approaches to the problem of discovering the undetected failures and ended up concentrating on the confidence interval of the failure rates. For instance, we have observed X failures after performing N tests. If the sample size is sufficient, we can approximate the binomial distribution with normal distribution and calculate the confidence interval for X

$$X \pm N * \frac{z_{1-\alpha/2}\sqrt{p(1-p)}}{\sqrt{N}}, \tag{2}$$

where $p = \frac{X}{N}$ is the failure rate and $z_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile of normal distribution with parameter $1 - \alpha$ determining the confidence level. For example, if $\alpha = 0.05$, the number of failures lies between this confidence interval with 95% probability.

The decisions that are taken based on the testing results, such as marketing strategies and additional testing decisions, do not take the confidence interval into account. The device is either compatible or not. This is not necessarily a valid approach, because there will always be statistical variance in the results, and this variance is mainly connected to the sample size. However, we also will consider the observed failures as accurate results from the tests performed up to some point, and then estimate the failure rate for the unperformed tests within the limits of statistical confidence.

Suppose that the maximum number of tests that we can perform is M. We can estimate that we would detect failures with the failure rate p and thus would find $X = Mp$ failures if we perform all those M tests, according to the results estimated from the sample data. Now let us assume we want to do only N tests. We can again estimate that we would detect $X_N = Np$ failures. We want to estimate the risk of not performing the remaining $M - N$ tests. How many failures can we estimate that we would see with the whole testing set M? The result can be calculated using formula (2) for confidence interval. If we say that the observed failure rate p_{obs} is the lower limit of some actual failure rate p_{act},

$$p_{obs} = p_{act} - z_{1-\frac{\alpha}{2}} * \frac{\sqrt{p_{act}(1 - p_{act})}}{\sqrt{N}},$$

we can solve the expected failure rate

$$p_{act} = \frac{p_{obs}}{a},$$

where

$$a = 1 + \frac{\frac{z}{p_{obs}} - \sqrt{\left(\frac{z}{p_{obs}}\right)^2 + 4M * \frac{z}{p_{obs}} - 4M}}{2M},$$

and $z = z_{1-\frac{\alpha}{2}}$. Figure 3 presents the failure rate as a function of the sample size. Not only does the confidence level quickly become much smaller than with small sample size, but also the validity of the normal distribution approximation gets better.

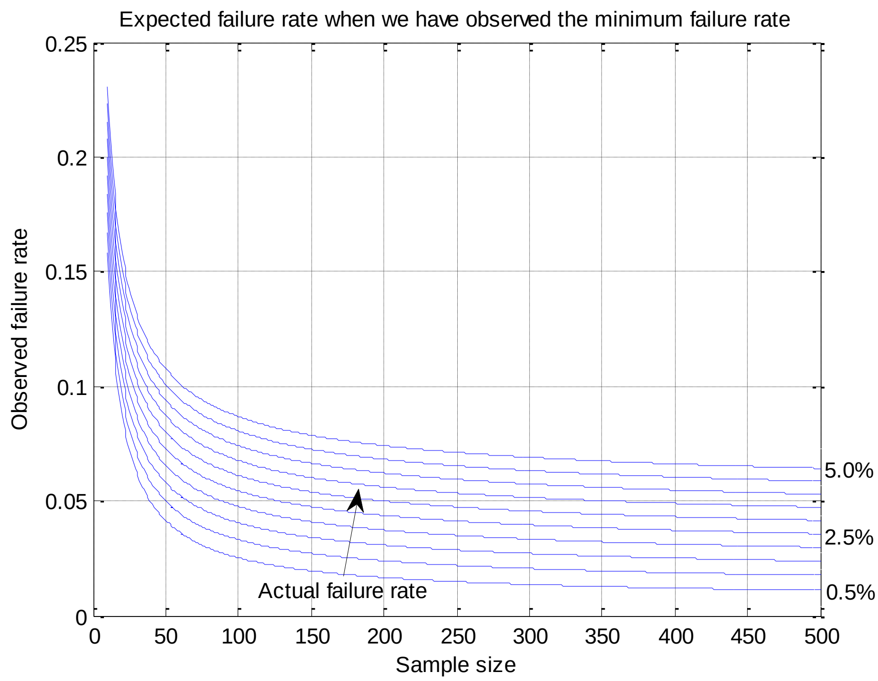


Figure 3 Actual failure rate when calculated with the observed failure rate.

From the figure we can see that for example if we have observed 5% failure rate with 300 tests we can expect the actual failure rate to be still about 7%.

Now that we have determined the actual failure rate we expect to see after N tests, we can calculate the expected number of failures found with $M - N$ tests. The number of undetected failures U is

$$U = Np_{obs} + (M - N)p_{act} - Mp_{obs} = (M - N)(p_{act} - p_{obs}). \quad (3)$$

3.4 Determining the compatibility of a device

In addition to the number of undetected failures, we are also interested in what do failures mean in terms of compatibility. As a result, we need to determine the limits within which the product is compatible with a device. The company did not have specific limits for compatibility so we estimated the compatibility limits for each category from a given dataset.

The compatibility limits should be stated as a number of allowed failed tests per testing category. Let l_j be the limit for testing category j. For each test that we perform, there is some failure probability p_j for testing category j. From the binomial distribution we see that if we perform n_j test cases with failure rate p_j, the expected count of failed tests is $E = n_j * p_j$. When the limit for declaring the device compatible with the product is l_j, we can calculate the probability that one device passes all the compatibility tests

$$P(device\ compatible) = P(category_1 = pass) * P(category_2 = pass) * P(category_3 = pass) = \prod_{j=1}^{3} P(X_j < l_j * n_j), \quad (4)$$

where X_j is the number of failed tests in testing category j. Here we have also assumed that different categories are independent from each other.

The probability that the device passes the compatibility test in testing category j can be calculated using equation (1)

$$P(X_j < l_j * n_j) = 1 - \sum_{m=0}^{n_j - l_j * n_j} \binom{n_j}{m} p_j^{\ n_j} (1 - p_j)^{n_j - m}. \quad (5)$$

We can also calculate the probability that the device passes the compatibility test with the expected failure rate p_{act}.

3.5 Optimal sample size

When we have determined the expected failure rates, we can estimate the optimal number of tests. The optimum in this case would be the point where the increase in the number of test cases is more expensive than taking the risk of not observing the failures. The analysis from this point onwards is more risk analysis than statistical estimation.

To determine the cost of testing, say that the unit cost of a test is c, which is equal for each testing category. In addition there is a purchasing cost C_i for both device categories, $i \in \{1,2\}$. Let N_i be the number of devices tested and M_i the target number of devices. The goal is that the savings from reducing the testing sample are higher than the expected cost due to an incompatible device that we did not detect. Let L_i further be the cost caused for the company when a device turns out to be incompatible. The optimal situation would be that the cost of testing would be lower than the cost of not testing

$$\sum_i \sum_j N_i n_j c + \sum_i N_i C_i + \left\{ \sum_i (M_i - N_i) L_i \prod_{j=1}^{3} P\left(X_j(p_{obs,i,j}) < l_j * n_j\right) \right\} < \sum_i (M_i - N_i) L_i \left(\prod_{j=1}^{3} P\left(X_j(p_{obs,i,j}) < l_j * n_j\right) - \prod_{j=1}^{3} P\left(X_j(p_{act,i,j}) < l_j * n_j\right) \right), \quad (6)$$

where the left side is the direct costs we have from testing the N_i devices and the accepted risk of incompatible devices and the right side is the cost of undetected incompatible devices. Here, we have assumed that the cost of incompatibility is the same for each testing category. We can write equation (6) as a cost function for both device categories separately

$$f_i(N) = (nc + C_i)N + (M_i - N)L_i\left(2P_{obs,i}(N) - P_{act,i}(N)\right), \quad (7)$$

where $n = \sum_j n_j$ and $P_{obs}(N)$ and $P_{act}(N)$ total compatibility estimates from equation (4) with failure rates p_{obs} and p_{act} respectively. Minimizing this cost function for both categories will give the optimal sample size. The minimum can be found by determining the zero-point of the derivative. We will not find the analytical minimum here, but refer to numerical calculations in the results section.

3.6 Analysis tools

In the different phases of this project we used different computer tools. The original data was in Excel-files where we made some modifications. In the next phase, we used QlikView version 9.0 for group-wise pass rate and volume calculations. QlikView is a tool designed for combining data from various sources and formats and also for visualization. With QlikView it was easy to combine all the test data into one database. Because Qlikview is more of a database calculation tool than single equation calculation tool, we used Matlab to compute the final results.

4 Results

The following results, estimated from the test results of a product family, are based on the theoretical discussion in previous section. The calculations for undetected failures per testing category are based on test counts shown in Table 4. These are the average counts for each testing category within the product portfolio.

Table 4 Number of test cases in each testing category.

Testing category	Number of test cases
Category 1	14
Category 2	15
Category 3	61

4.1 Failure rates for the testing categories

We determined both failure rates p and pass rates for each group (Table 5). Pass rate is defined as $1 - p$.

Table 5 Failure rates and pass rates.

Testing\Device	Device 1		Device 2	
	Pass rate	Failure rate	Pass Rate	Failure rate
Category 1	98.3%	1.7%	99.6%	0.4%
Category 2	97.1%	2.9%	98.1%	1.9%
Category 3	97.7%	2.3%	97.0%	3.0%

The failure rates are, as expected, quite close to zero. Thus the normal distribution approximation should be considered with caution. Furthermore, failure rates seem to depend at least on the testing category but also slightly on the tested device. Table 6 summarizes the results from a two-way variance analysis. The p-value in the last column tells that the device groups do not have statistically significant differences and the testing categories have statistical differences only at 8% significance level. The interaction of the categories does not have statistically significant differences. This means that we should consider the devices purely as one group. However, after discussions with the company we calculated the results for each group separately because of the nature of the groups.

Table 6 Results from two-way variance analysis.

Source of variation	SS	df	MS	F	p
Device	0.63	1	0.63	0.09	0.7606
Testing	37.16	2	18.58	2.79	0.0772
Interaction	17.96	2	8.98	1.35	0.2745
Error	199.57	30	6.65		
Total	255.32	35			

4.2 Undetected failures

We performed the calculations with a reference point of 200 devices. Expert from the target company estimated that this would be a number of devices that we would never be testing because of testing costs but it presents well the total device population. The number of failures we would detect with the failure rates from Table 5 are shown in Table 7. Table also takes into account the number of test cases performed normally in each testing category (Table 4).

Table 7 Expected number of failures with 200 tested devices.

Testing Category	Device 1	Device 2
Category 1	48	11
Category 2	87	57
Category 3	281	366

With equation (3) we calculated the estimated number of failures that we would possibly be able to notice if we test less than 200 devices. Figure 4 shows the number of undetected failures when the number of tested devices approaches 200 devices. We have rounded the failure counts to the closest integer for clarity.

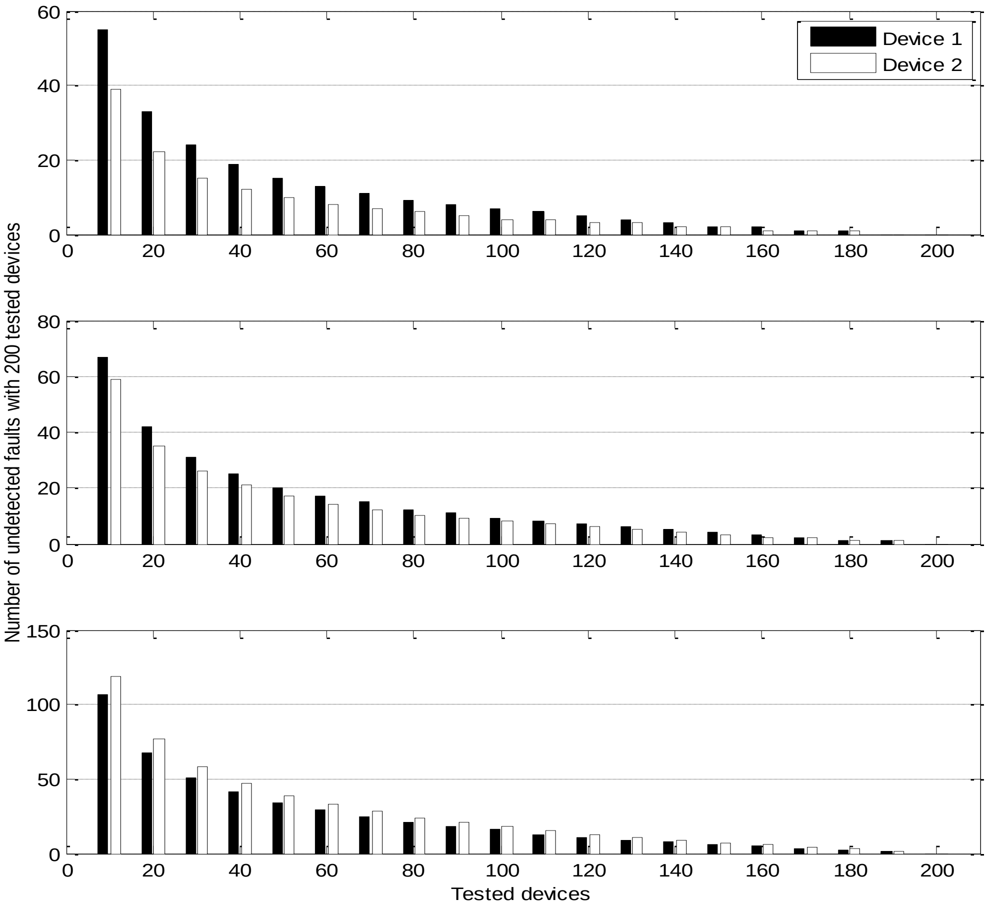
Testing Category	Number of undetected failures when the comparison device sample includes 200 devices.
1	
2	
3	

Figure 4 Number of undetected failures.

The number of failures does not correlate linearly with compatibility, but it still can tell us how many incompatible devices can be undetected in the worst case. For instance, if our assumption of independency of test cases is incorrect, but the number of failures still follows the binomial distribution with our calculated failure rates, we would see 11 failures in testing category 1 for device 2 when we test 200 devices. But if we test only 20 devices, we should be prepared to see additional 22 failures with those 11 failures for 200 devices.

4.3 Compatibility effects

The idea of following the undetected failures raises a question what the failures mean in terms of compatibility. To answer this, we have limits for compatibility. Table 8 shows the estimated compatibility limits in percentages. We could only derive common results for the devices due to lack of data. But this should not be an issue as the devices should have equal compatibility limits when resolving global compatibility.

Table 8 Compatibility limits for testing categories. The limit tells the percentage of tests that the device needs to pass to be defined as compatible.

Testing category	Compatibility limit
Category 1	l_1 = 99.0%
Category 2	l_2 = 95.8%
Category 3	l_3 = 89.9%

Now we can calculate the accepted number of failed tests for each testing category and based on that calculate the pass rates for both device categories. Table 9 shows the results.

Table 9 Expected compatibility pass rates for example test volumes and compatibility limits.

Category	Test parameters		Expected pass rates	
	Number of tests	**Accepted number of fails**	**Device 1**	**Device 2**
Category 1	14	0	78.66%	94.54%
Category 2	15	1	93.12%	96.78%
Category 3	61	6	99.95%	99.77%
Total	**90**	**7**	**73.21%**	**91.29%**

As can be seen from Table 9 testing category 1 dominates the compatibility decision and has the lowest categorical pass rate in both device groups. Figure 5 shows how the compatibility estimation develops as the number of tested devices increases. Note that we calculated the compatibilities using the estimated actual failure rate p_{act} and there hasn´t been any reference sample size. But we can see from the graph that when the sample size is about 200, the compatibility probability has already quite well set to a value and does not increase much. The figure shows also the compatibility pass rates from Table 9 (straight lines). Appendix A contains compatibility estimations for each testing category.

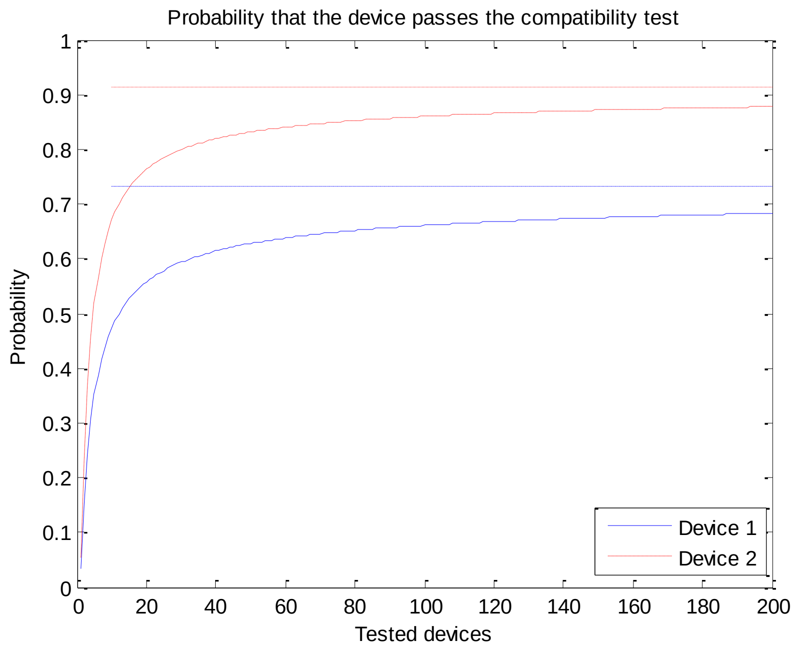


Figure 5 Compatibility test pass rate for device groups.

4.4 Optimal costs

To calculate the optimum, we used the same reference population 200 devices as in Section 4.2. Figure 6 shows an example graph of the cost functions. The testing for device 1 with these parameters is the cheapest when we test 27 devices. The cost for device category 2 is minimized when we test 16 devices respectively.

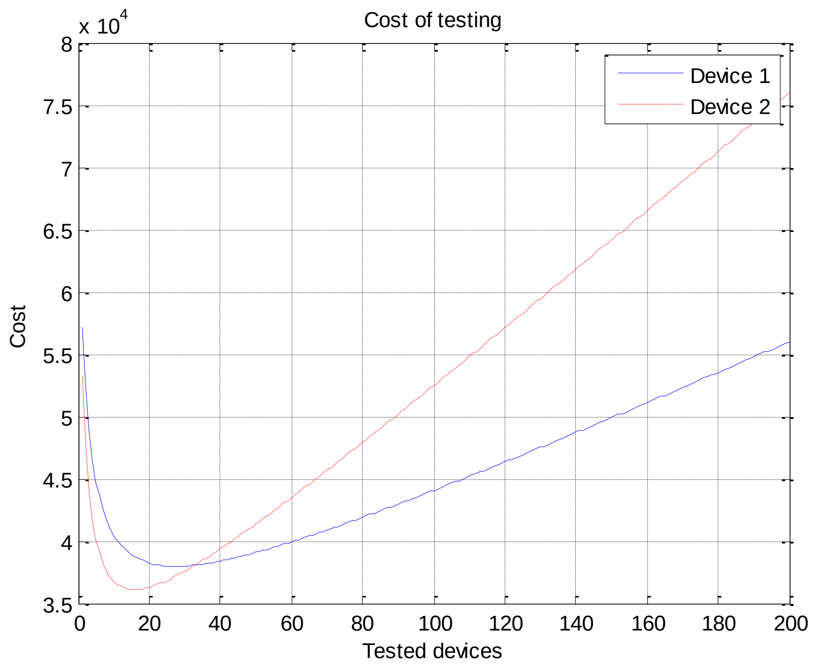


Figure 6 Cost of testing. Estimation parameters c = 2, C1 = 100, C2 = 200, L1 = 200, L2 = 150.

We note that the optimal number of tested devices differs considerably when the cost parameters change. If a minimum is found within reasonable limits, the optimum is often between 15 and 50 devices. Appendix B contains more figures with different parameters. The testing parameters in the graphs differ greatly because we do not have knowledge of the actual costs related to the incompatibility.

5 Discussion and Conclusions

In this estimation we had to make many assumptions and simplifications, some of them very reasonable, some less so. The purpose of this study was to bring insight to the estimation of the testing reliability.

One simplification was to consider all test cases in one testing category, as well as all tested devices, independently from each other. This may not be a completely correct assumption, because some of the devices can be more prone to errors as well as some tests might be harder than others and the tested product may not have the tested functionality. However, this study focused on one of the last testing rounds before the product launch, and most of the critical errors have already been found and fixed. This suggests that the purpose of the tests is to find the global compatibility estimate for the tested product and to locate the failures rising from statistical variance in tests.

Another simplification was made in the estimation of compatibility limits. The company does not have such limits and the decisions are mainly based on expert opinions. We suggested implementing such limits which already gave rise to discussions in the company. Pre-determined compatibility limits would make analytical decision making easier in the future.

This study also largely neglected the results of the variance analysis when despite of the results we estimated the results for both device groups separately. The analysis could have been done for a combined group. However, analysis for both device categories separately gave possibly better understanding of the testing situation. Another disregarded theoretical assumption was the applicability of the normal distribution approximation for binomial distribution. However, our Monte Carlo simulations during the project showed that the normal approximation did not widen the confidence limits and thus did not weaken the accuracy of our analysis.

Despite of the imperfections we reached interesting results. However, when studying the results, the reader should remember that if we apply these results to a new testing scenario, the failure rates and testing category test counts need to be adjusted to fit the testing in question. The results are best applicable when we already have an estimate on the expected failure rate. Moreover, to further improve the analysis, the expected pass rates could be derived from beta distribution, which would mean that the expected number of failures follows beta-binomial distribution. This would bring better global solution to the sample size estimation.

Another point of interest for further study could be to simulate how the compatibility changes when we change the number of test cases. This could require a different viewpoint, because adding test cases is not the same as adding devices to the testing sample. Adding test cases would mean that the tested device should be able to fulfill the functionality and although we assume that the test cases are independent, adding one could ruin this assumption. Studying each test case could bring more insight on the factors affecting the test results and make the estimations more accurate.

6 References

[1] A.J. Wheeler, A.R. Ganji, Introduction to Engineering Experimentation, 3rd Edition, 2010, School of Engineering San Francisco State University, Pearson.

[2] P.D. T. O'Connor, Practical Reliability Engineering, 4th Edition, 2002, John Wiley & Sons Ltd, Chichester.

[3] D.J. Smith, Reliability Maintainability and Risk, 6th Edition, 2001, Butterworth-Heinemann, Oxford.

[4] J. S. Oakland, Statistical Process Control, 5th Edition, 2003, Butterworth-Heinemann, Oxford.

[5] P. Laininen, Todennäköisyys ja sen tilastollinen soveltaminen, 1998, Oy Yliopistokustannus / Otatieto, Helsinki.

[6] I. Mellin, Todennäköisyyslaskenta ja tilastotiede: Kaavat, 2006, Material for applied probability calculations, Aalto University, School of Science.

[7] M. Modarres, Risk Analysis in Engineering: Techniques, Tools and Trends, 2006, Taylor and Francis, Boca Raton (FL).

[8] A. Salo, A. Käki, Risk analysis, Lecture material, 2011, Aalto University, School of Science.

Appendix A - Compatibility rate

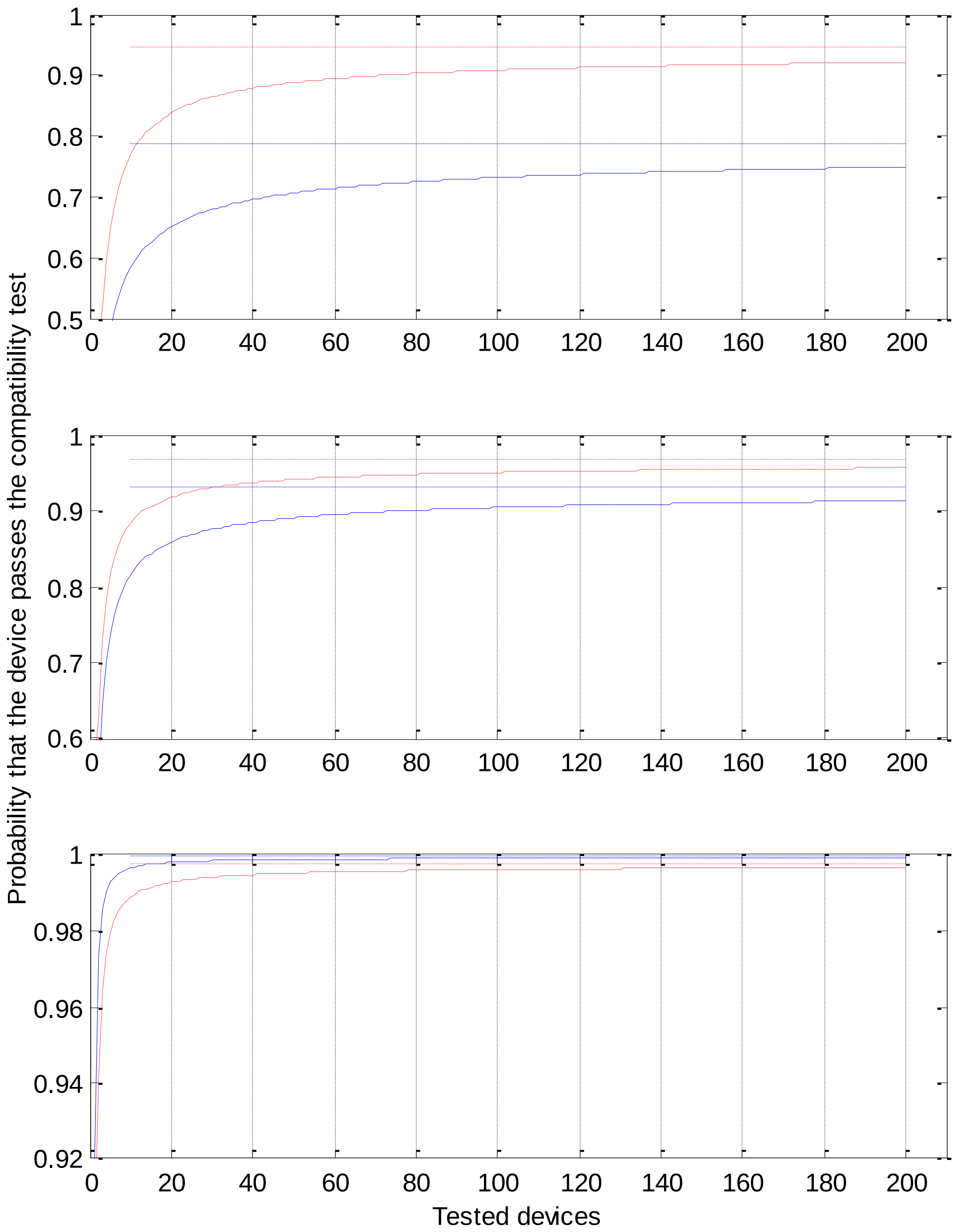
Testing Category	Pass rate of the compatibility test with the estimated actual failure rate
1	
2	
3	

Figure 7 Compatibility pass rate estimation. Solid lines represent device 1 and dashed lines device 2. Straight lines show the compatibility rates calculated with the observed failure rate.

Appendix B – Cost figures

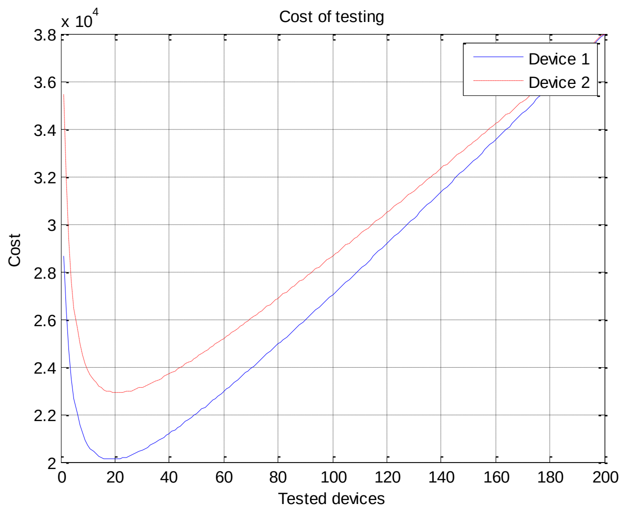


Figure 8 Testing costs. Estimation parameters c = 1, C = 100 and L = 100 for both devices.

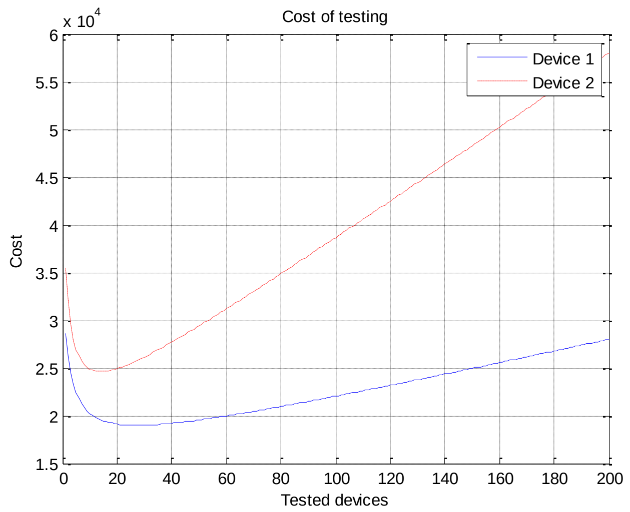


Figure 9 Testing costs. Estimation parameters c = 1, C1 = 50, C2 = 200, L = 100 for both devices.

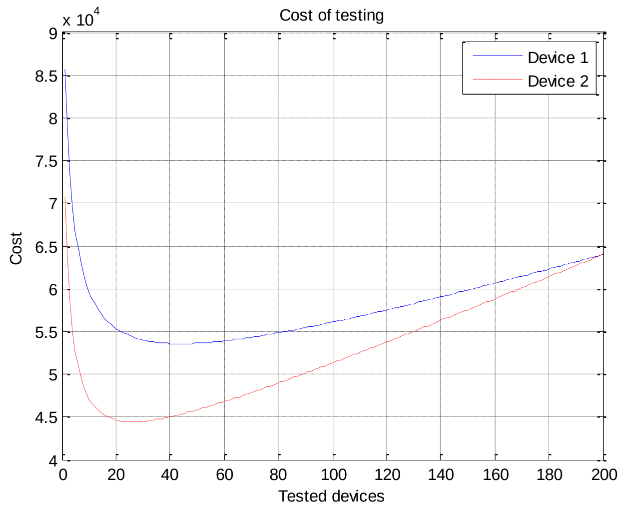


Figure 10 Testing costs. Estimation parameters c = 3, L1 = 300, L2 = 200, C = 50 for both devices.

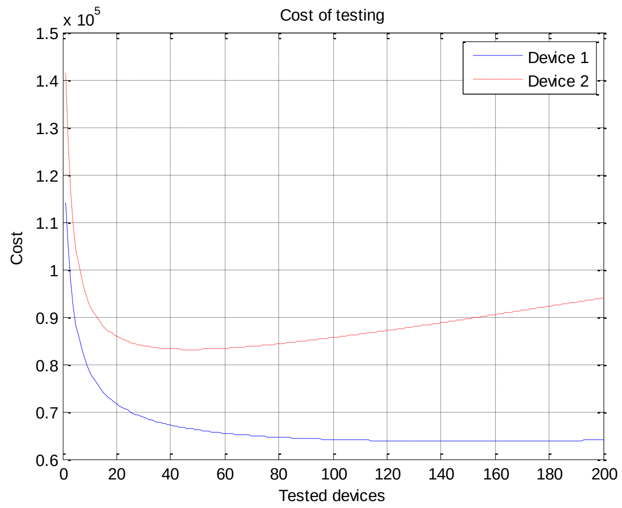


Figure 11 Testing costs. Estimation parameters c = 3, C1 = 50, C2 = 200, L = 400 for both devices.