Using Operating Characteristic (OC) Curves to Balance Cost and Risk

Best Practice

Authored by: Lenny Truett, Ph.D. STAT T&E COE

The goal of the STAT T&E COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.AFIT.edu/STAT.

STAT T&E Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary

The operating characteristic (OC) curve is the primary tool in lot acceptance sampling plans (LASP). They allow data from a sample to be used to draw conclusions about the lot as a whole with defined risks. Although LASP may not be used in DoD testing, the same concept can be applied to reliability or performance when the response is expressed as a percentage (e.g. the system must successfully complete the task at a rate of 80%). One limitation of this methodology is that it only applies to a task that is performed repetitively at the same conditions because it requires all of the samples to be taken from the same population. OC curves allow the test planners to easily see how changing different criteria that define required performance and level of risk impacts the required amount of testing. Therefore, they are an ideal way to balance cost and risk. Understanding OC curves can also be a practical way to better understand alpha, beta, delta, and sample size for many other types of tests. They also clearly demonstrate how not considering delta can lead to a flawed perception of risk.

Keywords: Operating Characteristic curves, OC curves, acceptance testing, consumer's risk, producer's risk, AQL, RQL

Introduction

Operating Characteristic (OC) curves are widely used in industry for lot acceptance sampling plans (LASP). The x-axis is typically the percent defective, and the y-axis is the probability that the lot will be accepted. The user defines the Acceptable Quality Limit (AQL) and the Rejectable Quality Limit (RQL), and the producer's risk and the consumer's risk. The OC curve is generated by determining a sample size and an allowable number of failures or defects. Figure 1 is an OC curve with an AQL of 0.9 (percent defect of 0.1), a RQL of 0.8, producer's risk of 0.1, consumer's risk of 0.1, sample size of 86 and acceptance number of 12 (maximum number of failures allowed in the sample).

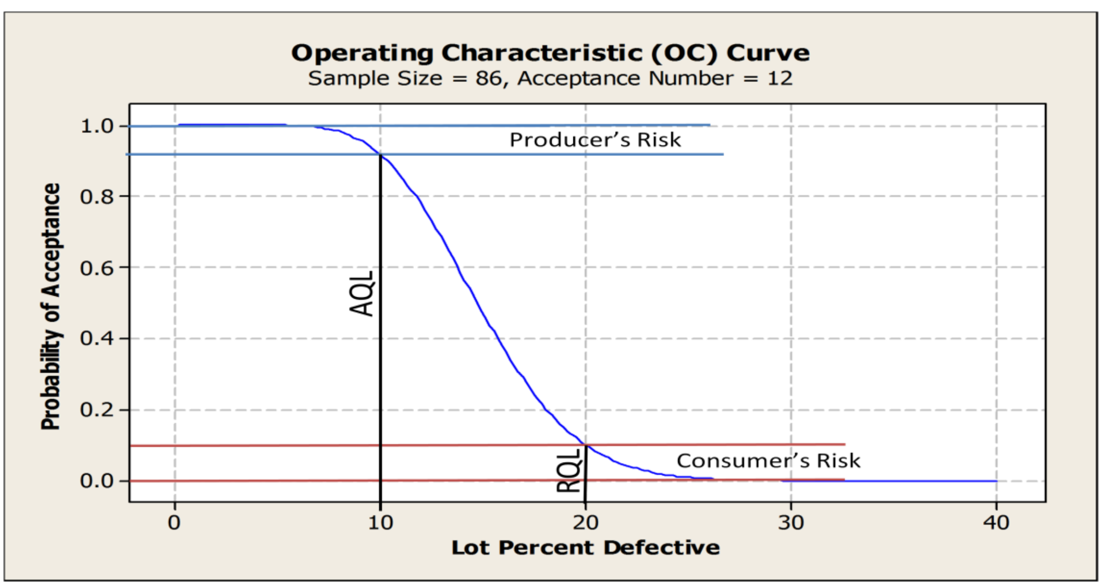

Figure 1: Sample OC curve

Each of the criteria can have a profound effect on the others and understanding how they interact is critical for developing a sampling plan that balances cost versus risk. While your initial reaction may be that this is not applicable to DoD testing, the same principles for creating OC curves can be useful for certain types of DoD testing. Also, OC curves are a practical graphic method to understand criteria alpha, beta, delta, sample size and how they relate to power and confidence.

Understanding the Criteria

Setting Limits

The Acceptable Quality Limit (AQL) is a percent defective that is the requirement for the quality of the producer's product. The producer would like to design a sampling plan such that there is a high probability of accepting a lot that has a defect level less than or equal to the AQL. Or stated another way, "What level of performance do I want to pass."

The Rejectable Quality Limit (RQL) is a maximum percent defective that would be unacceptable to the consumer. The consumer would like the sampling plan to have a low probability of accepting a lot with a defect level as high as the RQL. Or stated another way, "What level of performance do I want to fail." While not specifically mention in OC planning, the difference between the AQL and RQL is delta. OC curves help define and visualize a meaningful delta because it considers what value should pass, and what value should fail, instead of abstractly setting the difference between them at 1 or 2 standard deviations.

Setting Risks

Understanding the risks is important to the process of developing OC curves. Notice that risks is plural. There are two type of risk that should be considered, but they do not necessarily have to be equal. In LASP, the system is assumed to pass the AQL, so if the system truly is at or above the AQL, but the sample has more than the allowable number of defects, this is a type I error. In LASP, the type I error rate (or alpha) is associated with producer's risk. However, alpha is not always the producer's risk, and this will be discussed later.

If true system performance does not meet the RQL, but the number of defects in the in sample is equal to lower than the acceptance number, this is a type II error. In LASP, the type II error rate (or beta) is associated with consumer's risk. However, beta is not always the consumer's risk, and this will be discussed later.

Ideally alpha would be set at 0.05 but not more than 0.10. Many test in the DoD use alpha as high as 0.20 because of costs or other constraints. This might be required by budget or time constraints, but this means there is up to a 20% chance that you will reject a system that exactly meets the AQL. If I was a producer that delivered a system that met the requirement, I would be uncomfortable knowing that I might fail the test 1 out of 5 times. Any time constraints lead to accepting an alpha higher than 0.10

leadership should be made aware of the risk. Furthermore, they should be informed of the cost to lower alpha to an acceptable level.

Ideally beta would be 0.10 or less but 0.20 is also common. To assess the risk of passing a system that does not meet the RQL, the consequences of incorrectly passing the system should be considered. Another way to decrease beta is to increase delta, but this would increase the AQL and/or lower the RQL and may result in unacceptable limits. So when considering "What do I want to pass and fail" you should also consider how much risk you are will to accept at that level.

Sample Size and Allowable Failures

The last two criteria to consider in developing OC curves are the sample size (*n*), and the number of allowable failures or defects(*c*). As *n* increases the slope of the curve will increase, allowing the AQL and RQL to be closer together or consumer's and producer's risk will be lowered. As *c* increases the curve will move to the right.

Drawing and Interpreting OC Curves

Creating a Sampling Plan

It is possible to generate OC curves in Excel, but this paper will not focus on the derivation required to do this. There are many other software tools that can generate OC curves, but this paper will use Minitab 16 Statistical Software for all examples because it is easy to use and flexible enough most requirements. The **Acceptance Sampling by Attribute** function is under the **Stat** tab and **Quality Tools** sub menu. The input screen is shown in Figure 2.

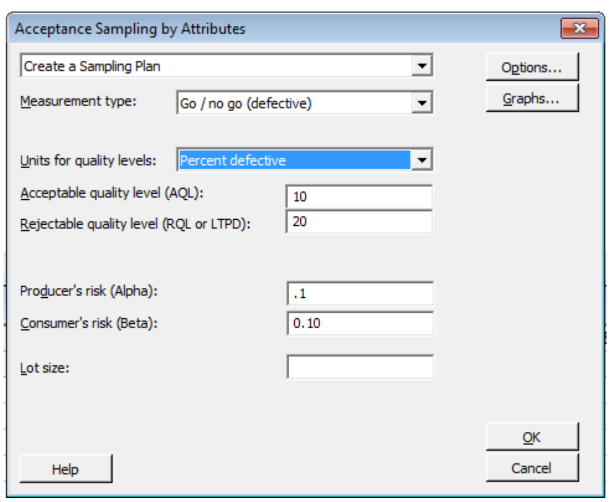

Figure 2: Minitab Dialog Screen

This section will show you how to create a sampling plan in Minitab. We will focus on the Go / no go (defective) measurement type, but there is also the ability to select the number of defects. We will also use the percent defective default, but other options include the proportion defective or defectives per million. The first two inputs boxes are the AQL and RQL in percent defective. Therefore enter 10 for a 0.90 proportion and 20 for a 0.80 proportion. Here alpha and beta are set to 0.1 and 0.1 and the results are shown in Figure 1. The program will also output the results in tabular form as shown in Table 1.

```
Acceptance Sampling by Attributes

Measurement type:  Go/no go
Lot quality in percent defective
Use binomial distribution to calculate probability of acceptance


Acceptable Quality Level (AQL)          10
Producer's Risk (Alpha)                 0.1

Rejectable Quality Level (RQL or LTPD)  20
Consumer's Risk (Beta)                  0.1


Generated Plan(s)

Sample Size        86
Acceptance Number  12

Accept lot if defective items in 86 sampled <= 12;  Otherwise reject.


  Percent  Probability  Probability
Defective    Accepting    Rejecting
       10        0.914        0.086
       20        0.099        0.901
```

Table 1: Sample Minitab Output

The sample size given is the smallest *n* that will meet or exceed all of the requirements. In this case the actual alpha is 0.086 and the actual beta is 0.901. For this sampling plan, we would conclude that the system was acceptable if we observed 12 or less defects (or failures). But remember, even if you choose to accept the performance, there is still risk to the producer (a sample that fails a system that truly meets the AQL) defined by alpha and the consumer (a sample that passes a system that is at or below the RQL) defined by beta. A good sampling plan does not eliminate the risks; it just balances the risks versus the cost in an objective way. Also, the only decision that should be made using a sampling plan is to accept or reject the system. No other conclusions should be inferred even if the results are far above the AQL or far below the RQL.

To see how to balance cost versus risks, let's look at two examples. For both cases we want the system performance to be 0.90. In one case we can only test 50 times, and in the other case we can test 200 times.

Given the limited number of tests available in the first case, we will have to relax some of our criteria. We can increase our risk, or lower our RQL. Increasing alpha and beta to 0.15 would require 59 runs with 8 acceptable defects. Decreasing the RQL to 0.75 but keeping alpha and beta at 0.10 would require 40 runs with 6 allowable failures. The OC curves for both of these cases are shown in Figure 3.

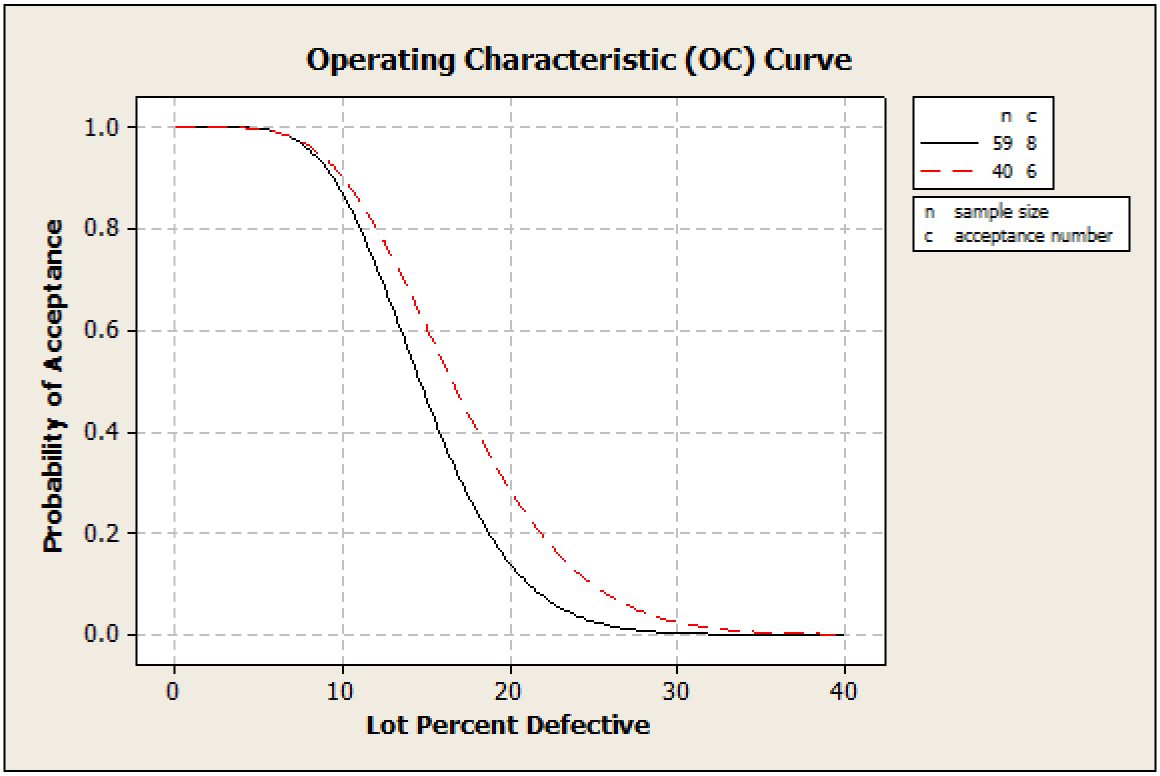

Figure 3: OC Curves for *n*=59 *c*=8 and *n*=40 and *c*=6

While these sampling plans do offer alternatives neither seems ideal because one requires 9 more tests than available, and the other does not use all of the available testing and may be giving you less information than possible. A method for better optimizing the OC curve will be given in the next section.

If we assume that we have more resources available, we can tighten some criteria. If we reduce alpha and beta to 0.05, this results in n=139 and c=19. If we increase the RQL to 0.85 this results in n=288 and c=35. These OC curves are shown in Figure 4. Again, neither of these answers appears to be optimized for 200 tests. Possible recommendations could be to reduce testing to 139, if the current RQL is acceptable, or increase the amount of testing to 288 if the raising the RQL to 0.85 justifies the additional cost. There is no right answer; this is simply a tool to balance cost and risk. The impact of changing alpha beta and the delta is different for each system under test.

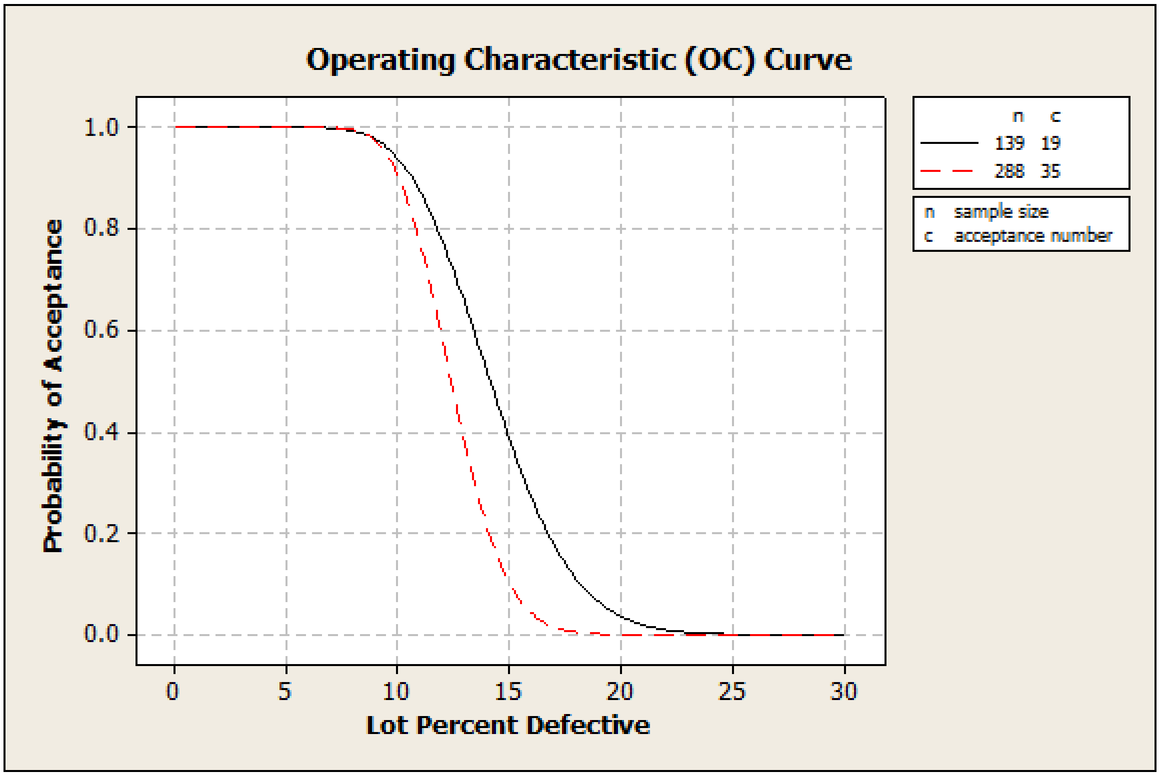

Figure 4: OC Curves for *n*=139 *c*=19 and *n*=288 and *c*=35

Comparing User Defined Sampling Plans

In the previous section, we attempted to develop sampling plans with the constraints of 50 and 200 tests by changing the input criteria. Another method to develop OC curves available in Minitab is to compare user defined sampling plans. This option is available under the first dropdown arrow in the dialog box. This will change the input to look like Figure 5.

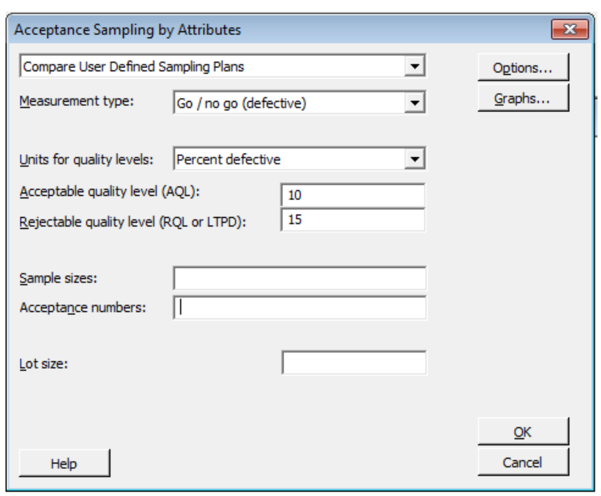

Figure 5: Compare User Defined Sampling Plans Input Screen

This tool is very useful because it allows you to plot and analyze several sampling plans at one time. You can enter multiple values for *n* in the sample size box and input the corresponding value for *c* in the acceptance numbers block. Going back to our previous example, if we really want to define the best plan possible for exactly 50 tests, we can enter 50 multiple times for *n* and vary *c*. These results are shown in Figure 6 and Table 2.

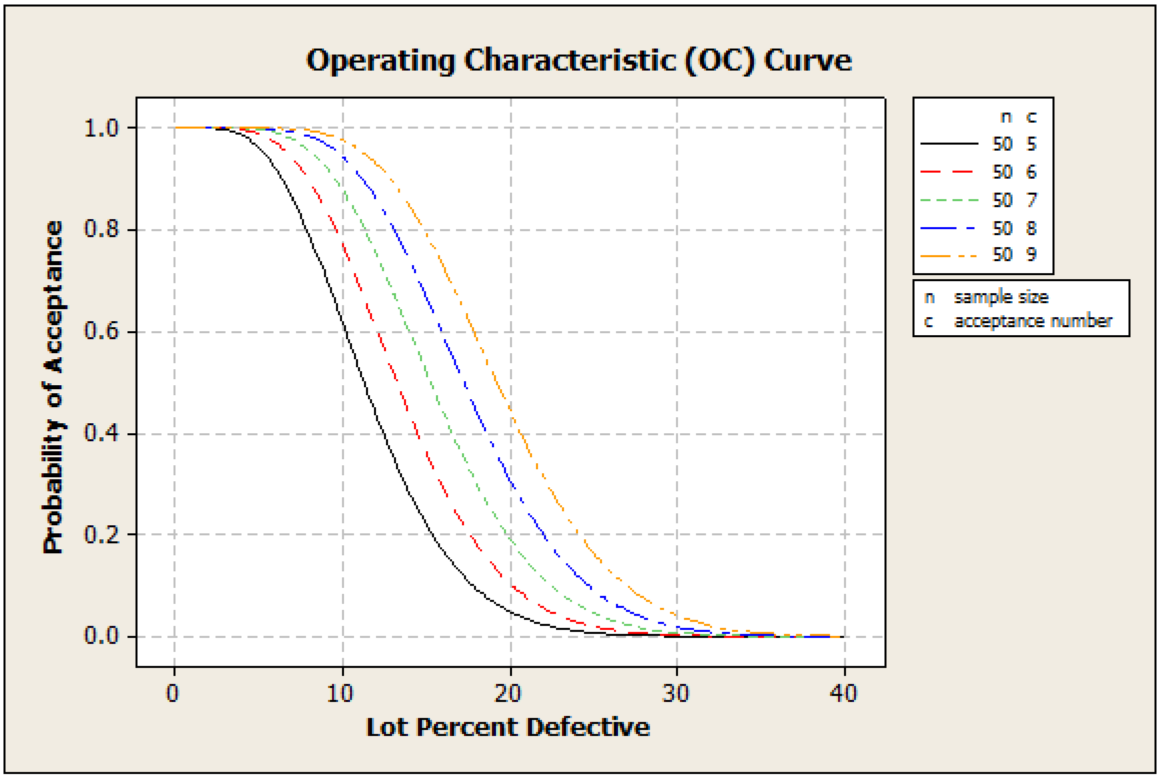

Figure 6: Compare User Defined Sampling Plans *n*=50

Acceptance Sampling by Attributes

Measurement type: Go/no go
Lot quality in percent defective
Use binomial distribution to calculate probability of acceptance

Acceptable Quality Level (AQL) 10
Rejectable Quality Level (RQL or LTPD) 20

Compare User Defined Plan(s)

Sample Size (n)	Acceptance Number (c)	Percent Defective	Probability Accepting	Probability Rejecting
50	5	10	0.616	0.384
50	5	20	0.048	0.952
50	6	10	0.770	0.230
50	6	20	0.103	0.897
50	7	10	0.878	0.122
50	7	20	0.190	0.810
50	8	10	0.942	0.058
50	8	20	0.307	0.693
50	9	10	0.975	0.025
50	9	20	0.444	0.556

Accept lot if defective items in n sampled <= c; Otherwise reject.

Table 2: Compare User Defined Sampling Plans *n*=50

For this example, let’s assume that we will keep the RQL at 0.80. For c=6 the consumers risk is only 0.103, but the producers risk is 0.23. For c=7, the consumer’s risk increases to 0.19, but the consumer’s risk in only 0.122. For this example of a fixed AQL, RQL and *n*, the biggest question is where to put the greater risk, on the consumer or producer.

For the case with 200 tests, the OC curves are shown in Figure 7. Both the green line for c=26 and the blue line for c=28 have a producer’s risk less than 0.1 and a very low consumers risk for an RQL of 0.8. Two possible conclusions from this graph are you could perform less testing or increase your RQL to approximately 0.83 or 0.84. You could get the exact values for alpha and beta by going back into the compare user defined sampling plans input screen and changing the RQL to 0.83 and 0.84 and rerunning the results. Again, there is no right answer and each case is different.

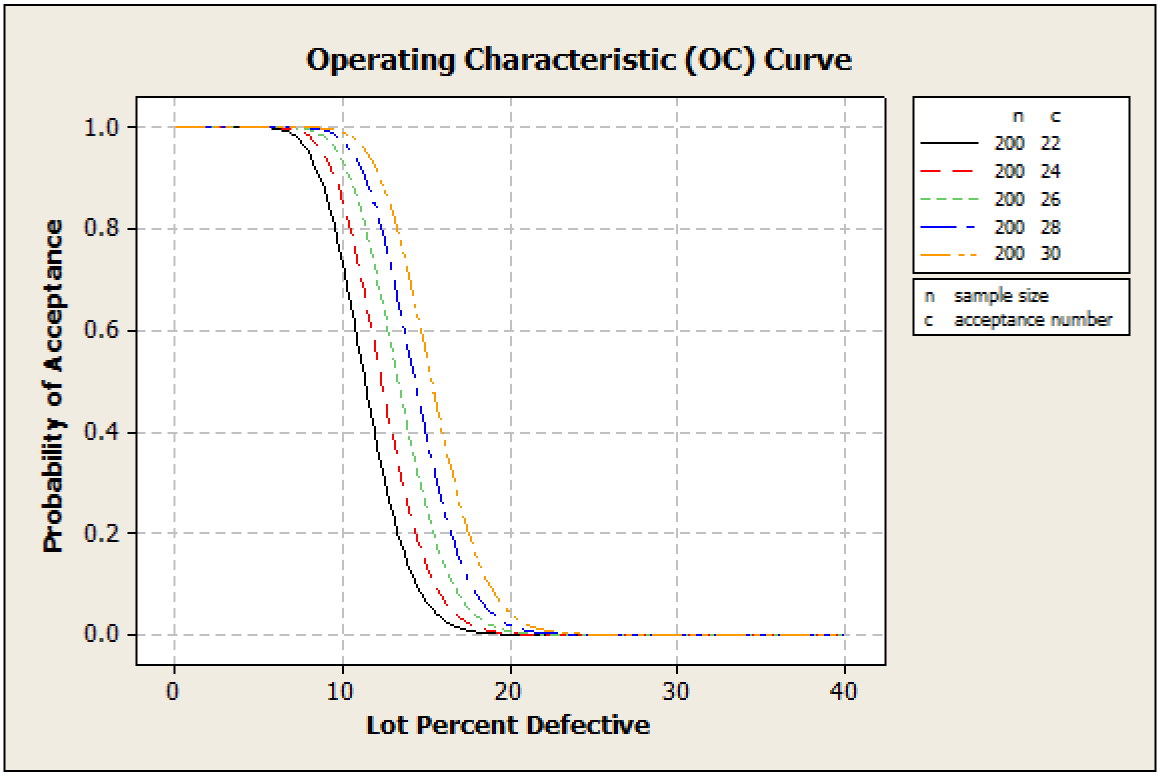

Figure 6: Compare User Defined Sampling Plans *n*=200

Applying OC Curves to DoD Requirements

Traditional DoD View of Requirements

In LASP, the lot is assumed to meet the AQL, and will only be rejected if there is evidence that it is bad. In this case the requirement is the AQL and the null hypothesis is that the proportion is >= AQL. Alpha is the chance of a type I error (saying the system is bad when it is not) and is therefore the producer's risk. Beta, or the chance of a type II error, is the chance of accepting a system that is below the RQL (saying the system is good when it is bad) and is therefore the consumer's risk.

When trying to verify requirements in DoD, the traditional approach has been to assume that the system does not meet the requirement unless there is evidence to reject this assumption. Also, there is often no consideration of a meaningful delta. OC curves are a good tool for graphically showing the issues with this approach.

Under the null hypothesis that the proportion is <= the requirement, the requirement is set to the RQL. Even when there is an Objective (O) and Threshold (T) value given, testing is primarily design to evaluate the Threshold value. Since there is only one value given, the AQL becomes the same as the RQL. Also since we are looking for data to reject the RQL, alpha, or the type I error rate becomes the consumer's risk, and beta becomes the producer's risk. This can cause a deal of confusion if the null and alternative hypotheses are not clearly stated. With no delta, the producer's risk + consumer's risk = 1. This means that if one risk is low, the other risk is high. This is demonstrated by the OC curve in Figure 7 for *n*=50 and *c*=1 and *c*=2.

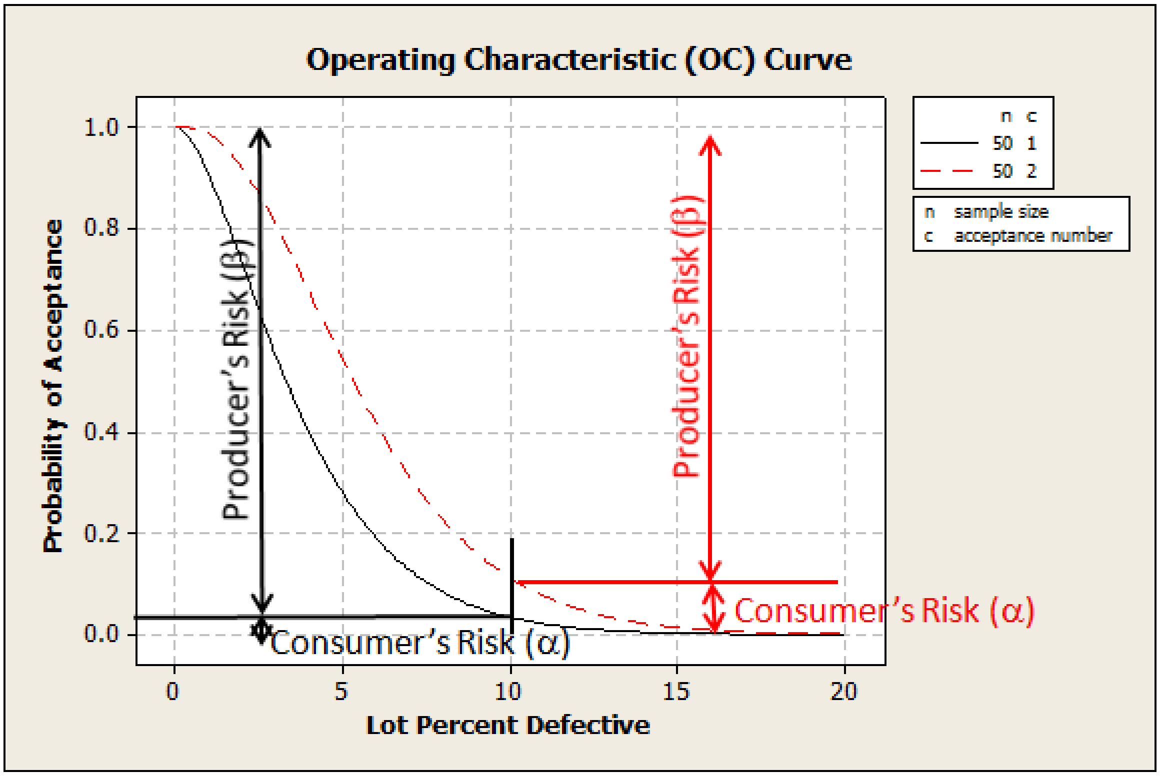

Figure 7: Compare User Defined Sampling Plans *n*=50

Another danger illustrated by this chart is not examining the effects of the criteria. If you require RQL=AQL, alpha=0.1 and *n*=50. You would get only one allowable failure or defect. For *c*=1 alpha=0.034 but for *c*=2, alpha= 0.112. In both cases beta is 1-alpha and is very high.

Another important conclusion from this chart is that even if the vendor delivers a system that exceeds the requirement, he would still have a low probability of it being accepted. If the vendor delivered a system actually only had 0.05 failures or defects, for *c*=2, the probability of acceptance is only 0.541 and for *c*=1, the probability of acceptance falls to 0.279. The sampling plan created without consideration of delta is a very different from the one in the previous section that allowed 6 or 7 failures or defects, and it has very serious impacts on the probability of good system being accepted.

Alternative View of Requirements

If we only consider one-level requirements (threshold=objective) it is not intuitive whether the assumption should be that the system will pass or the system will fail. Furthermore, being restricted to only one level, either the producer or consumer is going to take most of the risk. The lack of a defined, meaningful delta is prevalent in DoD programs and leads to a flawed process for balancing cost and risk. We should not think that a requirement is evaluated via a mechanical process that results in **THE** answer. Instead, when we examine a requirement, we should think in terms of AQL (what do I want to pass) and RQL (what do I want to fail) and to help construct appropriate testing requirements and that considers delta. This process also helps to clarify the issue of determining what the null hypothesis should be. Then we can examine various levels of producer's and consumer's risk and balance that against the total amount of testing required. Finally, all of the options can be presented to leadership in a clear manner that explains the costs versus the risks instead of simple reporting that we followed the process and produced the answer.

Summary and Conclusion

OC curves are useful in evaluating repetitive testing events when the response is expressed as a proportion. They allow the users to easily examine the effects of the input criteria and help to balance cost and risk. They are also helpful in understanding alpha, beta, delta, and sample size because the changes are displayed graphically and many different values can be plotted as once. For DoD testing, the most valuable lesson to learn from OC curves is the importance of defining an AQL (what do you want to pass) and a RQL (what do you want to fail). When this is not done, delta becomes zero, and there are serious impacts to the test design that leads to a flawed process for balancing cost and risk.