A Probabilistic Approach to Determining the Number of Widgets to Build in a Yield-Constrained Process

Timothy P. Anderson
The Aerospace Corporation

Introduction
Many cost estimating problems involve determining the number of Widgets to build in a yield-constrained manufacturing process, when it takes, on average, \( n \) attempts to produce \( m \) successes (\( m \leq n \)). Examples include computer chips, focal plane arrays, circuit boards, FPGAs, etc. The simplistic approach to this problem is to multiply the number of Widgets needed, \( m \), by the expected number of attempts needed to produce a single success, \( n \). For example, if a contractor reports that it takes, on average, 10 attempts to build one working Widget, and if four such Widgets are needed for a space-borne application, then the simplistic approach would be to plan for \( 4 \times 10 = 40 \) units, and estimate the cost accordingly. However, if the cost analyst uses the simplistic approach, he/she is likely to be disappointed, as the probability that 40 attempts will actually produce 4 working Widgets is only about 57%. Consequently, there is a 43% probability that 40 attempts will be insufficient. In fact, if the analyst wants to have, say, 80% confidence that four working Widgets will be available, then he/she should plan for 54 attempts! Obviously, this could have a huge impact on the cost estimate.

The purpose of this research paper is to describe the nature of the problem, to model the problem as a Negative Binomial random variable, and to develop the necessary thought process that one must go through in order to adequately determine the number of widgets to build given a desired level of confidence. This will be of great benefit to cost analysts who are in the position of estimating costs when certain hardware elements behave as described previously. The technique will also be very useful in cost uncertainty analysis, enabling the cost analyst to determine the appropriate probability distribution for the number of widgets needed to achieve success in their programs.

Example
Consider an example that was recently experienced by the author in which an independent cost estimate was being performed for a space-borne system that required four identical Widgets per platform. The unit cost of the Widgets was estimated using an engineering buildup method. For the purposes of this paper, we will assume the unit cost is $500K per Widget whether it meets specifications or not, and with little or no cost improvement for subsequent units (i.e., 100% learning curve). That is, the $500K cost includes the fabrication of the Widget as well as testing to ensure that the Widget meets specifications (i.e., it works).

It was known that there were yield issues in this Widget manufacturing process such that, on average, 10 Widgets had to be manufactured for every deliverable, fully functioning Widget. The question posed to the author at the time was, “how many Widget builds should be included in the independent cost estimate to ensure that a total of four Widgets could be
delivered?” The obvious, though simplistic, approach was to use the “expected value” approach, estimating the total number of Widgets by multiplying the number needed with the number required to achieve a success. That is, if four are needed, and the expected yield is 1 out of 10, then one would expect to have to build 4 x 10 = 40 individual units in order to achieve the desired outcome, with a resulting cost estimate of 40 x $500K = $20M.

However, upon further investigation, looking at the estimate from a cost-risk perspective, it became apparent that the “expected-value” approach carried a substantial amount of uncertainty. We could, for example, get lucky and meet our goal with only half that number. On the other hand, we could fall upon bad times, in which 40 units still would not be enough. So, the author looked at how to model this manufacturing situation from a probabilistic perspective. Considering the number of Widgets to build as a random variable, with a rate of one “success” for every 10 “trials,” the obvious choice was to model the number of successes as a Negative Binomial random variable. As a result, it was determined that the probability of obtaining four suitable Widgets from a batch of 40 attempts was about 57%. Moreover, if we wanted to have, say, 80% confidence that we could achieve 4 successes, then we needed to plan to build 54 units!

The remainder of this paper discusses the methodology used to arrive at this conclusion, describes the negative binomial distribution and how it applies to this problem, and uses the method to (1) determine an appropriate number of widgets to plan for, (2) model the number of widgets as a probability distribution for use in a cost risk analysis, and (3) describe the use of the approach in calculating the most probable cost (MPC) in a source selection activity.

The Negative Binomial Distribution

The Negative Binomial probability distribution is a discrete probability distribution that models the number of trials, \( n \), needed to achieve a specified number of successes, \( m \) \((m \leq n)\), under a key set of assumptions, and is built up from first principles as described below.

**First, The Bernoulli Random Variable**

Suppose a Widget manufacturing process is such that it takes, on average, \( n \) attempts (trials) in order to produce one satisfactory Widget (a success). One can then view each attempt as a Bernoulli random variable with probability of success, \( p \):

\[
\begin{align*}
p &= 1/n
\end{align*}
\]

provided the trials have these critical properties:

1. The result of each trial is classified as either a success, or a failure;
2. The probability, \( p \), of success is the same in every trial; and
3. The trials are independent – the outcome of one trial has no influence on later outcomes.

**Next, The Binomial Random Variable**

Discrete probability theory enables us to extend the concept of Bernoulli random variables to count the number of successes, \( m \), in \( n \) repeated Bernoulli trials, each with probability of
success, $p$. Since it is a function of Bernoulli trials, it follows that a binomial experiment is one that possesses the following properties:

1. The experiment consists of $n$ identical trials.
2. Each trial results in one of two outcomes — Success, or Failure
3. The probability of success on a single trial is $p$, and remains the same from trial to trial. Consequently, the probability of failure is $(1 - p) = q$.
4. The trials are independent.
5. The random variable of interest is $Y$, the number of successes observed during the $n$ trials.

The probability mass function (PMF) of the binomial distribution is given as follows. A random variable $Y$ is said to have a binomial distribution based on $n$ trials with success probability $p$ if and only if:

$$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad y = 0, 1, 2, \ldots, n, \quad \text{and} \quad 0 \leq p \leq 1$$

where

$$\binom{n}{y} = \frac{n!}{y! (n-y)!}$$

The mean and variance of the binomial distribution are listed below.

$$\mu = E(Y) = np$$
$$\sigma^2 = Var(Y) = np(1 - p)$$

Unfortunately, the binomial random variable does not address the problem at hand, since it counts the number of successes in $n$ trials, when what we desire is the number of trials needed to achieve $m$ successes. But, it represents the next step in the buildup to the negative binomial distribution.

**Then, the Geometric Random Variable**

The geometric distribution is defined for an experiment that is very similar to the binomial experiment. It is also concerned with identical and independent trials, each of which can result in either a success or a failure. The probability of success is $p$, and is constant from trial to trial. However, instead of counting the number of successes that occur in $n$ trials, the geometric random variable $Y$ represents the number of the individual trial on which the first success occurs. Thus, the experiment consists of a series of trials, and concludes with the first success.

The PMF of the geometric distribution is given as follows. A random variable $Y$ is said to have a geometric probability distribution if and only if:

$$P(Y = y) = (1 - p)^{y-1}p, \quad y = 1, 2, 3, \ldots, \quad 0 \leq p \leq 1$$

The mean and variance of the geometric distribution are listed below.

$$\mu = E(Y) = \frac{1}{p}$$
As before, the geometric random variable does not quite address the problem at hand, since it counts the number of trials needed to achieve the first success, when what we desire is the number of trials needed to achieve \( m \) successes. However, this is our last stop on the road to development of the negative binomial distribution.

**Finally, The Negative Binomial Random Variable**

At last, we can extend the concept of the geometric random variable to count the number of trials needed to achieve the \( r^{th} \) success \((r > 1)\).

The negative binomial distribution is defined for an experiment that is very similar to the geometric experiment. Again, it is concerned with identical and independent trials, each of which can result in either a success or a failure. The probability of success on any given trial is \( p \), and is constant from trial to trial. However, instead of considering the number of the trial on which the first success occurs, the negative binomial random variable \( Y_r \) is the number of the individual trial on which the \( r^{th} \) (second, third, fourth, etc.) success occurs. Thus, the experiment consists of a series of trials, and concludes with the \( r^{th} \) success.

The PMF of the negative binomial distribution is given as follows. A random variable \( Y_r \) is said to have a negative binomial probability distribution if and only if:

\[
P(Y_r = y) = \binom{y - 1}{r - 1} p^r (1 - p)^{y - r}, \quad y = r, r + 1, r + 2, \ldots, 0 \leq p \leq 1
\]

The mean and variance of the negative binomial distribution are listed below.

\[
\mu = E(Y_r) = \frac{r}{p}
\]
\[
\sigma^2 = Var(Y_r) = \frac{r(1 - p)}{p^2}
\]

And, the cumulative distribution function of \( Y_r \) is given as follows:

\[
P(Y_r \leq y) = \sum_{i=r}^{y} \binom{i - 1}{r - 1} p^r (1 - p)^{i - r}
\]

**Example:**

Using the previous example, we have a situation in which it takes 10 attempts to produce one functional Widget. Thus, \( p = 1/10 = 0.1 \), and is assumed to be independent from trial to trial. From the definition above, the expected value of the number of attempts needed to achieve four functional Widgets is

\[
\mu = E(Y_4) = \frac{4}{0.1} = 40
\]
What is the probability that we will achieve four functional Widgets in exactly four attempts?

\[ P(Y_4 = 4) = \binom{3}{3} (0.1)^4(0.9)^0 = 0.0001 \]

What is the probability that we will achieve four functional Widgets in exactly 10 attempts?

\[ P(Y_4 = 10) = \binom{9}{3} (0.1)^4(0.9)^6 = 0.0045 \]

What is the probability that we will achieve four functional Widgets in exactly 40 attempts?

\[ P(Y_4 = 40) = \binom{39}{3} (0.1)^4(0.9)^{36} = 0.0206 \]

Each of these probabilities is very small, so it probably doesn’t make sense to base a cost estimate on a specific number of attempts. That is, if we were to estimate the cost based on the expected value, 40, we would have a cost estimate of $20M with an associated probability of 0.0206. Why would anyone want to commit $20M on such a small likelihood of success?

Of course, in reality, the manufacturing process would stop once the 4th success was achieved. This could happen at almost any time. It could happen after the 4th attempt, though that is highly unlikely. Or it could happen after the 33rd attempt with substantially higher likelihood. So, a more relevant question is, “What is the probability that we will achieve success in \( n \) or fewer attempts?”

Example:

What is the probability that we will achieve four functional Widgets in \( no \ more \ than \) four attempts?

\[ P(Y_4 \leq 4) = \sum_{i=4}^{4} \binom{i-1}{3} (0.1)^4(0.9)^{i-4} = 0.0001 \]

Note: you may have noticed that this is a silly question! In this case, we would need to have four successes in a row, so the question is the same as \( P(Y_4 = 4) \), and, of course, the answer is the same too!

\[ P(Y_4 = 4) = \binom{3}{3} (0.1)^4(0.9)^0 = 0.0001 \]

What is the probability that we will achieve four functional Widgets in \( no \ more \ than \) 10 attempts?

\[ P(Y_4 \leq 10) = \sum_{i=4}^{10} \binom{i-1}{3} (0.1)^4(0.9)^{i-4} = 0.0128 \]

What is the probability that we will achieve four functional Widgets in \( no \ more \ than \) 40 attempts?
Now we see that if we plan for 40 attempts, then there is nearly a 58% chance that we will have our four successes sometime within those 40 trials. Unfortunately, we just don’t know on which trial we will be able to stop. At the same time, it is quite possible that 40 trials are not enough! We still have more than a 42% chance that it will take more than 40 trials in order to achieve four functional Widgets! What is a cost analyst to do?!

Determining the Appropriate Number of Widgets to Build

What the cost analyst would really like to know is, “On which trial will we achieve the \( r^{th} \) success?” But, as we have just shown via the negative binomial distribution, the number of the trial on which the \( r^{th} \) success will occur is a random variable, and cannot be known with certainty. However, there are a few good guesses that we can make. One way to do this would be to select the most likely trial number. Another would be to select the average trial number, and a third would be to select the trial number that corresponds to a given level of confidence.

The Most Likely Trial Number

Consider the PMF of the Negative Binomial distribution for our previous example shown below.

This is the PMF for values of \( y \) from 4 to 100.

Notice that the PMF is maximized at a value near 30, which is less than the expected value of 40. In fact it can be shown that the negative binomial distribution’s PMF is maximized at:
In this example, the PMF is maximized when \( y = 31 \).

\[
y_{mode} = \left[ 1 + \frac{r - 1}{p} \right]
\]

Therefore, using the most likely trial number approach, we can say that the most likely trial number on which the 4\(^{th}\) success will occur is trial number 31, and thus estimate the cost of building 31 Widgets. The drawback to this method, however, is that there is still a very low probability that it will take exactly 31 trials to achieve four successful Widgets, and the probability that we will achieve four successes at some point up to and including the 31\(^{st}\) trial is also significantly small.

\[
P(Y_4 = 31) = \binom{30}{3} (0.1)^4 (0.9)^{27} = 0.0236
\]

\[
P(Y_4 \leq 31) = \sum_{i=4}^{31} \binom{i-1}{3} (0.1)^4 (0.9)^{i-4} = 0.3762
\]

However, there is nearly a 38% chance that we will be successful within 31 trials, so it may be worth the risk. In this case, our cost estimate would be $500K \times 31 = $15.5M.

**The Average Trial Number**

One might choose to use this approach in order to be more confident that a sufficient number of trials will be performed in order to achieve four successful Widgets. We have been here before. The average trial number is simply the mean, or expected value, of the negative binomial random variable.

\[
\mu = E(Y_r) = \frac{r}{p}
\]

For our Widget example, this equates to 40 trials, as shown previously. Using this approach we can say that the weighted average number of trials in which we can achieve four successful Widgets is 40, and thus estimate the cost of building 40 Widgets. Again, however, there is still a very low probability that it will take exactly 40 trials to achieve four successful Widgets, and the probability that we will achieve four successes at some point up to and including the 40\(^{th}\) trial is also fairly small.

\[
P(Y_4 = 40) = \binom{39}{3} (0.1)^4 (0.9)^{36} = 0.0206
\]

\[
P(Y_4 \leq 40) = \sum_{i=4}^{40} \binom{i-1}{3} (0.1)^4 (0.9)^{i-4} = 0.5769
\]
However, there is nearly a 58% chance that we will be successful within 40 trials, and our cost estimate in this case would be $500K x 40 = $20M.

**The Trial Number That Corresponds to a Given Level of Confidence**

So far we have seen the *most likely* and *average* trial number approaches, each corresponding to levels of confidence that seemed to be too low for comfort. In the *most likely* trial number approach, the resulting level of confidence for 31 trials was 37.62%, and in the *average* trial number approach, the resulting level of confidence for 40 trials was 57.69%.

A third approach may be to specify a level of confidence in advance, and then determine the number of trials needed to achieve the $r^{th}$ success based on that confidence level. Examples might be 30%, 50%, 80%, and so on.

**Example**

Using our Widget example, suppose management is comfortable with an 80% confidence level that the number of Widgets built will deliver four successes. One can refer to the cumulative distribution function (CDF) below to determine the number of trials necessary.

Here we can see that the number of trials that corresponds to the 80$^{th}$ percentile is 54 trials. Another way to view this is to develop a table of probabilities such as that shown below.
Again we see that 54 trials will give us a level of confidence of 80%. However, this will require a lot more money in the budget, as $500K \times 54 = $27M. Perhaps a more reasonable level of confidence would be 50%, in which case we would need to plan for 37 trials, or $18.5M.

So far we have shown that it is impossible to determine, with precision, the exact number of trials needed to produce \( m \) successful widgets, and that the best we can do is to make an educated guess when estimating the number of trials necessary. Perhaps a more interesting proposition would be to use the uncertainty discussed above in a cost risk analysis. Recall that when performing a cost risk analysis, we are ultimately attempting to capture the estimating uncertainty in a cost probability distribution. As part of this process, we model the uncertainty of both the CERs and the CER input variables. Since quantity is an input variable at some point in the estimating process, then it stands to reason that quantity can be modeled as a random variable. Therefore, where the quantities of Widgets that are yield constrained are concerned, the quantity can be modeled as a negative binomial random variable.

### Number of Widgets as a Random Variable in a Cost Risk Analysis

As should be apparent by this point, it is difficult to arrive at a “best answer” when determining how many Widgets will be required. We have considered the “most likely” value, the “average” value, and the “number that corresponds to a given level of confidence.” These are useful methods if one is trying to plan a budget or develop a point cost estimate. But, when developing an independent cost estimate, it is far more interesting to understand the range of possible outcomes for use in a cost risk/uncertainty analysis.

For someone who performs cost risk analyses, it is preferable to model the “number to build” as a random variable rather than to use a deterministic value derived through any of the means described previously.

**Example**

Suppose a certain WBS element in a cost estimate is “1.2.4 Widgets.” Suppose also that the (fictional) cost estimating relationship (CER) for Widgets (whether they work or not!) is:
In other words, we have a CER that estimates the cost of an individual Widget as a function of its area in square inches. As is typical of CERs, this one also has uncertainty measured by the standard percent error (SPE). In keeping with our original example, we assume that there is no significant learning curve, so the cost of the $n^{th}$ Widget is the same as the cost of the first Widget. This is not necessarily a realistic assumption though.

In addition, suppose we know that it takes, on average, 10 attempts to build one functional Widget, and again, we need four functional Widgets. The typical cost risk approach is to model (a) the uncertainty of the independent variable (sq. in.), (b) the uncertainty of the CER, and (c) the uncertainty of the number of Widgets that need to be built in order to result in four functional Widgets. Let us suppose that the area is fixed, so there is no need to model that uncertainty. Then, using, say, @RISK, one would model the CER uncertainty as (usually) a lognormal distribution, and the number required, $M$, as a negative binomial distribution.

$$
\text{Widget Cost} \sim \text{Lognormal} \left( \mu = 57.17 + 0.3(Area\text{ (sq. in.)})^{1.82}, \sigma = 22\% \right)
$$

$$
M \sim \text{Neg. Bin.} (r = 4, p = 0.1)
$$

If the area is 55.12 square inches, then the spreadsheet implementation would look like the following:
The resulting point estimate is $20.0M (FY11) for 40 attempts at building the Widgets. But now take a look at the probability distribution that results from running a Monte Carlo simulation.

With this result we see the combined probability distribution in which both the numbers of attempts necessary, and the cost of each attempt, are modeled as random variables. This distribution represents a convolution of the negative binomial quantity and the lognormal CER, and has the typical long right tail that we would expect.

By examining this distribution, we see that there is a 58.4% chance that the cost to build four functioning Widgets will not exceed $20.0M. But we also see that there is substantial likelihood of costing more than that, with even $50M not being out of the realm of possibility. However, using this approach we can also see the range of possible costs. More interesting yet, we can see below that there is an 80% chance that the cost to build four functioning Widgets will not exceed $27.6M.
Recall that when we assumed a fixed cost of $500K per Widget, the 80% confidence level result was to build 54 Widgets at a cost of $27.0M. The cost distribution gives a reasonably consistent answer, but does not require us to worry about “how many to build.” This distribution includes the combined effects of uncertainty in the cost of each Widget and the uncertainty in the number to build. If we simply budget $27.6M, we have an 80% chance that we will meet our goal.

**Modeling Uncertainty in a Source Selection Evaluation**

Finally, we consider the use of the negative binomial distribution as one means of determining the most probable cost (MPC) in a government source selection. Here, one must be careful. In an industry cost proposal, especially when two or more competitors exist, an offeror’s proposed cost is likely to be lower than it should be. The reason for this is that, in a competition, offerors are motivated to bid the lowest price possible while still being plausibly reasonable and realistic. The usual result is that the ultimate winner of the competition has bid a price that is too low, and must rely on inevitable subsequent engineering changes in order to recoup the “real” cost of the contract. Therefore, in a government source selection, the government computes an MPC that theoretically represents the “real” cost of the contract, so that it will be prepared for future cost increases.

Continuing on with the earlier example, suppose an offeror in a competitive source selection bids a price in which it is assumed that four functional Widgets can be produced in a minimum amount of trials. It is unlikely that the offeror would be bold enough to assume it will take only four trials, with success on each trial. But it would not be out of the realm of possibility for the offeror to bid, say, the 20th percentile solution. Furthermore, suppose the offeror is bidding an optimistic price for the Widgets, for example, $400K. This would correspond to 24 trials, yielding a cost estimate of $24 × $400K = $9.6M for the four Widgets. The offeror would, of course, come up with some plausible sounding justification for being able to produce the four Widgets in only 24 trials, and for the discounted price of a build. But, using the assumptions in our example, there is only about a one in five chance that the offeror would be successful given 24 attempts.
Now look at this from the government cost evaluator’s perspective. For the MPC, the cost evaluator should refer to the techniques described previously in determining the appropriate number of Widgets that will need to be built. In addition, the cost evaluator would tend to use the more realistic price of $500K per build. Moreover, to be consistent, since the MPC is the “most probable cost,” the cost evaluator should use the “Most Likely Trial Number” method.

**Example**

The offeror has bid a price that corresponds to 24 builds at a price of $400K per build, for a total price of $9.6M. Now, for the purposes of the MPC, the government cost evaluator assumes the more realistic price of $500K per build, and assumes the “most likely trial number” is

\[ y_{\text{mode}} = \left\lfloor 1 + \frac{r - 1}{p} \right\rfloor = \left\lfloor 1 + \frac{4 - 1}{0.1} \right\rfloor = 31 \]

The resulting “most probable cost” for the Widgets is 31 × $500K = $15.5M. As shown previously, however, the resulting probability that 31 trials will be enough is just a little over 37%. Of course, calculation of an MPC is not actually constrained to be the “most probable” cost, but can actually be defined as needed. Therefore, for a given source selection activity, the source selection team can choose to use one of the other “more probable” methods, such as the “average trial number” method or the “trial number that corresponds to a given level of confidence” method.

**Summary and Conclusions**

This paper has described the nature of the problem of determining the number of Widgets to build in a yield-constrained manufacturing process. The number of Widgets to build was modeled as a negative binomial random variable. The example of a Widget manufacturing process was used to illustrate how to determine the appropriate number of Widgets to plan for under various assumptions. We have concluded that there is no “best answer” to the question, but we were able to develop the methodology needed to arrive at alternative answers under the assumptions of the “most likely” number, the “average” number, and the number that corresponds to a “given level of confidence.”

Then, we discussed how a cost analyst could use the methods described to model the number of Widgets as a random variable in a cost risk analysis, and we also discussed how to apply the methods in a government source selection evaluation – with the goal of determining the government’s most probable cost position.

This paper will be of use to cost analysts who are in the position of estimating costs when certain hardware elements are yield-constrained, and to those who use cost uncertainty analysis techniques. In addition, the paper will be useful to source selection cost evaluators in support of MPC development.

**Potential Future Research**

We have avoided the question of what to do if the manufacturing process “improves” over time, meaning that the probability of success increases with increased numbers of trials.
Future research on this topic should include a study of how to determine the number of Widgets to build if the success probability is not constant.

**Disclaimer**

All trademarks, service marks, and trade names are the property of their respective owners.