Center for Quality and Productivity Improvement
University of Wisconsin
610 Walnut Street
Madison, Wisconsin 53705

(608) 263-2520
(608) 263-1425 FAX
quality@engr.wisc.edu

Report No. 152

2^{k-q} Experiments with Binary Responses: Sampling Until a Fixed Number of Defectives is Observed

Søren Bisgaard and Ilya Gertsbakh

August 1997

CQPI
University of Wisconsin

Report No. 152

Report Series
in Quality and Productivity

2^{k-q} Experiments with Binary Responses: Sampling Until a Fixed Number of Defectives is Observed

Søren Bisgaard

Center for Quality and
Productivity Improvement and
Department of Industrial Engineering
University of Wisconsin-Madison

Ilya Gertsbakh

Ben Gurion University, Beer-Sheva, Israel

ABSTRACT

In this article we will discuss the use of two-level factorial and fractional factorial experiments with binary responses (defectives/non-defectives) where the purpose is to reduce the rate of defectives. Contrary to a traditional fixed sample size scheme, we will consider one where each factorial combination is sampled until a fixed number of defectives is observed. The total number of items until that occurs is then used as the response. Such an Inverse Binomial sampling scheme has several practical and economical benefits that will be discussed. For the design of experiments based on this idea, we provide a methodology for choosing the necessary number of defectives r, to detect a given change in the probability of producing a defective unit with fixed levels of Type I and Type II errors.

KEYWORDS: Inverse Binomial Sampling, Quality Improvement Experiments, Factorial Experiments, Sample size estimation.

2^{k-q} Experiments With Binary Responses: Sampling Until A Fixed Number Of Defectives Is Observed

Søren Bisgaard and Ilya Gertsbakh

In this article we will discuss the use of two-level factorial and fractional factorial experiments with binary responses (defectives/non-defectives) where the purpose is to reduce the rate of defectives. Contrary to a traditional fixed sample size scheme, we will consider one where each factorial combination is sampled until a fixed number of defectives is observed. The total number of items until that occurs is then used as the response. Such an Inverse Binomial sampling scheme has several practical and economical benefits that will be discussed. For the design of experiments based on this idea, we provide a methodology for choosing the necessary number of defectives r, to detect a given change in the probability of producing a defective unit with fixed levels of Type I and Type II errors.

Introduction

Customers of industrial products increasingly expect to receive mass produced items with guaranteed levels of defectives in a range below 100 parts per million (ppm). One example where such stringent requirements are imposed is in the production of cathode-ray tubes (CRT's) for televisions or computer displays. Achieving such low levels of defectives is usually a technological challenge of extraordinary proportions. It frequently requires close collaboration between research, production and the quality engineering department. A potent tool in such efforts is the use of factorial experiments.

In this article we will consider the situation where the only reliable response is a binary defective or non-defective. This is often the case for the production of CRT's, micro chips, metal castings, plastic moldings and a vast number of other products. For such production processes, low levels of defectives are traditionally achieved with control charts by controlling key process parameters within narrow limits. The hope is that if these input parameters are held at consistent levels, it will guarantee consistency in the quality of the output product.

The success of such an approach is based on at least two assumptions. One is that the right set of parameters is being monitored and controlled. A second assumption is that the nominal settings of the process parameters are targeted at their optimal (or satisfactory) levels so that defectives are produced only because of variability in these process parameters around the nominal values. That both of these assumptions are satisfied, or nearly so, is seldom if ever known in practice. In fact, it would appear safer to assume that they are not. Consequently it ought to be standard practice to experiment with processes to screen out the most important factors influencing the rate of defectives, and when those have been identified, engage in optimization experiments to determine the best, or at least a set of more satisfactory, parameter settings to be used as target values for subsequent process control.

Experimenting as we are suggesting with only a binary response $x_i = (0,1)$ = (non-defective, defective) is inherently difficult. A binary response has a very low information content. Nevertheless, it is fairly common in industry that factorial experiments with binary responses are set up such that for each of the N factorial trials, a set of n products is run off and the proportion defectives $\hat{\theta}_j = n^{-1}\sum_{i=1}^{n} y_{ij}$ is used as the response in the subsequent analysis. Experiments of this kind have lately been discussed by Bisgaard and Fuller (1995a) who provided simple tables for determining an appropriate *fixed sample size n* for each factorial combination in two-level factorial designs. In another article focusing on the analysis of such experiments, Bisgaard and Fuller (1995b) further discussed the use of variance stabilizing transformations. Since this fixed sample size approach provides important clues to how an inverse binomial sampling scheme can be constructed, we will briefly review this material in the next section.

The Fixed Sample Size Approach

Following Bisgaard and Fuller (1995a), suppose a $2^{k-q} = N$ fractional factorial experiment is set up where k is the number of factors, q the degree of fractionation and N the number of factorial combinations tested. Further let the probability of producing a defective unit at the factorial combination $\mathbf{x}_j$ be approximated locally by

$$\theta_j = \theta_0 + b_A x_A^{(j)} + \ldots + b_G x_G^{(j)} + b_{AB} x_A^{(j)} x_B^{(j)} + \ldots \qquad (1)$$

where $\mathbf{x}_j = (x_A^{(j)}, \ldots, x_G^{(j)})$ is the row vector of ± 1's of the design matrix corresponding to the *jth* factorial combination, θ_o is the mean probability of producing a defective, and $b_A, \ldots, b_G, b_{AB} \ldots$ are the coefficients, or half of the factorial effects, reflecting the influence of the corresponding factors.

Suppose a certain factor, say A, is *active* and that it produces a change Δ in θ_j when $x_A^{(j)}$ is changed from -1 to $+1$. When testing whether A is statistically significant based on data from an experiment, there is always the possibility that we do not detect that A is significant. Thus we commit a Type II error. Similarly there is a possibility of detecting an effect when none is present. In that case we would commit a Type I error. As a basis for determining the sample size, n, for each of the N factorial combinations, we choose two numbers α, the probability of committing a Type I error, and β the probability of committing a Type II error when the true effect is Δ.

Before we proceed let us note that Bisgaard and Fuller's (1995a) derivation, as well as ours shown below, are based on using a variance stabilizing transformation for the response. An alternative would be to use a Generalized Linear Model (GLIM), see e.g. McCullagh and Nelder (1989), a common approach to the analysis of binary data, that uses transformations of the data that are not variance stabilizing. However, the proofs based on standard least squares theory presented below for calculating the stopping rule depends on several intermediate steps, in particular the Normal approximations, on the constant variance property of the response. The final result pertaining to the stopping rule we obtain will, however, apply to Inverse Binomial sampling in general regardless of how the data will be analyzed. Hence the user should feel free to use GLIM or any other appropriate method in the subsequent analysis.

Bisgaard and Fuller (1995a) in their proof argued that when using a variance stabilizing transformation, the estimated effects will be more or less Normally distributed with constant variance as long as the factor changes only create moderate changes in the defective probability. Suppose the change Δ of the defective probability on the transformed scale is denoted by δ. Applying Normal distribution theory, they then showed that the critical value c could be chosen by balancing the risks α and β of making the two decision errors. Specifically they required that if $\delta > 0$ the critical value, c , should satisfy the two criteria

$$\begin{array}{ll} (i) & c = \sigma z_{1-\alpha/2} \\ (ii) & c = \delta - \sigma z_{1-\beta} \end{array} \qquad (2)$$

where z_ϕ is the standardized Normal deviate corresponding to the ϕ's quantile of the unit Normal distribution and σ the standard error of the estimated effect. (An equivalent requirement is imposed if $\delta > 0$.) Equating (*i*) and (*ii*) in (2) they showed that

$$\sigma = \frac{\delta}{z_{1-\alpha/2} + z_{1-\beta}}. \qquad (3)$$

In the Binomial case by substituting the appropriate relations into (3), they then provided an estimate of the necessary sample size as

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{N\delta^2}. \qquad (4)$$

Based on (4), Bisgaard and Fuller (1995a) provided tables for finding the minimal number n to guarantee the prescribed a and b values for a given design size N, the hypothetical mean response probability q_o and the significant level D of an active factor. For example, suppose $q_o = 0.1$, $D = 0.05$, $N = 8$, $a = 5\%$, $b = 10\%$ then their Table 1a suggests a sample of size $n = 186$ for each of the eight experimental trials.

The reason Bisgaard and Fuller (1995a) suggested using a variance stabilizing transformation in their derivation is that the natural estimator of θ_j is $\hat{\theta} = \sum_{i=1}^{n} y_{ij} / n$. Its variance is $Var\{\hat{\theta}_j\} = \theta_j(1-\theta_j)/n$ which depends on θ_j, the very parameter we hope to change in the experiment. That creates a dilemma! If we are successful in our experimental effort we will change θ_j. Hence we will violate the least squares assumption of constant variance of the response for all the trials. If we instead use a variance stabilizing transformation $\varphi(\hat{\theta}_j)$, in this case for *binomial* sampling $\varphi(\hat{\theta}_j) = \arcsin\sqrt{\hat{\theta}_j}$, then the variance will be approximately the same for each trial even when we are successful in changing the defect rate θ_j.

Critique of the Fixed Sample Size Approach

Although the fixed sample size approach is widely used, and works well in many cases, especially when the defect probabilities are relatively large, it has several essential drawbacks. One is that sample sizes and thus the duration

of the experiment might be prohibitively large for small defect probabilities. This issue was especially made transparent by the tables produced by Bisgaard and Fuller (1995a). Unfortunately as we will indicate below, there is not much we can do about that. Another drawback, not completely unrelated to the first, is that the experimenter has little control over how many defectives are produced during the experiment (except for the upper limit nN). Because of the large sample sizes required when experimenting with processes that have relatively small defect rates, this may lead to the production of a large number of defectives if, for example, a certain factor combination causes an undesirable large increase in the value of θ_j.

It is for this latter reason that we will suggest a different sampling strategy. At each design point, rather than fixing the sample size, we suggest continued sampling until a prescribed number of defectives r have accumulated. Hence we measure the response as the total number of products y produced until reaching r defectives. This implies that we leave the duration of each trial "loose", but fix the number of defectives that are produced during the trials.

Inverse Binomial sampling as this is called, has several desirable properties. Suppose, for example, that a particular factor combination increases the defective probability θ_j to an undesirable level. Rather than producing a large number of defectives for this combination as in the fixed sample size approach, we will instead quickly terminate that run before producing too much scrap, and move on to explore hopefully better factor combinations. On the other hand, if a factor combination produces more desirable conditions, then the production at that combination will continue for a long while before r defectives are produced. However, that should usually not be a problem. All along we will be producing good products. Consequently the experimenter can assure production management, often leery about having any experiments conducted on ongoing production processes, that little time and material is wasted. Good products are being made during the experiment. Thus only a minimal time will be spent on unfavorable scrap producing conditions, and we can guarantee that only a fixed predetermined number (Nr) of defectives, usually not much different from what would be produced if the process was left alone, will be produced.

The Inverse Binomial Sampling Scheme

We will now more formally discuss how we can set up an Inverse Binomial sampling scheme as described above. Of primary importance from a design of experiments standpoint is how to determine r, the number of defectives to be observed for each of the N trials in a 2^{k-q} design. As in the Binomial sampling scheme, this will be determined based on the number of factorial trials in the experiment N, a previous estimate of the defective percentage at the center of the design θ_o, the minimum change in the defective rate Δ the experimenter wants to detect, and the probability of Type I and II errors α and β.

Before we proceed with the formal test criteria, we first need to consider the appropriate estimator for θ, the probability of producing a defective unit. If the individual defectives are independent Bernoulli random variables with probability θ of producing a defective, and a sample of size Y is needed to observe exactly r defectives then a uniform minimum variance unbiased (UMVU) estimator of θ is given by

$$\hat{\theta} = \frac{r-1}{Y-1}. \tag{5}$$

for $r \geq 2$ (see e.g. Kotz and Johnson, 1983 and our Appendix 1). Note incidentally that the maximum likelihood estimator of θ is r/Y and thus differs from (5).

When the true θ lies in the range [0, 0.7], a remarkably "stable" variance stabilizing transformation developed by Anscombe (1948) is

$$\begin{aligned} \varphi(\hat{\theta}) &= \sinh^{-1}\sqrt{\frac{Y-r+3/8}{r-3/4}} \\ &= \sinh^{-1}\sqrt{\frac{\hat{\theta}^{-1}(r-1)+11/8-r}{r-3/4}} \end{aligned} \tag{6}$$

Note that a larger Y implies a smaller $\hat{\theta}$. We prefer a transformation $\varphi(\hat{\theta})$ that is a strictly increasing function of $\hat{\theta}$. Thus we will switch the sign on the transformation and express (6) in terms of $\hat{\theta}$ as

$$\begin{aligned} \varphi(\hat{\theta}) &= -\sinh^{-1}\sqrt{\frac{Y-r+3/8}{r-3/4}} \\ &= -\log\left[\frac{\sqrt{(r-1)/\hat{\theta}-r+1.375}+\sqrt{(r-1)/\hat{\theta}+0.625}}{\sqrt{r-0.75}}\right] \end{aligned} \tag{7}$$

Choosing the Stopping Rule r

For the determination of the appropriate stopping rule, r, a central role is played by some results proven by Bisgaard and Fuller (1995a, pp. 346-347) which we will provide here in a somewhat generalized form. Let $\varphi(\hat{\theta})$ be a variance stabilizing transformation for $\hat{\theta}$ and assume that the function $\varphi(\theta)$ is differentiable with respect to θ. Further suppose a factor, say A, is changed from its low to its high level, and thereby induces a change from $-\Delta/2$ to $\Delta/2$ of

r	$b(r)$	r	$b(r)$	r	$b(r)$
2	1.315	22	0.222	60	0.131
3	0.853	24	0.212	62	0.129
4	0.659	26	0.203	64	0.126
5	0.553	28	0.193	66	0.124
6	0.485	30	0.189	68	0.123
7	0.436	32	0.182	72	0.119
8	0.401	34	0.176	76	0.116
9	0.372	36	0.171	80	0.113
10	0.349	38	0.167	84	0.110
11	0.329	40	0.161	88	0.108
12	0.313	42	0.158	92	0.105
13	0.298	44	0.154	96	0.103
14	0.286	46	0.151	100	0.101
15	0.275	48	0.148	110	0.096
16	0.265	50	0.144	120	0.092
17	0.256	52	0.141	130	0.088
18	0.249	54	0.138	140	0.085
19	0.241	56	0.136	150	0.082
20	0.234	58	0.133	160	0.080

Table 1. *Values of the stopping rule r and the corresponding factors b(r).*

the response $\hat{\theta}$ around its mean value θ_o. The effect of A is then Δ on this scale. However, on the transformed scale this same change, denoted by δ, is

$$\delta = \varphi(\theta_0 + \Delta/2) - \varphi(\theta_0 - \Delta/2). \tag{8}$$

Expanding in a first order Taylor series we find that

$$\begin{aligned}\delta &= \varphi(\theta_0 + \Delta/2) - \varphi(\theta_0 - \Delta/2) \\ &\cong \varphi(\theta_0) + (\Delta/2)\varphi'(\theta_0) - [\varphi(\theta_0) + (-\Delta/2)\varphi'(\theta_0)] \\ &= \Delta \cdot \varphi'(\theta_0).\end{aligned} \tag{9}$$

Bisgaard and Fuller (1995) then essentially proved the following proposition.

Proposition 1: In an N run two-level factorial design, if the true effect of a factor measured on the original scale is Δ, then the estimate of this effect on the transformed scale, $Effect_\varphi$, is approximately Normally distributed with $E\{Effect_\varphi\} = \delta \cong \Delta \cdot \varphi'(\theta_0)$ and

$$Var\{Effect_\varphi\} \cong 4[\varphi'(\theta_0)]^2 V\{\hat{\theta}\}/N. \tag{10}$$

Further, if $\Delta = 0$ then the estimated effect on the transformed scale is approximately Normally distributed with $E\{Effect_\varphi\} = 0$ and the same variance as in (10).

As indicated above, the proof of this proposition is essentially provided by Bisgaard and Fuller (1995a). They showed the result for the specific case of the $\varphi(\hat{\theta}) = \arcsin(\sqrt{\hat{\theta}})$ transformation appropriate in the fixed sample size Binomial situation. However, it is relatively straightforward to see that the entire proof will hold for any differentiable transformation. We will therefore not repeat the somewhat lengthy proof.

Proposition 2. In an N run two-level factorial design, if the true effect of a factor measured on the original scale is Δ, then the sampling plan with α and β can be determined as a solution to

$$V\{\hat{\theta}\} = \frac{\Delta^2 N}{4(z_{1-\alpha/2} + z_{1-\beta})^2} \tag{11}$$

Proof: Combining (10), (9) and (3) we have

$$4[\varphi'(\theta_0)]^2 V\{\hat{\theta}\}/N = \frac{\Delta^2[\varphi'(\theta_0)]^2}{(z_{1-\alpha/2}+z_{1-\beta})^2}.$$

This expression gives the desired result after some simple reductions.

Notice that (11) is a somewhat surprising result. The particular transformation used does not play any role since $\varphi'(\theta)$ cancels out. Thus all we need to know is the variance of the estimator before the transformation.

If we use the estimator given by (5), we show in Appendix 1 that bounds for $V\{\hat{\theta}\} = \sigma_{\hat{\theta}}^2$ suggests that a good approximation is

$$\sigma_{\hat{\theta}} = b(r)\theta\sqrt{1-\theta} \qquad (12)$$

where $b(r)$ is a constant that depends on the particular value r.

In deriving the approximation (12) we may choose $b(r)$ such that we have a perfect fit at $\theta_o = 0.1$ which would appear to be in the middle of the most important range of θ for practical applications. This approximation turns out to be very accurate for $\theta_0 \in [0.02, 0.4]$ for $r \geq 10$. Table 1 provides a summary of the constants needed to determine r.

With this approximation we can now insert (12) into the general expression (11) which then gives

$$b(r) = \frac{\Delta\sqrt{N}}{2(z_{1-\alpha/2}+z_{1-\beta})\theta_0\sqrt{1-\theta_0}}, \qquad (13)$$

Since everything on the right hand side of (13) can be determined by the experimenter prior to experiment, we can compute the value $b(r)$ needed to satisfy. Next we can use Table 1 to find the particular r that produces a $b(r)$ close to the number produced by (13). That particular r is then the stopping rule we are seeking.

In some cases we might need to determine values outside the range of Table 1 and for $0 < \theta_0 \leq 0.02$. To do so, notice from (A1.4) that for small θ_0 and large r, $V\{\hat{\theta}\} \cong \theta^2(1-\theta)/(r-2)$. Now inserting this expression into (11) we get

$$r \approx \frac{4(z_{1-\alpha/2}+z_{1-\beta})^2\theta_0^2(1-\theta_0)}{\Delta^2 N} + 2 \qquad (14)$$

Thus (14), an explicit formula for the stopping rule, can be used for small defect rates where $r > 160$. Some examples presented in the next section will illustrate our approach.

Examples of the Determination of the Stopping Rule

In this section we will present four examples of how to determine the stopping rule r for some practical cases.

Example 1: Suppose we want to run an 8-run two-level factorial experiment to reduce the defective probability θ and that the underlying model is assumed to be of the form (1) with θ_o=0.1, estimated from control chart studies prior to the experiment. Further, suppose we want to be able to detect factor changes of $\Delta = 0.05$ and use the conventional $\alpha = 5\%$ and $\beta = 10\%$. To determine the stopping rule r, we proceed as follows:

Step 1: Determine $b(r)$ from equation (13). For this example we get

$$b(r) = \frac{0.05\sqrt{8}}{2(1.96+1.28)\cdot 0.1\sqrt{1-0.1}} = 0.230$$

Step 2: Enter Table 1 with the above $b(r)$ and find the closest corresponding r. The closest value to 0.230 is $b(20)$ =0.234. Thus $r = 20$ so for each trial we have to sample until we observe 20 defectives.

To further illustrate this approach, we have simulated data for an eight run 2^{7-4} design with a base probability of defectives $\theta_o = 0.1$ at the design center point, two active factors B and C with B=C=$\Delta/2$=0.025 and r=20. The design and the simulated transformed responses (multiplied by 1,000) using the modified Anscombe's transformation (7) are all given in Table 2.

The seven effects, calculated the usual way but based on the transformed data using either (6) or (7) are shown as a Normal Plot in Figure 1. From this figure we see that B and C are clearly identified as being significant.

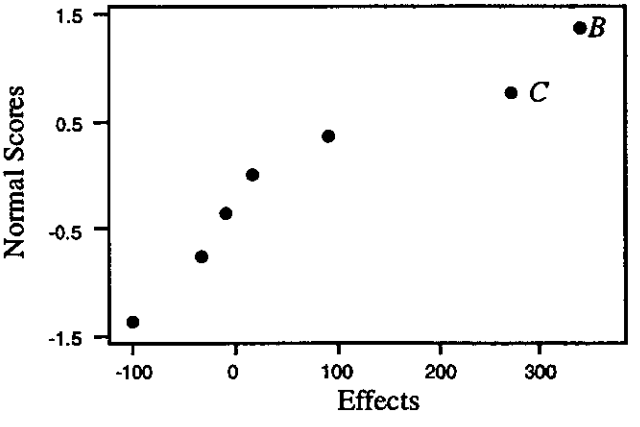

Figure 1. Normal Plot of effects for data in example 1.

Run	A	B	C	D	E	F	G	θ_i	Y	$1000\times\varphi(\hat{\theta})$
1	−	−	−	+	+	+	−	0.05	264	−1980
2	+	−	−	−	−	+	+	0.05	402	−2200
3	−	+	−	−	+	−	+	0.10	171	−1750
4	+	+	−	+	−	−	−	0.10	179	−1780
5	−	−	+	+	−	−	+	0.10	235	−1920
6	+	−	+	−	+	−	−	0.10	168	−1750
7	−	+	+	−	−	+	−	0.15	107	−1450
8	+	+	+	+	+	+	+	0.15	106	−1500

Table 2. *A 2^{7-4} design with simulated data.*

It is interesting to note that this Inverse Binomial sampling scheme on average will require

$$M=\sum_{i=1}^{N}\frac{r}{\theta_i}\approx 1,867$$

observations, and would produce exactly $rN=20\times 8=160$ defectives. Let us compare this with the standard Binomial sampling scheme. From Table A1 of Bisgaard and Fuller (1995a) we find that the sample size under the same circumstances should be $n=186$. Thus, the total number of observations will be $n\cdot N=1488$, while the expected number of defectives will be $nN\theta_0\approx 149$, with a standard deviation of

$$\sigma^*=\sqrt{n}\sqrt{\textstyle\sum_{i=1}^{N}\theta_i(1-\theta_i)}=11.5$$

Thus comparing the two approaches there does not appear to be a gain over the standard Binomial sampling. However, taken relative to the expected sample size the defective rate for the Inverse Binomial Sampling is 160/1867=0.0857 as compared to 149/1488 = 0.1001 for the Binomial sampling scheme.

That the plans are very similar under "normal conditions" when we compare them using the expected number of observations should not be surprising. They are essentially based on the same binary information and have essentially identical likelihood functions. Thus by the likelihood principle, see e.g. Cox and Hinkley (1974, pp. 39-40) they provide the same inference. However, as we have pointed out before, the primary benefit of the Inverse Binomial plan is the protection it affords against producing large numbers of scrap, and the quick switching away from "bad" conditions.

We will now illustrate this property. Suppose one of the eight conditions in the factorial design results in a much higher defective rate, say $\theta_8=0.5$ and that the remaining seven conditions deviate only very little from the average of $\theta_0=0.1$. In that case the number of defectives will under the Inverse Binomial Sampling scheme continue to be $rN=20\times 8=160$, but for the standard fixed sample Binomial, the expected number of defectives would be $\sum_{i=1}^{8}n\theta_i=168(0.5+7\cdot 0.1)\approx 202$. Now let us put these numbers in perspective. If we had left the process alone for a period similar to what it would take to produce $8\times 186=1488$ products as prescribed in the Binomial sampling plan, we would, with a defective rate of $\theta_0=0.1$, produce an average of $1488\times 0.1\approx 149$ defectives. Thus what we gain by using an Inverse Binomial sampling plan is a protection against large number of defectives during the experiment if one or more factorial conditions turn out to be "bad". The cap on the number of defectives this affords, which is roughly about the same size as what would have been produced anyway if we left the process alone, is the main reason why we think the Inverse Binomial scheme in some cases might be a useful alternative to Binomial sampling.

Example 2. Suppose $\alpha=5\%$, $\beta=10\%$, $N=16$, $\theta_0=0.3$ and $\Delta=0.1$. Let us determine the stopping rule r and compare the inverse Binomial sampling with the fixed sample size Binomial sampling. To find the stopping rule, we proceed again as follows.

Step 1. Find $b(r)$ from (13):

$$b(r)=\frac{0.1\sqrt{16}}{2\cdot 0.3\sqrt{0.7\cdot 3.24}}=0.246.$$

Step 2. Enter Table 1 and find an r that produces a $b(r)$ close to 0.246. In this case $r=18$.

For comparison with a similar Binomial plan, suppose only one factor is active at $\Delta=0.1$. The expected sample size is then for the Inverse Binomial plan

$$M = \sum_{i=1}^{N} \frac{r}{\theta_i} = \frac{N}{2}\frac{r}{\theta_0 - \Delta/2} + \frac{N}{2}\frac{r}{\theta_0 + \Delta/2}$$
$$= 8\frac{18}{0.3 - 0.05} + 8\frac{18}{0.3 + 0.05} = 987$$

and the total number of defectives in the experiment is $rN = 18 \cdot 16 = 288$. For a corresponding Binomial plan, we find from Table 1a of Bisgaard and Fuller (1995a) with the values for N, θ_0, Δ, α and β as above, that the individual trial sample size should be $n = 58$. Thus the total sample size will be $nN = 58 \cdot 16 = 928$ and the expected number of defectives will be $nN\theta_0 = 928 \cdot 0.3 = 278$.

Again the plans are similar under "normal" circumstances. However, should an unexpected large two factor interaction effect appear and make four runs much different than the others, say $\theta_4 = \theta_8 = \theta_{12} = \theta_{18} = 0.8$, and the remaining as assumed on average about 0.3, then the Binomial sample scheme will have an expected number of defectives of $\sum_{i=1}^{16} n\theta_i = 58(4 \cdot 0.8 + 12 \cdot 0.3) \approx 394$ or about 37% more than under the Inverse Binomial Sampling scheme.

Example 3. Suppose we want an experiment with $\alpha = 5\%$, $\beta = 10\%$, $N = 16$, $\theta_0 = 0.0001$ and $\Delta = 0.0001$. In that case we are outside the range where Table 1 is applicable. We therefore turn to equation (14) and find that the stopping rule is

$$r \approx \frac{4(1.96 + 1.28)^2 (0.001)^2 (1 - 0.001)}{(0.0001)^2 16} + 2 \approx 262.$$

Example 4. In this last example we will show the real strength of the Inverse Binomial scheme which is when a process is already fairly well optimized into the below 1% range, but nevertheless need further reductions. This would for example be the case for high volume processes. Now suppose $\alpha = 5\%$, $\beta = 10\%$, $N = 16$, $\theta_0 = 0.01$ and $\Delta = 0.0025$. To find the stopping rule r for the Inverse Binomial scheme, we proceed again as follows.

Step 1. Find $b(r)$ from (13):

$$b(r) = \frac{0.0025\sqrt{16}}{2 \cdot 0.01\sqrt{0.99 \cdot 3.24}} = 0.155.$$

Step 2. Enter Table 1 and find an r that produces a $b(r)$ close to 0.155. In this case $r = 43$. Thus the total number of defectives in the experiment is $rN = 43 \cdot 16 = 688$.

For a corresponding Binomial plan, we find from Bisgaard and Fuller (1995a) with the values for N, θ_0, Δ, α and β as above, that the individual trial sample size should be $n = 4146$. Thus the total sample size will be $nN = 4146 \cdot 16 = 66,336$ and the expected number of defectives will be $nN\theta_0 = 66,336 \cdot 0.01 = 663$ if the changes provoked by the experimental trials are small and the overall defective average is 0.01. However, should again an unexpected two factor interaction effect make four runs much different from the others, say $\theta_4 = \theta_8 = \theta_{12} = \theta_{16} = 0.2$, and the remaining as assumed on average about 0.01, then the standard Binomial sample scheme will have an expected number of defectives of $\sum_{i=1}^{16} n\theta_i = 4,446(4 \cdot 0.2 + 12 \cdot 0.01) \approx 3,763$ or about 6 times as many as the Inverse Binomial Sampling scheme.

Discussion and Conclusion

Using standard Binomial sampling, the tables produced by Bisgaard and Fuller (1995a) demonstrated clearly that when the probability of defectives is small, the sample size n for each of the factorial combinations becomes unmanageably large. For example, for a sixteen run experiment, if the average defective probability is about $\theta_0 = 0.01$ or 1%, and we wish to detect a change of $\Delta = 0.001$ or 0.1%, then we need a sample size for each run of n ≈ 26,000. Such large sample sizes may on face value look unrealistic. In particular management will frequently display reluctance to allow such large experiments because of fear of the large costs involved. Unfortunately, Inverse Binomial sampling is not going to reduce the expected length of the experiments under normal circumstances. The likelihood principle prevents that, as we have also seen in the examples in the past section. In fact, using (14) we find that r=260 and the expected run length is about 26,000.

One of the key reasons, in our experience, that management is reluctant to allow any experimentation on ongoing processes and in particular large ones, is their fear that even minor modifications to the process parameters might create a disastrous number of defectives. However, from a rational point of view such an inclination to experimentation is counterproductive. It does not allow for quality problems to be solved; a 1% defective rate for a high volume process can be a serious problem. Thus as Box (1966) has said "to find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it)." If management wants to reduce the defectives, they must therefore acknowledge that experiments often will be needed — even extremely large ones when the defect rate is very small. However, their concerns should also be addressed. Their fears are real and understandable. Thus we think that if the quality engineer can guarantee management that no more than a fixed number of defectives will be produced, that might provide enough of a compelling argument to sway them to initiate an experiment.

In the sixteen run inverse Binomial experiment proposed above with $r = 260$ for a process currently running at an average of 1 % defectives, that would mean a total of 16 ×

260 = 4160 defectives produced. That might sound as a large number, but is no more ($16 \times 26{,}000 \times 0.01 = 4160$) than what would have been produced had the process been left alone during a similar period of time. On the other hand, if a Binomial experiment was run and one or several factorial combinations resulted in a large increase in the defect level, then with a fixed individual sample size of 26,000 that would quickly translate into a huge scrap heap on the production floor.

What we have proposed in this article is an idea based on sound statistical principles that to a large extent will address the fear management might have of bungling up a process that is already running well. The proposed Inverse Binomial sampling method will allow for the experimental determination of increasingly better process conditions when the response is binary and the defective rate small. One major benefit of this variable sampling scheme, as opposed to a fixed, is that only few defectives are produced because the switching rule will quickly send the process away from unfortunate factorial combinations that have a propensity for producing defectives, toward better combinations. Moreover, the experimental process can be run as a Evolutionary Operation process, see Box and Draper (1969).

Appendix 1: A Few Useful Facts About Inverse Binomial Sampling (IBS)

Let $X_1, X_2, \ldots$ be a sequence of independent identically Bernoulli random variables with $\Pr\{X_i = 1\} = \theta, i = 1, 2, \ldots$ Inverse Binomial sampling (IBS) is then a sampling scheme which stops at random after Y trials when the total number of defectives reaches a prescribed number r. The random variable Y has a so-called Negative Binomial Distribution given by

$$Pr(Y = m) = \binom{m-1}{r-1}\theta^r(1-\theta)^{m-r}, n = r, r+1, \ldots \quad \text{(A1.1)}$$

and zero otherwise. Note that the random variable Y as defined here is the total number of trials to observe r defectives and not as defined by some authors, the number of successes until r defectives have occurred.

As shown e.g. by Kotz and Johnson (1983) a uniform minimum variance unbiased (UMVU) estimator of θ is $(r \geq 2)$

$$\hat{\theta} = (r-1)/(Y-1). \quad \text{(A1.2)}$$

Thus by direct application of the definition for the variance of a discrete random variable, we have that

$$Var\{\hat{\theta}\} = \sum_{m=r}^{\infty}\left(\frac{r-1}{m-1} - \theta\right)^2 \cdot \binom{m-1}{r-1}\theta^r(1-\theta)^{m-r}. \quad \text{(A1.3)}$$

Unfortunately, we know of no closed-form formula for $Var\{\hat{\theta}\}$. However, the following set of inequalities provided by Mikulski and Smith (1976)

$$\frac{\theta^2(1-\theta)}{r} \leq Var\{\hat{\theta}\} \leq \frac{\theta^2(1-\theta)}{r-2+\theta} \leq \frac{\theta^2(1-\theta)}{r-2} \quad \text{(A1.4)}$$

can help provide approximations. For example, by subtracting the two outer quantities from each other, it is readily seen that the error of replacing (A1.3) by $\theta^2(1-\theta)/(r-2)$ is smaller than $\theta^2(1-\theta)/(r-2)$ and hence of the order of r^2.

Appendix 2: A Variance Approximation for the Unbiased Estimator for Small r

From the sharper inequality of (A1.4), since θ is additive in the denominator, it appears that $\sigma_{\hat{\theta}} = Var\{\hat{\theta}\}^{1/2}$ for small θ and even for small r and, of course better for large r's, can be approximated by a function of form

$$\sigma_{\hat{\theta}} \equiv b(r)\theta\sqrt{1-\theta} \quad \text{(A2.1)}$$

where $b(r)$ is some constant depending on the specific r (e.g. we will ignore the influence of θ since it is very small). Next we select $b(r)$ such that (A2.1) equals the true $\sigma_{\hat{\theta}}$ given by (A1.3) for $\theta = 0.1$, a value that seems right in the middle of the interval for potential use and should we go much below then the influence of ignoring θ will be even smaller. Thus by calibrating the b(r)'s to give exact equality with (A1.3) for $\theta = 0.1$ we get a set of constants $b(r)$ as shown in Table 1.

Numerical investigations have shown that $\sigma_{\hat{\theta}}$ given by (A2.1) and the $b(r)$'s given by Table 1 approximates the true $\sigma_{\hat{\theta}}$, calculated from (A1.3), very accurately in the range $\theta \in [0.02, 0.4]$, i.e. for practically all values of θ of interest, and for a wide range of r-values. For example, for $r = 20$, (A2.1) produces at $\theta = 0.4$ a relative error of 1.6%; for $r = 60$, the same error is 0.5%; for $\theta = 0.04$ the relative errors are 0.4% and 0.1%, respectively. Moreover, it is clear from (A1.4) that the larger r, the better is the approximation, and that the table values will converge relatively quickly to $b(r) \approx 1/\sqrt{r-2}$.

Acknowledgment

S. Bisgaard's work on this research was supported by the National Science Foundation (NSF) Grant No. DMI 9500140 and Grant No. EEC 8721545. I. Gertsbakh was supported by a sabbatical stipend from Ben Gurion University.

References

Anscombe, F. J. (1948), "The Transformation of Poisson, Binomial and Negative Binomial data." *Biometrika*, Vol. 35, pp. 246-254.

Bisgaard, S. and Fuller, H. T. (1995a), "Sample Size Estimates for 2^{k-p} Designs with Binary Responses," *Journal of Quality Technology*, Vol. 27, No. 4, pp. 344-354.

Bisgaard, S. and Fuller, H. T. (1995b), "Analysis of Factorial Experiments with Defects and Defectives as the Responses," *Quality Engineering*, Vol. 7, No. 2, pp. 429-443.

Box, G. E . P. (1966), "Use and Abuse of Regression," *Technometrics*, Vol. 8, No. 4, pp. 625-629.

Box, G. E. P. and Draper, N. R. (1969), *Evolutionary Operation*, New York: John Wiley and Sons.

Cox, D. R. and Hinkley D. V. (1974, *Theoretical Statistics*, London: Chapman and Hall.

Kotz, S. and Johnson, N. (1983), *Encyclopedia of Statistical Sciences*, Vol. 4, p. 250, Editors-in-Chief, Wiley & Sons, 1983.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, Second Edition, New York: Chapman and Hall.

Mikulski, P. W. and Smith, P. J. (1976), "A Variance Bound for Unbiased Estimation in Inverse Sampling", *Biometrika*, Vol. 63, No. 1, pp. 216-217.