A Stopped Negative Binomial Distribution

Michelle DeVeaux[a], Michael J. Kane[a,*], Daniel Zelterman[a]

[a] *Department of Biostatistics*
School of Epidemiology and Public Health
Yale University, New Haven, CT

Abstract

This paper introduces a new discrete distribution suggested by curtailed sampling rules common in early-stage clinical trials. We derive the distribution of the smallest number of independent and identically distributed Bernoulli trials needed in order to observe either s successes or t failures. The closed form expression for the distribution as well as the compound distribution are derived. Properties of the distribution are shown and discussed. A case study is presented showing how the distribution can be used to monitor sequential enrollment of clinical trials with binary outcomes as well as providing post-hoc analysis of completed trials.

Keywords: discrete distribution, curtailed sampling

1. Introdution and Motivation

Consider the design for a Phase II clinical trial that investigates the efficacy of iniparib in patients with breast cancer gene-associated (BRCA) ovarian cancer [**??**]. In the first stage, 12 patients were treated with iniparib (BSI-201/SAR240550). The outcome was defined by Response Evaluation Criteria in Solid Tumor (RECIST). There were two endpoints: terminate or enroll additional patients at the end of this stage. If at least two of these patients respond

[☆] This research was supported by grants R01CA131301, R01CA157749, R01CA148996, R01CA168733, and PC50CA196530 awarded by the National Cancer Institute, and support from the Yale Comprehensive Cancer Center.

[*] Corresponding author
Email addresses: michelle.deveaux@yale.edu (Michelle DeVeaux), michael.kane@yale.edu (Michael J. Kane), daniel.zelterman@yale.edu (Daniel Zelterman)

favorably, then the trial will proceed to the next stage where additional patients were able to be treated. If fewer than two respond then the trial would have been terminated.

The maximum sample size was 12 but the number of patients necessary to reach any endpoint could be less. Our goal is to describe the distribution of the enrollment size of the design for planning purposes and to use this distribution to estimate the success probability. If all 12 patients are enrolled at once, as in the classic design, then the sample size is 12. However, in most clinical trials, the patients are enrolled sequentially, often with one patient's outcome realized before the next one enters the trial. In the present example, observing two successful patients allows us reach one endpoint so the sample required could be as small as two. So, 12 might not be necessary. Similarly 11 observed treatment failures also ends the stage. This sampling mechanism, in which the experiment ends as soon as any of the endpoints is reached, is call *curtailed sampling*. Under curtailed sampling the range of the sample size is between two and 12.

We assume each of the patient outcomes can be modeled as independent, identically distributed Bernoulli(p) random variables. The realization of the trial is a sequence of these random variables that stops when either a specified number of success or failures is reached. In the previous example suppose two successes were reached after enrolling 10 patients (one in the third step and one at the 10^{th}). The sample path is illustrated in Fig. 1. The vertical axis denotes the number of successful outcomes. The horizontal axis counts the number of patients that have been enrolled. The horizontal and vertical boundaries represent endpoints for the trial.

The next section of this paper introduces our notation and basic results including the density of the distribution along with a description of it's relation to other distributions. Sections 2 and 3 derive the distribution based on a defined Bernoulli process and give some basic properties of the distribution. Section 4 derives the compound distribution using a Beta prior. Section 5 develops a post-hoc analysis of a completed trial. Section 6 is devoted to discussion and conclusions. The simulations, model fitting routines, and visualizations

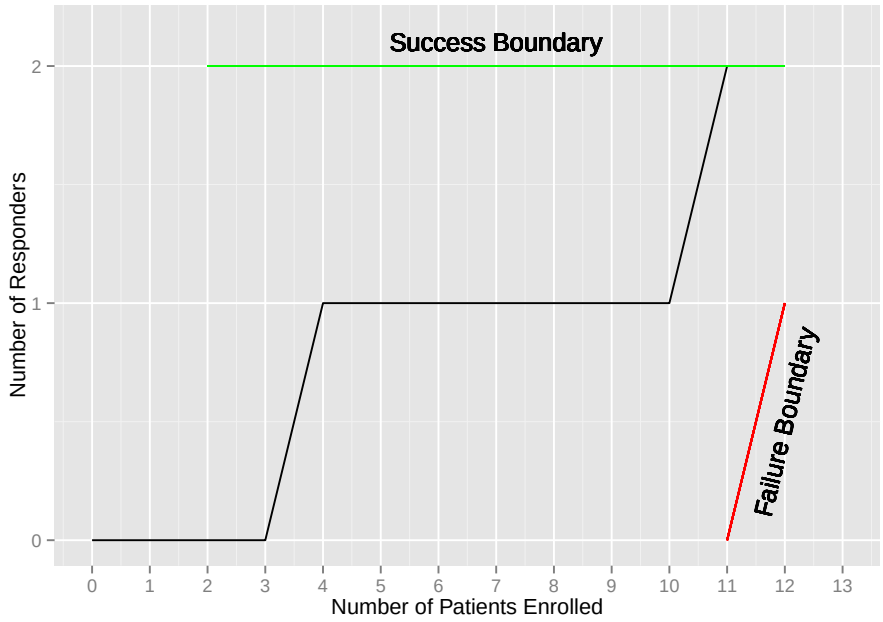

Figure 1: A hypothetical realization of the breast cancer trial.

presented in this paper were generated using the snb package [**??**] for the R programming environment [**??**]

2. Notation and Summary of Distributional Results

Let $b_1, b_2, \ldots$ denote a sequence of independent, identically distributed, Bernoulli random variables with $\mathbb{P}[b_i = 1] = p$ and $\mathbb{P}[b_i = 0] = 1 - p$, for probability parameter $0 \leq p \leq 1$. In the clinical trial setting $b_i = 1$ corresponds to a successful outcome. Let s and t be positive integers. Define the stopped negative binomial (SNB) random variable Y as the smallest integer value such that $\{b_1, \ldots, b_Y\}$ contains *either* s successes *or* t failures. That is, either $\sum_i^Y b_i = s$ or $\sum_i^Y 1 - b_i = t$.

The distribution of Y has support on integer values in the range

$$\min(s, t) \leq Y \leq s + t - 1$$

and it is distributed as

$$\mathbb{E}\ I_{\{Y=k\}} = S(k,p,s)\ I_{\{s\leq k\}} + S(k,1-p,t)\ I_{\{t\leq k\}} \tag{1}$$

where $I_{\{f\}}$ is one if f is true and zero otherwise and

$$S(k,p,s) = \binom{k-1}{s-1} p^s (1-p)^{k-s} \tag{2}$$

S is the negative binomial probability In 1 we can see that the mass at any point on the support of the distribution may come from $S(p)$; $S(1-p)$; or both of their probabilities.

To show the result in Eqn. 1 consider the process $\mathbf{X} = \{X(k) : k = 0, 1, ...\}$ with $X(0) = 0$ and

$$X_{k+1} = X_k + b_{k+1}\ I_{\{k-t<X_k<s\}}.$$

The process can be conceptualized as a series of coin flips that stops when either s successes or t failures are reached. At each step a coin is flipped. If it is heads, the process advances one diagonally in the positive horizontal and vertical direction in Fig. 1. Otherwise, it advances in the positive horizontal direction only. When the process stops either $X_k = s$ or $X_k = k - t$.

Proposition 1. *The distribution of the stopping time* $\underset{k}{\operatorname{argmin}} \{X_k \geq s \cup X_k \leq k - t\}$ *is equal to the SNB and is given in (1).*

Proof. The probability that a given realization of $\mathbf{X}$ reaches s at the kth outcome is the probability that, at time $k-1$ there are $s-1$ successful outcomes and $k-s$ unsuccessful outcomes multiplied by the probability of a success at time k (2). The probability $X_k = k - t$ is the probability that, at outcome $k-1$ there are $k-t$ successful outcomes and $t-1$ unsuccessful outcomes multiplied by the probability of an unsuccessful outcome at time k. Finally, we need to show that the sum

$$R = \sum_{k=s}^{s+t-1} S(k,p,s) + \sum_{k=t}^{s+t-1} S(k,1-p,t) \tag{3}$$

$$= \sum_{k=s}^{s+t-1} \binom{k-1}{s-1} p^s (1-p)^{k-s} + \sum_{k=t}^{s+t-1} \binom{k-1}{k-t} p^{k-t} (1-p)^t \tag{4}$$

is equal to one.

Substitute $i = k - s$ in the first summation and $j = k - t$ in the second. Then R can be written as the cumulative distribution function of two negative binomial distributions

$$R = \sum_{i=0}^{t-1} \binom{i+s-1}{i} p^s (1-p)^i \; + \sum_{j=0}^{s-1} \binom{j+t-1}{j} p^j (1-p)^t. \tag{5}$$

Let $\mathcal{I}_p(s,t)$ be the *regularized incomplete beta function*, which is also one minus the c.d.f. of the negative binomial distribution. This function satisfies $\mathcal{I}_p(s,t) = 1 - \mathcal{I}_{1-p}(t,s)$ [**??**]. Then it can be proven that 1 is a valid probability mass by showing that the probability mass function sums to one.

$$\begin{aligned} R &= \sum_{i=0}^{t-1} \binom{i+s-1}{i} p^s (1-p)^i + \sum_{j=0}^{s-1} \binom{j+t-1}{j} p^j (1-p)^t \\ &= 1 - \mathcal{I}_p(s,t) + 1 - \mathcal{I}_{1-p}(t,s) \\ &= 1. \end{aligned}$$

60 □

Proposition 2. *Let S be distributed SNB with parameters p, s, and t. Then the moment generating function (MGF) of S is*

$$\mathbb{E}\; e^{xS} = \frac{pe^x}{1-qe^x} \mathcal{I}_{qe^x}(s,t) + \frac{qe^x}{1-pe^x} \mathcal{I}_{pe^x}(s,t) \tag{6}$$

for $q = 1 - p$ when $x \leq -\log(q)$ and $x \leq -\log(p)$.

Proof. By definition, the MGF of the SNB is:

$$\mathbb{E}\; e^{xS} = \sum_{k=s}^{s+t-1} \binom{k-1}{s-1} p^s q^{k-s} e^{kx} + \sum_{k=t}^{s+t-1} \binom{k-1}{k-t} p^{k-t} q^t e^{kx}$$

and this can be rewritten as:

$$\mathbb{E}\; e^{xS} = \sum_{k=s}^{s+t-1} \binom{k-1}{s-1} (pe)^{sx} (qe^x)^{k-s} + \sum_{k=t}^{s+t-1} \binom{k-1}{k-t} (qe^x)^t (pe^x)^{k-t}. \tag{7}$$

Taking the first summation in Equation 7:

$$\begin{aligned} \sum_{k=s}^{s+t-1} \binom{k-1}{s-1} (pe)^{sx} (qe^x)^{k-s} &= \left(\frac{pe^x}{1-qe^x}\right)^s \sum_{k=s}^{s+t-1} \binom{k-1}{s-1} (qe^x)^{k-s} (1-qe^x)^s) \\ &= \left(\frac{pe^x}{1-qe^x}\right)^s \mathcal{I}_{qe^x}(s,t). \end{aligned}$$

Since the incomplete beta function has support on zero to one $qe^x \leq 1$. This implies that $x \leq -\log(q)$.

A similar expression can be derived using the same calculation with the constraint that $x \leq -\log(p)$. The result follows from the afore mentioned property of the regularized incomplete beta function. □

The probability mass function of Y has a variety of shapes for different choices of the parameters (s, t, p). These shapes are illustrated in Fig. 2.The SNB is related to the negative binomial distribution. Specifically, if t is large then the $Y - s$ has a negative binomial distribution with

$$\mathbb{P}\{Y = s + j\} = \binom{s+j-1}{s-1} p^s (1-p)^j$$

for $j = 0, 1, \ldots$. A similar statement can be made when s is large and t is small.

For the special case of $s = t$, the distribution of Y is the riff-shuffle, or minimum negative binomial distribution [**??**]. Similar derivations of the closely-related maximum negative binomial discretedistributions also appear in [**??**, **??**]. The maximum negative binomial is the smallest number of outcomes necessary to observe at least c successes *and* c failures, but the SNB givethe number of coin flips to observe *either* s heads or t tails.

3. Connection to the Binomial Tail Probability

4. The Posterior Distribution

Let us assume that p has a Beta distribution, with constant prior parameters α and β. Then the closed form posterior distribution of the SNB is

$$\begin{aligned} f\left(k|\mathbf{b},\alpha,\beta\right) = &\binom{k-1}{s-1} \frac{B\left(\alpha+s, k-s+\beta\right)}{B(\alpha,\beta)}\, I_{\{s\leq k\leq s+k-1\}} + \\ &\binom{k-1}{k-t} \frac{B\left(\alpha+k-t, t+\beta\right)}{B(\alpha,\beta)}\, I_{\{t\leq k\leq s+k-1\}} \end{aligned} \tag{8}$$

where B is the Beta function and $\mathbf{b}$ is the realization of Bernoulli data.

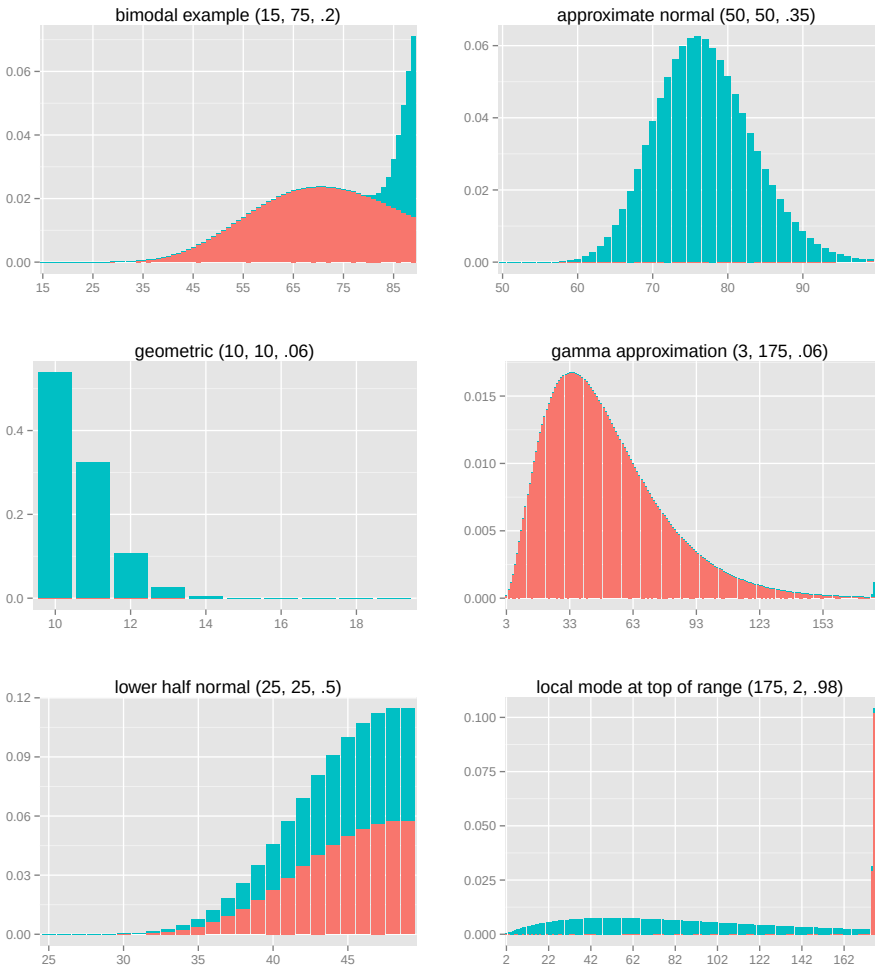

Figure 2: Different shapes of the SNB distribution with parameters (s, t, p), as given. Red indicates mass contributed by hitting s, teal indicates mass contributed by hitting t.

Proposition 3. *The posterior PMF of the Stopped Negative Binomial distribution with a Beta(α, β) prior is:*

$$f(k|s,t,\alpha,\beta) = \binom{k-1}{s-1}\frac{B(\alpha+s,k-s+\beta)}{B(\alpha,\beta)}\ I_{\{s\leq k\leq s+k-1\}} + \binom{k-1}{k-t}\frac{B(\alpha+k-t,t+\beta)}{B(\alpha,\beta)}\ I_{\{t\leq k\leq s+k-1\}} \tag{9}$$

where B is the Beta function.

Proof. For notational simplicity, assume that $s,t \leq k \leq s+t-1$. When this is not the case appropriate terms should be removed as dictated by the indicator functions.

$$\begin{aligned} f(k|s,t,\alpha,\beta) &= \frac{1}{B(\alpha,\beta)}\int_0^1 \binom{k-1}{s-1}p^{\alpha+s-1}(1-p)^{k-s+\beta-1} + \\ &\qquad \binom{k-1}{k-t}p^{k-t+\alpha-1}(1-p)^{t+b-1}\,dp \\ &= \frac{1}{B(\alpha,\beta)}\binom{k-1}{s-1}\int_0^1 p^{\alpha+s-1}(1-p)^{k-s+\beta-1}\,dp + \\ &\qquad \frac{1}{B(\alpha,\beta)}\binom{k-1}{k-t}\int_0^1 p^{k-t+\alpha-1}(1-p)^{t+b-1}\,dp \end{aligned}$$

The result follows by applying the definition of the Beta function to the integral terms. □

Eqn. 9 suggests an alternative construction for Phase II clinical trials. Patients could be enrolled sequentially until there are enough patients to guarantee the desired level of power or until a specified number of adverse events are recorded. In cases such as these the endpoints s and t are known and new outcomes can be incorporated into the estimate of p to provide updated odds of successful enrollment. Later it will be shown that estimates of the distribution of p can even be used after an endpoint has been reached to determine the amount of uncertainty in the endpoint.

5. Discussion and Conclusion

We have presented a new discrete distribution by curtailed sampling rules common in early-stage clinical trials, which we refer to as the Stopped Negative

Binomial distribution. The distribution models the stopping time of a sequential trial where the trial is stopped when a number of events are accumulated. The posterior distribution was derived for the case when the event probability p has a Beta distribution. Using a trial description from `clinicaltrials.gov` we showed how the SNB is an integral part of trial monitoring, and post-hoc analysis and can be used in a framework alternative to the Simon two-stage optimal design while providing better estimates of the event probability along with quantifying the uncertainty associated with the estimates. As a result, we were able to show fewer patient enrollments and achieve the same estimates, when added in a hypothetical, but representative, clinical trial.

Current work focuses on the generalization of the distribution and the application in other areas of clinical trials. Adverse outcomes in particular are another area that could greatly benefit both from the presented distribution and its framework. For these trials monitoring may need to be performed not only on the outcome of an intervention but also on the safety of the trial. This especially true in areas like late-stage cancer treatments where the treatment is harsh and patients may be forced to drop out as a result of the side-effects of the treatment; a matter independent of the outcome. Other application areas include providing designs that allows clinicians to balance uncertainty and success probability with a minimum number of patient enrollees.

Here are two sample references: [**?** **?**].

References