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 novel 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 its characteristics and properties are explored.

Keywords: discrete distribution, curtailed sampling

1. Introdution and Motivation

Consider a prototypical Phase II single-arm clinical trial. 12 patients are enrolled and treated. If at least 2 of those patients respond favorably, then the trial will proceed to the next stage. If fewer than two respond then the trial would have been terminated.

The maximum sample size is 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. This distribution will be used for planning purposes. If all 12 patients are enrolled at once, as in the classic design, then the sample size is

[☆]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)

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 - 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.

Assume each of patient outcome can be modeled as an independent, identically distributed Bernoulli(p) random variable. The trial is realized as 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 rest of this paper derives the distribution for the number of patients needed to end a trial and explores the characteristics and properties of this distribution. The next section 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 derives the distribution based on a defined Bernoulli process and give some basic properties. Section 3 provides a connection to the Binomail tail probability. Section 4 derives the posterior distribution using a Beta prior. Section 5 provides a brief discussion on the use of the distribution in clinical trials along with future avenues for generalization.

2. Distributional Result

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

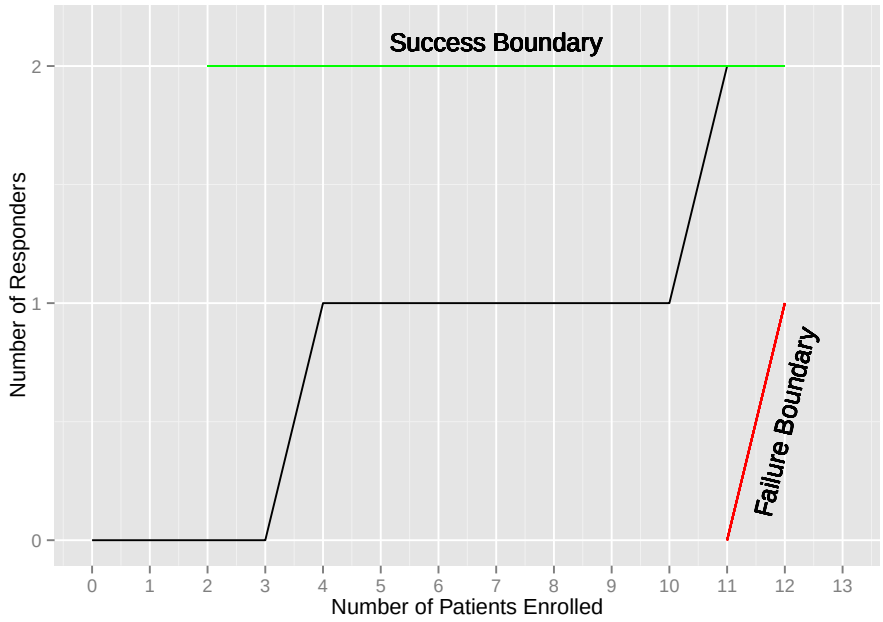

Figure 1: A hypothetical realization of a trial.

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

To prove 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. Otherwise, it advances in the positive horizontal direction only. The process stops when 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 Eqn. 1.*

Proof. The probability 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.

$$S(k, p, s) = \binom{k-1}{s-1} p^s (1-p)^{k-s} \tag{3}$$

The probability a given realization reaches $k - t$ 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.

$$S'(k, p, t) = \binom{k-1}{k-t} p^{k-t} (1-p)^t \tag{4}$$

It can be shown that $S(k, p, s) = S'(k, 1-p, s)$ by realizing that $\binom{k-1}{k-s} = \binom{k-1}{s-1}$.

To show that sum of Eqn. 3 and 4 sum to one over their support

$$R = \sum_{k=s}^{s+t-1} S(k, p, s) + \sum_{k=t}^{s+t-1} S(k, 1-p, t) \tag{5}$$

$$= \sum_{k=s}^{s+t-1} \binom{k-1}{s-1} p^s (1-p)^{k-s} + \sum_{k=t}^{s+t-1} \binom{k-1}{k-t} p^{k-t} (1-p)^t \tag{6}$$

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

$$R = \sum_{i=0}^{t-1} \binom{i+s-1}{i} p^s (1-p)^i \; + \sum_{j=0}^{s-1} \binom{j+t-1}{j} p^j (1-p)^t. \tag{7}$$

Let $\mathcal{I}_p(s,t)$ be the *regularized incomplete beta function* and recall this function satisfies $\mathcal{I}_p(s,t) = 1 - \mathcal{I}_{1-p}(t,s)$ [1].

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

□

3. Shape and Basic Properties

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 [2]. Similar derivations of the closely-related maximum negative binomial discrete distributions also appear in [3] and [4]. 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.

4. Connection to the Binomial Tail Probability

Proposition 2. *Let Y be distributed as SNB(p, s, t) and let B be distributed Binomial with size $n = s + t - 1$ and success probability p. Then*

$$\mathbb{P}[B \geq s] = \mathbb{P}[Y \leq n | \#success = s]. \tag{8}$$

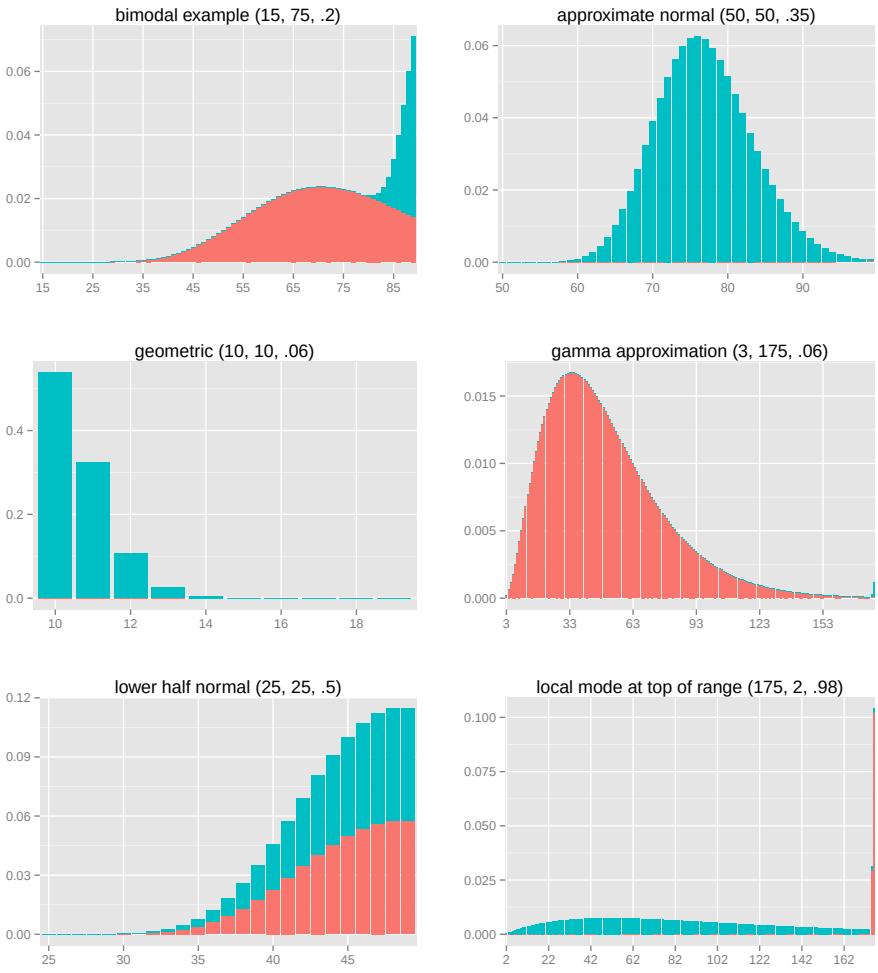

Figure 2: Different shapes of the SNB distribution with parameters (s, t, p), as given. Red indicates mass contributed by hitting s, teal indicates mass contributed by hitting t.

That is, the probability that the number of successes is at least s in the Binomial model is the same that the trial stops with s successes in the SNB model.

Proof. The Binomial tail probability is

$$\begin{aligned}\mathbb{P}[B \geq s] &= \sum_{k=s}^{s+t-1} \binom{n}{k} p^k (1-p)^{n-k} \\ &= 1 - \sum_{k=0}^{s-1} p^k (1-p)^{n-k} \\ & 1 - \mathcal{I}_{1-p}.\end{aligned}$$

Next, consider the SNB probability

$$\mathbb{P}[Y \leq n | \#\text{success} = s] = \sum_{k=s}^{s+t-1} \binom{k-1}{s-1} p^s (1-p)^k - s.$$

Let $i = k - s$ then using the fact that $\binom{i+s-1}{s-1} = \binom{i+s-1}{i}$ the summation can be rewritten as

$$\mathbb{P}[Y \leq n | \#\text{success} = s] = \sum_{i=0}^{t-1} \binom{i+s-1}{i} p^s (1-p)^i \tag{9}$$

$$= 1 - \mathcal{I}_{1-p}(t, s). \tag{10}$$

65 □

5. The Moment Generating Function

Proposition 3. *Let S be distributed SNB with parameters p, s, and t. Then the moment generating function (MGF) of S is*

$$\mathbb{E}\ e^{xS} = \frac{pe^x}{1 - qe^x} \mathcal{I}_{qe^x}(s, t) + \frac{qe^x}{1 - pe^x} \mathcal{I}_{pe^x}(s, t) \tag{11}$$

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 can be rewritten as:

$$\mathbb{E}\ e^{xS} = \sum_{k=s}^{s+t-1} \binom{k-1}{s-1}(pe)^{sx}(qe^x)^{k-s} + \sum_{k=t}^{s+t-1} \binom{k-1}{k-t}(qe^x)^t(pe^x)^{k-t}. \quad (12)$$

Taking the first summation in Equation 12:

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

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. □

6. 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

$$\begin{aligned}f\left(k|\mathbf{b},\alpha,\beta\right) = \binom{k-1}{s-1}\frac{B\left(\alpha+s,k-s+\beta\right)}{B(\alpha,\beta)}\ I_{\{s\leq k\leq s+k-1\}} + \\ \binom{k-1}{k-t}\frac{B\left(\alpha+k-t,t+\beta\right)}{B(\alpha,\beta)}\ I_{\{t\leq k\leq s+k-1\}}\end{aligned} \quad (13)$$

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

Proposition 4. *The posterior PMF of the Stopped Negative Binomial distribution with a Beta(α, β) prior is:*

$$\begin{aligned}f(k|s,t,\alpha,\beta) = \binom{k-1}{s-1}\frac{B\left(\alpha+s,k-s+\beta\right)}{B(\alpha,\beta)}\ I_{\{s\leq k\leq s+k-1\}}\ + \\ \binom{k-1}{k-t}\frac{B\left(\alpha+k-t,t+\beta\right)}{B(\alpha,\beta)}\ I_{\{t\leq k\leq s+k-1\}}\end{aligned} \quad (14)$$

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. □

7. 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.

References

[1] M. Abramowitz, I. A. Stegun, Handbook of mathematical functions: with formulas, graphs, and mathematical tables, Vol. 55, Courier Corporation, 1964.

[2] V. Uppuluri, W. Blot, A probability distribution arising in a riff-shuffle, Random Counts in Scientific Work, 1: Random Counts in Models and Structures (1970) 23–46.

[3] Z. Zhang, B. A. Burtness, D. Zelterman, The maximum negative binomial distribution, Journal of Statistical Planning and Inference 87 (1) (2000) 1–19.

[4] D. Zelterman, Discrete distributions: applications in the health sciences, John Wiley & Sons, 2005.