Almost-Binomial Random Variables

Peter Thompson (thompsop@wabash.edu), Wabash College, Crawfordsville, IN 47933

The parameter n in the binomial distribution is a positive integer. What happens if we allow it to be any positive number? The new distribution we get is not only easy to work with but is useful in approximation situations that beginning statistics students could encounter. In addition, applications involving the distribution provide a good source for undergraduate research problems.

We begin by setting $g(x) = \binom{n}{x} p^x (1-p)^{n-x}$, where $\binom{n}{x} = \frac{n(n-1)\cdots(n-x+1)}{x!}$. When $p < 0.5$, we observe that many of the features of the binomial distribution still hold. In particular, $\sum_{x=0}^{\infty} g(x) = 1$, which follows from the binomial series,

$$\left(1 + \frac{p}{1-p}\right)^n = \sum_{x=0}^{\infty} \binom{n}{x} \left(\frac{p}{1-p}\right)^x,$$

$$\sum_{x=0}^{\infty} x g(x) = np, \quad \sum_{x=0}^{\infty} (x - np)^2 g(x) = np(1-p),$$

and

$$\eta(t) = \sum_{x=0}^{\infty} t^x g(x) = (1 - p + pt)^n, \quad \text{for } |t| < 0.5/p.$$

The function $g(x)$ is not a probability density function (*pdf*) because $\binom{n}{x}$ alternates between negative and positive values for $x > [n] + 1$. With this in mind, we define the *pdf* of the *almost-binomial*(n, p) distribution to be

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, [n], \\ 1 - \sum_{x=0}^{[n]} \binom{n}{x} p^x (1-p)^{n-x}, & x = [n] + 1, \end{cases}$$

where $[n]$ denotes the greatest integer less than or equal to n.

We observe that when $p < 0.5$, the sum of the terms beyond $x = [n]$ in each series above tends to be extremely small. In fact, $|\sum_{x=[n]+1}^{\infty} g(x)| < |g([n] + 1)| < p^{[n]}$. Consequently, if $p < 0.5$ and X has an *almost-binomial*(n, p) distribution, $E(X) \approx np$ and $\mathrm{Var}(X) \approx np(1-p)$, and the probability generating function (*pgf*) for X, $\eta(t)$, satisfies $\eta(t) \approx (1 - p + pt)^n$, $|t| < 0.5/p$.

Example 1. For $n = 5.5$ and $p = 0.4$, we get the following *pdf* of the almost-binomial(5.5, 0.4) distribution.

x	0	1	2	3	4	5	6
$f(x)$	0.06023	0.22085	0.33128	0.25766	0.10736	0.02147	0.00115

Note that $np = 2.2$ and $E(X) = 2.2008$, and $np(1-p) = 1.32$ and $\mathrm{Var}(X) = 1.3205$.

Almost-binomial(n, p) distributions form a rich two-parameter family of discrete distributions where the mean exceeds the variance, and they are useful in a variety of

modeling situations. Next, we show how they can be used to model sums of independent random variables.

Suppose that $X_1, X_2, \ldots, X_k$, are independent Bernoulli(p_i) random variables (referred to as *Poisson trials*). Let $X = \sum_{i=1}^{k} X_i$ (the distribution of X is called the *Poisson-binomial* distribution). We can approximate probabilities concerning X as follows. We set

$$np = \mu = \sum_{i=1}^{k} p_i, \quad \text{and} \quad np(1-p) = \sigma^2 = \sum_{i=1}^{k} p_i(1-p_i),$$

and get

$$n = \frac{\left(\sum_{i=1}^{k} p_i\right)^2}{\sum_{i=1}^{k} p_i^2} \quad \text{and} \quad p = \frac{\sum_{i=1}^{k} p_i^2}{\sum_{i=1}^{k} p_i}.$$

Then we use the almost-binomial(n, p) distribution to approximate probabilities involving X.

Example 2. A machine has 20 components that operate independently. Suppose the probability that the ith component fails is $p_i = \frac{i}{100}$. Let X be the total number of components that fail. For beginning students, the exact distribution of X is difficult to deal with, especially if they only have a hand calculator. However, almost-binomial approximations are straightforward. Calculating n and p as above, we get $p = 0.13667$ and $n = 15.366$. The next table gives the almost-binomial approximations to $P(X = x)$, denoted AlBin(x), and (for comparison purposes) the true values for $P(X = x)$.

x	0	1	2	3	4	5	6	7	8
AlBin(x)	0.10455	0.25431	0.28917	0.20395	0.09981	0.03592	0.00982	0.00208	0.00034
$P(X = x)$	0.10432	0.25452	0.28946	0.20385	0.09961	0.03585	0.00985	0.00211	0.00036

A theoretical justification for almost-binomial approximations is given in [**4**]. (See also [**1**, pp. 188–191], for a similar theoretical discussion of binomial approximations.) There we can find the following result as well as a comparison of almost-binomial, binomial, and Poisson approximations.

Theorem. Suppose $X = \sum_{i=1}^{k} X_i$, where $X_1, X_2, \ldots, X_k$, are independent Bernoulli random variables with parameters p_i. Let $p = \frac{\sum_{i=1}^{k} p_i^2}{\sum_{i=1}^{k} p_i}$, and $n = \frac{(\sum_{i=1}^{k} p_i)^2}{\sum_{i=1}^{k} p_i^2}$, and let $\tilde{P}$ be a random variable for which $P(\tilde{P} = p_i) = \frac{p_i}{\sum_{i=1}^{k} p_i}$. Then for $A \subset \{0, 1, \ldots, [n]\}$,

$$|P(X \in A) - P(\text{AlBin}(n, p) \in A)| \leq \frac{4}{1-p}\,\text{Var}(\tilde{P}) + \frac{k - [n] - 1}{([n]+1)(1-p)} P(X \geq [n]+2).$$

I have directed two undergraduate research projects related to the almost-binomial distribution. The first intern (320 hours), Abu Jalal, developed error bounds for almost-binomial approximations to probabilities involving sums of independent hypergeometric random variables [**3**]. The second intern (160 hours), Don Claycomb, is currently finishing work on practical (non-theoretical) improvements to the error bound given

in the theorem. This theorem is rather conservative and one can generally get much smaller bounds on the errors in approximation that hold for intuitively worst-case situations—thus, arguably everywhere.

There are other opportunities for undergraduate research that faculty members could craft. One interesting problem would be to work out the details of testing a null hypothesis that data is from a Poisson distribution by comparing the Poisson fit to the data with that of the best fitting almost-binomial distribution. If one suspected the data would be under-dispersed relative to the Poisson distribution (before collecting data), a one-sided test of this sort may be an interesting competitor to the Poisson dispersion test.

Other possible problems would include looking at some of the material in [2] that used the Poisson-binomial distribution, namely logistic regression and conditional Bernoulli models. The almost-binomial approximation to the Poisson-binomial may be of interest here.

References

1. A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation*, Oxford University Press, 1992.
2. S. Chen, and J. Liu, Statistical applications of the Poisson-binomial and conditional Bernoulli distributions, *Statistica Sinica* **7** (1997) 875–892.
3. A. Jalal, Error bounds involving almost-binomial approximations of hypergeometric probabilities, *Pi Mu Epsilon Journal*, **11** (2001) #4 187–193.
4. P. Thompson, Almost-binomial random variables, technical report, Wabash College, 1999.

———o———

The Roots of a Quadratic

Leonard Gillman (len@math.utexas.edu), The University of Texas, Austin, TX 78712

In a recent discussion among several math teachers, both college and precollege, someone remarked that we do students a disservice when we let them "solve" a quadratic equation by means of the formula without having them check their answers. The obvious question at this point is just how the students are expected to do the checking. Most of the group agreed that substituting into the quadratic is too hard, at least for beginning students, especially when complex numbers are involved.

Regarding this last assertion, observe that there is no need to use the i-notation in order to do the checking. Consider the equation $f(x) = 0$, where

$$\begin{aligned} f(x) &= ax^2 + bx + c \\ &= 2x^2 - 5x + 7. \end{aligned} \tag{1}$$

I include enough detail to illustrate the mechanism. Introduce the symbol

$$u = \sqrt{b^2 - 4ac} = \sqrt{-31}, \text{ and note that } u^2 = -31.$$

The solutions given by the quadratic formula are $\frac{5 \pm u}{4}$. For the solution $s = \frac{5+u}{4}$, say, we get

$$f(s) = 2\left(\frac{5+u}{4}\right)^2 - 5\left(\frac{5+u}{4}\right) + 7$$