Deriving size-biased distributions

Susanna Makela

February 14, 2017

Negative Binomial

Suppose we have a superpopulation of random variables $X_i, i = 1, 2, \ldots$, where $X_i \overset{iid}{\sim} NegBin(k, p)$:

$$Pr(X_i = x) = \binom{x+k-1}{x} p^k (1-p)^x,$$

with $k > 0$, $p \in (0, 1)$, and $x = 0, 1, 2, \ldots$. Under this parameterization, X is the number of failures we would see in a sequence of Bernoulli trials with probability of success p before seeing a prespecified number of successes, k.

Now consider sampling (without replacement) from this superpopulation with probability proportional to size (PPS) sampling. (Here we are assuming a superpopulation in part to avoid complications of sampling with replacement.) Using I_i as the binary random variable indicating whether X_i is included in the sample, PPS sampling implies that $Pr(I_i = 1|X_i = x) \propto x$. The distribution of X in our sample is therefore size-biased (reference Patil and Rao here), meaning that we are more likely to observe larger values of X in our sample, compared to the superpopulation. The distribution of sizes in our sample, denoted X_i^*, is then

$$\begin{aligned}
Pr(X_i^* = x) &= Pr(I_i = 1|X_i = x)Pr(X_i = x) \\
&\propto x\binom{x+k-1}{x} p^k (1-p)^x \\
&= \frac{x\binom{x+k-1}{x} p^k (1-p)^x}{\sum_{x=0}^{\infty} x\binom{x+k-1}{x} p^k (1-p)^x} \\
&= \frac{x\binom{x+k-1}{x} p^k (1-p)^x}{\mathbb{E}[X_i]} \\
&= \frac{x\binom{x+k-1}{x} p^k (1-p)^x}{(1-p)k/p} \\
&= \frac{x(x+k-1)!}{x!\,(k-1)!\,k} p^{k+1}(1-p)^{x-1} \\
&= \frac{(x+k-1)!}{(x-1)!\,k!} p^{k+1}(1-p)^{x-1} \\
&= \frac{((x-1)+(k+1)-1)!}{(x-1)!\,k!} p^{k+1}(1-p)^{x-1} \\
&= \binom{(x-1)+(k+1)-1}{x-1} p^{k+1}(1-p)^{x-1} \\
&= Pr(W = x-1),
\end{aligned}$$

where $W \sim NegBin(k+1, p)$. In other words, the sizes X_i^* in our PPS sample are distributed as $X_i^* = 1 + W_i$, where $W_i \overset{iid}{\sim} NegBin(k+1, p)$.

Negative Binomial Parameterizations in R and Stan

In R, the negative binomial is parameterized the same way as above:

$$Pr(X = x) = \binom{x+k-1}{x} p^k (1-p)^x$$

for $x = 0, 1, 2, \ldots$, $k > 0$, and $0 < p \leq 1$. Here x is the number of failures that occur in a sequence of Bernoulli trials before a specified number of successes, k, is achieved. Under this parameterization,

$$\mathbb{E}[X] = \frac{k(1-p)}{p}$$

and

$$Var[X] = \frac{k(1-p)}{p^2}.$$

Stan has two parameterizations for the negative binomial distribution. The one we use parameterizes the distribution in terms of $\mu > 0$ and $\phi > 0$:

$$Pr(X = x) = \binom{x+\phi-1}{x} \left(\frac{\mu}{\mu+\phi}\right)^x \left(\frac{\phi}{\mu+\phi}\right)^\phi.$$

Under this parameterization,

$$\mathbb{E}[X] = \mu$$

and

$$Var[X] = \mu + \frac{\mu^2}{\phi}.$$

To convert between the two parameterizations, we can equate the means and variances and solve for μ and ϕ in terms of k and p. Solving for μ is trivial:

$$\mu = \frac{k(1-p)}{p}.$$

Solving for ϕ, we see that

$$\begin{aligned}
\mu + \frac{\mu^2}{\phi} &= \frac{k(1-p)}{p^2} \iff \\
\mu + \frac{\mu^2}{\phi} &= \frac{\mu}{p} \iff \\
\frac{\mu}{\phi} &= \frac{1}{p} - 1 \iff \\
\frac{\mu}{\phi} &= \frac{1-p}{p} \iff \\
\frac{\phi}{\mu} &= \frac{p}{1-p} \iff \\
\phi &= \frac{\mu p}{1-p} \iff \\
\phi &= k.
\end{aligned}$$

So, a random variable X_i that is distributed $NegBin(k, p)$ under the R parameterization would be distributed as $NegBin(\mu, \phi)$ under the Stan parameterization, where

$$\mu = \frac{k(1-p)}{p} \qquad \text{and} \qquad \phi = k.$$

We know that the size-biased sample sizes X_i^* are distributed as $X_i^* = 1 + W_i$, where $W_i \sim NegBin(k+1, p)$ under the R parameterization. If we were to write $NegBin(k+1,p)$ in the Stan parameterization as $NegBin(\mu^*, \phi^*)$, what are μ^* and ϕ^* in terms of μ and ϕ? The expression for ϕ^* is trivial, since

$$\phi = k \implies \phi^* = k + 1.$$

The expression for μ^* is

$$\begin{aligned} \mu^* &= \frac{(k+1)(1-p)}{p} \\ &= \frac{k(1-p)}{p} + \frac{(1-p)}{p} \\ &= \mu + \frac{\mu}{\phi}. \end{aligned}$$

We can then write the distribution for the size-biased sample sizes X_i^* as

$$X_i^* = 1 + W_i, \quad W_i \sim NegBin(\mu^*, \phi^*),$$

where

$$\mu^* = \mu + \frac{\mu}{\phi} \qquad \text{and} \qquad \phi^* = \phi + 1.$$

Lognormal

We can also consider a continuous distribution for the superpopulation. Suppose the size variables X_i are distributed lognormally in the superpopulation: $X_i \sim LogNormal(\mu, \sigma^2)$. If we then do PPS sampling (without replacement), what is the distribution of observed sizes X_i^*? The derivation in this case is quite a bit longer than in the negative binomial case, but we include it below for completeness. Using I_i to denote the indicator of X_i being included in the sample, we have

$$\begin{aligned} p(X_i^*) &= p(I_i = 1 | X_i) p(X_i) \\ &\propto x \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left(-\frac{1}{2\sigma^2} \left(\log(x) - \mu\right)^2\right) \\ &= \frac{x \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left(-\frac{1}{2\sigma^2} \left(\log(x) - \mu\right)^2\right)}{\int_0^\infty x \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{x} \exp\left(-\frac{1}{2\sigma^2} \left(\log(x) - \mu\right)^2\right) dx} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\log(x) - \mu\right)^2\right)}{\mathbb{E}[X_i]} \\ &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\log(x) - \mu\right)^2\right)}{\exp(\mu + \sigma^2/2)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(\log(x) - \mu\right)^2 - (\mu + \sigma^2/2)\right) \end{aligned} \tag{1}$$

To simplify the exposition, we'll consider the term inside the exponential separately.

$$
\begin{aligned}
-&\frac{1}{2\sigma^2}\left(\log(x)-\mu\right)^2-(\mu+\sigma^2/2) \\
&= -\frac{1}{2\sigma^2}\left(\log(x)-(\mu+\sigma^2)+\sigma^2\right)^2-(\mu+\sigma^2/2) \\
&= -\frac{1}{2\sigma^2}\left(\left(\log(x)-(\mu+\sigma^2)\right)^2+2\sigma^2(\log(x)-(\mu+\sigma^2))+\sigma^4\right)-(\mu+\sigma^2/2) \\
&= -\frac{1}{2\sigma^2}\left(\log(x)-(\mu+\sigma^2)\right)^2-(\log(x)-(\mu+\sigma^2))-\sigma^2/2-(\mu+\sigma^2/2) \\
&= -\frac{1}{2\sigma^2}\left(\log(x)-(\mu+\sigma^2)\right)^2-\log(x)+(\mu+\sigma^2)-(\mu+\sigma^2) \\
&= -\frac{1}{2\sigma^2}\left(\log(x)-(\mu+\sigma^2)\right)^2-\log(x)
\end{aligned}
$$

Substituting this last expression back into the exponential in (1), we get

$$
\begin{aligned}
&\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(\log(x)-\mu\right)^2-(\mu+\sigma^2/2)\right) \\
&\quad= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(\log(x)-(\mu+\sigma^2)\right)^2-\log(x)\right) \\
&\quad= \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{x}\exp\left(-\frac{1}{2\sigma^2}\left(\log(x)-(\mu+\sigma^2)\right)^2\right),
\end{aligned}
$$

which we recognize as the density of a lognormal distribution with parameters $\mu+\sigma^2$ and σ^2. The sampled sizes X_i^* are thus distributed as $X_i^* \sim LogNormal(\mu+\sigma^2, \sigma^2)$.