Andrew Gard - equitable.equations@gmail.com

Computing discrete probabilities in R

R comes with four built-in functions for each major discrete probability distribution, including the binomial, geometric, and negative binomial. The format of the commands are the same for every distribution.

- pmfs start with the letter "d".
- cdfs start with the letter "p".
- inverse cdfs start with the letter "q".
- functions to generate random values start with the letter "r"

The rest of the function name is a shortened version of the distribution name.

Distribution	Probability	Cumulative	Inverse	Random	$P(X = x)$
$\text{Bin}(n, p)$	`dbinom`	`pbinom`	`qbinom`	`rbinom`	$\binom{n}{x} p^x q^{n-x}$
$\text{Geom}(p)$	`dgeom`	`pgeom`	`qgeom`	`rgeom`	$q^{x-1} p$
$\text{NB}(r, p)$	`dnbinom`	`pnbinom`	`qnbinom`	`rnbinom`	$\binom{x+r-1}{r-1} p^r q^x$
$\text{Hyp}(N_1, N_2, n)$	`dhyper`	`phyper`	`qhyper`	`rhyper`	$\frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N_1+N_2}{n}}$
$\text{Pois}(\lambda)$	`dpois`	`ppois`	`qpois`	`rpois`	$\frac{\lambda^x e^{-\lambda}}{x!}$

Example. Let $X \sim \text{Bin}(12, .3)$.

- Compute $P(X = 4)$, $P(X \leq 4)$, and $P(X \geq 4)$.
- What is the 80^{th} percentile of this distribution?
- Simulate 1000 random draws from this distribution.

When $X \sim \text{Bin}(12, .3)$,

$$
\begin{aligned}
P(X = 4) &= \texttt{dbinom}(4, 12, .3) \approx .231 \\
P(X \leq 4) &= \texttt{pbinom}(4, 12, .3) \approx .724 \\
P(X \geq 4) &= 1 - \texttt{pbinom}(3, 12, .3) \approx .507
\end{aligned}
$$

The 80^{th} percentile of a distribution is the smallest value in the support that has cumulative probability greater than 80%. It's typically denoted $\pi_{.80}$.

$$\pi_{.80} \;=\; \texttt{qbinom}(.80, 12, .3) \;=\; 5$$

We can confirm this using the cdf.

$$\texttt{pbinom}(5, 12, .3) \approx .88 \quad \text{but} \quad \texttt{pbinom}(4, 12, .3) \approx .72$$

All of these functions are *vectorized*, meaning they can compute multiple values at once. For instance, the following code gives all the cumulative probabilities of Bin(12, .3).

```
> pbinom(0:12, 12, .3)
 [1] 0.01384129 0.08502505 0.25281535 0.49251577 0.72365547 0.88215126 0.96139916
 [8] 0.99051063 0.99830834 0.99979362 0.99998459 0.99999947 1.00000000
```

We generate random values in Bin(12, .3) using the `rbinom` command.

```
set.seed(0)
rbinom(100, 12, .3)
```

```
 [1] 6 3 3 4 6 2 6 6 4 4 1 2 2 4 3 5 4 4 8 3 5 6 2 4 2 3 3 0 3 5 3 3 4 4 2 5 4
[38] 5 2 5 3 5 4 5 4 4 5 1 3 5 4 3 5 3 2 1 2 3 4 4 3 6 3 3 3 4 3 3 5 1 5 3 5 3
[75] 3 3 6 5 3 5 6 3 4 3 3 5 2 4 2 2 2 2 1 4 5 5 5 3 3 5
```

These values range from 0 to 8 and have a mean of 3.64, which is close to the expected value of $\mu = (12)(.3) = 3.6$.

Example. Every day, a new author has a 10% chance of selling a copy of their ebook. Assume that purchases are independent of one another and that there are never multiple sales in a single day.

- Find the probability that the first purchase occurs on the 10^{th} day.
- Find the probability that the second purchase occurs during the second week (days 8-14).

Let X be the number of days that pass before the first purchase, and Y the total number of days before the second. Then X has a geometric distribution and Y has a negative binomial distribution. In each case, $p = .10$.

Warning! The R functions for these distributions refer to the number of *failures* that occur before the last success, not the total number of trials. With this convention, we need to compute

- $P(X = 9)$ when $X \sim \text{Geom}(.10)$.
- $P(6 \leq Y \leq 12)$ when $Y \sim \text{NB}(2, .10)$.

In Geom(.10),

$$P(X = 9) = \texttt{dgeom}(9, .10) \approx .039$$

The probability that exactly 9 days elapse before the first sale on day 10 is about 4%.

In NB$(2, .10)$,

$$P(6 \leq Y \leq 12) = \texttt{pnbinom}(12, 2, .10) - \texttt{pnbinom}(5, 2, .10) \approx .266$$

The probability that the second sale occurs in the second week is about 27%.