pmultinom

Alexander Davis

April 24, 2018

Contents

1 Readme

pmultinom is a library for calculating multinomial probabilities. The probabilities that can be calculated include the multinomial cumulative distribution function:

$$P(N_1 \leq u_1, N_2 \leq u_2, \cdots, N_k \leq u_k)$$

In this case the usage would be `pmultinom(upper=us, size=n, probs=ps, method="exact")` where `us` is the vector containing $u_1, u_2, \cdots, u_k$, and `n` and `ps` are the parameters of the multinomial distribution. This usage is analogous to the use of `pbinom`. Another important case is the probability of seeing more than some minimum number of observations in each category:

$$P(N_1 > l_1, N_2 > l_2, \cdots, N_k > l_k)$$

In this case the usage would be `pmultinom(lower=ls, size=n, probs=ps, method="exact")` where this time `ls` is the vector containing $l_1, l_2, \cdots, l_k$. Notice that in this case these are greater than signs, not greater than or equal signs. This is analogous to the usage of `pbinom` with `lower.tail=FALSE`. With some creativity, these can be adapted to calculate the probability that the maximum or minimum of a multinomial random vector is a given number, or that a given category will be the most or least observed. `pmultinom` also supports a more general usage, in which both lower and upper bounds are specified:

$$P(l_1 < N_1 \leq u_1, l_2 < N_2 \leq u_2, \cdots, l_k < N_k \leq u_k)$$

In this case the usage would be `pmultinom(lower=ls, upper=us, size=n, probs=ps, method="exact")`

2 Example application

The following calculation is inspired by the 2018 paper Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing by Casasent et al. We will focus on one patient in the study, Patient 6. This patient had invasive ductal carcinoma, a form of breast cancer. 204 of these invasive cells were isolated, and their DNA was sequenced individually. On the basis of the different genetic mutations in each cell, the cancer cells were divided into 5 groups, called "subclones". There were 43 cells in the first subclone, 20 in the second, 82 in the third, 17 in the fourth, 5 in the fifth, plus 37 stromal cells (a "stromal" cell is a non-cancerous cell mixed in to the tumor). This vector of numbers can be regarded as the outcome of a draw from a multinomial distribution. The sample size parameter is $n = 204$. The category probabilities are unknown, but can be estimated from this observation to be the number of cells in each subclone, divided by the sample size. This estimation ignores the possibility of additional, unobserved clones, but nevertheless this is the estimate that will be used in this example.

The question we will answer is: in a tumor with these subclone frequencies, is 204 cells enough to reliably observe all five of these subclones? We will requre that 2 cells be sequenced from each subclone. This is a multinomial probability of the type that `pmultinom` can calculate.

So, the first calculation: what is the probability of observing at least 2 cells from each of the subclones? First, we need to define the parameters of the multinomial distribution.

```
> library(pmultinom)
> ncells <- 204
> subclone.freqs <- c(43, 20, 82, 17, 5, 37)/ncells
```

The vector `subclone.freqs` contains the relative frequency of each subclone, plus one value which represents the frequency of stromal cells. Next, we need to define how many cells we want from each category:

```
> target.number <- c(2, 2, 2, 2, 2, 0)
```

It's 2 for every subclone of the cancer, and 0 for the stromal cells, which are not important to observe since they are not cancerous and have no detectable genetic mutations. Before we can use these numbers, however, since `pmultinom` uses less than signs for lower bounds, we need to subtract one:

```
> lower.bound <- target.number - 1
```

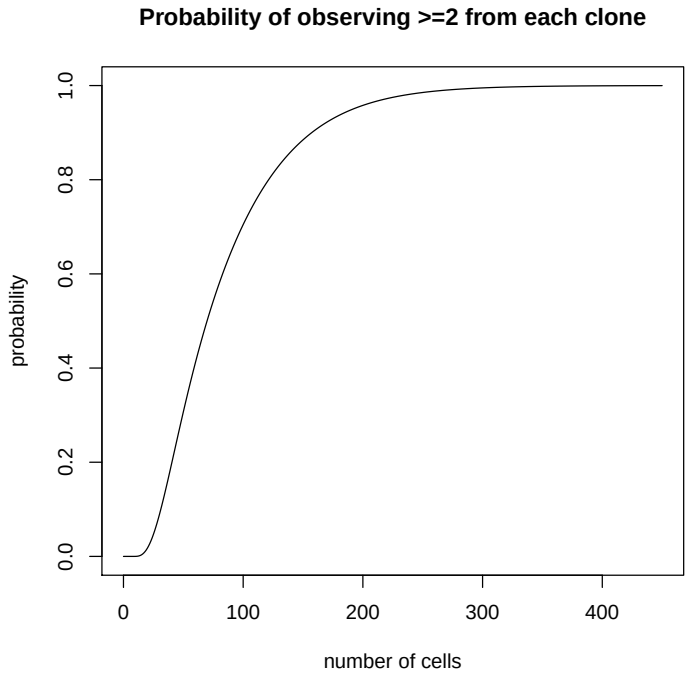
Now, we can use `pmultinom` to calculate the probability of observing at least three cells from each subclone, with the number of cells that were actually sequenced.

```
> pmultinom(lower=lower.bound, size=ncells, probs=subclone.freqs, method="exact")

[1] 0.9612182
```

We can also calculate this probability for all possible sample sizes. Since `pmultinom` is vectorized over `size`, this is done by inputting a vector of sizes as the `size` argument:

```
> xvals <- 0:450
> path <- pmultinom(lower=lower.bound, size=xvals, probs=subclone.freqs, method="exact")
> plot(xvals, path, type='l',
+      main="Probability of observing >=2 from each clone",
+      xlab="number of cells", ylab="probability")
```



In addition, using `invert.pmultinom`, we can calculate how many cells would be required to achieve a 95% probability of observing 2 from each subclone:

```
> invert.pmultinom(lower=lower.bound, probs=subclone.freqs,
+                  target.prob=.95, method="exact")

[1] 192
```