Stat 610 Lab 4/Homework 7

Due Wednesday, November 5, 11:59pm

Assignment

You will experiment with fitting distributions to some baseball statistics. In particular, we will look at how well the number of home runs a player hits in a season can be modeled by the Poisson and negative binomial distributions.

1. The data are available in the `Lahman` package. Install with `install.packages('Lahman')`, and then get the data using `library(Lahman)` and `data(Batting)`. The `data` command will add a data frame called `Batting` to your environment.

   We will be interested in modeling the number of home runs hit by each player in 2018. These are available with `subset(Batting, yearID == '2018')$HR`. You should save them as something, e.g. `hrs <- subset(Batting, yearID == '2018')$HR`.

2. Fitting a Poisson by eye: Suppose that we are modeling `hrs` as coming from a Poisson distribution. Make a plot like the one shown in the notes (https://jfukuyama.github.io/teaching/stat610/notes/lecture16.html#(20)) that shows the empirical and expected fraction of the time we observe `hrs = i`, for `i` between 0 and 30. For the "expected" part of the plot, you will need to choose what mean parameter to pick for the Poisson. Experiment with a couple and pick the one that you like the best.

3. "Best-fitting" Poisson: It turns out that both method of moments and maximum likelihood estimation tell you that the "best-fitting" Poisson will be the one whose mean parameter matches the empirical mean (the mean of `hrs`).

   Make the same sort of plot you did in the previous question, but with the mean of the Poisson set to be the mean of `hrs`. What do you think of the fit?

4. Negative binomial by eye: The negative binomial, like the Poisson, is a distribution that is supported on $0, 1, 2, \ldots$. However, it has two parameters instead of just one, meaning that its variance can be different from its mean (for the Poisson distribution the mean equals the variance).

   There are a number of different parameterizations of the negative binomial, but for the purposes of this exercise the easiest is probably the one given at https://mc-stan.org/docs/2_20/functions-reference/nbalt.html. The parameter μ at the linked site is called `mu` in the `dnbinom` function in R, and the parameter ϕ at the linked site is called `size`. μ gives the mean of the distribution, and the variance is $\mu + \mu^2/\phi$. Note that since ϕ is always positive, the variance of a negative binomial distribution is always larger than its mean, and how much larger is determined by ϕ.

   Make the same sort of plot as in question 2, using the `dnbinom` function, setting `mu` to equal the empirical mean, and experimenting with different values of `size`.

5. Negative binomial by method of moments: Remember that in method of moments, we choose parameters for the distribution by matching the theoretical moments of the distribution to the empirical moments of the data. Find parameters μ and ϕ (`mu` and `size`) so that the mean and variance of a NegativeBinomial(μ, ϕ) distribution match the empirical mean and variance of `hrs`. Make the empirical/expected plot again, and compare the fit of this negative binomial to the Poisson fit in question 3.

6. Negative binomial by maximum likelihood. Use `fitdistr` (in the `MASS` package) to find the maximum likelihood estimates of `mu` and `size`. Make the empirical/expected plot one last time, and compare the fit of this negative binomial to the method of moments negative binomial fit.

Submission parameters

– Submit an `.Rmd` and `.pdf` or `.html` file with your answers to the questions, your code, and a description of what it is doing.