MAST30027: Modern Applied Statistics

Assignment 3, 2019

Due: 9:00am Monday Oct 14th

This assignment is worth 13% of your total mark. To get full marks, show your working including derivation and the R code you use.

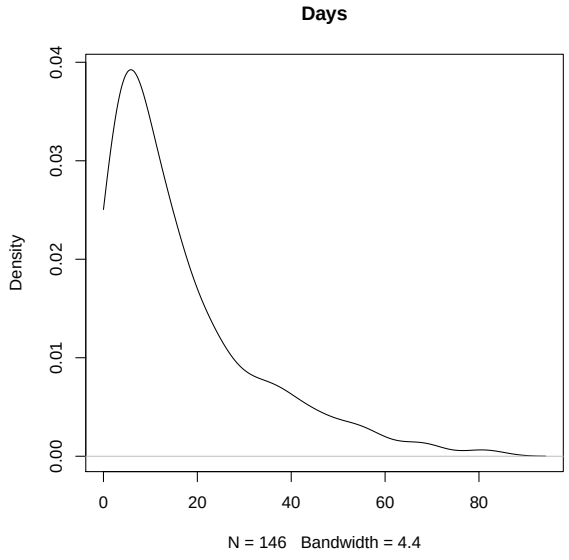
1. The negative binomial is often used for count data when the Poisson distribution is too restrictive, for example when the variance is noticably larger than the mean (overdispersion). For $p \in [0,1]$ and $r \in [0,\infty)$ we define the nbinom(r,p) mass function on $\{0,1,2,\ldots\}$ by

$$f(k|r,p) = \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)}(1-p)^r p^k.$$

Note that this definition does not require r to be integer valued.

We are going to use the negative binomial distribution to model data on absenteeism. The `quine` data frame from the `MASS` package has 146 rows and 5 columns. Children from Walgett, New South Wales, were classified by Culture, Age, Sex and Learner status and the number of days absent from school in a particular school year was recorded.

```
> library(MASS)
> data(quine)
> plot(density(quine$Days, from=0), main="Days")
```



(a) Suppose that we fix $r = 1.5$.
Find the MLE of p, using the `quine` data.

(b) Again fixing $r = 1.5$, suppose that p has a beta(1/2, 1/2) prior distribution.
Find the posterior distribution of p, conditioned on the `quine` data.

(c) The posterior mean is slightly smaller than the MLE estimate.
Will this always be the case? Why?

(d) We will now suppose that r is not known.

Let r have an exponential prior with mean 1.5, and p still have a beta(1/2, 1/2) prior. Write $\mathbf{y} = (y_1, \ldots, y_n)$ for the observed data ($n = 146$). Show that the posterior of r satisfies

$$f(r|\mathbf{y}) \propto \left[\prod_i \frac{\Gamma(y_i + r)}{\Gamma(r)}\right] \beta(y_\cdot + 1/2, nr + 1/2)e^{-(2/3)r}, \tag{1}$$

where $y_\cdot = \sum_i y_i$.

2. Gamma random variables can be used to simulate chi-square, t, F, beta, and Dirichlet distributions, as well as being useful in their own right. Hence it is important to be able to generate gamma r.v.s as efficiently as possible. In this assignment we investigate a popular algorithm due to Marsaglia and Tsang, for $\alpha \geq 1$.[1]

(a) Show that if $X \sim \text{gamma}(\alpha, 1)$ then $X/\lambda \sim \text{gamma}(\alpha, \lambda)$.

(b) Show that if X has density $h(x)^{\alpha-1}e^{-h(x)}h'(x)/\Gamma(\alpha)$ then $Y = h(X) \sim \text{gamma}(\alpha, 1)$.

You may assume that h is strictly increasing and maps the range of X onto $[0, \infty)$.

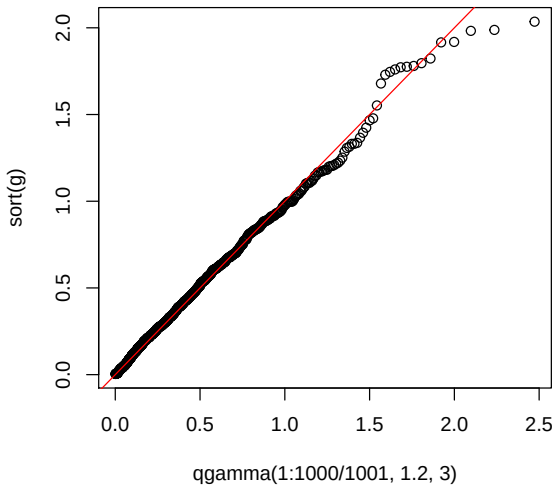
(c) Given $\alpha \geq 1$ put $d = \alpha - 1/3$ and $c = 1/\sqrt{9d}$ and then define $h : [-1/c, \infty) \to [0, \infty)$ by $h(x) = d(1 + cx)^3$. You can easily check that for this choice of h we have

$$\begin{aligned} h(x)^{\alpha-1}e^{-h(x)}h'(x) &\propto \exp(g(x)) \quad \text{where} \\ g(x) &= d\log((1 + cx)^3) - d(1 + cx)^3 + d. \end{aligned}$$

Less easily checked, but also true, is that $\exp(g(x)) \leq \exp(-x^2/2)$ on $[-1/c, \infty)$. (Checking those two facts is not required for the assignment 3.) Use these facts and the results from (a) and (b) to come up with an algorithm for simulating from gamma(α, λ). You may assume that you can already simulate from the standard normal distribution. Code up your algorithm and use it to generate 1000 gamma$(1.2, 3)$ pseudo-random variables. Demonstrate that your algorithm is working using a q-q plot.

The following R commands show how to make q-q plot for 1000 samples generated by g <- rgamma(1000, 1.2, 3). For your assignment, instead of rgamma(1000, 1.2, 3), you should write your own code which implements your algorithm.

```
> g <- rgamma(1000, 1.2, 3)
> plot(qgamma(1:1000/1001, 1.2, 3), sort(g))
> abline(0, 1, col="red")
```



[1] G. Marsaglia and W.W. Tsang, A simple method for generating gamma variables. ACM Trans. Math. Software, 26:363–371, 2000.