Notes on the Negative Binomial distribution for word occurrences

Edoardo Airoldi, Carnegie Mellon University (eairoldi@cs.cmu.edu)

The Negative-Binomial distribution can be obtained from expansion of $(Q-P)^{-r}$, where $Q=(1+P)$, $P>0$, and and r is positive real. Note that P need not be in $(0,1)$. The probability distribution is then

$$P(X=x) \;=\; \binom{r+x-1}{r-1}\left(1-\frac{P}{Q}\right)^{r}\left(\frac{P}{Q}\right)^{x}, \quad x\geq 0.$$

In this parameterization, mean $=rP$, and variance $=rP(1+P)$.

Airoldi et al. (2005) set $r=\kappa$, $P=\omega\delta$ and $Q=(1+P)=(1+\omega\delta)$. Keeping r instead of κ, we obtain

$$\begin{aligned} P(X=x) &= \binom{r+x-1}{r-1}\left(1-\frac{\omega\delta}{1+\omega\delta}\right)^{r}\left(\frac{\omega\delta}{1+\omega\delta}\right)^{x} \\ &= \binom{r+x-1}{r-1}(1+\omega\delta)^{-(x+r)}\,(\omega\delta)^{x}, \quad x\geq 0, \end{aligned}$$

where $\omega>0$, $\delta>0$, $r>0$. In order to expose the connections between Poisson and Negative-Binomial, we introduce a redundant parameter $\mu=r\delta$. The Negative Binomial with parameters $(r,\omega\mu,\omega\delta)$ reduces to a Poisson with parameter $(\omega\mu)$ as $r\to\infty$ and $\delta\to 0$, for a fixed μ. The Negative-Binomial with parameters $(r,\omega\mu,\omega\delta)$ has mean $=\omega\mu$, and variance $=\frac{\mu}{\delta}\omega\delta(1+\omega\delta)=\omega\mu(1+\omega\delta)$, that is, the same mean as the corresponding Poisson, reduced limit of the Negative-Binomial, with an extra variability factor, $(1+\omega\delta)$. Thus the parameter δ is referred to as the extra-Poissonness parameter.

The standard parameterization,

$$Pr(X=x) \;=\; \binom{r+x-1}{r-1}p^{r}\,(1-p)^{x}, \quad x\geq 0,$$

is obtained by setting $p=\left(1-\frac{P}{Q}\right)=\left(1-\frac{P}{1+P}\right)$, since $\frac{P}{Q}\in(0,1)$.

The parameter ω is used to condition on the size of the documents in the $(r,\omega\mu,\omega\delta)$ parameterization. Let us consider the Poisson case, where the rate $\lambda=\omega\mu$, and let us assume that our observations are number of times a certain word occurs in a set of documents, with possibly different word-lengths. The new parameterization $\omega\mu$ breaks the rate into two parts: μ, which is the rate of occurrence of the word under study, say, in a thousand (or ℓ) consecutive words of text, i.e., the length of the reference text in terms of number of words; and ω, which is the length of a document expressed as a pure number, multiple of the word-length of the reference text, e.g., $\omega=1.67$ for a text 1670 word long if the reference text is a thousand words, i.e., $\ell=1000$. This allows us to express the rate λ as the rate of occurrence of a word in a text of length equal to that of the reference text, μ, conditionally on the desired, or observed, length of the text, ω, expressed as a multiple of the word-length of the reference text.

References

E.M. Airoldi, A.G. Anderson, S.E. Fienberg, K.K. Skinner. Who wrote Ronald Reagan's radio addresses? *Bayesian Analysis*. To appear in 2005.

N.L. Johnson, S. Kotz, A.W. Kemp. *Univariate Discrete Distributions*. Wiley, 1993.

F. Mosteller, D.L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

F. Mosteller, D.L. Wallace. *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers*. Springer-Verlag, 1984.