Poisson Distributions

Poisson distributions are a family of probability distributions commonly used for modelling count data. Examples of count data include:

- Number of children delivered in one day at each of 10 hospitals.
- Number of mosquito bites received in an hour by a participant after being treated with an insect spray

Because a count cannot be less than zero, Poisson distributions are defined only on the positive reals and are skewed rather than symmetric. In addition, any Poisson distribution has the property that its mean is equal to its variance.

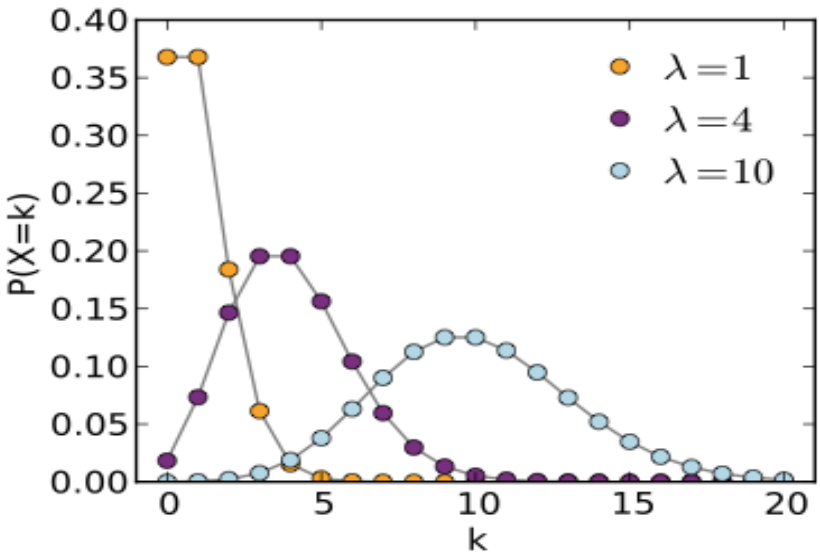
The family of Poisson distributions is parametrised by mean value, μ. The formula is:

$$P(x, \mu) = \frac{e^{-\mu}\mu^{x}}{x!}$$

where

- e: A constant equal to approximately 2.71828. (e is the base of the natural logarithm system.)
- μ: The mean number of successes that occur in a specified region.
- x: The actual number of successes that occur in a specified region.
- P(x,μ): The Poisson probability that exactly x successes occur in a Poisson experiment, when the mean number of successes is μ.

The curves for various values are given below (here λ is used for the mean instead of μ, image courtesy of Wikepedia “Poisson Distribution”).



How to choose between a Poisson model and another model for a response variable

There are various models that may be appropriate for count data. This include:

Normal distribution—may be used if all of the counts are fairly large, such as, number of red blood cells in a cubic mm of blood.

Poisson distribution—should be used if counts range from zero up, and the mean of the sample is similar to the variance (say, within a factor of 2).

Negative binomial distribution—a larger class of distributions (which includes the Poisson distributions) that should be used if the counts range from near zero up and the variance of the sample is larger than the mean. Negative binomial distributions have two parameters, mean μ and spread, often denoted by r. As r gets larger, the distribution approaches the Poisson distribution with mean μ. Lower values of r correspond to higher variance distributions. Generally in a negative binomial model, regression is done on the mean, that is, predictor variables are assumed to change the mean of the distribution but not the spread. However, it is also possible to set up a negative binomial model in which it is assumed the predictors change the spread parameter and keep the mean fixed.

Zero truncated models—Either a Poisson or negative binomial model may be modified to a zero truncated model, which is used in cases where zero counts will not appear. A common example is to model length of hospital stay for patients, where patients are only included in the study if they have stayed at least one night in hospital.

Zero inflated or Hurdle models—Either Poisson or negative binomial models may also be modified to zero inflated or hurdle models, which are two models for situations in which zeros are overrepresented. These are based on an assumption that some mechanism separates those units where counts may be greater than zero from those units where counts are always equal to zero. An example might be "number of convictions" measured on men between the ages of 18 and 22. Most men will have none, so there will be many more zeros than predicted by either a regular Poisson or negative binomial model. Although the technical forms of zero inflated and hurdle models are different, in practice they tend to deliver similar results. (See, eg, Zorn, "An Analytic and Empirical Examination of Zero-Inflated and Hurdle Poisson Specifications", Sociologial Methods and Research, Feb 1992, vol 26 no. 3, 368-400.)