# Frequency Distributions

## Negative Binomial

The R function dnbinom for the negative binomial uses two parameterizations. Style 1 is:

$$p(y|n,p) = \frac{\Gamma(y+n)}{\Gamma(n)y!}p^n(1-p)^y$$

This has mean $\mu = n(1-p)/p$ and variance $n(1-p)/p^2 = \mu/p$. The ratio of the third central moment to the variance (second central moment) is $\mu_3/\mu_2 = (2-p)/p = 2/p - 1$. For frequency distributions this ratio is often simpler than the skewness $= \mu_3/\mu_2^{3/2}$.

Style 2 makes $\mu$ a parameter. Consider another parameter $\sigma = (1-p)/(p\mu)$. Then $\mu + \sigma\mu^2 = \mu/p$, so is the variance. The dnbinom function sets a parameter $size = 1/\sigma$ for this style. In the first style, $n$ is called $size$ . Also, $p = 1/(1+\sigma\mu)$, so the third moment comes out with $\mu_3/\mu_2 = 2/p - 1 = 1 + 2\sigma\mu$.

The two styles are not different for a single sample being fit by a distribution. The parameters map into each other and give exactly the same fitted distribution. The difference comes when they are used as residual distributions for a model fit to a number of cells. Then typically one parameter is held constant across all the cells and the other fit for each cell. For style 1, $p$ is usually held constant and $n_j$ fit by cell. Then the variance of any cell is $\mu_j/p$. For style 2, it is natural to fit $\mu_j$ by cell and fix $\sigma$. This variance has to pick up random fluctuations from the cell mean as well as estimation error for the mean. The bigger variances for larger cells in style 2 generally work better for this.

In style 1 it is possible to fix $n$ and fit $p_j$ by cell. Then

$$\mu + \frac{\mu_j^2}{n} = n\frac{1-p_j}{p_j} + n\frac{(1-p_j)^2}{p_j^2} = n\frac{1-p_j}{p_j}\left[1 + \frac{1-p_j}{p_j}\right] = n\frac{1-p_j}{p_j^2} = Variance_j$$

Thus this gives style 2 with $\sigma = 1/n$. The two styles thus come down to the choice of which parameter to fix.

## Poisson Inverse Gaussian

This distribution is a mixture of the Poisson by the inverse Gaussian, so is more skewed than the negative binomial, which is the Poisson mixed by a gamma. It was introduced to the actuarial literature in G. E. Willmot (1987) and a similar article Dean, Lawless, and Willmot (1989). The formulas below come from Rigby, Stasinopoulos, and Akantziliotou (2008) and the documentation for the inverse Gaussian in the R package gamlss.dist. The probability function is:

$$p(y|\mu,\sigma) = \frac{2\sqrt{\alpha}}{\pi}\frac{\mu^y e^{1/\sigma}K_{y-0.5}(\alpha)}{(\alpha\sigma)^y y!}$$

where $\mu, \sigma$ are positive parameters, $K$ is the modified Bessel function of the second kind (also called the third kind, but archaic), and $\alpha = \sqrt{2\sigma+1}/\sigma$. It has mean $\mu$ and variance $\mu + \sigma\mu^2$, which match those of the second style of the negative binomial.

The skewness comes from $\mu_3/\mu_2 = 1 + 2\sigma\mu + (\sigma\mu)^2/(1+\sigma\mu)$, so is higher than the same quantity for the negative binomial by the last term.

A first style form of the Poisson-inverse Gaussian can be obtained by setting $p = 1/(1+\sigma\mu)$. This gives variance $= \mu/p$. Then $\mu_3/\mu_2 = 2/p - 1 + (1-p)^2/p$, and the ratio of the Poisson-inverse Gaussian skewness to that of the negative binomial is $1 + (1-p)^2/(2-p) = [3 - 3p + p^2]/[2-p]$.

## Generalized Poisson

This distribution, also called the Lagrangian Poisson, includes the Poisson as well as less dispersed distributions, similar to the binomial, and more dispersed distributions, similar to the negative binomial. One way to parameterize it is with two parameters $\alpha > 0$ and $\beta > 0.5$. It needs a scaling function $g(\alpha, \beta)$ that is identically 1 for $\beta > 1$ and usually is close to 1 in any case. Also when the variance is less than the mean ($\beta < 1$), the probability is zero over some cutoff count value, as it is for the binomial. The probability function is produced by

$$\frac{p(y|\alpha,\beta)}{g(\alpha,\beta)} = \frac{\alpha}{y!}(\alpha + y - y/\beta)^{y-1} e^{y/\beta - \alpha - y}$$

The probability is 0 in case the exponentiated term is positive, so $p(y) = 0$ if $y/\beta - y > \alpha$, which needs $\beta < 1$. The Poisson is the case with $\beta = 1$.

The mean and variance are given by $\alpha\beta$ and $\alpha\beta^3 = \beta^2\mu$. $\mu_3/\mu_2 = 3\beta^2 - 2\beta$. To compare this with the negative binomial with the same mean and variance (which requires $\beta > 1$), solve for $\alpha\beta = n(1-p)/p$ and $\beta^2\mu = \mu/p$. This gives $\beta^2 = 1/p$, so for the generalized Poisson, $\mu_3/\mu_2 = 3/p - 2/\sqrt{p}$. Dividing this by $2/p - 1$ shows that the ratio of the generalized Poisson skewness to that of the negative binomial is $[3 - 2\sqrt{p}]/[2-p]$. This is greater than 1, but less than the ratio for the Poisson-inverse Gaussian. The ratios are compared in Figure 1.

A good resource for this distribution is the R package RMKdiscrete. It includes calculation of $g(\alpha, \beta)$, which is often ignored. Basically this is what it has to be for the probabilities to sum to 1. This could be done by adjusting the probability at zero, but that would change the moments. Usually the distribution is defined in terms of $\rho = 1 - 1/\beta$, which is in $(-1, 1)$. That makes the probability function a little simpler, but still $\beta$ is used instead of $\rho$ in the moments. With this, $p(y) = 0$ for $\rho < 0, -\rho y > \alpha, \rho < -\alpha/y$.

It is also possible to define a style 2 GP, with Variance $= \mu + \sigma\mu^2$. Given $\sigma, \mu$, set $\beta^2 = 1 + \sigma\mu$, and $\alpha = \mu/\beta$. Then Variance $= \mu + \sigma\mu^2 = \beta^2\mu$, and $\mu_3/\mu_2 = 3(1+\sigma\mu) - 2\sqrt{1+\sigma\mu}$. Note that $1 + \sigma\mu =$ Variance/mean.

## Sichel

The Sichel is a three-parameter extension of the Poisson-inverse Gaussian, with another shape parameter $\nu$ that can be any real number. It is the Poisson mixed by the generalized inverse Gaussian. The Poisson-inverse Gaussian is the case $\nu = -0.5$. The negative binomial is a limiting case. Using the notation from the Poisson-inverse Gaussian section, set $c = K_{y-0.5}(1/\sigma)$ and now let $\alpha = \sqrt{2\sigma/c + 1}/\sigma$. Then the probability function is:

$$p(y|\mu,\sigma,\nu) = \frac{(\mu/c)^y K_{y+\nu}(\alpha)}{y!(\alpha\sigma)^{y+\nu} K_\nu(1/\sigma)}$$

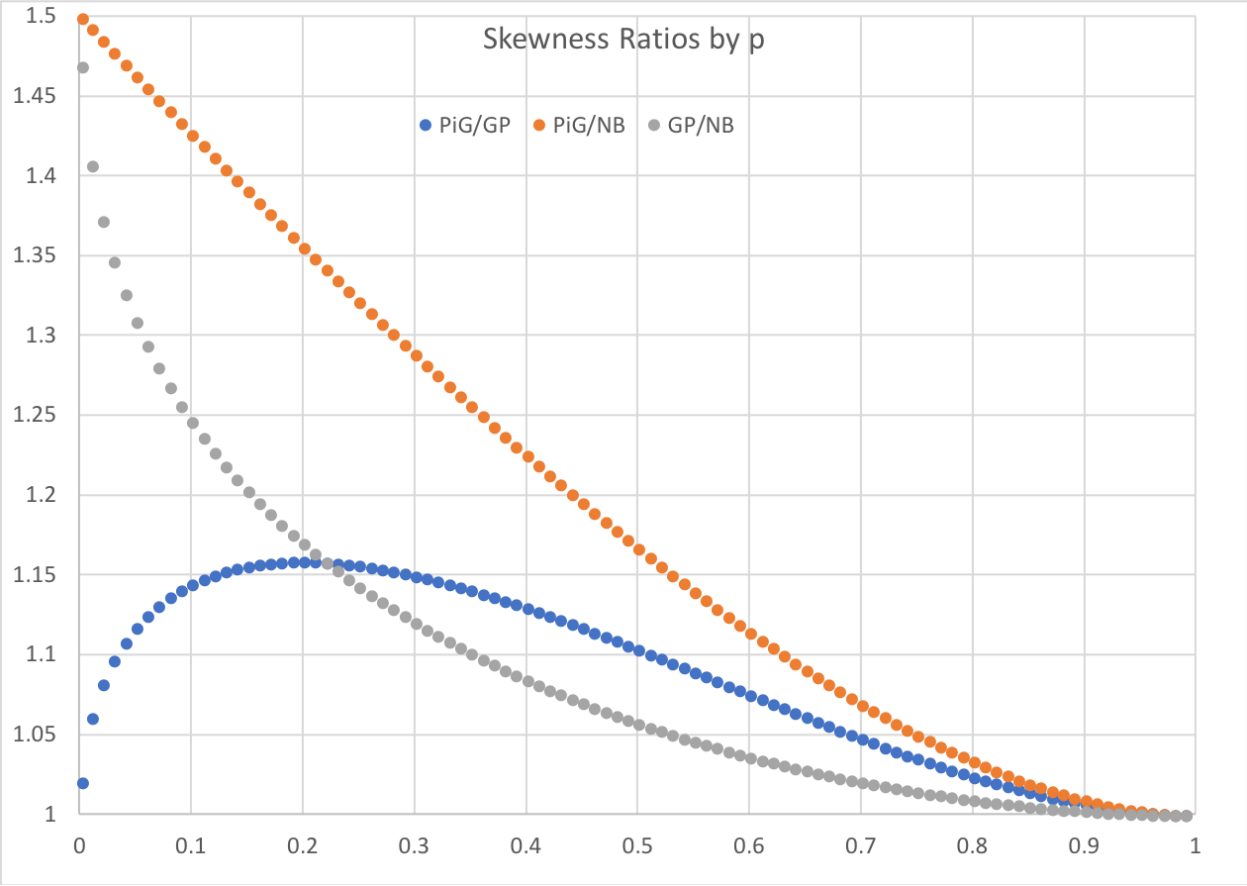


Figure 1: Skewness Ratios for Negative Binomial, Generalized Pareto and Poisson-Inverse Gaussian

Table 1: Claims Reported by Lag

| Lag: | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1978 | 368 | 144 | 16 | 12 | 8 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1979 | 393 | 191 | 25 | 10 | 7 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1980 | 517 | 151 | 28 | 9 | 2 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1981 | 578 | 185 | 25 | 8 | 8 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1982 | 622 | 254 | 29 | 6 | 6 | 2 | 4 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1983 | 660 | 206 | 49 | 17 | 4 | 5 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1984 | 666 | 243 | 39 | 22 | 11 | 5 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1985 | 573 | 234 | 28 | 16 | 17 | 10 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1986 | 582 | 266 | 53 | 12 | 3 | 6 | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1987 | 545 | 281 | 62 | 10 | 12 | 7 | 3 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1988 | 509 | 220 | 32 | 12 | 8 | 4 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 1989 | 589 | 266 | 43 | 27 | 5 | 4 | 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1990 | 564 | 210 | 49 | 18 | 12 | 7 | 6 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 1991 | 607 | 196 | 29 | 22 | 12 | 13 | 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1992 | 674 | 203 | 23 | 17 | 15 | 9 | 6 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1993 | 619 | 169 | 29 | 12 | 12 | 4 | 5 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1994 | 660 | 190 | 20 | 9 | 9 | 4 | 8 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1995 | 660 | 161 | 41 | 12 | 7 | 5 | 9 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

This can be more or less skewed than the Poisson-inverse Gaussian, but is more skewed than the negative binomial having the same mean and variance. The mean is just $\mu$ and the variance is $\mu + g_1\mu^2$, where $g_1 = 2\sigma(\nu+1)/c + 1/c^2 - 1$. If $\nu$ is not too far from -0.5, $c$ is close to 1, and $g_1$ is close to $\sigma$.

The skewness is more complicated. Set $g_2 = 2\sigma(2+\nu)/c^3 + 4\sigma^2(1+\nu)(2+\nu)/c^2 + 1/c^2 - 1$. Then $\mu_3 = \mu + 3\mu^2 g_1 + \mu^3(g_2 - 3g_1)$. Note that $g_1, g_2$ are shape constants, not involving $\mu$, so would not change by cell if only $\mu_j$ varies by cell.

## Example

Taylor (2000), Appendix B3.1, has a claim count reporting dataset. See Table 1. I estimated a row-column factor model for that, fitting piecewise-linear curves to the log of the factors by row and column using Bayesian shrinkage on the slope changes, with residuals from the frequency models discussed above. This is undoubtedly an overly-simplistic model, as there appears to have been a shift towards earlier reporting, but it will serve to illustrate some of the issues. The fitting was done in the R BayesianTools package, which can do Metropolis sampling from any model with the likelihood defined using R functions. The rows here are calendar years of report, so this is actually a separation model, as in Verbeek (1972).

The row-factor column model with piecewise-linear curves and Bayesian shrinkage on slope changes can be set up with the observations (270 of them here) in a column and a design matrix with dummy variables for the row and column parameters. Here those parameters are slope changes, which are actually the second differences of the log factors. Numbering rows and columns starting with 1, the slope changes keep adding up for subsequent cells. There is a constant term in the model that is not shrunk, so row 1 and column 1 both get zero log factors. The row 2 slope change is the log factor for all the cells in row 2. Then it is added to the row 3 slope change to get the log factor for row 3. Then the row 2 and 3 slope changes are both added to

Table 2: Penalized Likelihood for Several Distributions

| Model | NB1 | GP1 | NB2 | GP2 | PIG2 | Sichel |
|---|---|---|---|---|---|---|
| Loo | -531.2 | -530.3 | -526.5 | -526.4 | -526.3 | -526.1 |
| Penalty | 12.3 | 11.8 | 9.7 | 9.7 | 9.7 | 9.0 |

this sum and the slope change for row 4 to get the log factor for row 4, etc. For a cell from row $i$, the dummy variable for row $u$ gets value max(0,1+i-u), which is the number of times the row $u$ parameter is added up for row $i$. The columns work the same.

Shrinking the slope changes towards zero smooths out the piecewise-linear curves. If a slope-change parameter is zero, the previous slope continues at that point. Bayesian shrinkage accomplishes this by giving the slope changes mean-zero priors. I used the double-exponential prior, which has a parameter $s$ representing it's variance. The goodness of fit is indicated by penalized loglikelihood, where the penalty comes from a cross-validation approach. The model is fit on the dataset excluding one point successively for each point in the sample, and the loglikelihood is computed on the omitted point. The sum of these gives the leave-one-out penalized loglikelihood. I'll just call that loo. It is usually negative, and the higher the better.

Loo can be computed for various values of $s$, but the fully Bayesian approach is to put a prior on $s$ as well. I usually put a uniform prior on $log(s)$ and adjust the endpoints if the range of the resulting posterior distribution of $log(s)$ is pushing up against an endpoint. In practice, doing this seems to come up with values of $s$ close to what optimizing loo would do.

For a model with a lot of factors – this one has 14 column and 17 row factors to estimate – a good number of them get close to zero. Eliminating those variables from the model, so making them exactly zero, often improves loo as well. But it is not easy to tell which ones to eliminate, as they can work together in groups. A practical way to start off the process is to use classical lasso for the logs of the claims counts with just a normal distribution assumption. The R package glmnet has a function cvfit that uses a different form of cross validation to pick a range of smoothing constants. Lasso pushes some of the variables to exactly zero, and it is good at finding groups of variables that work together. An output from it, cvfit$lambda.min, estimates the smoothing that allows the most variables to remain in the model that is consistent with its cross-validation standards. I use the variables from that as a starting point for Bayesian shrinkage. Usually a few more will end up near zero. I eliminate those as well as long as it does not degrade loo to do so.

The result here left only two slope changes for the years – years 2 and 8 – and five for the reporting periods – periods 2,3,4,5, and 7. (Subtract 1 from period to get lag.)

Figure 2 shows the logs of the resulting piecewise-linear curves for the row and column factors from the best-fitting frequency distribution model, discussed below.

Table 2 shows the loo measure and the parameter penalty for this model fit with several frequency distributions for residuals. The style 2 models provide better fits than the style 1 models. The penalty has to do with how well the model fits the left out points. The Sichel model, with the most parameters, has the best loo and the lowest penalty. It may have done that with more shrinkage. Still, there is little difference in the goodness of fit among the style 2 models. Still, if range estimates are important in an application, the best-fitting model is still probably advantageous.

The Sichel fit had $\nu = -0.5829$, which is close to the -0.5 value assumed by the PiG. As a result, $c = 0.9979$ is

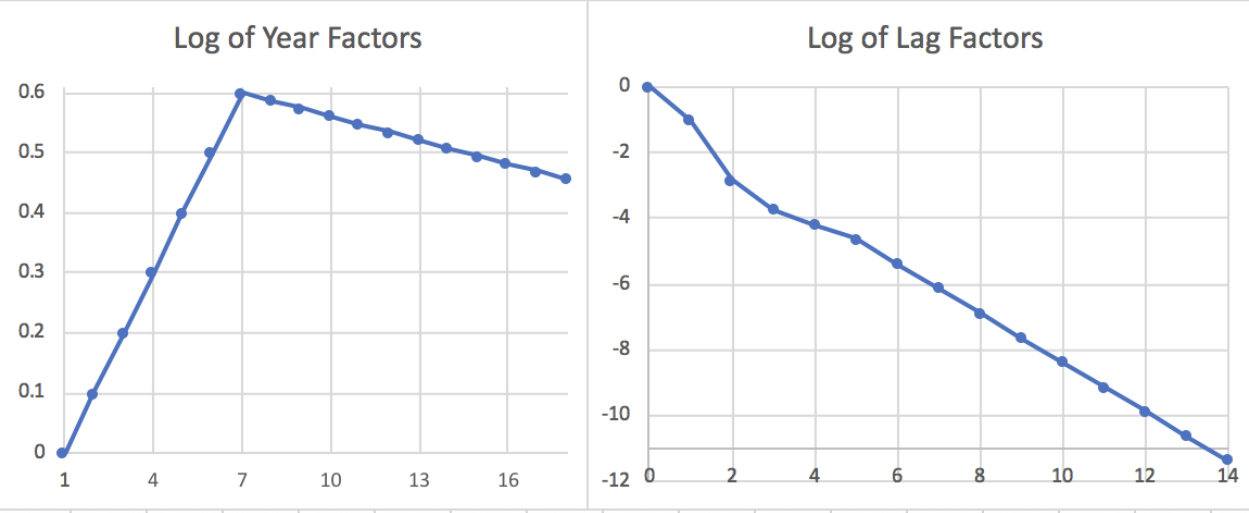


Figure 2: Log of Year and Lag Factors

Table 3: 1/s and Skewness for Selected Cell Means

| Model | 1/s | Mean 0.5 | Mean 5 | Mean 50 | Mean 500 |
|---|---|---|---|---|---|
| NB2 | 42.0 | 1.439 | 0.523 | 0.320 | 0.309 |
| GP2 | 42.6 | 1.439 | 0.524 | 0.343 | 0.389 |
| PiG2 | 42.5 | 1.439 | 0.528 | 0.386 | 0.443 |
| Sichel | 39.2 | 1.424 | 0.487 | 0.355 | 0.451 |

close to 1, and $1/g_1 = 39.20759$ is close to $1/\sigma = 39.20845$. Recall that the variance for the Sichel is $\mu + g_1\mu^2$. Table 3 shows the skewness for each of the style 2 fits at the mean parameters for hypothetical means of 0.5, 5, 50, and 500, which are all in the range of the cell means for this data. Also shown is $1/s$, which is $1/g_1$ for the Sichel and $1/\sigma$ for the other distributions.

The table illustrates the greater shape flexibility of the Sichel. Since $1/s$ is lower, the variances are higher than for the other distributions, but the skewnesses are lower for the smaller means, and higher for the larger means.

# References


Dean, C., J.F. Lawless, and G.E. Willmot. 1989. "A Mixed Poisson-Inverse-Gaussian Regression Model." *The Canadian Journal of Statistics* 17:2: 171–81.

Rigby, R.A., D.M. Stasinopoulos, and C. Akantziliotou. 2008. "A Framework for Modelling Overdispersed Count Data, Including the Poisson-Shifted Generalized Inverse Gaussian." *Computational Statistics and Data Analysis* 53: 381–93.

Taylor, Gregory. 2000. "Loss Reserving – an Actuarial Perspective." *Springer US*.

Verbeek, H. G. 1972. "An Approach to the Analysis of Claims Experience in Excess of Loss Reinsurance." *Astin Bulletin* 6: 195–202.

Willmot, Gordon E. 1987. "The Poisson-Inverse Gaussian Distribution as an Alternative to the Negative Binomial." *Scandinavian Actuarial Journal* 1987(3-4): 113–27.