Summary of the article 'Statistics of weighted Poisson events and its applications'

Esben Larsen Rasmussen and Lukas Fabian Ehrke

INTRODUCTION

The article introduces a new approximation for the distribution of the sum of random weights. The weight distribution is unknown. It s expected that the sum of the weights should follow a compound Poisson distribution (CPD), which is normally approximated using a normal distribution. The authors propose an estimation using a scaled Poisson distribution. The approximation can be used when estimating parameters. Additionally, they introduce a new Poisson bootstrap method which can be used to estimate confidence intervals.

SCALED POISSON DISTRIBUTION

The article starts out reviewing the CPD, and the expectation values for the mean (μ), variance (σ^2), skewness (γ_1) and excess (γ_2, also called kurtosis). They then examine the special case where all the Poisson processes of the CPD have the same probability. Having reviewed the CPD, the authors define the scaled Poisson distribution (SPD). The mean value of the SPD ($\tilde{\lambda}$) is defined as:

$$\tilde{\lambda} = \lambda \frac{E(w)^2}{E(w^2)} = \mu \frac{E(w)}{E(w^2)} = \frac{\mu}{s} \tag{1}$$

where λ is the average of the CPD, w are the weights, μ is the average of the weighted distribution, and $s = E(w^2)/E(w)$ is a scale factor depending only on the mean and variance of the weights and not on the individual weights. They then define a scaled variable $\tilde{x} = s\tilde{n}$, where $\tilde{n}$ is taken from a Poisson distribution with an average of $\tilde{\lambda}$. The last part ensures that $E(\tilde{x}) = E(x) = \mu$, and that $var(\tilde{x}) = var(x) = \sigma^2$. The skewness and the excess are given by $\tilde{\gamma}_1 = 1/\tilde{\lambda}^{1/2}$ and $\tilde{\gamma}_2 = 1/\tilde{\lambda}$. Afterwards, they derive an inequality for the ratio between the CPD and SPD ($\gamma_1/\tilde{\gamma}_1, \gamma_2/\tilde{\gamma}_2$) for both the skewness and the excess, showing that both ratios are equal to or greater than 1, with the equality holding if all weights are the same. Again this shows that the skewness and the excess is estimated better for the SPD than for the normal distribution, where both of these values are zero. However, both values being zero is not a desirable attribute when estimating a CPD, as they are inherently skewed. It should be noted the SPD approximation is only valid for positive weights.

Comparing the CPD, SPD and Normal Distribution

To compare the skew, the excess of the SPD, and CPD, one million events are simulated for different weight distributions. The number of weights is thereby randomly taken from a Poisson distribution with a mean of fifty. In table 1 the values are given. The weight distribution from which is sampled is given in the first column. For the first three rows a uniform distribution in the intervals $[0, 1]$, $[1, 2]$, and $[2, 3]$ are used. The next weight distribution is given by an exponential function. The following weight distribution is a truncated normal distribution with a mean and variance equal to one. In the last two rows the weight is either one or ten. The probability of assigning one as a weight is indicated in the parentheses. The second column gives the mean of the SPD calculated by equation 1. The last four columns give mean values of the skew and the excess of the CPD and SPD respectively. It is noted that for the exponential weighted distribution the difference between the CPD and SPD values are highest, as expected by the authors, since the SPD approximation becomes worse if the weights differ a lot. For each weight distribution the SPD is more similar to the CPD than the normal distribution.

Type of weight	$\tilde{\lambda}$	γ_1	γ_2	$\tilde{\gamma}_1$	$\tilde{\gamma}_2$
$u[0, 1]$	37.50	0.184	0.036	0.163	0.027
$u[1, 2]$	48.21	0.149	0.023	0.144	0.021
$u[2, 3]$	49.34	0.144	0.021	0.142	0.020
$\exp(-w)$	25.00	0.300	0.120	0.200	0.040
$\mathcal{N}_t(1, 1)$	36.48	0.199	0.045	0.166	0.027
1 ($p = 0.5$), 10	29.94	0.197	0.039	0.182	0.033
1 ($p = 0.8$), 10	19.01	0.299	0.092	0.229	0.052

Figure 1: Comparison between skewness and excess between SPD approximation and CPD

POISSON BOOTSTRAP

This method gives an estimate of parameter uncertainties or confidence intervals. It can be used for a set of observations x_i, $i = 1, 2, \ldots, n$, where each observation corresponds to one event with weight x_i. The observed value is given by $x_{obs} = \sum_i x_i$. For the Poisson bootstrap each value x_i is taken n_i times, where n_i is a Poisson distributed number with mean equal to one. In the method proposed in the article no results $x = \sum_i x_i n_i$ are thrown away. Nevertheless, by repeating this many times the confidence intervals can be estimated from the distribution of x.

APPLICATIONS

The authors look into some of the applications of the approximation by an SPD and the bootstrap method. Both methods can be used when the weight distribution is not known. Therefore the mean and variance of the weights are approximated by the empirical values.

Parameter Estimation

The first application is parameter estimation. Simulated weighted data and experimentally observed data in a histogram are used to accomplish this. In each bin j, there are m_j observed events, where m_j follows a Poisson distribution. For the simulated data each bin contains $x_j = \sum_i^{n_j} w_{ij}$, where for each generated event a weight is associated. The weight is dependent on the parameter that is to be estimated. To estimate the parameter the authors propose a Least Squares fit. For B bins the χ^2 expression is given by:

$$\chi^2 = \sum_j^B \frac{(cm_j - x_j)^2}{\delta_j^2} = \sum_j^B \frac{(cm_j - \sum_i^{n_j} w_{ji})^2}{\delta_j^2} \qquad (2)$$

where c is a normalization constant to account for different numbers of observed and generated events. The value of the denominator δ_j^2 must be estimated. This can only be done by first estimating the mean μ, which is assumed to be the mean of the two summations in the numerator. For the normal approximation the mean is estimated by:

$$\hat{\mu}_N = \left(\frac{cm}{c^2 m} + \frac{\sum w}{\sum w^2}\right) / \left(\frac{1}{c^2 m}\frac{1}{\sum w^2}\right) \qquad (3)$$

where the bin index j is suppressed. For the SPD approximation the mean can be estimated by optimizing a likelihood function. The result is:

$$\hat{\mu}_{SPD} = cs\frac{\tilde{n} + m}{c + s} \qquad (4)$$

where $\tilde{n} = x/s$. Then δ^2 can be approximated by:

$$\hat{\delta}^2_{SPD} = cs\,(\tilde{n} + m) \qquad (5)$$

The results can be seen in figure 2, where λ_n and λ_m are the expected values of events for the simulated data and the observed data respectively. The different weight distributions follow the same naming scheme as in figure 1. The mean, μ, should be estimated by the two methods. The following columns give the mean and the RMS of the estimates for the SPD and the normal approximation respectively. It is noticeable that the mean for the SPD is closer to the CPD for all simulations and that its standard deviation is smaller compared to the normal approximation.

Comparison of the estimates $\hat{\mu}_{SPD}$ from the SPD approximation and $\hat{\mu}_N$ from the normal approximation to the nominal mean value μ.

λ_n	λ_m	Weight	μ	$\hat{\mu}_{SPD}$	σ_{SPD}	$\hat{\mu}_N$	σ_N	$\frac{\hat{\mu}_{SPD}-\mu}{\mu}$	$\frac{\hat{\mu}_N-\mu}{\mu}$
20	20	$\exp(-x)$	20	19.98	3.68	19.10	3.84	0.001	0.045
10	10	$\exp(-x)$	10	9.73	2.64	9.11	2.81	0.027	0.089
10	50	$\exp(-x)$	10	9.88	1.38	9.58	1.59	0.012	0.042
20	50	$\exp(-x)$	20	19.84	2.61	19.46	2.74	0.008	0.027
50	50	$\exp(-x)$	50	49.78	5.79	49.12	5.91	0.004	0.013
10	10	$\mathcal{N}_t(1,1)$	12.88	12.78	3.13	12.05	3.30	0.008	0.068
20	20	$\mathcal{N}_t(1,1)$	25.75	25.67	4.40	24.93	4.53	0.003	0.032
20	50	$\mathcal{N}_t(1,1)$	25.75	25.69	3.22	25.27	3.31	0.002	0.019
10	10	$u[2,3]$	25.00	25.00	5.61	23.74	5.87	0.000	0.050
20	20	$u[2,3]$	50.00	50.00	7.94	48.74	8.13	0.000	0.025
50	50	$u[2,3]$	125.00	125.01	12.54	123.75	12.67	0.000	0.010

Figure 2: Comparison of the two estimates $\hat{\mu}_{SPD}$ and $\hat{\mu}_N$ to the nominal value μ

Approximation of Confidence Intervals

If there are n observations with weight w_i, the confidence intervals can be approximated by performing a Poisson bootstrap and taking integrals of the final distribution of x. An improvement to this method is proposed: The n_i are not taken from a Poisson distribution with mean equal to one, but with a mean equal to μ. The mean μ is chosen such that if one performs the bootstrap, the fraction α of the outcomes is below the observed value x_{obs}. The value $x_{obs} \cdot \mu$ gives the limit for the $1 - \alpha$ interval. For a number of observations n, where n is taken from a Poisson distribution with mean fifty, and a uniform weight distribution between zero and one, the authors get an observed value of $x_{obs} = 22.01$. In figure 3 confidence intervals for both methods are given. The second line gives the confidence limits for the first method, and the last one for the improved method. It is evident that the confidence limits for the improved method are shifted to higher values, which is expected.

α	0.01	0.05	0.10	0.1585	0.8415	0.90	0.95	0.99
CL	13.8	16.0	17.2	18.2	25.8	26.9	28.5	31.4
CL*	14.4	16.5	17.6	18.5	26.2	27.3	28.9	32.1

Figure 3: Confidence limits estimated by the two methods

SUMMARY

A compound Poisson distribution describes the sum of random weights. The number of weights is hereby Poisson distributed. It is shown that an approximation by a scaled Poisson distribution reproduces the higher moments better than a normal approximation. The SPD approximation also performs better than a normal approximation with regards to parameter estimation. In addition, a special bootstrap method has been introduced.

Using this one can estimate confidence intervals.