American Journal of Biostatistics 2 (2): 56-60, 2011
ISSN: 1948-9889
©2011 Science Publication
doi:10.3844/ajbssp.2011.56.60 Published Online 2 (2) 2011 (http://www.thescipub.com/ajbs.toc)

Is Poisson Dispersion Diluted or Over-Saturated? An Index is Created to Answer

Ramalingam Shanmugam

School of Health Administration,
Texas State University-San Marcos, San Marcos, TX 78666, USA

ABSTRACT

A prelude to interpret a pattern in the repeating incidences is to identify the underlying frequency distribution of the collected data. A case in point is the Poisson distribution which is often selected for medical count data such as gene mutations, medication error and number of ambulatory pickups in a day. A requirement for the Poisson distribution is that the variance ought to be equal to the mean. The variance signifies the volatility in the occurrences. An implication is that the volatility becomes more when the average incidence is higher. When this requirement of the functional equivalence of the Poisson mean and variance is breached, the data deviates from a Poisson distribution. How could a data analyst recognize and point out to the medical team the dilution level of the requirement in their data? For this purpose, a simple and easier geometrical approach is developed in this article and illustrated with several historical data sets in the literature.

Keywords: Data Mean and Variance, Correlation, Healthcare Management, Shifting Angle

1. INTRODUCTION

What is Poisson distribution? A genesis of Poisson distribution with a misnomer is intriguing. About 119 years ago, it was first introduced by de Moivre not by a French Probabilist Poisson, though the distribution is named Poisson. The Poisson distribution is frequently employed to explain uncertainty in count data such as the medication errors, adverse events, radioactive decay, traffic congestion, molecular mutations, ambulatory pickups of patients from their home as long as the data are about rarity (Cameron and Trivedi, 1986; Dalal *et al*., 1989; Davutyan, 1989; Deb and Trivedi, 1997; Winkelmann and Zimmermann, 1994; Thakur *et al*., 1980). For a chance mechanism to be governed by a Poisson distribution, three are assumptions which should be validated. The chance for any rare event to occur is proportional to the length of the time interval which is usually an infinitely small, the chance, for two or more rare events to simultaneously occur in a smaller time interval is slim and what happens in one time interval is stochastically independent of what happens in any other non-overlapping time interval.

A random variable Y in a background with all above three assumptions is called Poisson type. The probability mass function of Poisson random variable is Equation 1:

$$\text{poisson}(y,\lambda) = e^{-\lambda}\lambda^{y} / y! \qquad (1)$$

where, y = 0, 1, 2,…, a collection of observables is and $0<\lambda<\infty$ is the parameter space. The Poisson distribution is a member of the mean exponential family.

The Poisson probability model is popularly used to describe rare events such as arrival patterns in a queuing system, the number of decaying atomic elements in particle physics, the number of cancerous cells, the number of failing units in reliability discussions, the number of financial risk applicants seeking credit card, the number of false claims in auto insurance, the number of virus in toxicology studies, the number of foreign genes in bacteriology investigation, the number of traffic accidents in a highway during a time interval, the number of epileptic seizures in a patient and the number of cholera cases in a family during an epidemic among others. A unique property of Poisson distribution is the equality of mean and dispersion. That is, var (Y) = λ = E (Y). In real life data, this unique Poisson property is not completely seen for a variety of reasons. Either an over or under dispersion in Poisson type data is noticed. To fix this breach of the requirement, statisticians seek modified Poisson distribution. After considering a gamma probability density for the intensity parameter, the Poisson distribution gets convoluted into a version called inverse binomial distribution. In spite of this remedial approach, even the inverse binomial distribution has been found to be poorly fitting many Poisson type data. In this remedial process, an incidence rate restricted Poisson distribution was introduced. Still, the one often wonders about what causes all versions of Poisson distribution to fail to fit their data. Is the rarity of the event doubtful? Or, is it the lack or dilution of unique Poisson property of equal mean and dispersion? Statisticians experience a frustration to understand the poor Poisson fit. How could a data analyst recognize and point out to the medical team a breach of the requirement in a data? For this purpose, a simple and easier geometrical approach is developed in this article and illustrated with several historical data sets in the literature. This geometric view of over/under Poisson dispersion as introduced and explained in this article would help to capture the breaching level in Poisson type data of real life scientific, social, economic, finance, medical, engineering, business, public health and industrial data from the literature are considered in the illustration.

1.1. Geometric View of Over/Under Poisson Dispersion

Consider a random sample $y_1, y_2, ..., y_n$ of observations from a Poisson distribution in (1) with the incidence parameter 0<λ<∝. Let their sample mean and dispersion be $\bar{y} = \sum_{i=1}^{i=n} y_i / n$ and $s_y^2 = \sum_{i=1}^{i=n} (y_i - \bar{y}) / (n-1)$ respectively. It is well known that $\bar{y}$ and s_y^2 are unbiased estimator of their population counterparts μ and σ_y^2 respectively. The unique property of equal mean and dispersion is echoed in the mapping of dispersion in terms of the mean by a bisecting straight line OD passing through the coordinate (0,0) at an angle equal to 45° in **Fig. 1**.

When an over or under dispersion prevails, the straight line passes through the coordinate (0, 0) but at an angle larger (in the case of over dispersion as in **Fig. 1** or smaller (in the case of under dispersion as in **Fig. 2** than 45°. In other words, the line OD passing through the origin signifies perfect Poissonness in the data. For that, the point B should coincide with the point D. When the points B and D coincide, there is no dilution of Poisson dispersion and it means β is zero.

When β ≠ 0, the Poisson dispersion is saturated or diluted in the sense of unique property. An over dispersion is synonymous with an angle β>45°. In the case of under dispersed data, the point D would be below the bisecting diagonal line OD at an angle β>45°. Now, using the trigonometric formula:

$$\tan(A \pm B) = \frac{\tan A \pm \tan B}{1 \mp \tan A \tan B}$$

It is easy to notice that the length BD is Equation 2:

$$\tan\beta = \frac{|S_y^2 - \bar{y}|}{S_y^2 + \bar{y}} \qquad (2)$$

which is less than one when angle β>45°. The ratio in (1) can be called percent dilution index of Poisson dispersion. In the case of over dispersion, the dispersion s_y^2 is larger than $\bar{y}$. In the case of under dispersion, the dispersion s_y^2 is less than $\bar{y}$. Using the well-known Pythagoras theorem and geometric concepts for trigonometric formulas, it is easy to obtain that the length of BD, OD and OB are respectively $s_y^2 - \bar{y}, \bar{y}\sqrt{2}$ and $\sqrt{s_y^4 + \bar{y}^2}$. Also:

$$\begin{aligned}\text{Cos}(\beta) &= [1 + \tan^2(\beta)]^{-1/2} \\ &= \frac{(s_y^2 + \bar{y})^2}{[(s_y^2 + \bar{y})^2 + (|s_y^2 - \bar{y})|)^2]^{1/2}}\end{aligned}$$

Hence Equation 3:

$$\beta = \text{Cos}^{-1}\left(\frac{(s_y^2 + \bar{y})^2}{[(s_y^2 + \bar{y})^2 + (|s_y^2 - \bar{y})|)^2]^{1/2}}\right) \qquad (3)$$

When equality of mean and dispersion exists (that is, $s_y^2 - \bar{y} = 0$) as required in Poisson data, the expressions (2) and (3) imply that β = Cos^{-1} (1) = 45°. Otherwise, >45° in the case of over dispersion **Fig. 1** and β >45° in the case of under dispersion **Fig. 2**.

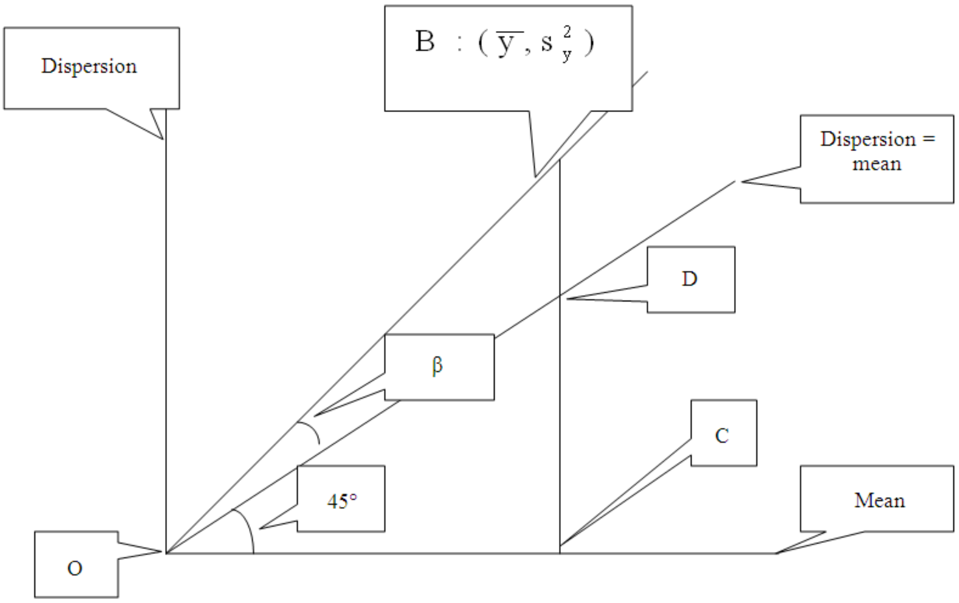

Fig. 1. Geometry of over-dispersion

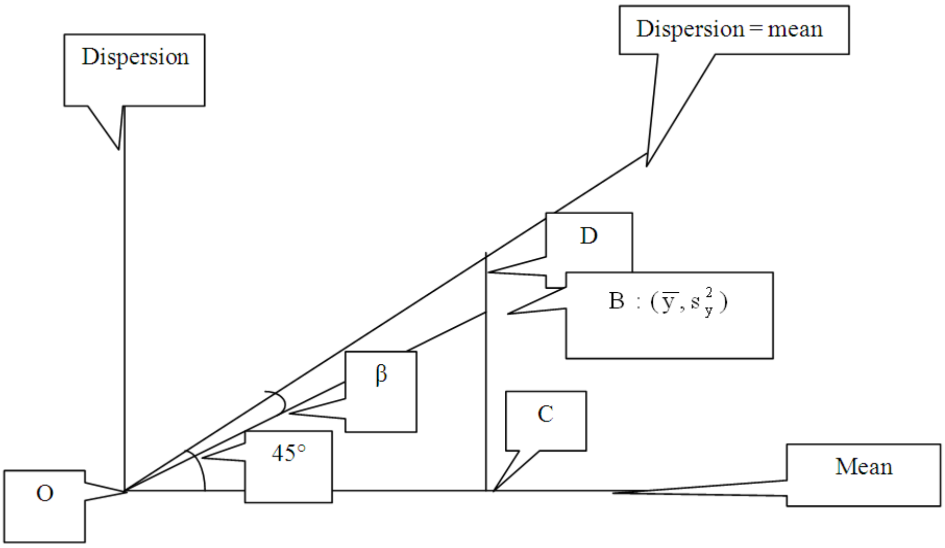

Fig. 2. Geometry of under-dispersion

Notice that a regression like relation exists between dispersion and mean with slope (= regression coefficients). The slope is than (45°± β). In the case of over dispersion, the sign is (+) and in the case of under dispersion, the sign is (-). But, it is so that than (45° = 1) in the case of equality of mean and dispersion as required in Poisson data with β = 0. Also, the regression coefficient is linked to the correlation coefficient. This means that the estimate of the correlation under Poisson equal dispersion is:

$$\hat{\rho}_{(\bar{Y},S_Y^2)} = C\hat{o}rr(\bar{Y},S_Y^2 \,|\, S_Y^2 = \bar{Y}) = 1$$

However, under over/under dispersion, an estimate of the correlation is Equation 4:

Table 1. Over-dispersion Poisson dilution and the correlation between the sample mean and dispersion

Data variable	Sample mean $\bar{y}$	Sample dispersion S_y^2 (dispersion>mean)	Angle, $\hat{\beta}$ using (2) in degrees	Poisson dilution index using (2) (%)	Correlation $\hat{\rho}_{(\bar{Y},S_Y^2)}$ using (6)
# mutations seen in a length of 1089 sites of amino acids	1.190	1.590	8.18	14	0.1
# contract strikes in US manufacturing companies	5.500	13.400	22.68	41	0.17
# ships damaged by waves	10.200	236.500	42.53	91	0.87
# patents in German companies	304.600		44.00	99	0.99
# failed U.S. banks during 1947-81	6.300	11.800	16.90	30	0.48
# airplane accidents in Canada during 1974-88	0.013	0.125	39.06	81	0.72
# yeast cells	0.680	0.790	4.27	7	0.06
# soldiers killed in Prussian army	0.700	0.760	2.35	4	0.03
# O-rings with thermal distress at a given Fahrenheit temperature and pounds per square inch pressure in NASA flights with failures during 1981-86	0.390	0.430	2.79	4	0.04
# times parasite visited without any attack before the third host attacked it	2.100	4.700	20.92	38	0.12
# doctors visit according to austrailan health survey during 1977-78	0.302	0.637	19.63	35	0.08
# daily traffic accidents in virginia state during January 1, 1969 through October 31, 1970	0.860	0.979	3.70	6	0.05

Table 2. Under-dispersion Poisson dilution and the correlation between the sample mean and dispersion

Data variable	Sample mean $\bar{y}$	Sample dispersion S_y^2 (dispersion>mean)	Angle, $\hat{\beta}$ using (2) in degrees	Poisson dilution index using (2) (%)	Correlation $\hat{\rho}_{(\bar{Y},S_Y^2)}$ using (6)
(dispersion<mean) # children in Germany in a family where the mother's age is between 40-65	2.10	1.70	6.00	10	0.08
# hospital admissions due to acute poisoning during full moon	0.92	0.66	9.30	16	0.12
# suicides per year in 18 states of Germany by Von Bortkiewicz	1.85	0.11	41.59	88	0.15
# hospital visits according to national medical expenditure Survey of 1987-88 of n = 4406 respondents	1.50	0.56	24.52	45	0.21
# tram accidents by n = 134 drivers over the years 1965-1970	2.75	0.35	37.74	77	0.20

$$\hat{\rho}_{(\bar{Y},S_Y^2)} = \hat{Corr}(\bar{Y},S_Y^2 \mid S_Y^2 \neq \bar{Y}) = \frac{\sqrt{\hat{var}(\bar{Y})}(\hat{slope})}{\sqrt{\hat{var}(S_Y^2)}} \quad (4)$$

Substituting in (4) the $\hat{slope} = \frac{S_y^2}{\bar{y}}$, the results in (5) and (6), an expression for the estimate of the correlation coefficient in (7) is obtained. Notice that Equation 5 and 6:

$$\hat{Var}(\bar{Y} \mid Poisson) = \frac{\sigma_Y^2}{n} \quad (5)$$

And:

$$\hat{Var}(S_Y^2 \mid Poisson) \approx \frac{|\mu_{4,Y} - \sigma_Y^2|}{n} \quad (6)$$

where, $\mu_{4,Y} \approx \sigma_y^2 + \mu$ is the fourth central Poisson moment. After algebraic simplifications, the expression for an estimate of the correlation coefficient is Equation 7:

$$\hat{\rho}_{(\bar{Y},S_Y^2)} = \hat{Corr}(\bar{Y},S_Y^2 \mid Poisson) \approx \max\left(\frac{|\bar{y} - S_y^2|}{S_y^2 + \bar{y}} \sqrt{\left|\frac{\bar{y} - |\bar{y} - S_y^2|}{\bar{y} + |\bar{y} - S_y^2|}\right|}, 1\right) \quad (7)$$

In this approach, a visual meaning is involved and provided. The clients are at ease to quickly grasp the Poisson fit or it's lacking. The visualization seems to be the most effective way of learning complicated Poisson dispersion versus mean. In the next section, several well-known Poisson data are examined in terms of the formulas in (1) through (7).

1.2. Illustrations

The data sets in illustration include the number of genetic mutations, the number contract strikes in US manufacturing industries, the number of ships damaged due to sea waves, the number patents obtained in German companies, the number of airplane accidents in Canada, the number of failed US banks, the number of children per German family where mothers' age fall in 40-65 years, the number of daily hospital admissions due to poisoning during full moon season, the number of yeast cells as reported by William Gosset, the number of soldiers killed in Prussian army, the number of O-rings with thermal stress in NASA space flights, the number of yearly suicides in 18 states of Germany, the number of parasites visited before the attack by the host, the number of hospital visits by Australians and US citizen, the number of daily tram accidents in Yugoslavia and the daily number of accidents in Virginia State of the United States. Their sample mean and dispersions are displayed in **Table 1** for over dispersed Poisson data and in **Table 2** for under dispersed Poisson data. The angle, percent of dilution index from Poisson dispersion and the correlation between the sample average and dispersion of the data are estimated and displayed using expressions (1) through (6) in **Table 1** for over and in **Table 2** for under dispersion.

2. CONCLUSION

Biostatisticians and medical researchers have been puzzled when they experience the lack of fit of the data on rare incidences by the Poisson distribution. A root-cause of it is the existence of disparity between the mean and dispersion in the data. The disparity is recognized as over-saturated when the dispersion is more than the mean and as diluted when the dispersion is less than the mean. In either situation, there is a need to quantify the level of disparity in the data. Approching the quantification geometrically, the level of disparity could be indexed as demonstrated in this article. In some data, the dilution is severe but mild in other data. Likewise, the saturation is stronger in some data but mild in other data. The index is expressed as a percentage so that different data sets could be compared with one another. More often, the medical researches require the comparison of different drugs or the comparison of how differently patients perform under the same medication.

3. REFERENCES

Cameron, A.C. and P.K. Trivedi, 1986. Econometric models based on count data. Comparisons and applications of some estimators and tests. J. Applied Econ., 1: 29-53. DOI: 10.1002/jae.3950010104

Dalal, S.R., E.B. Fowlkes and B. Hoadley, 1989. Risk analysis of the space shuttle: Pre-challenger prediction of failure. J. Am. Statist. Assoc., 84: 945-957.

Davutyan, N., 1989. Bank failures as poisson variates. Econ. Lett., 29: 333-338. DOI: 10.1016/0165-1765(89)90212-7

Deb, P. and P.K. Trivedi, 1997. Demand for medical care by the elderly: A finite mixture approach. J. Applied Econ., 12: 313-326. DOI: 10.1002/(SICI)1099-1255(199705)12:3<313::AID-JAE440>3.0.CO;2-G

Thakur, C.P., P.N. Sharma and H.S. Akhtar, 1980. Full moon and poisoning. British Med. J., 281: 1684-1684.

Winkelmann, R. and K.F. Zimmermann, 1994. Count data models for demographic data. Math. Populat. Stud., 4: 205-221. PMID: 12287090