# Taylor's power law and the statistical modelling of infectious disease data

Doyo Gragn Enki
Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, UK

Angela Noufaily, Paddy Farrington, Paul Garthwaite
The Open University, Milton Keynes, UK

Nick Andrews, Andre Charlett
Public Health England, London, UK

Correspondence to:

Paddy Farrington, Department of Mathematics and Statistics, The Open University, Milton Keynes MK7 6AA, UK. Email: paddy.farrington@open.ac.uk

**Abstract**

Surveillance data collected on several hundred different infectious organisms over twenty years have revealed striking power relationships between variance and mean in successive time periods. Such patterns are common in ecology, where they are referred to collectively as Taylor's power law. In this paper, these relationships are investigated in detail, with the aim of exploiting them for the statistical modelling of infectious disease data. We confirm the existence of variance to mean power relationships, with exponent typically between 1 and 2. We investigate skewness to mean relationships, which are found broadly to match those expected of Tweedie distributions, and thus confirm the relevance of the Tweedie convergence theorem in this context. We argue that variance to mean power laws, when present, should inform statistical modelling of infectious disease surveillance data, notably in descriptive analysis, model building, simulation and interval and threshold estimation, the latter being particularly relevant to outbreak detection.

Key words: exponential dispersion model, infectious disease, power law, surveillance, Taylor's Law, Tweedie family.

# 1 Introduction

This paper is devoted to an exploration of variance to mean relationships in surveillance data on infectious diseases, focusing on implications for statistical modelling. The paper stems from a recent epidemiological study that suggested clear power relationships between mean and variance (Enki *et al* 2013). The present paper is devoted to a further exploration of these patterns, and how they might be exploited to improve the modelling of infectious disease data.

Enki *et al* (2013) undertook a review of the UK statistical surveillance system for outbreak detection, encompassing the 3,303 different organism types reported to the UK LabBase database over the 20 year period 1991 to 2011. Our investigations revealed striking linear relationships between the logarithm of the variance and the logarithm of the mean of the number of organisms detected in successive six-monthly time periods, which appears to be present for virtually all of the organism types for which sufficient data were available. Such variance to mean power laws have previously been observed for measles and whooping cough (Keeling and Grenfell 1999). Our data suggest they are ubiquitous for infectious diseases, whatever the mode of transmission.

It is desirable to use statistical models that correctly account for the variance to mean relationship (see for example McCullagh and Nelder 1989, pages 328-332). Thus, investigating the variance to mean relationship is a useful preparatory step for statistical modelling: for example, a Poisson or quasi-Poisson model might be indicated if the variance is proportional to the mean. In the case of exponential family models, the variance function largely determines the distribution. In turn, quantiles of this distribution may be important in some applications, notably for determining threshold values, as needed for detecting outbreaks of infectious diseases. These considerations apply to individual data sets; in our case we have access to data on hundreds of different organisms. The existence of power variance to mean relationships for virtually *all* these organisms is of intrinsic interest, may help in elucidating underlying mechanisms, and should inform the statistical modelling framework used to make inferences.

Power variance to mean relationships are instances of Taylor's power law, reviewed in Kendal (2004). Taylor's law originally related to the observation that populations within territories partitioned into quadrats exhibit a variance to mean power relationship of the form

$$\mathrm{var}(X) = \phi E(X)^p$$

where $X$ is the population count within a quadrat (Taylor 1961). Various mechanisms, discussed in Kendal's review, have been advanced to explain such empirical power laws, which have been observed in very different contexts. For example, they have been shown to result from certain types of birth and death processes (Anderson *et al* 1982, Keeling 2000). A further suggestion is that Taylor's law is the empirical manifestation of the asymptotic limiting behaviour of exponential dispersion models, which in turn is linked to the properties of the Tweedie family of distributions (Kendal 2004, Kendal and Jørgensen 2011).

The paper has three aims: first, to investigate in greater detail and with greater rigour the empirical evidence for power laws in observed variance to mean relationships in infectious disease surveillance data; second, to explore whether these data conform to the limiting behaviour of exponential dispersion models; and third, to seek to exploit the observed variance to mean relationships for the purposes of statistical modelling. We do not seek to investigate the underlying mechanisms which generate power laws. In Section 2, we describe the data and present the evidence for a variance to mean power law. In Section 3, we briefly review the Tweedie family of exponential dispersion models. In Section 4 we investigate the evidence that our data conform (in an asymptotic sense) to this family. In Section 5 we discuss the use of the power law and Tweedie models for infectious disease modelling, notably outbreak detection. These methods are applied to surveillance data in Section 6. The paper ends with a discussion of the potential and limitations of this approach in Section 7.

## 2 Variance-mean power relationships in infectious disease data

We used weekly counts of infectious organisms reported to Public Health England's LabBase database between week 1 of 1991 and week 26 of 2011. This is a computerized database of reports of infectious disease organisms identified in biological specimens (taken from samples of blood, faeces, or urine) collected by laboratories in England, Wales and Northern Ireland. Our data were arranged by week of collection of the specimen from which the infection was identified. More details of the data may be found in Enki *et al* (2013).

We explored the variance to mean relationship in successive 6-month periods. First, we de-seasonalised the data where appropriate. This was done by fitting a quasi-Poisson generalized additive model with log link, smooth trend and a seasonal factor with levels $\gamma_j$, $j = 1, \ldots, 12$. The de-seasonalised data are then obtained from the weekly counts $y_i$ as follows:

$$z_i = y_i \exp(\bar{\gamma} - \gamma_{s(i)})$$

where $\bar{\gamma}$ is the average of the $\gamma_j$ and $s(i)$ is the seasonal level for week $i$. We only applied this seasonal adjustment for 1015 organisms with nonzero counts in every season; for other organisms, too sparse to apply a meaningful seasonal adjustment, we used $z_i = y_i$. We then grouped the data in 41 6-monthly periods. Within each calendar year there are two such periods, indexed by $k = 1$ for the period January to June, and $k = 2$ for July to December.

If Taylor's law is deemed to apply to the weekly counts $y_i$ with

$$\log\{\text{var}(y_i)\} = \alpha + p \log\{E(y_i)\}$$

then

$$\log\{\text{var}(z_i)\} = \alpha_{s(i)} + p \log\{E(z_i)\}$$

with $\alpha_{s(i)} = \alpha + (2 - p)(\bar{\gamma} - \gamma_{s(i)})$. Now let $W_{jk}$ denote the set of weeks $i$ within period $k$ of year $j$, and set $z_{jk} = \sum\{z_i : i \in W_{jk}\}$. Provided that the means of the de-seasonalised data are deemed to be constant within 6-month periods, and weekly counts are independent, then

$$\log\{\text{var}(z_{jk})\} = \alpha_k + p \log\{E(z_{jk})\}$$

where $\alpha_k = \sum\{\alpha_{s(i)} : i \in W_{jk}\}$. This equation describes a linear relationship with two intercepts, corresponding to $k = 1, 2$. When the data are sparse, and are not de-seasonalised, we just used one intercept.

Perry (1981) has shown that simple regression of the logarithm of the empirical variance against the logarithm of the empirical mean (as used in Enki *et al* 2013) produces a negatively biased estimate of $p$ (that is, $\hat{p}$ is too low), and suggests that a gamma generalised linear model be used instead. This produces the same point estimates as the approach of Jørgensen *et al* (2011) using unbiased estimating equations.

Accordingly, we applied the gamma model to the 1737 different organism types with sufficient data to estimate the regression parameters, with two intercepts when de-seasonalized, and a single intercept when not.

In all cases, the empirical relationship between log variance and log mean is linear. Figure 1 shows the plots, together with the gamma regressions, for six common organisms. The full set of 1737 plots is in the Supplementary Materials. Figure 2 shows caterpillar plots of values of $\hat{p}$, along with approximate 95% confidence intervals. The variances of the estimates $\hat{p}$ were obtained using the scale parameters for the gamma regression models based on Pearson's chi-square statistic, which are consistent provided that the $p$ have been consistently estimated (McCullagh and Nelder 1989, page 296). On the left is the caterpillar plot for all 1737 organisms. The strange appearance of the leftmost tail of the plot is the result of sparse organism counts: when there is a single count in a six-month period, then the estimated mean and variance are identical. In the extreme case where this is true of all six-monthly periods, the dispersion (and hence the width of the confidence interval) is zero. The caterpillar plot on the right of Figure 2 is restricted to the 374 organisms for which the count in every six-month period is at least 2. These caterpillar plots show that most values of $p$ lie between 1 and 2, and few other than those affected by the sparseness issue just described have 95% confidence intervals entirely located outside this range.

We repeated the analysis with an annual rather than 6-monthly grouping of the data (and consequently a single intercept). The slopes are similar to those obtained with 6-monthly periods, as shown by the scatterplot in Figure 3, except for the

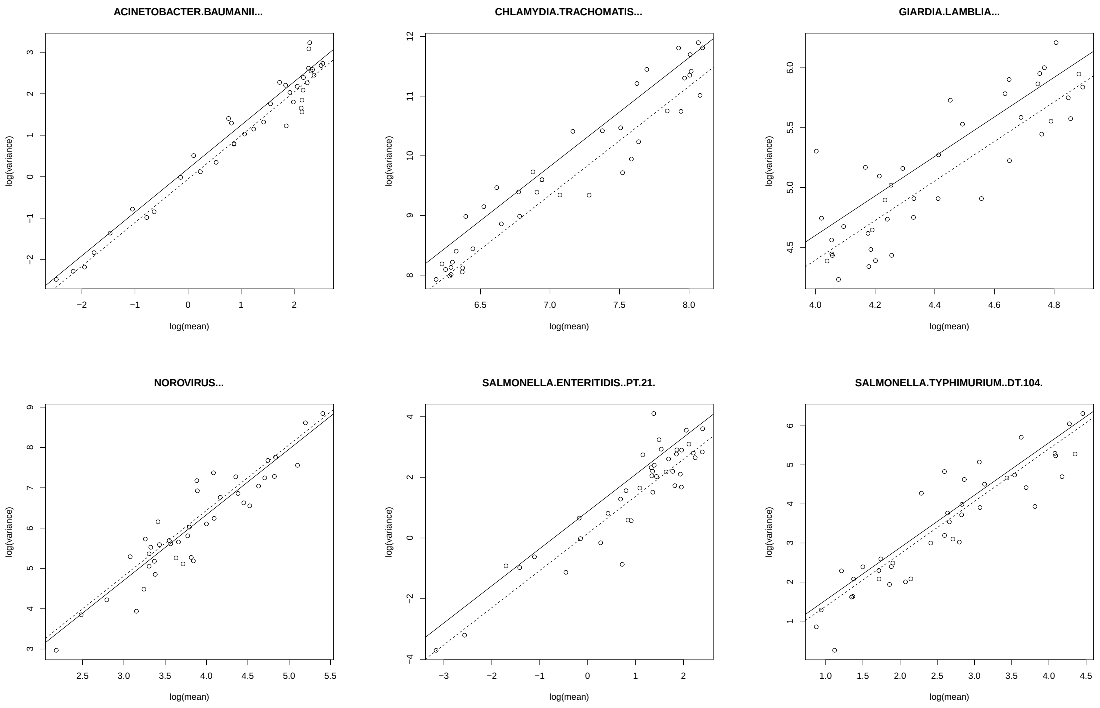

Figure 1: *Empirical relationship between log variance and log mean for six organisms: observed values (dots) and regression lines (with different intercepts for different seasons).*

organisms with larger slopes. We suspect that residual seasonality may contribute to inflating the variances in the analysis based on annual data. These results confirm the striking linear relationships between log variance and log mean observed by Enki *et al* (2013). These are strongly suggestive of power variance-to-mean relationships, with exponent typically lying between 1 and 2. These observations lend empirical support for Taylor's law over an extremely wide range of different micro and macro-organisms.

In a further exploration of the data, we plotted the estimated power parameters $p$ against the logarithm of the median of the 41 seasonally-adjusted counts; organisms for which the median of the seasonally-adjusted counts was less than or equal to 0.05 were grouped into a single category with median 0.05 (the median value of $p$

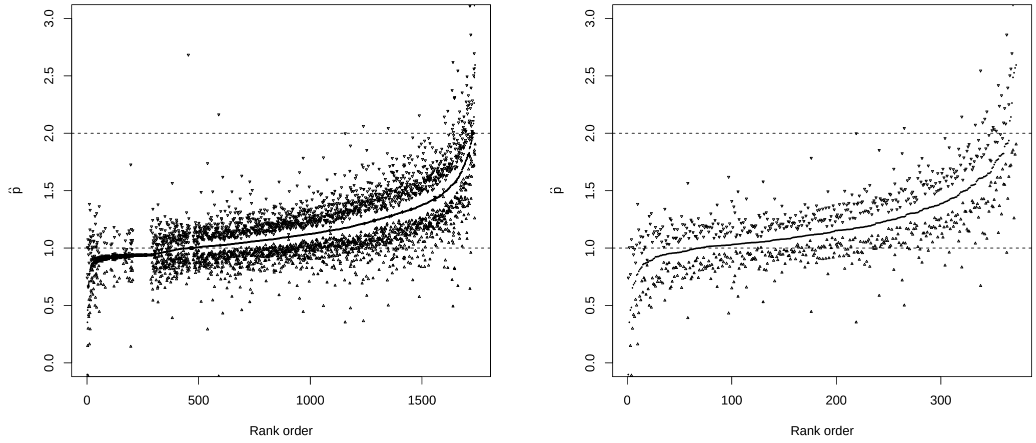

Figure 2: *Caterpillar plot of estimated values of slopes and corresponding 95% confidence limits. Left: all organisms; Right: excluding sparse organisms.*

for these was 1.07). The plot is shown in Figure 4. The loess curve superimposed on the scatterplot suggests that the power $p$ tends to increase slightly with the median adjusted count, though the spread of values is considerable at all medians. At low median adjusted counts $p$ is centrally located a little above 1, corresponding to the quasi-Poisson model. The four outliers with values of $p$ close to or above 2.5 are Rotavirus, *Cryptosporidium* Sp, Herpes simplex virus untyped and *Escherichia coli* untyped. There is also a single outlier with a negative $p$, corresponding to *Neisseria meningitidis* type B (types A and C had values of $p$ close to 1). Table 1 gives the values of $p$ for some specific organisms (in some cases averaged over the values for distinct subtypes). There is no evident relationship with the mode of transmission.

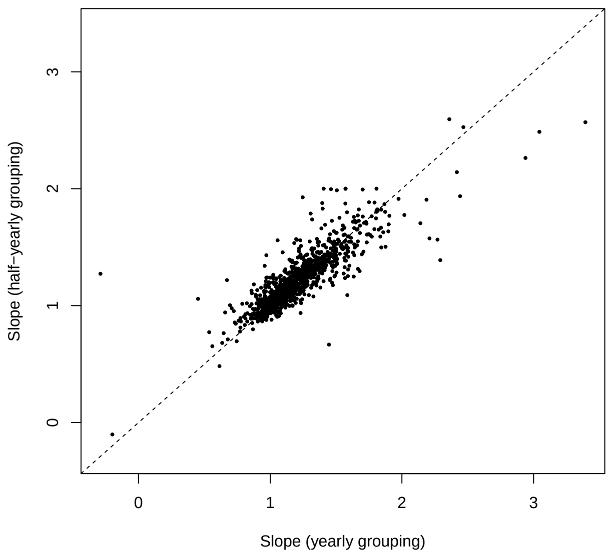

Figure 3: *Scatter plot of estimated slopes: bi-annual vs annual grouping.*

# 3 Tweedie models, asymptotics and scaling

Exponential dispersion models have densities (or probability mass functions) of the form

$$p(y; \theta, \phi) = c(y; \phi) \exp[\{\theta y - \kappa(\theta)\}/\phi],$$

where $\theta$ is the canonical parameter, $\phi > 0$ is the scale parameter, and $\kappa$ is the cumulant function for the underlying probability measure. The expectation of $Y$ is $E(Y) = \mu = \kappa'(\theta)$, and its variance is $\text{var}(Y) = \phi V(\mu)$ where $V(\mu)$ is the unit variance function. Within this class, the Tweedie models are those with power variance function $V(\mu) = \mu^p$ for $p \in (-\infty, 0] \cup [1, \infty)$; there are no such exponential dispersion models with $p \in (0, 1)$. For more details, see Jørgensen (1997).

Of particular interest to us are the Tweedie models with $p \geq 1$. The exponent $p = 1$ corresponds to (scaled) Poisson models, values $p \in (1, 2)$ to compound Poisson-gamma models, namely Poisson mixtures of gamma densities, which have an atom at 0 and are continuous above 0, while $p = 2$ corresponds to gamma models. The

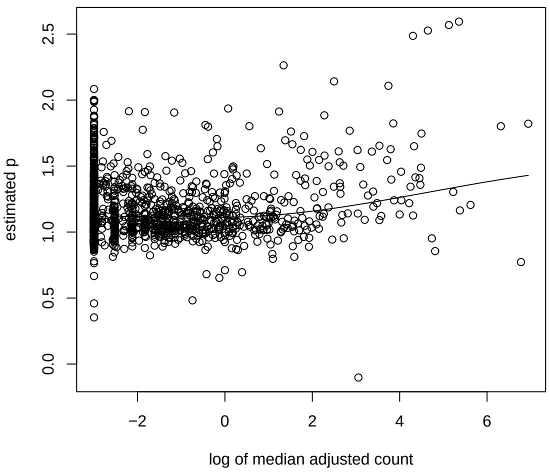

Figure 4: *Scatter plot of estimated powers against logarithm of median seasonally-adjusted count. The full line is a loess curve.*

values $p > 2$ correspond to stable distributions, including for instance the inverse Gaussian distribution ($p = 3$). The Tweedie models with $p$ close to 1 are multimodal (Dunn and Smyth 2005).

With the sole exception of the case $p = 1$, Tweedie models are continuous above 0, and hence are not suitable for direct modelling of infectious disease counts. They are nonetheless highly relevant owing to the Tweedie convergence theorem. This states that a random variable $cY$ where $Y$ belongs to an exponential dispersion family whose variance function $V(\mu) \sim c_0\mu^p$ as $\mu$ tends to zero (or infinity), converges in distribution to the Tweedie model with power parameter $p$ as $c$ tends to 0 (or infinity) (Jørgensen 1997, pages 148-149). Thus, a wide range of exponential dispersion models may be approximated by Tweedie models. It has been suggested that this theorem provides the theoretical basis underpinning the empirical evidence for Taylor's law (Kendall 2004). This contention will be discussed briefly in Section 7.

Tweedie models are attractive also by their scaling properties: it may be shown

Table 1: Estimated values of $p$ for selected organisms

| | | | |
|---|---|---|---|
| *Bordetella pertussis* | 1.06 | *Mycobacterium tuberculosis* | 1.13 |
| *Campylobacter jejeuni* | 1.12 | Norovirus | 1.63 |
| *Chlamydia trachomatis* | 1.81 | *Pseudomonas aeruginosa* | 1.65 |
| *Clostridium difficile* | 1.24 | Rubella virus | 1.37 |
| Cytomegalovirus | 1.36 | *Salmonella enteritidis* | 1.27 |
| *Escherichia coli* O157 | 1.27 | *Salmonella typhimurium* | 1.21 |
| Epstein Barr Virus | 1.14 | *Staphylococcus aureus* | 1.80 |
| Herpes Simplex Virus 1 | 1.36 | *Streptococcus* A | 1.31 |
| Herpes Simplex Virus 2 | 1.49 | *Streptococcus* B | 1.61 |
| Influenza A virus | 1.64 | *Taenia* | 1.26 |
| Influenza B virus | 1.81 | *Toxoplasma* | 1.21 |
| Measles virus | 1.11 | Varicella Zoster Virus | 1.19 |

that the Tweedie models correspond precisely to those exponential dispersion models with $V(1) = 1$ which are closed under scale transformations, that is, for which $V(c\mu) = g(c)V(\mu)$ for some function $g$, for all $c > 0$. Such a property is natural for counts of infectious diseases: if $Y$ is the number of cases per time unit, the statistical properties of $Y$ should not depend in fundamental respects on the time unit chosen. Note that this scaling invariance only applies to the power parameter $p$: unsurprisingly, the scale parameter $\phi$ is scale-dependent.

Let $\mathrm{Tw}_p(\mu, \phi)$ denote the Tweedie model of power $p$ with mean $\mu$ and scale parameter $\phi$. If $Y \sim \mathrm{Tw}_p(\mu, \phi)$ then $\kappa_2 = \mathrm{var}(Y) = \phi\mu^p$. We shall also be interested in its skewness. The third central moment of $Y$ is $\kappa_3 = p\phi^2\mu^{2p-1}$ and its skewness is

$$\rho_3 = \frac{\kappa_3}{\kappa_2^{3/2}} = p\phi^{\frac{1}{2}}\mu^{\frac{p}{2}-1}.$$

These relationships are preserved when iid variables are aggregated. Thus, suppose that $Y_i \sim \mathrm{Tw}_p(\mu, \phi)$, $i = 1, \ldots, n$, are independent and let $Z = Y_1 + Y_2 + \cdots + Y_n$. Then $Z \sim \mathrm{Tw}_p(\mu_Z, \phi_Z)$ where $\mu_Z = n\mu$ and $\phi_Z = n^{1-p}\phi$. For example, the skewness of $Z$ is $p\phi_Z^{1/2}\mu_Z^{p/2-1}$. Thus, aggregating independent counts of disease in successive equal time periods will preserve the key power relationships. Since for many of the infections we consider the prime mode of transmission, especially in the absence of an outbreak, is not directly from person-to-person (but is due to food contamination, or is of zoonotic origin - such cases often being described as 'sporadic' owing to the lack of clear epidemiological linkage between them), successive counts can often reasonably be considered to be independent (or only weakly dependent). The practical implication is that power relationships may validly be examined by aggregating data over several time units (in our case, weeks).

# 4 Evidence for convergence to Tweedie models

In this section we examine further evidence that Tweedie models provide an adequate description of count data for infectious diseases. Clearly, Tweedie models other than the Poisson (corresponding to $p = 1$) can only be approximate, since they are continuous above zero. It has already been established in Section 2 that our infectious disease data exhibit the variance to mean relationship typically expected of Tweedie models. Here we consider the empirical skewness to mean relationship, to examine whether it displays the Tweedie power relationship elucidated in Section 3.

We shall use the following sample skewness coefficient (Joanes and Gill 1998) for a sample of size $n$:

$$\hat{\rho}_3 = \frac{\sqrt{n(n-1)}}{(n-2)} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^3}{\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\}^{3/2}}.$$

If in a period of $n$ weeks there is just a single weekly count of 1, the others being zero, then $\hat{\mu} = n^{-1}$ and $\hat{\rho}_3 = n^{1/2}$. Thus, for rare organisms in which the count is 0 or 1 in every six-month period, the Tweedie skewness to variance relationship is exactly satisfied with $p = 1$ and $\phi = 1$. If Taylor's law is deemed to apply to the weekly de-seasonalised data $z_i$ , then

$$\log\{\rho_3(z_i)\} = \alpha_{s(i)} + (\frac{p}{2} - 1)\log\{E(z_i)\}$$

where $s(i)$ is the seasonal level for week $i$. Accordingly we fitted a normal errors generalised linear model to the empirical skewnesses, with logarithmic link and the linear predictor just described. When the data were not de-seasonalised (owing to sparseness) we just fitted a single intercept.

There is considerable scatter in the estimated skewness coefficients, making an assessment of the skewness to mean relationship more difficult than for the variance to mean relationship, a difficulty compounded by the curvilinear form of the theoretical relationship. Nevertheless, for many organisms the asymptotic relationship implied by the Tweedie model does appear to hold, as exemplified by the six organisms displayed in Figure 5. The full set of 1724 plots (comprising the organisms for which sufficient data were available) is in the Supplementary Materials. Figure 6 shows caterpillar plots for the values of $p$ estimated from the skewness to mean relationship, with approximate 95% confidence limits. These values typically lie between 0.5 and 2.5. The confidence intervals are quite wide, however, and are affected by the sparseness of the data for some organisms. Sample skewness evaluated from small samples may be prone to bias. Accordingly, we repeated the analyses with intervals of 1 year, rather than 6 months, so as to increase the typical sample size from which the sample skewness is calculated. We also applied a first-order bias correction to the skewness coefficient (Pewsey 2005):

$$\rho_3 = \frac{\sqrt{n(n-1)}}{n-2}\Big\{\beta_1 + \frac{3}{8n}(\beta_1(7+5\beta_2) - 4\beta_3)\Big\} + o(n^{-1}),$$

where

$$\beta_k = \frac{\mu_{k+2}}{\mu_2^{(k+2)/2}}, \qquad \mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^k.$$

Figure 7 shows the corresponding caterpillar plots for $p$ with annual groupings, with and without bias correction. Increasing the sample size narrows the confidence

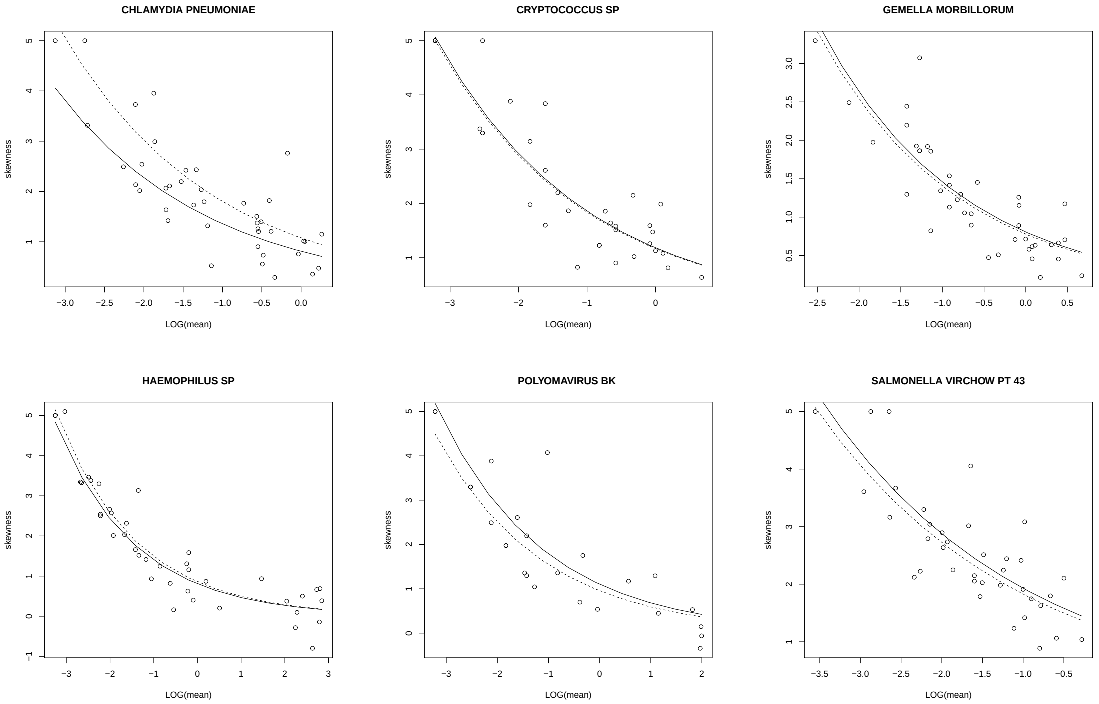

Figure 5: *Empirical relationship between skewness and log mean for six organisms (in 6-monthly periods): observed values (dots) and regression curves (with different intercepts for different seasons).*

intervals and brings most values $p < 1$ closer to 1. Applying the skewness correction produces a marked increase in the estimated values of $p$.

Note that if the sample skewness has relative bias that is constant as the mean $\mu$ varies, then the estimate of $p$ obtained by regressing the sample skewness against $\log(\mu)$ will be largely unaffected, owing to the log-linear relationship between them; a constant relative bias will be reflected in the estimated intercept. Thus, what matters primarily for our purposes is the dependence of the relative bias on the mean, rather than the magnitude of this bias.

To investigate these effects in more detail, and their possible impact on our observations, we undertook simulations of the sample skewness $\hat{\rho}_3$ from Tweedie distributions, using sample sizes $n = 26$ and $n = 52$, corresponding to typical six

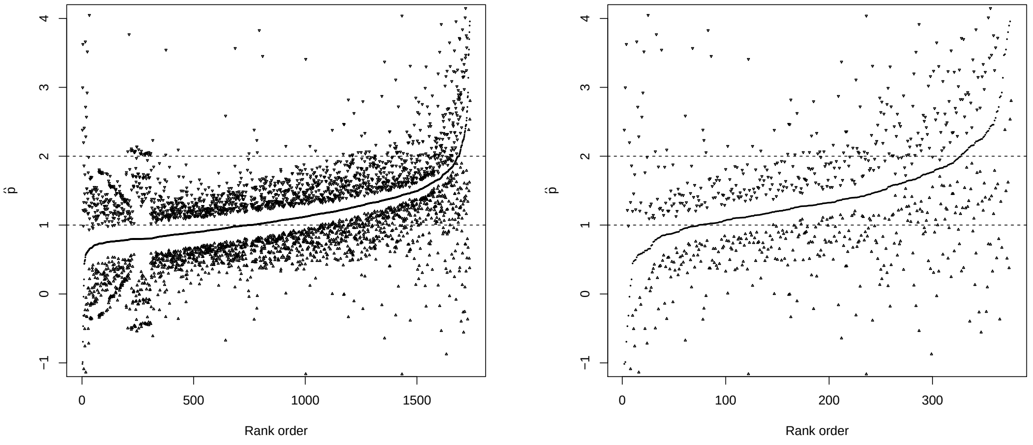

Figure 6: *Caterpillar plot of estimated values of* $p$ *and corresponding 95% confidence limits based on log(mean) vs skewness. Left: all organisms; Right: excluding sparse organisms.*

monthly and annual groupings, respectively. We also investigated the impact of the first-order bias correction to the skewness coefficient. Each value was obtained from 100 000 runs, using the R package `tweedie` (Dunn 2014). The relative biases in $\hat{\rho}_3$ for $\phi = 1$ are in Table 2; a broadly similar dependence of relative bias on the mean was observed with $\phi = 0.5$ and $\phi = 2$ (not shown). The simulation results show that skewness is generally estimated with negative bias, as suggested by Joanes and Gill (1998). As expected, the relative bias is less in absolute value with greater sample size, and (generally but not universally) with bias correction. However, for the uncorrected skewness, the bias becomes more negative as $\mu$ increases for $p$ close to 1. Thus, for values of $p$ close to 1, the trend in the plot of skewness against $\log(\mu)$ will tend to be too steeply negative, resulting in an underestimate of $p$. For larger values of $p$, applying a bias correction increases the dependence of the relative bias on the mean (while reducing its absolute value). Thus, applying a skewness correction is likely to reduce the bias in $p$ for $p$ close to 1 (thus increasing the estimated $\hat{p}$) but will make matters worse for larger $p$. The magnitude of the bias for the estimated

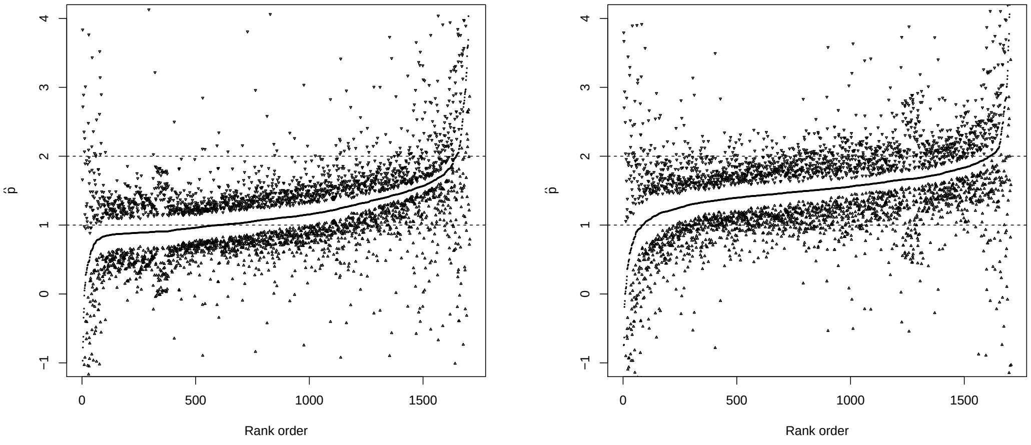

Figure 7: *Caterpillar plot of estimated values of p and corresponding 95% confidence limits based on log(mean) vs skewness for annual data. Left: without bias correction; Right: with bias correction.*

variance is very much smaller than for the skewness (results not shown).

Figure 8 shows a scatterplot of the values of $p$ estimated from the variance to mean relationship against those estimated from the skewness to mean relationship, using 6-monthly groupings. The correlation is 0.69. Similar results were obtained with annual groupings (correlation 0.74). The insights provided by the simulations help explain the appearance of Figure 8, and lead us to conclude that the imperfect correspondence of the values of $p$ estimated from the variance and from the skewness are due, in large part, to mean-dependent bias in the estimation of the skewness, especially for low $\mu$.

Overall, these results lend some support to the view that the empirical variance to mean power relationship reflects asymptotic convergence to Tweedie distributions, at least insofar as third moments are concerned. This would suggest, in turn, that Tweedie distributions may perhaps be used to approximate the distributions of infectious disease counts.

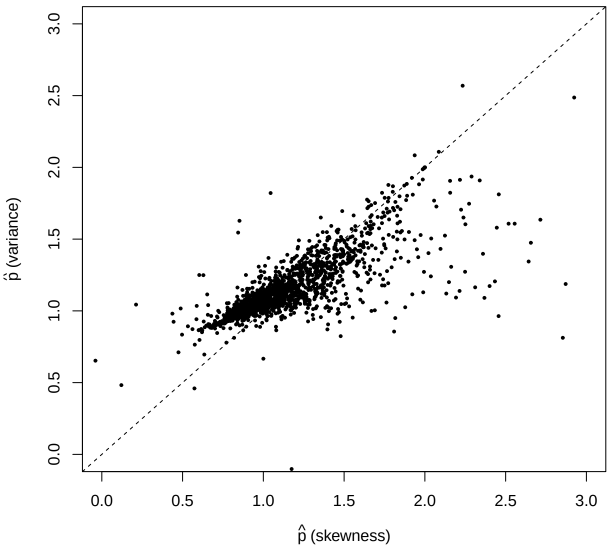

Figure 8: *Scatter plot of estimated slopes: variance-based vs skewness-based, with 6-monthly groupings.*

# 5 Power laws and Tweedie models for infectious disease modelling

Typically, descriptive analyses of counts of infectious diseases from surveillance systems have been based on the Poisson or negative binomial distributions, the latter resulting from gamma mixtures of Poisson distributions to produce overdispersion models in which the variance is proportional to the mean (McCullagh and Nelder 1989, page 199). Where surveillance of relatively common infections is involved, overdispersion relative to the Poisson is common, perhaps owing to variations in reporting rates from different locations. It is convenient in such circumstances to use a quasi-Poisson model, with dispersion estimated from the data.

The preceding sections indicate that such models may not be appropriate for infectious disease counts, if examination of the variance to mean relationship suggests that the variance is proportional to a power $p$ of the mean with $p > 1$. Tweedie models with $p > 1$ are not discrete, and hence not directly appropriate. One option is to use discrete counterparts of the Tweedie model, such as Tweedie-Poisson mixtures or Hinde-Demétrio models (Jørgensen 1997, pages 165-170, Kokonendji *et al* 2004), for which the unit variance function is approximately proportional to $\mu^p$ for large $\mu$. A simpler option is to adopt a quasi-Tweedie modelling approach, in which the variance is (at least approximately) proportional to a power of the mean with exponent $p \geq 1$, but without explicitly specifying the likelihood. Quasi-Tweedie models can be fitted within the quasi-likelihood generalised linear modelling framework, using the following quasi-deviance contribution for the count in week $i$ (McCullagh and Nelder 1989, page 327) when $p \neq 1, 2$:

$$d_i = 2\Big\{\frac{1}{1-p}\Big({y_i}^{2-p} - y_i{\mu_i}^{1-p}\Big) + \frac{1}{p-2}\Big({y_i}^{2-p} - {\mu_i}^{2-p}\Big)\Big\}.$$

This requires the power parameter $p$ to be fixed. In a first stage, $p$ is estimated using the gamma regression method described in Section 2. If the data are too sparse to allow this, or if $\hat{p} < 1$, then take $p = 1$ and fit a quasi-Poisson model; otherwise, set $p = \hat{p}$. Dunn and Smyth (2005) argue that assuming that $p$ is fixed has little impact on inferences for the Tweedie model, owing to the fact that $p$ is orthogonal to $\mu$ and $\phi$. We surmise that the same applies for quasi-Tweedie models.

Tweedie models are likely to be useful in other ways relevant to infectious diseases. For example, the evaluation of statistical techniques for infectious disease surveillance is often based on simulations using Poisson or negative binomial distributions. A more realistic option, especially for more common infections, may be to simulate data from Tweedie distributions, rounding the results to the nearest integer to obtain counts.

Finally, a key advantage of the quasi-Tweedie approach is that the true but unknown distribution, though not Tweedie, can nevertheless be approximated by a Tweedie density, by virtue of the Tweedie convergence theorem, at least when the means $\mu$ are large, which is the setting in which the approach is likely to be most fruitful (if counts are low, the appropriate limit when $\mu$ tends to zero is typically Poisson). The evaluation of Tweedie densities is discussed by Dunn and Smyth (2005, 2008). This circumvents a disadvantage of the quasi-likelihood approach: no distribution is presumed, which makes it difficult to obtain reliable prediction intervals for individual observations. This convergence property is likely to be particularly useful in regression-based outbreak detection, where the upper threshold estimated under the null hypothesis that no outbreak is occurring is obtained using a prediction interval. For example, the outbreak detection system in use since the early 1990s at Public Health England, and applied to the LabBase data described above, uses a quasi-Poisson regression method, with the upper prediction limits based on a normal approximation or the quantiles of the negative binomial distribution (Farrington *et al* 1996, Noufaily *et al* 2013). These limits are compared to those obtained with the asymptotic Tweedie distribution in the next section.

# 6 Application to infectious disease surveillance data

We compared the upper prediction limits obtained for several organisms using (a) normal approximations, (b) negative binomial quantiles, and (c) Tweedie approximations. The organisms were selected to obtain a range of values of the exponent $p$, which was estimated using the gamma model of Section 2 applied to adjacent six-monthly periods. We used upper 99.5% prediction limits: organisms with counts above these thresholds are deemed aberrant and undergo further investigations. The algorithm is designed to detect the following types of aberrations: sudden changes in level, sudden changes in trend, and sudden changes in the phase or amplitude of seasonal fluctuations.

For methods (a) and (b) we applied the algorithm of Noufaily *et al* (2013). In brief, to obtain the upper prediction limit at week $t$ a quasi-Poisson generalised linear model (with log link, linear time effect and factorial seasonal effect) is applied to the observed counts in recent years, but excluding the most recent six months, lest these counts are contaminated by a current outbreak. This gives estimates of the expected value $\mu_t$ at time $t$ and of the dispersion $\phi$, from which the limits (a) and (b) are derived from the 0.995-quantiles of the normal (after a suitable transformation) and negative binomial distributions. See Noufaily *et al* (2013) for details. The algorithm is implemented in the R package `surveillance` (Höhle 2007).

For method (c), we replaced the quasi-Poisson model with a quasi-Tweedie generalised linear model (with pre-estimated $p$) to obtain the expected value $\mu_t$ at $t$ and the dispersion $\phi$ (which is different from that obtained for methods (a) and (b)). The quasi-Tweedie model was fitted using the function `glm` within R (R development Core Team 2014); suitable code is provided in the Appendix. The upper prediction limit was then obtained as the 0.995 quantile of the Tweedie $\mathrm{Tw}_p(\mu_t, \phi)$ density, using the function `qtweedie` within R package `Tweedie` (Dunn 2014).

For all methods, we switched off the reweighting mechanism in the standard version of the outbreak detection algorithm. This mechanism downweights high baseline values, in order to reduce the impact of past outbreaks. It was switched off because it would need to be modified for the quasi-Tweedie model, thus complicating the comparison of the three methods.

For all organisms, the expected values obtained using the quasi-Poisson and quasi-Tweedie models are virtually identical. The upper prediction limits obtained using the normal approximation for the quasi-Poisson and negative binomial quantiles were also very close.

For organisms with $p$ close to 1, the upper prediction limits based on Tweedie quantiles were very similar to those obtained using the other two methods. Figure 9 shows the results for *Acinetobacter baumanii*, with $p = 1.05$, and Varicella zoster virus, with $p = 1.19$. Figure 10 shows the results for two organisms with intermediate values of $p$: Cytomegalovirus ($p = 1.36$) and *Chlamydia trachomatis* ($p = 1.82$). Here, the Tweedie upper prediction limits are generally higher than those for the quasi-Poisson, the difference increasing with the mean frequency and with $p$. For Cytomegalovirus, note that the system detects the abrupt change in level around

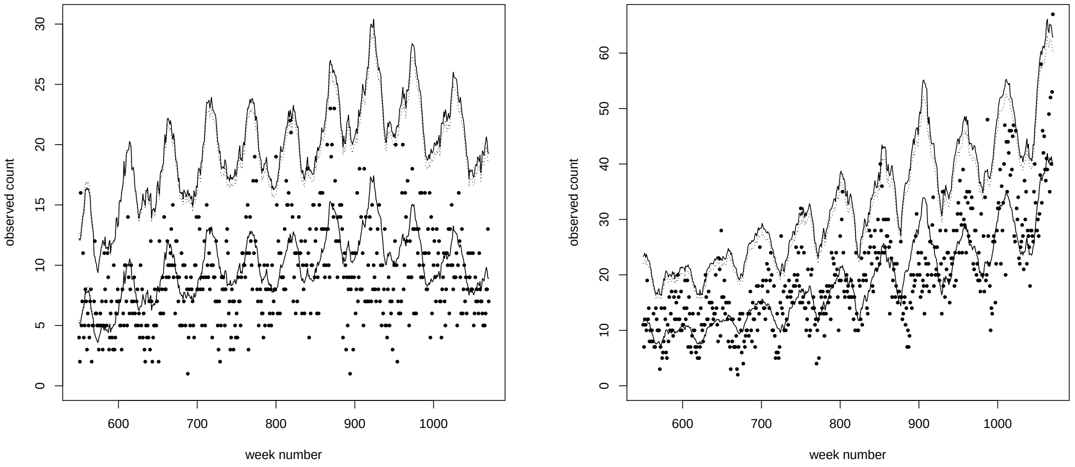

Figure 9: *Weekly count (dots), expected values and upper thresholds for quasi-Tweedie (full lines) and quasi-Poisson (dotted lines). Left: Acinetobacter baumanii; right: Varicella zoster virus.*

week 875, after which performance is degraded until the algorithm adjusts to the new level. Figure 11 shows the results for two organisms with high values of $p$: *Bacteroides fragilis* ($p = 2.14$) and *Cryptosporidium* Sp ($p = 2.49$). Here, the Tweedie upper prediction limits can be very different from those for the quasi-Poisson. Typically they are higher, except in the troughs of the seasonal cycles. The practical consequence of these observations is that, when $p$ is high, the false positive rate (that is, the proportion of non-outbreak weeks in which organism counts are flagged as aberrant) will tend to be too large using the standard outbreak detection algorithm, if as suggested in Section 4, the distribution of the organism of interest is well approximated by a Tweedie distribution. However, organisms with $p$ less than about 1.4 should not be unduly affected unless the mean frequency is very high.

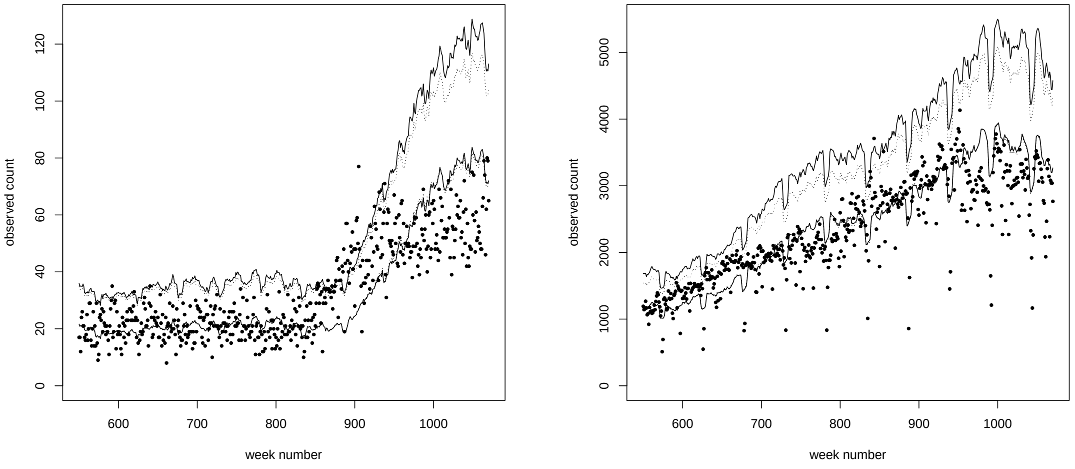

Figure 10: *Weekly count (dots), expected values and upper thresholds for quasi-Tweedie (full lines) and quasi-Poisson (dotted lines). Left: Cytomegalovirus; right: Chlamydia trachomatis.*

# 7 Discussion

Analysis of extensive surveillance data on a large number of distinct organisms provides strong evidence that the variance to mean relationship may be described by power laws in virtually all instances, the power parameter $p$ typically lying between 1 and 2, with some exceptions. Further investigation of skewness to mean relationships provides some evidence to suggest that they conform to Tweedie distributions, as predicted by the Tweedie convergence theorem. The strength of evidence is somewhat reduced by the problem of mean-dependent bias in estimating sample skewness. Nevertheless, we tentatively conclude that Tweedie distributions can be used to approximate distributions of counts.

We advocate the explicit use of observed variance to mean relationships, and the Tweedie convergence theorem, for four related purposes in infectious disease modelling. In the first instance, an examination of the variance to mean relationship, after removing seasonal variation, can provide useful information and ought

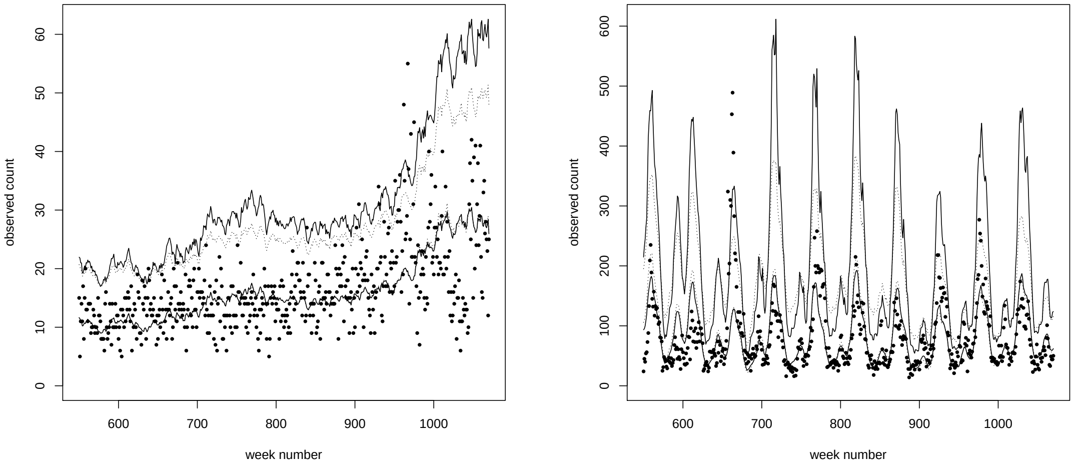

Figure 11: *Weekly count (dots), expected values and upper thresholds for quasi-Tweedie (full lines) and quasi-Poisson (dotted lines). Left: Bacteroides fragilis; right: Cryptosporidium Sp.*

to become a routine element of the descriptive statistical investigation of infectious disease surveillance data. Second, quasi-Tweedie generalised linear models (the term quasi-Tweedie solely denoting a power variance to mean relationship) should be more commonly used for model building purposes. Third, model evaluation using simulations based on Poisson and negative binomial models is likely to underestimate the extent of random variation and skewness in infectious disease counts; simulations using discretised Tweedie distributions might be more realistic. And finally, the calculation of thresholds and prediction intervals, notably in outbreak detection systems, may be more accurate if based on asymptotic Tweedie distributions, at least when $p$ is appreciably greater than 1.

For infections with $p$ close to 1, Poisson, quasi-Poisson or negative binomial models are likely to be adequate. However, for organisms with higher values of $p$, we have shown in Section 6 that there are benefits in adopting the methods we propose. While these may be impractical for the purposes of large multiple routine surveillance systems, more focused organism-specific surveillance systems may benefit from making use of the empirical relationships suggested by the data.

Our focus has been entirely on exploiting variance to mean relationships for statistical modelling purposes. We have not sought to explore the underlying mechanisms which produce such power relationships. Keeling and Grenfell (1999) show that simulations with their stochastic, pulsed realistic age-structured SEIR model produce power variance to mean relationships at higher incidences. Such mechanisms have been further explored in birth and death processes (Anderson *et al* 1982, Keeling 2000), though an entirely convincing explanation at the level of mechanism remains elusive. Faddy (1997) explicitly constructs a family of statistical models for count data based on birth processes, which exhibits a range of variance to mean behaviours. Kendal (2004) and Kendal and Jørgensen (2011) suggest that the Tweedie convergence theorem provides a purely mathematical explanation for such observed patterns, in terms of asymptotic limiting behaviour. However, the theorem presumes that the distribution to which the limiting process applies in some sense involves a variance to mean power relationship of order $p$ - and this, it seems to us, calls for a mechanistic rather than purely mathematical explanation. Nor is it clear what factors, other than mean frequency, govern the value of $p$.

Finally, it would be interesting to know whether the power relationships we have observed are also present in infectious disease data from other countries, or in data on non-infectious diseases. Contrasting data sets obtained in different contexts may help to clarify whether the observed patterns stem from underlying mechanisms, the specificities of different reporting systems, or purely stochastic effects.

# Acknowledgments

This research was supported by a project grant from the Medical Research Council and by a Royal Society Wolfson Research Merit Award to PF.

# References

Anderson R.M., Gordon D.M., Crawley M.J. and Hassell M.P. (1982) Variability in the abundance of animal and plant species. *Nature*, **296**, 245-248.

Dunn P.K. and Smyth G.K. (2005) Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing*, **15**, 267-280.

Dunn P.K. and Smyth G.K. (2008) Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Statistics and Computing*, **18**, 73-86.

Dunn P.K. (2014) `tweedie`: Tweedie exponential family models. R package version 2.2.1.

Enki D.G., Noufaily A., Garthwaite P.H., Andrews N.J., Charlett A., Lane C. and Farrington P. (2013) Automated biosurveillance data from England and Wales, 1991-2011. *Emerging Infectious Diseases*, **19**, 35-42.

Faddy M.J. (1997) Extended Poisson process modelling and analysis of count data. *Biometrical Journal*, **4**, 431-440.

Farrington C.P., Andrews N.J., Beale A.J., Catchpole M.A. (1996) A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society* Series A, **159**, 547-563.

Höhle M. (2007) Surveillance: an R package for the monitoring of infectious diseases. *Computational Statistics*, **22**, 571-582.

Joanes D.N. and Gill C.A. (1998) Comparing sample measures of skewness and kurtosis. *The Statistician*, **47**, 183-189.

Jørgensen B. (1997) *The Theory of Dispersion Models.* Chapman & Hall, London.

Jørgensen B., Demétrio C.G.B., Kristensen E., Banta G.T., Petersen H.C. and Delefosse M. (2011) Bias-corrected Pearson estimating function for Taylor's power law applied to benthic macrofauna data. *Statistics and Probability Letters*, **81**, 749-758.

Keeling M. and Grenfell B. (1999) Stochastic dynamics and a power law for measles variability. *Philosophical transactions of the Royal Society of London* series B, **354**, 769-776.

Keeling M.J. (2000) Simple stochastic models and their power-law type behaviour. *Theoretical Population Biology*, **58**, 21-31.

Kendal W.S. (2004) Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. *Ecological Complexity*, **1**, 193-209.

Kendal W.S. and Jørgensen B. (2011) Tweedie convergence: a mathematical basis for Taylor's power law, 1/f noise, and multifractality. *Physical Reviews* E, 84.066120.

Kokonendji C.C., Dossou-Gbété S. and Demétrio C.G.B. (2004) Some discrete exponential dispersion models: Poisson-Tweedie and Hinde-Demétrio classes. *SORT*, **28**, 201-214.

McCullagh P. and Nelder J.A. (1989) *Generalized Linear Models.* Second edition. Chapman & Hall, London.

Noufaily A., Enki D.G., Farrington P., Garthwaite P., Andrews N. and Charlett A. (2013) An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in Medicine*, **32**, 1206-1222.

Perry J.N. (1981) Taylor's power law for dependence of variance on mean in animal populations. *Applied Statistics*, **30**, 254-263.

Pewsey A. (2005) The large-sample distribution of the most fundamental of statistical summaries. *Journal of Statistical Planning and Inference*, **134**, 434-444.

R Development Core Team. (2014) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Taylor L.R.(1961) Aggregation, variance and the mean. *Nature*, **189**, 732-735.

# Appendix

The following R code may be used for fitting the quasi-Tweedie model, for a specified numerical value of $p$. First, set up the following list to specify the variance function.

```
varp<-list(
varfun=function(mu){mu^p},
validmu=function(mu){all(mu>0)},
dev.resids=function(y,mu,wt){
2*wt*((1/(1-p))*(y^(2-p)-y*mu^(1-p))+(1/(p-2))*(y^(2-p)-mu^(2-p)))},
initialize=expression(mustart<-ifelse(y==0,0.001,y)
)
```

The model is then fitted with a call to R function `glm`, using `family=quasi(link="log",variance=varp)`.

Table 2: Relative bias for sample skewness (uncorrected and bias-corrected) for different values of $\mu$ and $p$ for sample sizes $n = 26$ and $n = 52$, with $\phi = 1$

| | sample skewness uncorrected | | | | sample skewness bias corrected | | | |
|---|---|---|---|---|---|---|---|---|
| $n = 26$ | | | | | | | | |
| $\mu$ | 0.1 | 1 | 10 | 100 | 0.1 | 1 | 10 | 100 |
| $p = 1$ | 0.025 | -0.156 | -0.169 | -0.178 | -0.215 | -0.086 | -0.082 | 0.091 |
| $p = 1.2$ | -0.098 | -0.161 | -0.178 | -0.178 | -0.200 | -0.103 | -0.093 | -0.091 |
| $p = 1.4$ | -0.186 | -0.176 | -0.176 | -0.176 | -0.219 | -0.126 | -0.097 | -0.091 |
| $p = 1.6$ | -0.225 | -0.195 | -0.186 | -0.178 | -0.227 | -0.153 | -0.118 | -0.098 |
| $p = 1.8$ | -0.241 | -0.219 | -0.202 | -0.194 | -0.225 | -0.183 | -0.150 | -0.132 |
| $p = 2$ | -0.244 | -0.245 | -0.244 | -0.245 | -0.214 | -0.214 | -0.213 | -0.215 |
| $n = 52$ | | | | | | | | |
| $\mu$ | 0.1 | 1 | 10 | 100 | 0.1 | 1 | 10 | 100 |
| $p = 1$ | 0.038 | -0.084 | -0.097 | -0.105 | -0.077 | -0.022 | -0.022 | 0.028 |
| $p = 1.2$ | -0.071 | -0.089 | -0.096 | -0.094 | -0.089 | -0.030 | -0.023 | -0.017 |
| $p = 1.4$ | -0.120 | -0.101 | -0.097 | -0.097 | -0.106 | -0.045 | -0.026 | -0.022 |
| $p = 1.6$ | -0.143 | -0.114 | -0.106 | -0.100 | -0.112 | -0.059 | -0.039 | -0.027 |
| $p = 1.8$ | -0.152 | -0.133 | -0.122 | -0.112 | -0.110 | -0.080 | -0.062 | -0.056 |
| $p = 2$ | -0.151 | -0.153 | -0.152 | -0.153 | -0.100 | -0.103 | -0.101 | -0.102 |