Trends in disguise

Vytaras Brazauskas[1]
University of Wisconsin-Milwaukee

Bruce L. Jones[2]
University of Western Ontario

Ričardas Zitikis[3]
University of Western Ontario

Abstract. Human longevity is changing; but at what rate? Insurance claims are increasing; but at what rate? Are the trends that we glean from data true or illusionary? The shocking fact is that true trends might be quite different from those that we actually see from visualized data. Indeed, in some situations the upward trends (e.g., inflation) may even look decreasing (e.g., deflation). In this paper, we discuss this 'trends in disguise' phenomenon in detail and offer a way for estimating true trends.

Keywords and phrases: Deductible; Deflation; Inflation rate; Insurance claims; Polynomial Pareto distribution.

[1] Corresponding author: Department of Mathematical Sciences, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, Wisconsin 53201, U.S.A. E-mail address: vytaras@uwm.edu

[2] Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario N6A 5B7, Canada. E-mail address: jones@stats.uwo.ca

[3] Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario N6A 5B7, Canada. E-mail address: zitikis@stats.uwo.ca

1 Introduction

Trends of human longevity are changing (cf., e.g., Bebbington, Lai, and Zitikis, 2011; Green and Bebbington, 2013; references therein). Very likely, they are increasing, and might be even more increasing than we have thought so far (cf., e.g., Ediev, 2011, 2013).

Non-life insurance claims are increasing, and perhaps faster than we have observed them (e.g., Brazauskas, Jones, and Zitikis, 2009; Fackler, 2011).

Health care costs are increasing too, and at much higher rates than the general inflation (e.g., Committee on Post-Employment Benefit Plans, 2012).

Due to their importance, these issues have been actively researched and discussed by academics and practitioners (e.g., Brickmann, Forster, and Sheaf, 2005; Gesmann, Rayees, and Clapham, 2013). The current consensus is that the problem of estimating such trends with any degree of certainty is quite challenging. To illustrate the ongoing discussion, next is a quote from Brickmann, Forster, and Sheaf (2005) that eloquently summarizes the challenge:

> Claims inflation is one of the key assumptions used by non-life actuaries. An appreciation of claims inflation rates is needed in virtually all the areas that non-life actuaries get involved, including reserving, pricing, planning and capital modelling. For reserving, an understanding of historical rates of claims inflation is of primary importance, whereas for planning, and capital modelling, the actuary needs to understand the expected future rates. For pricing, both the past and future rates are needed – the former in order to restate the past years on a comparable basis to the current year, and the latter in order to project the results into the next policy year. Like it or not, non-life actuaries cannot get away from claims inflation! Which is a bit of a problem since it can be very difficult to measure. It is hard to accurately gauge the level claims inflation has been running at in each past year. This is because the truth is hidden in the claims data which is distorted by lots of other factors such as changes in the mix of business, changes in limits, deductibles and policy terms, and changes in settlement patterns. And, however difficult it is to put figures on past inflation rates, it is even harder to estimate the future level of claims inflation.

Indeed, the shocking fact is that in various phenomena the true underlying trends, which we cannot always observe, can be quite different from those that we actually see even when observations are affected by simple 'filters' such as truncation, in which case we only see an outcome if it is above a certain threshold (e.g., an insurance deductible) and do not see it otherwise. Think also of layers such as 'infant mortality' and 'senescent mortality' that have been of particular interest in the literature (e.g., Green and

Bebbington, 2013; references therein). We indeed find many layer-based considerations in diverse research areas, such as demography, finance, insurance, reliability engineering, and survival analysis.

To illustrate the phenomenon, in Figure 1.1 we have produced six plots: the three left-hand panels depict the yearly means and medians of *true-but-unobserved* insurance losses, and the three right-hand panels depict the corresponding yearly means and medians of *observed* insurance losses, that is, which are above a certain deductible. In Section 2 we shall give details on the underlying distributions, their parameters, and deductible values. At the moment, we only note that the assumed 'true' inflation rate is 10% per year (three left-hand panels) whereas the observed inflation rates are 0.76%, 0%, -1.05% (mean based), and 0.58%, 0%, -0.53% (median based), which are depicted in the three right-hand panels from top to bottom. Notice that the third inflation rate, which corresponds to the bottom right-hand panel, is even negative, let alone zero as is the case in the middle right-hand panel.

To gain additional preliminary intuition on the distributions used in Figure 1.1, in Figure 1.2 we have depicted the density plots of the three parent distributions (left-hand panel) alongside their corresponding three truncated versions (i.e, after deductibles) in the right-hand panel. One can easily see that while the three complete densities are markedly different, their truncated versions are almost indistinguishable. Note that data truncation, estimation of risk capital, and other related issues have also been actively researched in the literature on operational risk modeling (e.g., Cavallo, Rosenthal, Wang, and Yan, 2012; Opdyke and Cavallo, 2012).

The purpose of this paper is to discuss the aforementioned 'trends in disguise' phenomenon, and to also offer a practical and efficient way for estimating the true underlying trends, such as the true inflation rate of insurance claims. We have organized the rest of the paper as follows. In Section 2 we shall provide three distributions that illustrate different scenarios of trends in disguise. In particular, for a positive underlying trend, we shall explore the cases when (*i*) the observed trend is positive but smaller than the actual one, (*ii*) the observed trend is zero, and (*iii*) the observed trend is negative. Furthermore, in Section 3 we shall introduce a model and a statistical technique for finding the true trend, derive the likelihood function for estimating unknown parameters, and present a numerical example that deals with losses of an employee group under a prescription drug coverage. In Section 4 we shall summarize the findings of this paper and offer a few concluding remarks.

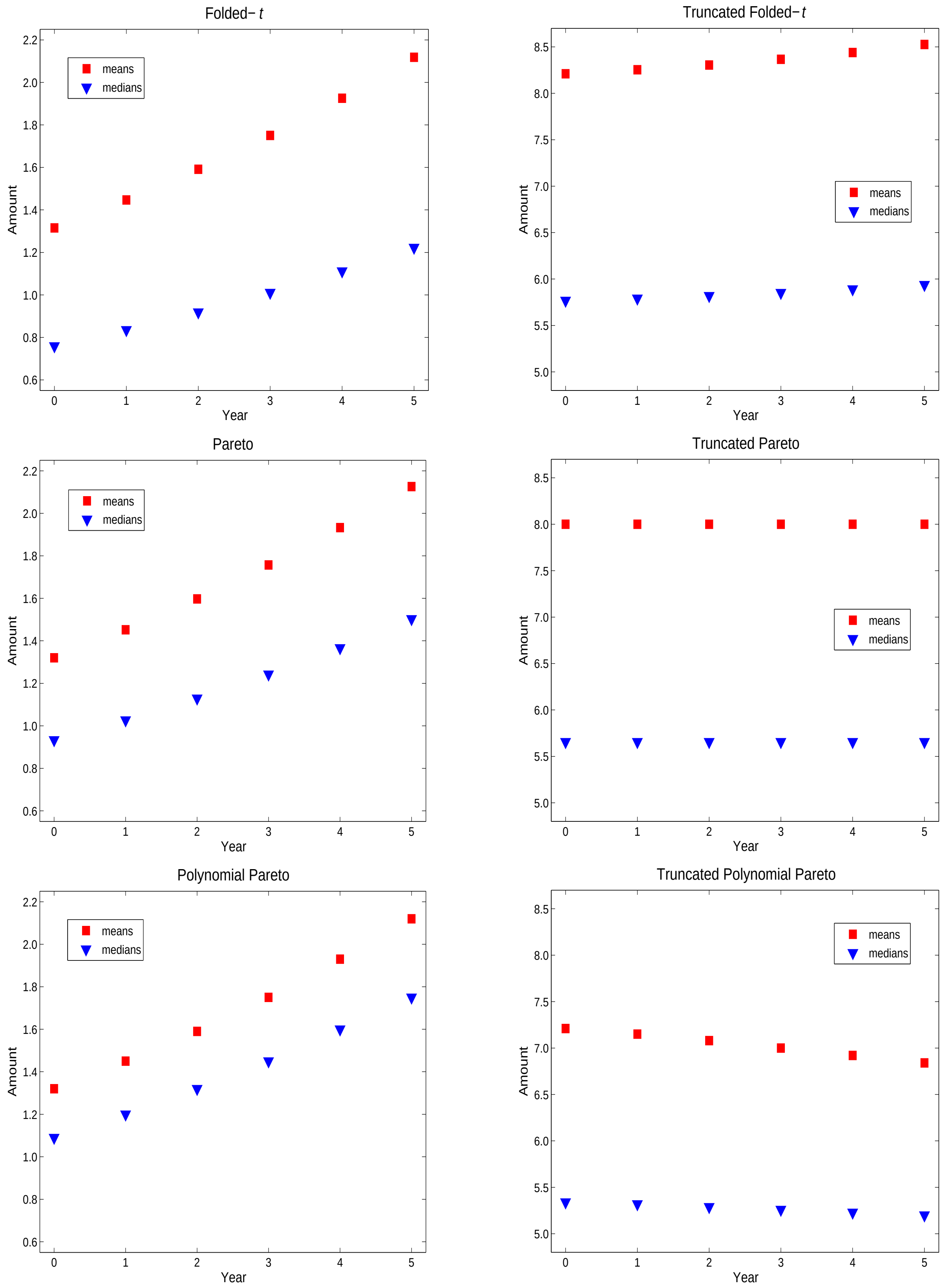

Figure 1.1: Trends in disguise: means and medians of *unobserved* (left-hand column) and *observed* (right-hand column) insurance losses from selected distributions.

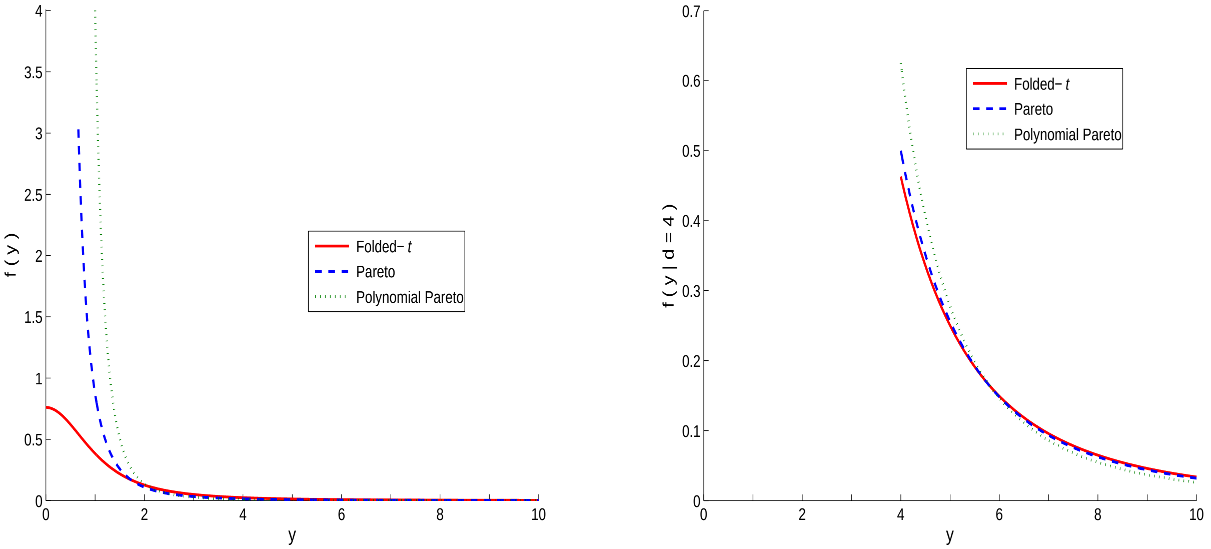

Figure 1.2: Complete probability density functions of the folded-t, Pareto, and Polynomial Pareto distributions (left-hand panel) and their truncated versions (right-hand panel).

2 How much can the true trends be disguised by the observed ones?

The immediate answer to the question posed in the title is that true and observed insurance losses (or whatever variables we might be analyzing) might exhibit trends that are comonotonic with different rates, or even exhibit anticomonotonic trends. Indeed, trends in observed variables can look slower than in the true ones (Subsection 2.1), may even completely disappear (Subsection 2.2), and most disturbingly, while trends in true unobserved variables might be positive, the trends of observed variables can be negative (Subsection 2.3). Moreover, as we shall see in the following three subsections, the loss models which our illustrations are based upon are attractive and practically sound parametric distributions of insurance claims.

2.1 Inflation can look diminished

Suppose that during each year $j \in \{1, \ldots, J\}$ for J consecutive years, we observe insurance claims that follow the folded-t distribution, which has been successfully used by Brazauskas and Kleefeld (2011, 2014) and Scollnik (2014) to model insurance data. In particular, the authors have demonstrated that this distribution well captures the highly-skewed and heavy-tailed nature of insurance claims. The survival function of the folded-t distribution is given by

$$S_{\mathrm{FT}(\nu,\,\sigma)}(x) = 2\, S_{\mathrm{T}(\nu)}(x/\sigma), \qquad x > 0, \tag{2.1}$$

where $S_{\mathrm{T}(\nu)}(x) = \int_x^\infty f_{\mathrm{T}(\nu)}(t)\,dt$ is the survival function of Student's t distribution with ν degrees of freedom, and its density $f_{\mathrm{T}(\nu)}$ is defined by

$$f_{\mathrm{T}(\nu)}(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{1}{2}\right)} \frac{1}{\sqrt{\nu}} \frac{1}{\left(1 + t^2/\nu\right)^{(\nu+1)/2}}, \qquad -\infty < t < \infty,$$

where Γ denotes the gamma function. (Note that $\nu > 0$ is not required to be integer.) As follows from Brazauskas and Kleefeld (2011, Section 2), the mean and the median of $X_{\mathrm{FT}(\nu,\sigma)} \sim S_{\mathrm{FT}(\nu,\sigma)}$ are given by the formulas

$$\mathbf{E}\big[X_{\mathrm{FT}(\nu,\sigma)}\big] = \sigma\sqrt{\nu}\,\frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{1}{2}\right)}, \qquad \nu > 1, \tag{2.2}$$

and

$$\mathbf{median}\big[X_{\mathrm{FT}(\nu,\sigma)}\big] = \sigma\,F^{-1}_{\mathrm{T}(\nu)}(0.75), \tag{2.3}$$

where $F^{-1}_{\mathrm{T}(\nu)}$ is the quantile function of Student's t distribution with ν degrees of freedom. Now we are ready to provide our first bit of information about Figure 1.1.

Note 2.1 Let $\nu = 2$, $\sigma_0 = 0.93$, and $\sigma_j = \sigma_0(1+r)^{j-1}$ with $r = 0.10$. With σ_j instead of σ on the right-hand sides of equations (2.2) and (2.3), the resulting means and medians of the losses for each of the six years $j = 1, 2, \ldots, 6$ are depicted in the top left-hand panel of Figure 1.1.

Furthermore, for the truncated (at d) and limited (at u) folded-t random variable, the survival function can be derived as follows:

$$\begin{aligned} S^{(d,u]}_{\mathrm{FT}(\nu,\sigma)}(x) &= \mathbf{P}\big[\min\{X_{\mathrm{FT}(\nu,\sigma)}, u\} > x \mid X_{\mathrm{FT}(\nu,\sigma)} > d\big] \\ &= \frac{\mathbf{P}\big[\min\{X_{\mathrm{FT}(\nu,\sigma)}, u\} > x,\ X_{\mathrm{FT}(\nu,\sigma)} > d,\ X_{\mathrm{FT}(\nu,\sigma)} > u\big]}{\mathbf{P}\big[X_{\mathrm{FT}(\nu,\sigma)} > d\big]} \\ &\quad + \frac{\mathbf{P}\big[\min\{X_{\mathrm{FT}(\nu,\sigma)}, u\} > x,\ X_{\mathrm{FT}(\nu,\sigma)} > d,\ X_{\mathrm{FT}(\nu,\sigma)} \le u\big]}{\mathbf{P}\big[X_{\mathrm{FT}(\nu,\sigma)} > d\big]} \\ &= \frac{\mathbf{P}\big[X_{\mathrm{FT}(\nu,\sigma)} > u\big]\,\mathbf{1}\{x < u\} + \mathbf{P}\big[x < X_{\mathrm{FT}(\nu,\sigma)} \le u\big]\,\mathbf{1}\{x < u\}}{\mathbf{P}\big[X_{\mathrm{FT}(\nu,\sigma)} > d\big]} \\ &= \frac{S_{\mathrm{T}(\nu)}(x/\sigma)}{S_{\mathrm{T}(\nu)}(d/\sigma)}\,\mathbf{1}\{x < u\}, \qquad x > d, \end{aligned} \tag{2.4}$$

where the indicator $\mathbf{1}\{x < u\}$ is equal to 1 when $x < u$ and 0 when $x \ge u$. (Note that $S^{(d,u]}_{\mathrm{FT}(\nu,\sigma)}(x) = 1$ for $x \le d$.) To derive formula for the mean of the truncated-and-limited folded-t random variable $X^{(d,u]}_{\text{FT}(\nu,\,\sigma)} \sim S^{(d,u]}_{\text{FT}(\nu,\,\sigma)}$, we use expression (2.4), integration by parts, and straightforward simplifications:

$$\begin{aligned}
\mathbf{E}\big[X^{(d,u]}_{\text{FT}(\nu,\,\sigma)}\big] &= \int_0^d S^{(d,u]}_{\text{FT}(\nu,\,\sigma)}\,dx \;+\; \int_d^u S^{(d,u]}_{\text{FT}(\nu,\,\sigma)}\,dx \;+\; \int_u^\infty S^{(d,u]}_{\text{FT}(\nu,\,\sigma)}\,dx \\
&= d \;+\; \frac{1}{S_{\text{T}(\nu)}(d/\sigma)} \int_d^u S_{\text{T}(\nu)}(x/\sigma)\,dx \;+\; 0 \\
&= \frac{S_{\text{T}(\nu)}(u/\sigma)}{S_{\text{T}(\nu)}(d/\sigma)} \cdot u \;+\; \frac{\sigma}{S_{\text{T}(\nu)}(d/\sigma)} \int_{d/\sigma}^{u/\sigma} t f_{\text{T}(\nu)}(t)\,dt \\
&= \frac{S_{\text{T}(\nu)}(u/\sigma)}{S_{\text{T}(\nu)}(d/\sigma)} \cdot u \;+\; \frac{\sigma}{S_{\text{T}(\nu)}(d/\sigma)}\,\frac{\nu}{\nu-1}\left[\left(1+\frac{(d/\sigma)^2}{\nu}\right) f_{\text{T}(\nu)}(d/\sigma)\right. \\
&\qquad \left. -\;\left(1+\frac{(u/\sigma)^2}{\nu}\right) f_{\text{T}(\nu)}(u/\sigma)\right], \qquad \nu > 1.
\end{aligned} \tag{2.5}$$

And expression for the median of the truncated-and-limited folded-t random variable $X^{(d,u]}_{\text{FT}(\nu,\,\sigma)} \sim S^{(d,u]}_{\text{FT}(\nu,\,\sigma)}$ follows directly from (2.4):

$$\mathbf{median}\big[X^{(d,u]}_{\text{FT}(\nu,\,\sigma)}\big] = \min\left\{u,\, \sigma\, F^{-1}_{\text{T}(\nu)}\Big(0.5 + 0.5 F_{\text{T}(\nu)}(d/\sigma)\Big)\right\}. \tag{2.6}$$

In particular, when $u = \infty$, then we call $X^{(d,\infty]}_{\text{FT}(\nu,\,\sigma)}$ and $S^{(d,\infty]}_{\text{FT}(\nu,\,\sigma)}$ the truncated (at d) folded-t random variable and survival function, respectively. The corresponding cdf $F^{(d,\infty]}_{\text{FT}(\nu,\,\sigma)}$ for various parameter values is depicted in Figure 2.1. The mean and the median of $X^{(d,\infty]}_{\text{FT}(\nu,\,\sigma)}$ can be derived from equations (2.5) and (2.6) by simply setting the parameter

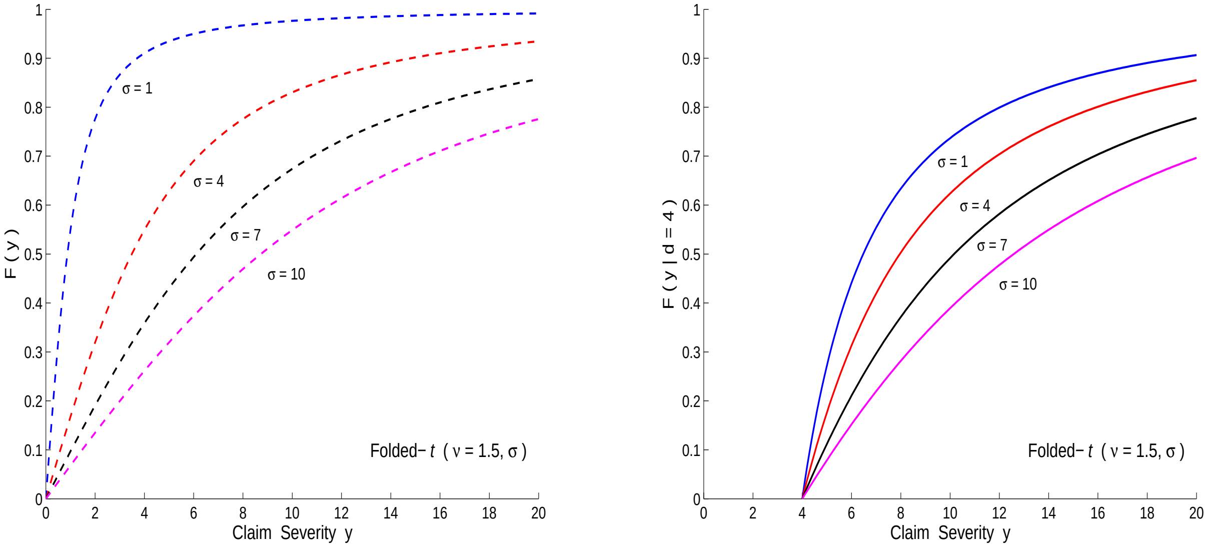

Figure 2.1: Complete-but-*unobservable* (left-hand panel) and truncated-and-*observable* (right-hand panel) folded-t distributions of inflated losses with $\nu = 1.5$ and various values of σ.

u to ∞, in which case we obtain the following equations:

$$\mathbf{E}\big[X^{(d,\infty]}_{\mathrm{FT}(\nu,\,\sigma)}\big] = \sigma\,\frac{\nu}{\nu-1}\left(1+\frac{(d/\sigma)^2}{\nu}\right)h_{\mathrm{T}(\nu)}(d/\sigma), \qquad \nu>1, \tag{2.7}$$

and

$$\mathbf{median}\big[X^{(d,\infty]}_{\mathrm{FT}(\nu,\,\sigma)}\big] = \sigma\,F^{-1}_{\mathrm{T}(\nu)}\Big(0.5+0.5F_{\mathrm{T}(\nu)}(d/\sigma)\Big), \tag{2.8}$$

where $h_{\mathrm{T}(\nu)} = f_{\mathrm{T}(\nu)}/S_{\mathrm{T}(\nu)}$ denotes the hazard rate function of Student's t distribution with ν degrees of freedom. Notice also that if we let $d=0$ in equations (2.7) and (2.8), they become (2.2) and (2.3), respectively.

Note 2.2 Let $\nu=2$, $\sigma_0=0.93$, and $\sigma_j=\sigma_0(1+r)^{j-1}$ with $r=0.10$. Furthermore, let $d=4$. Then with σ_j instead of σ on the right-hand sides of equations (2.7) and (2.8), the means and the medians of the truncated (at d) losses for each of the six years $j=1,2,\dots,6$ are depicted in the top right-hand panel of Figure 1.1, with the inflation rates being 0.76% (mean based) and 0.58% (median based).

2.2 Inflation can become invisible

Suppose that during each year $j\in\{1,\dots,J\}$ for J consecutive years, we observe insurance claims that follow the classical Pareto distribution of the first kind, also known as a single-parameter Pareto distribution in actuarial literature, which has the survival function

$$S_{\mathrm{P}(\alpha,\,\theta)}(x) = \left(\frac{\theta}{x}\right)^{\alpha}, \quad x\ge\theta, \tag{2.9}$$

where $\alpha>0$ and $\theta>0$ are the shape and scale parameters, respectively. Recall that the mean of this distribution is $\alpha\theta/(\alpha-1)$, assuming of course $\alpha>1$, and the median is $2^{1/\alpha}\theta$.

Note 2.3 Let $\alpha=2$, $\theta_0=0.66$, and $\theta_j=\theta_0(1+r)^{j-1}$ with $r=0.10$. With θ_j instead of θ, the resulting means and medians of the losses for each of the six years $j=1,2,\dots,6$ are depicted in the middle left-hand panel of Figure 1.1.

Furthermore, the truncated (at d) and limited (at u) Pareto random variable has the survival function (cf. equation (2.4))

$$S^{(d,u]}_{\mathrm{P}(\alpha,\,\theta)}(x) = \left(\frac{d}{x}\right)^{\alpha}\mathbf{1}\{x<u\}, \qquad x>d.$$

Hence, the corresponding mean and median are

$$\mathbf{E}\big[X^{(d,u]}_{\mathrm{P}(\alpha,\,\theta)}\big] = \frac{\alpha d}{\alpha-1} - \frac{u(d/u)^{\alpha}}{\alpha-1}, \quad \alpha>1, \tag{2.10}$$

and

$$\mathbf{median}\big[X^{(d,u]}_{\mathrm{P}(\alpha,\,\theta)}\big] = \min\big\{u,\, 2^{1/\alpha}d\big\}\,. \tag{2.11}$$

Note 2.4 Let $\alpha = 2$, $\theta_0 = 0.66$, and $\theta_j = \theta_0(1+r)^{j-1}$ with $r = 0.10$. Furthermore, let $d = 4$ and $u = \infty$. Since equations (2.10) and (2.11) are independent of θ_j's, we have constant means and medians of the truncated (at d) losses for each of the six years $j = 1, 2, \ldots, 6$. This fact is seen from the middle right-hand panel of Figure 1.1, where the inflation rate is 0%.

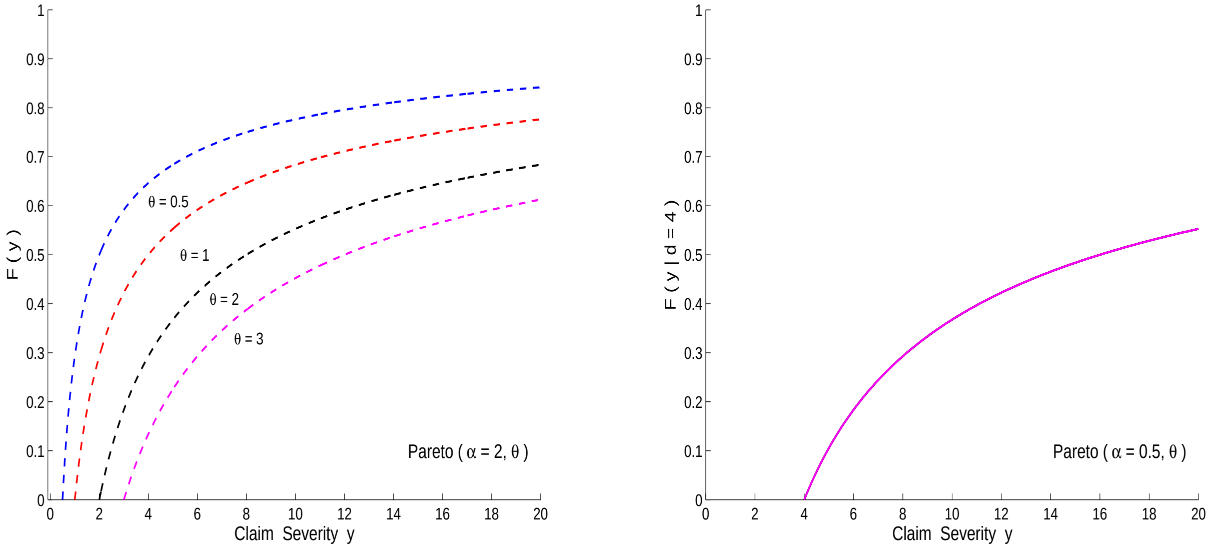

Figure 2.2: Complete-but-*unobservable* (left-hand panel) and truncated-and-*observable* (right-hand panel) Pareto distributions of inflated losses with $\alpha = 0.5$ and various θ.

2.3 Inflation can look even like deflation

Suppose that during each year $j \in \{1, \ldots, J\}$ for J consecutive years, we observe insurance claims that follow what we call the polynomial Pareto (PP) distribution, which has the survival function

$$S_{\mathrm{PP}(\alpha,\,\beta,\,\theta)}(x) = \left(\frac{\theta}{x}\right)^{\alpha} \exp\left\{\beta\left(\frac{1}{x} - \frac{1}{\theta}\right)\right\}, \quad x \geq \theta, \tag{2.12}$$

with parameters $\alpha > 0$, $\beta \geq 0$, and $\theta > 0$. Notice that when $\beta = 0$, then formula (2.12) reduces to (2.9) and thus the PP model becomes the standard Pareto one. We note in passing that we have created the PP distribution specifically for the purpose of illustrating the message stated in the title of this subsection. For other similar distributions, we refer to, e.g., Lai and Xie (2006), and Marshall and Olkin (2007).

Note 2.5 The hazard rate function of the polynomial Pareto distribution is

$$h_{\mathrm{PP}(\alpha,\,\beta,\,\theta)}(x) = \frac{\alpha}{x} + \frac{\beta}{x^2}, \quad x \geq \theta,$$

which can be viewed as a polynomial of the standard Pareto hazard rate function $1/x$, and this explains why we have decided to call this distribution 'polynomial Pareto.' Note also that h_{PP} is a special case of the class of rational hazard rate functions (cf., e.g., Bebbington, Lai, Murthy, and Zitikis, 2010).

Similarly to the previous subsection, the truncated (at d) and limited (at u) polynomial Pareto random variable has the survival function

$$S^{(d,u]}_{\mathrm{PP}(\alpha,\,\beta,\,\theta)}(x) = \left(\frac{d}{x}\right)^{\alpha} \exp\left\{\beta\left(\frac{1}{x} - \frac{1}{d}\right)\right\} \mathbf{1}\{x < u\}, \qquad x > d.$$

Unlike in the previous sections, the mean and the median of this distribution cannot be reduced to simple closed-form expressions, and thus we have to evaluate them numerically. Now we are ready to present our last bit of information about Figure 1.1.

Note 2.6 Let $\alpha = \beta = 2$, $\theta_0 = 1$, $\beta_j = \beta_0(1+r)^{j-1}$ and $\theta_j = \theta_0(1+r)^{j-1}$ with $r = 0.10$. The means and the medians of the ground-up losses for each of the six years $j = 1, 2, \ldots, 6$ are depicted in the bottom left-hand panel of Figure 1.1. Furthermore, with $d = 4$ and $u = \infty$, the means and the medians of the truncated (at d) losses for each of the six years $j = 1, 2, \ldots, 6$ are depicted in the bottom right-hand panel of Figure 1.1, with the inflation rates being -1.05% (mean based) and -0.53% (median based).

3 Estimating the true inflation

The results and discussions of the previous sections show that we cannot rely on eyeballing graphs for determining the true inflation rate – a model needs to be developed and appropriate statistical tools used. Bellow, we shall suggest an effective method for gleaning out true underlying inflation rates from data extracted from layers, such as those above a deductible or, more generally, above a specified threshold.

3.1 Model

Suppose that during each year $j \in \{1, \ldots, J\}$ for J consecutive years, random variables $Y_{j,1}, \ldots, Y_{j,N_j}$ manifest themselves but not all of them are actually seen. Namely, we can see only those $Y_{j,k}$'s whose values are in a certain layer, say in the interval $(d_j, u_j]$; we

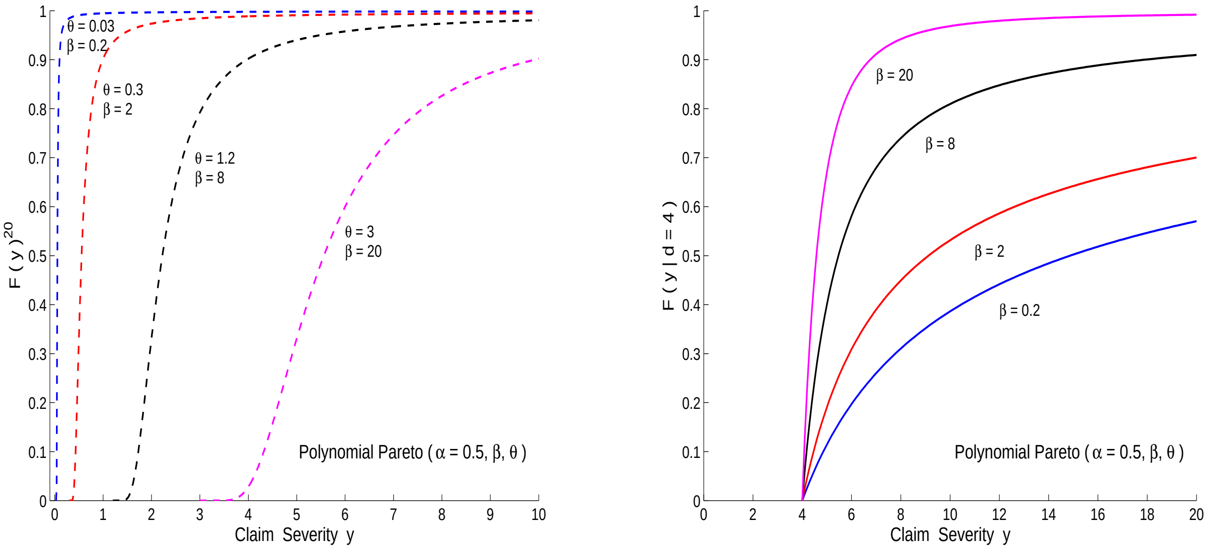

Figure 2.3: Complete-but-*unobservable* (left-hand panel) and truncated-and-*observable* (right-hand panel) Polynomial Pareto distributions of inflated losses with $\alpha = 0.5$ and various values of β and θ.

can think of the lower limit d_j as an insurance deductible and of the upper limit u_j as an insurance policy limit. Hence, we observe the random variables

$$X_{j,k} = \min\{Y_{j,k}, u_j\} \mid Y_{j,k} > d_j,$$

and there are $M_j = \sum_{i=1}^{N_j} \mathbf{1}\{Y_{j,k} > d_j\}$ of them with values either $Y_{j,k}$ or u_j depending on the magnitude of $Y_{j,k}$.

Denote the distribution of the counts N_j by $\mathcal{P}_j = \{P_j(n),\, n = 0, 1, 2, \ldots\}$. These integer-valued random variables N_j are not observable; only M_j's are observable. Denote the distribution of M_j by $\mathcal{Q}_j = \{Q_j(m),\, m = 0, 1, 2, \ldots\}$, which is given by the formula

$$Q_j(m) = \sum_{n=m}^{\infty} P_j(n) \binom{n}{m} S_j^m(d_j) F_j^{n-m}(d_j).$$

The following examples (cf., e.g., Klugman, Panjer, and Willmot, 2008) illustrate the distributions $\mathcal{P}_j$ and $\mathcal{Q}_j$:

(1) If $P_j(n) = \text{Poisson}(n \mid \lambda_j)$ for all $n = 0, 1, 2, \ldots$, where $\lambda_j > 0$ is a parameter, then $Q_j(m) = \text{Poisson}(n \mid \lambda_j S_j(d_j))$ for all $m = 0, 1, 2, \ldots$.

(2) If $P_j(n) = \text{NB}(n \mid \alpha_j, \theta_j)$ for all $n = 0, 1, 2, \ldots$, where $\alpha_j > 0$ and $\theta_j > 0$ are parameters, and NB stands for the negative binomial distribution, then $Q_j(m) = \text{NB}(n \mid \alpha_j, \theta_j S_j(d_j))$ for all $m = 0, 1, 2, \ldots$.

(3) If $P_j(n) = \int \text{Poisson}(n \mid \lambda)\, dG_j(\lambda)$ for all $n = 0, 1, 2, \ldots$, where G_j is the cdf of a random variable, then $Q_j(m) = \int \text{Poisson}(n \mid \lambda S_j(d_j))\, dG_j(\lambda)$ for all $m = 0, 1, 2, \ldots$.

Note 3.1 When G_j is degenerate at the point $\lambda_j > 0$, then (3) reduces to (1), and when G_j is the gamma distribution with parameters $\alpha_j > 0$ and $\theta_j > 0$, then (3) reduces to (2).

Assume that all the underlying random variables $Y_{j,k}$ are independent, have densities, and let, for each year $j \in \{1, \ldots, J\}$, the densities of the random variables $Y_{j,1}, Y_{j,2}, \ldots$ be f_j. The dependence of the density f_j on j stems from the underlying problem. Namely, we assume that there is a certain trend $\vec{\sigma} = (\sigma_1, \ldots, \sigma_J)$ from year to year (think of inflation) and thus there is an underlying cumulative distribution function (cdf) F_0 such that the cdf of $Y_{j,k}$ is equal to $F_0(y/\sigma_j)$, and thus

$$f_j(y) = \frac{1}{\sigma_j} f_0\left(\frac{y}{\sigma_j}\right),$$

where f_0 is the density of F_0. We also assume that $Y_{j,1}, Y_{j,2}, \ldots$ are independent of N_j, that is, the claim severities are independent of the frequencies. This is a common assumption in the classical actuarial literature (cf., e.g., Klugman, Panjer, and Willmot, 2008), but there are of course many research papers that depart from it (cf., e.g., Sendova and Zitikis, 2012; Li and Sendova, 2014; and references therein). Our task is to estimate the trend $\vec{\sigma}$, keeping in mind that we observe only $X_{j,k}$'s, but not $Y_{j,k}$'s.

3.2 Estimation

To estimate the trend $\vec{\sigma} = (\sigma_1, \ldots, \sigma_J)$, we use the maximum likelihood technique and thus need to derive the complete likelihood, which is

$$\begin{aligned} L &= \prod_{j=1}^{J} Q_j(m_j) \prod_{k=1}^{m_j} \left(\frac{f_j(y_{j,k})}{S_j(d_j)}\right)^{\mathbf{1}\{d_j < y_{j,k} < u_j\}} \left(\frac{S_j(u_j)}{S_j(d_j)}\right)^{\mathbf{1}\{y_{j,k} \geq u_j\}} \\ &= \prod_{j=1}^{J} Q_j(m_j) \frac{S_j(u_j)^{\sum_{k=1}^{m_j} \mathbf{1}\{y_{j,k} \geq u_j\}}}{S_j(d_j)^{m_j}} \prod_{k=1}^{m_j} f_j(y_{j,k})^{\mathbf{1}\{d_j < y_{j,k} < u_j\}}, \end{aligned} \tag{3.1}$$

where $S_j(y) = 1 - F_j(y)$, and m_j is the number of observed $y_{j,k}$'s, that is, all of those $y_{j,k}$'s that are above d_j.

While the second equation of (3.1) is a simple rearrangement of the terms in the previous expression, the first equation needs some explanation. We begin by noting that, conditionally on the number $M_j = m_j$ of observations, the inside product on the right-hand side of the first equation multiplies the likelihoods of all the observed values: if a value $y_{j,k}$ has been recorded, then its conditional on being above d_j likelihood is

$f(y_{j,k})/S_j(d_j)$, but if the value u_j has been recorded, then we do not know the actual value of the corresponding loss $y_{j,k}$, except that it is somewhere on or above u_j, and thus the conditional likelihood becomes $S_j(u_j)/S_j(d_j)$. Since the entire inside product is the conditional likelihood upon $M_j = m_j$, we turn it into an unconditional likelihood by multiplying it by the probability $Q_j(m_j)$ of $M_j = m_j$. Since the losses over the J years are independent by assumption, we obtain the complete likelihood by multiplying the individual $j = 1, \ldots, J$ likelihoods. This leads us to the complete likelihood given by formula (3.1).

Since we observe the random variables $X_{j,k} = \min\{Y_{j,k}, u_j\} \mid Y_{j,k} > d_j$, we next rewrite formula (3.1) in terms of $x_{j,k}$'s:

$$L = \prod_{j=1}^{J} Q_j(m_j) \frac{S_j(u_j)^{\sum_{k=1}^{m_j} \mathbf{1}\{x_{j,k}=u_j\}}}{S_j(d_j)^{m_j}} \prod_{k=1}^{m_j} f_j(x_{j,k})^{\mathbf{1}\{x_{j,k}<u_j\}}, \tag{3.2}$$

We shall use formula (3.2) in our illustrative example below.

3.3 Example

We illustrate the above ideas by considering losses of an employee group under a prescription drug coverage. Specifically, we investigate losses during the five consecutive years 2008–2012 for drugs prescribed to treat cancer. The data set is confidential, and we thus do not elaborate here on its source, nor provide further details, except that we do provide the necessary summary statistics that are sufficient for our following analysis and also for the appreciation of results.

We proceed as follows. Even though our data set is complete, and we do use it to check the accuracy of our findings based on the above described methodology, we start our considerations by first constructing a new data set by artificially imposing a deductible of 1,000 dollars on the original data. Table 3.1 provides summaries of the complete as well as of the truncated data that result from imposing the noted deductible. We

Table 3.1: Summary of cancer drug loss data.

Year	Exposure	Number of losses	Average of losses	Number of losses $> 1,000$	Average of losses $> 1,000$
2008	52,239	3,083	449.44	444	2,190.23
2009	52,950	2,968	510.64	445	2,363.51
2010	52,853	3,087	488.20	450	2,386.79
2011	52,158	3,208	496.55	454	2,521.35
2012	50,649	3,366	568.04	503	3,053.64

see that there are roughly 50,000 life-years of exposure in each calendar year, meaning that approximately 50,000 individuals are covered by the prescription drug insurance coverage. While the number of cancer drug losses seems high, we must recognize that a given individual may have multiple prescription drug losses in a year. Notice that a fairly small fraction of losses exceed 1,000, and that the average loss among losses above 1,000 is large due to the long tail of the loss distribution.

Histograms of the losses in each calendar year are given in the five panels of Figure 3.1, where we have found that it is more informative to examine histograms of the logarithms

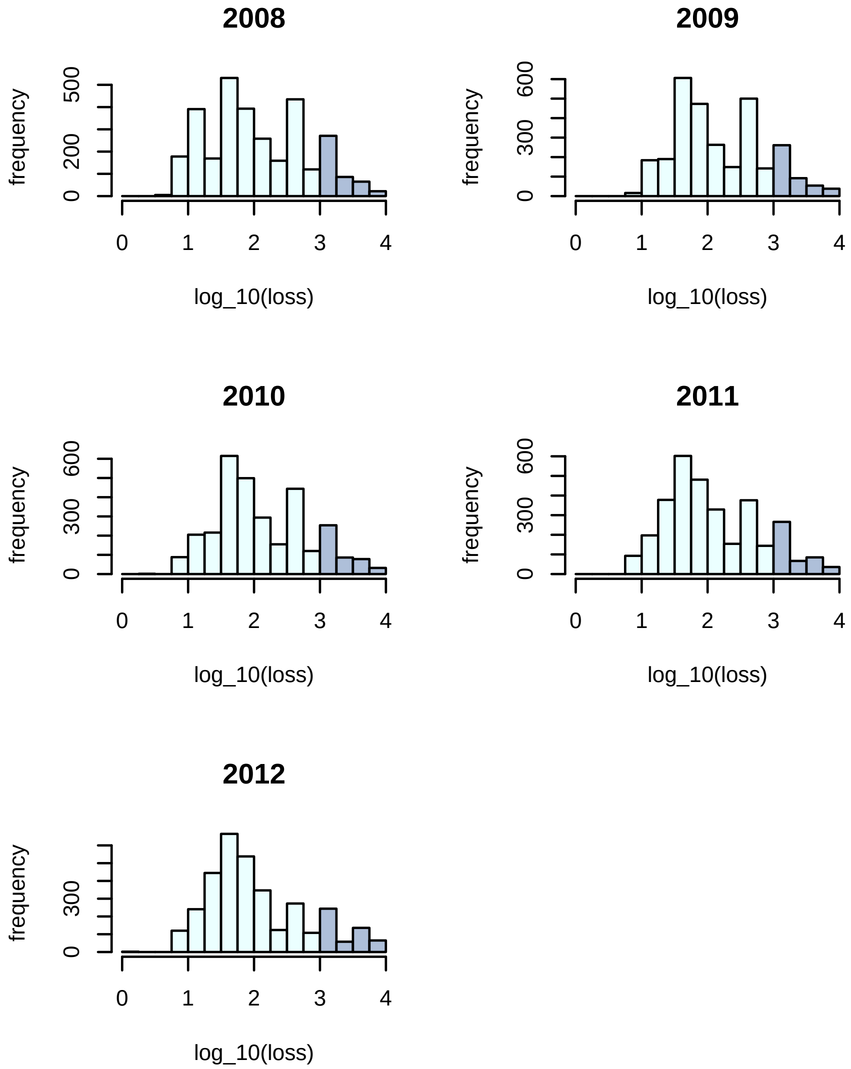

Figure 3.1: Histograms of the logarithms (base 10) of cancer drug losses for the five years from 2008 to 2012.

(base 10) of the losses than the histograms of the losses themselves. The distributions of the losses in excess of 1,000 are then delineated by the bars of the histograms that correspond to values greater than 3.

While the histograms provide information about the loss distributions, it is difficult to glean any information about the trend in loss amounts. We can calculate crude estimates of the inflation rate using the averages shown in Table 3.1 by the formula

$$\widehat{r}_{crude} = \left(\frac{\text{average loss in 2012}}{\text{average loss in 2008}}\right)^{1/4} - 1.$$

Using the averages of all loss amounts, the crude inflation rate estimate is 0.0450, but if we assume that we observe only the losses above 1,000, then the crude inflation rate estimate becomes 0.0757. We know, however, that this estimate is not an appropriate estimate because it does not correctly allow for the truncation.

In order to improve upon our estimate while still using only the observed losses that exceed 1,000, we employ the model developed above. Namely, we assume that in each year $j = 2008, \ldots, 2012$, the number of losses follows the Poisson distribution with the mean $\lambda_j = \lambda e_j$, where e_j is the exposure in the year j (see Table 3.1). Then the number of losses above 1,000 in the year j has a Poisson distribution with the mean $\lambda_j S_j(1,000)$. We further assume that losses in each year j are independent and identically distributed, and follow the folded-t distribution with parameters ν and $\sigma_j = \sigma(1+r)^{j-2008}$. We have four parameters to estimate: λ, r, ν, and σ. Using the likelihood technique described above, we obtain the estimates reported in Table 3.2. A few comments follow.

Table 3.2: Maximum likelihood estimates of parameters with approximate 95 percent likelihood ratio confidence intervals.

Parameter	Estimate	95% CI
λ	0.0355	(0.0341, 0.0370)
r	0.0411	(0.0269, 0.0556)
ν	1.61	(1.55, 1.67)
σ	520	(503, 538)

First, note that the model-based estimate 0.0411 of the true inflation rate is close to the ‘true’ crudely estimated rate 0.0450 which is based on the entire data set. Note also that the other crude estimate 0.0757, which is based on the truncated data, is quite different from the model-based estimate.

Of course, we should bear in mind that estimates depend on how well our model describes the loss distribution for each year. We can check this for losses above 1,000. Having said this, we want to point out that given the complete data set that we actually have, we do know the loss amounts even below 1,000, but in general we would not normally have any information about the losses below the deductible. Hence, our conclusions about loss trends would depend on some aspects of the model that may not be verifiable.

Next, our data on losses above 1,000 provide information about the product $\lambda S_j(1,000)$, but not about λ and $S_j(1,000)$ individually. Hence, our estimates of λ and $S_j(1,000)$ may be quite poor even if our estimates of the rate at which losses above 1,000 occur and of the conditional distribution of losses above 1,000 happen to be quite good.

Finally, it would appear at least in our real-life illustrative example that having data for five years is not enough to obtain a very good estimate of the trend: the year-to-year variation in the loss distribution is too great. Nevertheless, the illustrative example has successfully demonstrated that with good care of the matter, one can nevertheless extract a reasonably accurate estimate of the true inflation rate even from data layers, such as that above deductible.

4 Concluding notes

In this paper, we have studied the 'trends in disguise' effect that is caused by data truncation. In particular, it was shown that depending on the underlying probability distribution, the actual and observed trends can exhibit comonotonic (but with different rates) as well as anticomonotonic patterns. In addition, we have constructed a model and derived its likelihood function that can be successfully used for estimating the true underlying trends. A numerical illustration of the proposed methodology has been provided using a data set of losses of an employee group under a prescription drug coverage.

The paper also suggests several avenues for future research. For example, it would be of interest to examine large-sample asymptotic as well as small-sample based properties of the maximum likelihood estimator of the underlying trend, as well as of other parameters of interest. Sensitivity of the proposed model to the distributional assumptions should also be explored. Furthermore, it would be useful to develop robust model-fitting procedures, and to evaluate their performance. Answering these and other related questions, however, has been beyond the scope and space of the current paper, due to the inevitable technical complexity of considerations.

Acknowledgments

We thank the anonymous referee for insightful comments and suggestions that helped us to improve the paper. The first author gratefully acknowledges the hospitality of the

Department of Statistical and Actuarial Sciences at the University of Western Ontario, where all three authors met for two weeks in the fall of 2013 and started working on this project. The second and third authors also acknowledge the support of their research by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] Bebbington, M., Lai, C.D., Murthy, D.N.P., and Zitikis, R. (2011). Rational polynomial hazard functions. *International Journal of Performability Engineering*, 6, 35–52.

[2] Bebbington, M., Lai, C.D., and Zitikis, R. (2011). Modelling deceleration in senescent mortality. *Mathematical Population Studies*, 18, 18–37.

[3] Brazauskas, V., Jones, B.L., and Zitikis, R. (2009). When inflation causes no increase in claim amounts. *Journal of Probability and Statistics*, Article ID 943926, 1–10.

[4] Brazauskas, V. and Kleefeld, A. (2011). Folded- and log-folded-t distributions as models for insurance loss data. *Scandinavian Actuarial Journal*, 2011(1), 59–74.

[5] Brazauskas, V. and Kleefeld, A. (2014). Authors' reply to "Letter to the Editor: Regarding folded models and the paper by Brazauskas and Kleefeld (2011)" by Scollnik. *Scandinavian Actuarial Journal*, 2014(8), 753–757.

[6] Brickmann, S., Forster, W., and Sheaf, S. (2005). Claims inflation – uses and abuses. *GIRO Convention 2005*. Institute and Faculty of Actuaries, UK.

[7] Cavallo, A., Rosenthal, B., Wang, X., and Yan, J. (2012). Treatment of the data collection threshold in operational risk: A case study with the lognormal distribution. *Journal of Operational Risk*, 7(1), 3–38.

[8] Committee on Post-Employment Benefit Plans (2012). Health care trend rate. *Educational Note*, May 2012. Canadian Institute of Actuaries.

[9] Ediev, D.M. (2011). Life expectancy in developed countries is higher than conventionally estimated. Implications from improved measurement of human longevity. *Journal of Population Ageing*, 4, 5–32.

[10] Ediev, D.M. (2013). Decompression of period old-age mortality: when adjusted for bias, the variance in the ages at death shows compression. *Mathematical Population Studies*, 20, 137–154.

[11] Fackler, M. (2011). Inflation and excess insurance. *ASTIN Colloquium 2011* (Parallel Session 8). Madrid, Spain.

[12] Gesmann, M., Rayees, R., and Clapham, E. (2013). A known unknown. *The Actuary*, May 02, 2013. Institute and Faculty of Actuaries, UK.

[13] Green, R.M. and Bebbington, M.S. (2013). A longitudinal analysis of infant and senescent mortality using mixture models. *Journal of Applied Statistics*, 40(9), 1907–1920.

[14] Klugman, S.A., Panjer, H.H., and Willmot, G.E. (2008). *Loss Models: From Data to Decisions*, 3rd edition. Wiley, New York.

[15] Lai C.D. and Xie, M. (2006). *Stochastic Ageing and Dependence for Reliability.* Springer, New York.

[16] Li, Z. and Sendova, K.P. (2014). On a ruin model with both interclaim times and premiums depending on claim sizes. *Scandinavian Actuarial Journal* (in press).

[17] Marshall, A.W. and Olkin, I. (2007). *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families.* Springer, New York.

[18] Opdyke, J.D. and Cavallo, A. (2012). Estimating operational risk capital: The challenges of truncation, the hazards of MLE, and the promise of robust statistics. *Journal of Operational Risk*, 7(3), 3–90.

[19] Scollnik, D.P.M. (2014). Letter to the Editor: Regarding folded models and the paper by Brazauskas and Kleefeld (2011). *Scandinavian Actuarial Journal*, 2014(3), 278–281.

[20] Sendova, K.P. and Zitikis, R. (2012). The order-statistic claim process with dependent claim frequencies and severities. *Journal of Statistical Theory and Practice*, 6(4), 597–620.