Financial Statistical Journal (2020) Volume 3 Issue 1
doi:10.24294/fsj.v3i1.420

Empirical Analysis of Claims Development Trapezoids following Benford's Law

Jochen Heberle*, Tobias Gummersbach

University of Hamburg, Department of Business Administration, Von-Melle-Park 5, 20146 Hamburg, Germany

ABSTRACT

In this paper we make an empirical analysis of a wide range of claims development trapezoids following Benford's law. In particular we determine Benfors's law for different characteristic factors depending on claims development triangles/trapezoids. These characteristic factors are the cumulative claims payments, the incremental claims payments and the individual development factors. For each characteristic factor hypothesis testing is done for verifying/rejecting Benford's law.

***Keywords*:** Benford-law; claims reserving; run-off triangle/trapezoid; fraud detection

JEL Classification: C12, C46, G22

1. Introduction and motivation

In this work an empirical statistical analysis is done for an actuarial dataset. Therefore we use Benford's law for demonstration. Benford's law, named for physicist Frank Benford, who worked on the theory in 1938 (cf. Benford[2]) is the mathematical theory of leading digits.

In many data sets, the leading digits of numbers are distributed in a specific way, which Benford discovered. This specific way – the Benford law – is non-linear. In Benford's distribution it states that, for example, the digit "1" appears about 30 percent of the time as first digit. On the other hand the digit "9", as first digit, appears less than 5 percent of the time (cf **Figure 2**). An easy to understand example of this behaviour are house numbers: House numbers in streets begin with the "1", but not all streets have 20 or up to 90 house numbers. So the digit "1" is the most frequently used first digit, followed by the "2" and so on.Nowadays, Benford's law is used for example in:

- Accounting fraud detection (in 2001 accounting fraud was detected in the Enron Corporation);
- election data (in 2009 Benford's law was used to detect fraud in the Iranian elections);
- genome data.

In this work we analyse a set of claims development trapezoids following Benford's law. This work is done to determine the assumption that there exists characteristic triangle/trapezoid-factors following Benford's law. The basic idea behind this work is: If an actuary has got the knowledge that specific triangle/trapezoid-factors follow a given distribution (e.g. Benford distribution) he can check given development triangles/trapezoids against this distribution. Possible reasons for checking this can be:

- Determining the plausibility of the given triangle/trapezoid;
- detecting fraud in the given data (cf. Durtschi *et al*.[6] or Diekmann & Jann[5]);
- detecting outliers (this might be helpful for further analysis).

We do not specially focus on one of these items, so the analysis made in this paper is done on a general point of view. The verification that the given set of development trapezoids, respectively some characteristic factors, follows Benford's law is done with hypothesis testing. Therefore, we use the well known Kolmogorow-Smirnow-test (see for example Govindarajulu[10, pp.182–187]).

Most actuarial science papers deal with very limited datasets for example with only one development triangle (cf.

doi: 10.24294/fsj.v3i1.420

Mack[12], England & Verrall[7] or Merz & Wüthrich[13]) if the paper is a more "theoretical" one. Or they deal with a larger set of claims development triangles generated with some statistical methods such as bootstrapping for example (cf. England & Verrall[8], Pinheiro *et al.*[14] or Heberle *et al.*[11]). In fact these larger "observation"-datasets are not real datasets – they are mostly generated from a very limited dataset. The use of only one – or especially very limited – datasets reflects from the fact that larger datasets are not – or even not easily – available for most scientists.

The structure of the paper is as follows. In Section 2 some notation is introduced and Benford's law is presented. The characteristic factors, namely the cumulative and incremental claims payments and the individual development factors are also introduced in Section 2. Section 3 is the detailed empirical analysis with a dataset made available by GR-NEAM[1]. At the end a conclusion is given in Section 4.

2. Notation and Benford's law

For reasons of simplicity we only speak of development "triangles", but all formulas hold true for development trapezoids as well.

2.1 Notation

In the following we assume that we have N development triangles and that $C_{i,j}$ denotes the cumulative payments for accident year $i \in \{0, \ldots, I\}$ and development year $j \in \{0, \ldots, J\}$ for one given development triangle. With this notation, at time $t = I$ and for a given development triangle, we have observations

$$O = \{C_{i,j} \mid i + j \leq I\} . \tag{2.1}$$

Figure 1 shows a given development triangle at time t = I. The upper left part in this triangle is observable, while the lower right part is unobservable at time t = I.

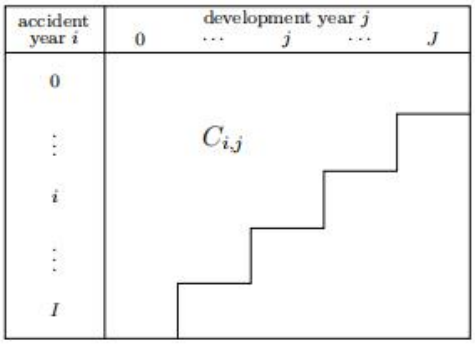

Figure 1; Observable cumulative payments *Ci,*j for a given development triangle at time $t = I$.

In the following we deal with some characteristic values which are observable or can be computed from the given sets of different observations O given by the *N* development triangles. These different characteristic values are:

1. Cumulative claims payments $C_{i,j}$ for each triangle for $i = 0, \ldots, I$ and $j = 0, \ldots, J$ with $i + j \leq I$.
2. Incremental claims payments Xi,j for each triangle for $i = 0, \ldots,$ I and $j = 0, \ldots, J$ with $i + j \leqslant I$.
3. Individual development factors *Fi,*j for each triangle for $i = 0, \ldots, I-1$ and $j = 0, \ldots, J - 1$ with $i + j \leqslant I-1$.

Remarks 2.1:

- The cumulative claims payments $C_{i,j}$ $(i + j \leq I)$ itself have not to be computed since these values are given in the upper left part of each development triangle (cf. equation (2.1)).
- The incremental claims payments $X_{i,j}$ as well as the individual development factors $F_{i,j}$ both generate "new" sets of observations, one set for every triangle.
- Each new set of observations containing the individual development factors $F_{i,j}$ is smaller than the corresponding set containing $C_{i,j}$ or $X_{i,j}$ (see equations (2.4), (2.5) and (2.6))
- The use of exact these three characteristic values is determined through a more or less excessive usage in the

[1] General Re-New England Asset Management, Inc.; Pond View Corporate Center; 76 Batterson Park Road; Farmington, CT 06032 USA.

literature.

Since the cumulative claims payments have to be given in our framework, the incremental claims payments as well as the individual development factors must be defined. These definitions are given below.

Definition 2.2 (Incremental claims payments for a single development triangle): For *a given development triangle and the corresponding observation set O the incremental claims payments are given by*

$$X_{i,j} = \begin{cases} C_{i,0} & \text{if } j = 0 \\ C_{i,j} - C_{i,j-1} & \text{otherwise} \end{cases} \tag{2.2}$$

for $i = 0, \ldots, I$ and $j = 0, \ldots, J$ with $i + j \leq I$.

With Definition 2.2 the cumulative claims payments $C_{i,j}$ for each triangle for i = 0, . . . , I and j = 0, . . . , J with $i + j \leq I$ can be written as

$$C_{i,j} = \sum_{k=0}^{j} X_{i,k}.$$

The individual development factors are given in the following definition.

Definition 2.3 (Individual development factors for a single development triangle): *For a given development triangle and the corresponding observation set O the individual development factors are given by*

$$F_{i,j} = \frac{C_{i,j+1}}{C_{i,j}} \tag{2.3}$$

for $i = 0, \ldots, I-1$ and $j = 0, \ldots, J-1$ with $i + j \leq I - 1$.

To analyse the diff erent characteristic factors, i.e. to analyse the cumulative claims payments, the incremental claims payments as well as the individual development factors, these datasets must be given in three vectors. Therefore, we write

$$v_C, \quad v_X, \quad v_F.$$

The vector v_C contains all observable cumulative claims payments over all development triangles, while v_X is the vector with the computed incremental claims payments and v_F is the vector with the computed individual development factors.

The dimensions of these vectors are:

$$\dim(v_C) = \left((I+1)(J+1) - \tfrac{1}{2}J(J+1)\right) N \tag{2.4}$$

$$\dim(v_X) = \left((I+1)(J+1) - \tfrac{1}{2}J(J+1)\right) N \tag{2.5}$$

$$\dim(v_F) = \left((I+1)(J+1) - \tfrac{1}{2}J(J+1) - (I+1)\right) N \tag{2.6}$$

2.2 Benford's law

Benford's law states that in many sources of data the leading digits are distributed in a specifig – non-uniform – way, the Benford distribution. The Benford distribution can be defined as follows (cf. Benford[2]).

Definition 2.4 (Benford distribution): *A set $A \subseteq R$ of real numbers satisfy Benford's law if the probability of the occurrence of the m-th significant decimal digit $d \in \{0, \ldots, 9\}$ of every number $0 \neq x \in A$ is given by*

$$P(D_m(x) = d) = \sum_{k=\lfloor 10^{m-2} \rfloor}^{10^{m-1}-1} \log_{10}\left(1 + \frac{1}{10k+d}\right)$$

Thereby, $D_m(x)$ $(x \neq 0)$ denotes the m-th decimal digit of x counted from the left and started with 1. The brackets $\lfloor \cdot \rfloor$ denotes Gaussian-brackets ("floor-function").

Remarks 2.5:

- For a more detailed explanation of Benford's law see Berger & Hill[3].
- There is a more general version of Definition 2.4 with a logarithm to a general base B (not to base 10), but in this

paper we are only working with base 10.

- Leading zeros are eliminated so that $D_1(x)$ 6= 0 for all 0 6= x ∈ A.
- $D_1(0)$ is not defined since the occurrence of an 0 at the first position is not possible (that is because of the elimination of leading zeros).

Example 2.6: Given

$$\sqrt{2} \approx 1.4142 \quad \text{and} \quad \pi^{-1} \approx 0.3183$$

the operator $D_m(x)$ works as follows:

$$D_1(\sqrt{2}) = D_1(-\sqrt{2}) = D_1(10\sqrt{2}) = 1, \quad D_2(\sqrt{2}) = 4, \quad D_3(\sqrt{2}) = 1,$$
$$D_1(\pi^{-1}) = D_1(10\pi^{-1}) = 3, \quad D_2(\pi^{-1}) = 1, \quad D_3(\pi^{-1}) = 8$$

In Table 1 and **Figure 2** the probabilities described in Definition 2.4 are displayed.

	probabilities (%)		
digit	**1st digits**	**2nd digits**	**3rd digits**
0	–	11.97	10.18
1	30.10	11.39	10.14
2	17.61	10.88	10.10
3	12.49	10.43	10.06
4	9.69	10.03	10.02
5	7.92	9.67	9.98
6	6.69	9.34	9.94
7	5.80	9.04	9.90
8	5.12	8.76	9.86
9	4.58	8.50	9.83
Σ	100.00	100.00	100.00

Table 1. Probabilities (in percent) for the first, second and third digits for the Benford distribution

Benford's law is often used only for the first and second digits. The reason is that the Benford distribution tends to the uniform distribution on {0, . . . , 9} exponentially fast if *m* increases (see Definition 2.4 or Diaconis[4]). prob.

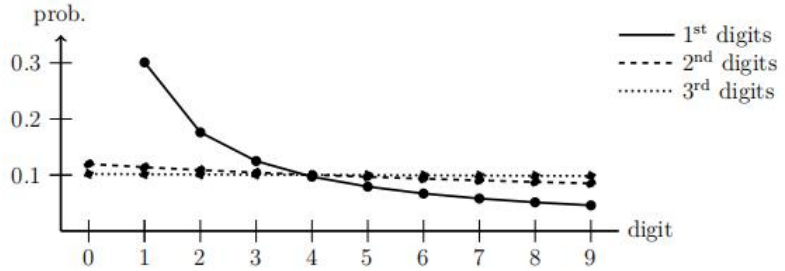

Figure 2; Probabilities for the first, second and third digits for the Benford distribution**.**

3. Empirical analysis

In our empirical analysis we use a dataset made available by GR-NEAM containing cumulative claims development trapezoids of diff erent property/casualty insurer and re-insurer. All these development trapezoids are "all lines of business" trapezoids.

After cleaning the dataset – which was in 99% just rejecting development trapezoids with at least one accident year only containing zeros – we got $N = 442$ claims development trapezoids. Each of these claims development trapezoids contains $I + 1 = 25$ accident years, beginning in 1987 and ending in 2011, and $J + 1 = 10$ development years. An

example development trapezoid is given in Table 2. Furthermore, we continue using i = 0, . . . , 24, respectively $j = 0, \ldots, 9$. So we are not using the "real" years $i = 1987, \ldots, 2011$.

The given dataset contains

$$\begin{aligned}\dim(v_C) &= \left((I+1)(J+1) - \tfrac{1}{2}J(J+1)\right) N \\ &= \left(25 \cdot 10 - \tfrac{1}{2} \cdot 9 \cdot 10\right) \cdot 442 = 90{,}610\end{aligned}$$

observations of cumulative claims payments. The same quantity of observations is given for the incremental claims payments $X_{i,j}$ $(i = 0, \ldots, 24, j = 0, \ldots, 9$ and $N = 442)$, i.e. $\dim(v_X) = 90{,}610$. The quantity of observations for the individual development factors $F_{i,j}$ $(i = 0, \ldots, 23, j = 0, \ldots, 8$ and $N = 442)$ is given by:

$$\begin{aligned}\dim(v_F) &= \left((I+1)(J+1) - \tfrac{1}{2}J(J+1) - (I+1)\right) N \\ &= \left(25 \cdot 10 - \tfrac{1}{2} \cdot 9 \cdot 10 - 25\right) 442 = 79{,}560\end{aligned}$$

In **Table 3** the empirical frequencies for the first three digits of the vectors v_C , v_X and v_F are compared with the corresponding theoretical frequencies given by the Benford

accident year i	development year j									
	0	1	2	3	4	5	6	7	8	9
0	367465	527971	624176	677167	703192	714530	717752	718465	719822	719773
1	369615	545329	646275	697780	721142	731072	733159	737003	736601	737023
2	399390	622015	735778	784595	805098	813733	819120	819384	819941	820156
3	440584	704845	825586	873817	891207	897340	899436	900247	901128	901373
4	557079	880136	1028915	1088000	1104958	1111641	1115078	1115998	1116756	1116950
5	497857	764992	885571	924559	938503	933479	945950	946455	946750	945191
6	550800	814439	921427	959968	973229	978664	980833	981584	982395	982704
7	598811	871903	985752	1026622	1039464	1043290	1046212	1048530	1048963	1049228
8	659893	962172	1078868	1120715	1132873	1137810	1139094	1140345	1140457	1140593
9	660413	940388	1058378	1097987	1112481	1117551	1119460	1120388	1120767	1120915
10	701999	1007601	1118124	1159094	1172886	1178060	1179940	1181337	1182220	1182495
11	780854	1092069	1204873	1248110	1260777	1265813	1267439	1268899	1269555	1269926
12	776920	1069427	1185378	1226759	1241176	1246444	1248875	1250062	1251023	1251501
13	815345	1119941	1238465	1281002	1299013	1304600	1306228	1307217	1309205	1309404
14	839782	1161931	1279122	1327983	1344901	1349389	1352101	1353418	1354340	1354757
15	837464	1155526	1277806	1332653	1353085	1360931	1363989	1364855	1365142	1365314
16	798695	1071472	1186893	1245486	1268734	1276423	1279313	1280708	1281527	
17	813122	1073297	1178935	1236942	1258510	1265235	1267888	1269579		
18	819662	1094312	1207496	1269654	1288589	1297201	1300637			
19	879315	1187815	1300015	1354009	1375906	1383696				
20	957476	1245155	1374119	1432848	1460084					
21	1012736	1324242	1442820	1501972						
22	978338	1288120	1418119							
23	1054954	1376934								
24	1067352									

Table 2. Example of one of the given development trapezoids

distribution (cf. Definition 2.4). Obviously, the empirical and theoretical frequencies are much closer to each other for the cumulative claims payments $C_{i,j}$ and for the incremental claims payments $X_{i,j}$ $(i = 0, \ldots, 24, j = 0, \ldots, 9$ and $N = 442)$ than for the individual development factors $F_{i,j}$ $(i = 0, \ldots, 23, j = 0, \ldots, 8$ and N = 442). **Figure 3**, 4 and 5 emphasize these observations.

In the next step hypothesis tests are made for the occurred empirical values against their theoretical ones using the well known Kolmogorow-Smirnow-test (K-S-test) which is almost one of the most popular goodness-of-fit tests. Since we are using the K-S-test for an underlying discontinuous distribution it is quite more difficult to compute exact p-values (cf. Gleser[9]). The R-package "dgo f" (cf. R Development Core Team[15] and Arnold & Emerson[1]) provides an exact computation of these p-values for small data-samples and a Monte-Carlo simulation of p-values for larger data-samples.

We test the null hypothesis

$$H_0 : F_{emp}(x) = F_{Benf}(x) \text{for all } x$$

against the alternative

$$H_1 : F_{emp}(x) \; 6= \; F_{Benf}(x) \text{for some } x.$$

frequencies (%) for $C_{i,j}$						
	1st digits		2nd digits		3rd digits	
digit	emp.	theo.	emp.	theo.	emp.	theo.
0	–	–	12.00	11.97	10.22	10.18
1	29.37	30.10	11.48	11.39	10.15	10.14
2	17.77	17.61	10.88	10.88	10.08	10.10
3	12.91	12.49	10.39	10.43	10.14	10.06
4	9.71	9.69	10.11	10.03	10.29	10.02
5	8.08	7.92	9.63	9.67	9.97	9.98
6	7.04	6.69	9.24	9.34	9.77	9.94
7	5.75	5.80	8.24	9.04	9.69	9.90
8	5.06	5.12	8.69	8.76	9.97	9.86
9	4.31	4.58	8.33	8.50	9.73	9.83

frequencies (%) for $X_{i,j}$						
	1st digits		2nd digits		3rd digits	
digit	emp.	theo.	emp.	theo.	emp.	theo.
0	–	–	11.88	11.97	10.25	10.18
1	30.59	30.10	11.35	11.39	10.15	10.14
2	17.63	17.61	10.97	10.88	10.20	10.10
3	12.48	12.49	10.43	10.43	10.19	10.06
4	9.74	9.69	10.00	10.03	10.03	10.02
5	7.74	7.92	9.62	9.67	10.14	9.98
6	6.70	6.69	9.27	9.34	9.90	9.94
7	5.72	5.80	9.11	9.04	9.54	9.90
8	5.06	5.12	9.01	8.76	9.70	9.86
9	4.33	4.58	8.36	8.50	9.90	9.83

frequencies (%) for $F_{i,j}$						
	1st digits		2nd digits		3rd digits	
digit	emp.	theo.	emp.	theo.	emp.	theo.
0	–	–	65.40	11.97	30.99	10.18
1	91.70	30.10	10.22	11.39	14.44	10.14
2	1.79	17.61	5.82	10.88	10.18	10.10
3	0.40	12.49	4.35	10.43	8.12	10.06
4	0.26	9.69	2.82	10.03	7.07	10.02
5	0.17	7.92	1.96	9.67	6.38	9.98
6	0.14	6.69	1.42	9.34	5.63	9.94
7	0.11	5.80	1.21	9.04	5.32	9.90
8	0.22	5.12	1.38	8.76	5.09	9.86
9	5.22	4.58	5.42	8.50	6.79	9.83

Table 3. Empirical and theoretical frequencies (in percent) for the first, second and third digits for cumulative claims payments $C_{i,j}$, for incremental claims payments $X_{i,j}$ and for the individual development factors $F_{i,j}$.

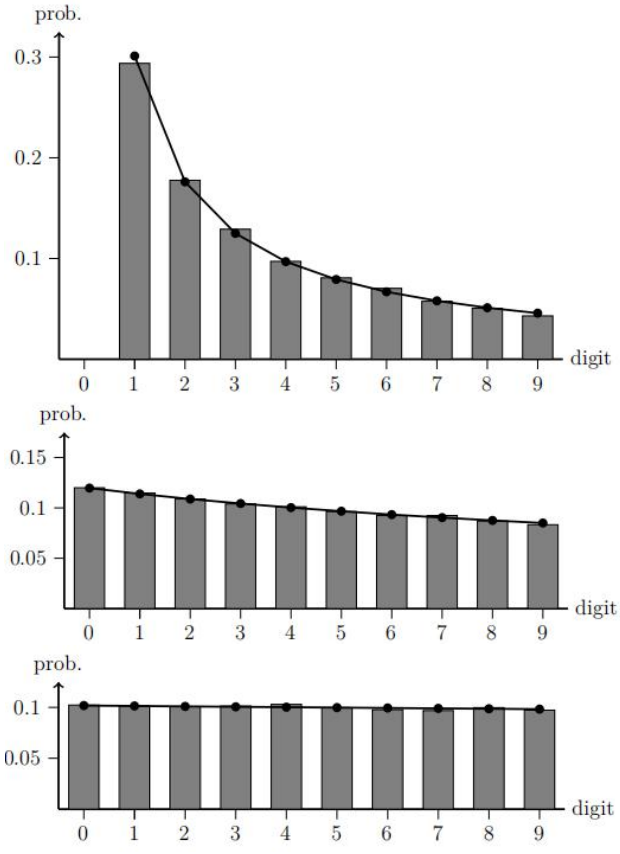

Figure 3; Empirical (bars) and corresponding theoretical (line) frequencies for the first, second and third digits for cumulative claims payments $C_{i,j}$.

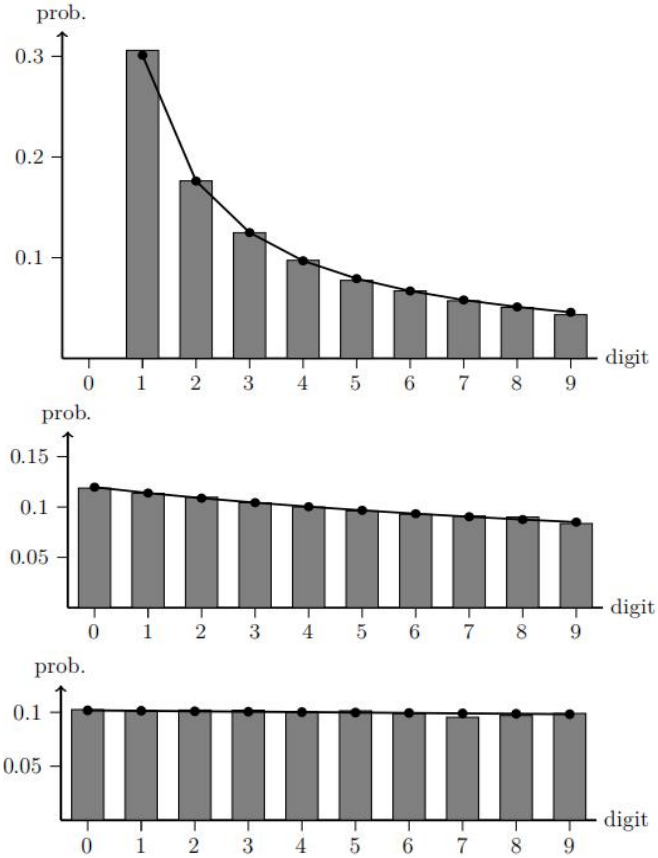

Figure 4. Empirical (bars) and corresponding theoretical (line) frequencies for the first, second and third digits for incremental claims payments $X_{i,j}$.

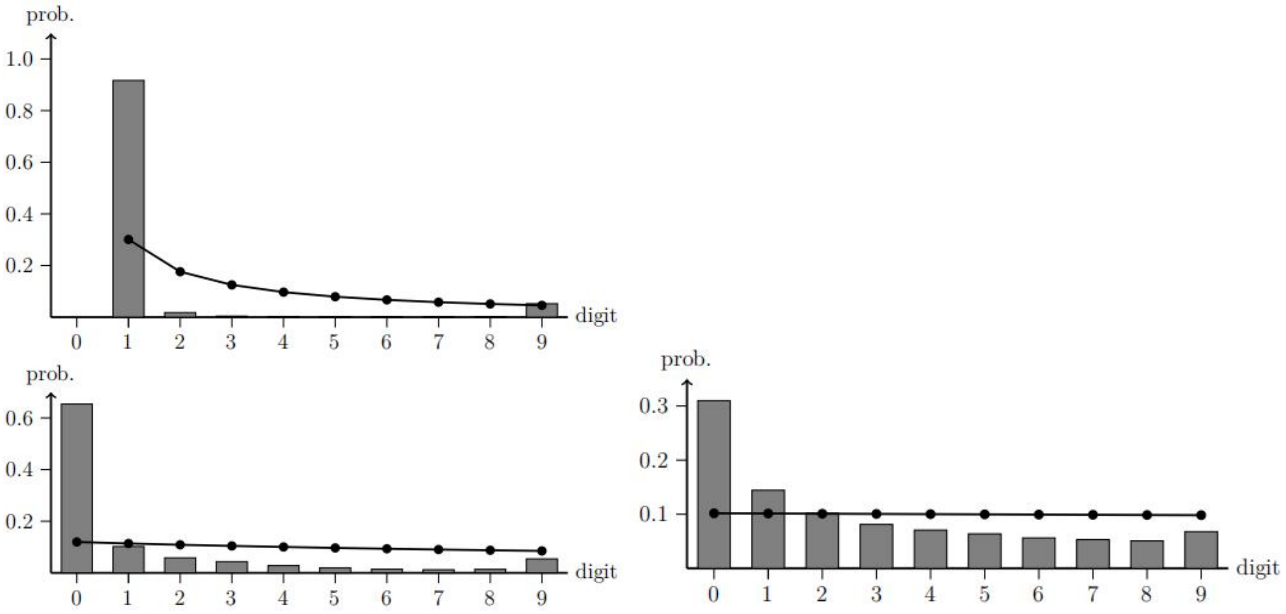

Figure 5; Empirical (bars) and corresponding theoretical (line) frequencies for the first, second and third digits for the individual development factors $F_{i,j}$.

The test criterion is given by:

$$D = \sup_x |F_{\text{emp}}(x) - F_{\text{Benf}}(x)|$$

Table 4 presents the results for the K-S-test, i.e. the values of the test criterion D and the corresponding p-values are listed. The results are the same than we got at first glance from **Figure 3, 4** and **5**.

	1st digits		2nd digits		3rd digits	
observations	D	p-value	D	p-value	D	p-value
C	0.1111	0.9991	0.1	0.9996	0.2	0.7487
X	0.1111	0.9991	0.1	0.9996	0.3	0.2705
F	0.6667	$1.819E{-4}$	0.7	$1.954E{-5}$	0.6	$5.682E{-4}$

Table 4. Values of test criterion D and corresponding p-values for the Kolmogorow-Smirnow goodness-of-fit test

One can see that the individual development factors $F_{i,j}$ for $i = 0, \ldots, I-1$ and $j = 0, \ldots, J-1$ with $i + j \leq I - 1$ do not fit to the Benford distribution for the first, second and third digits very well. Obviously, for the first digits the reason for this is quite clear. The fact that most individual development factors have got a leading "1" is because nearly all increments are larger than zero but not as high as the corresponding cumulative claims payments "near" the development year (while staying in the same accident year). This results in an individual development factor between 1 and 2. The following equations summarize this.

From equation (2.2) we get (for $j > 0$):

$$C_{i,j} = C_{i,j-1} + X_{i,j}$$

Together with equation (2.3) this becomes to:

$$F_{i,j} = \frac{C_{i,j+1}}{C_{i,j}} = \frac{C_{i,j} + X_{i,j+1}}{C_{i,j}}$$

Since in our dataset $0 < X_{i,j+1} < C_{i,j}$ holds true for most $i = 0, \ldots, I-1$ and $j = 0, \ldots, J-1$ with $i + j \leq I - 1$ it follows the result seen in the first plot of **Figure 5.**

4. Conclusion

In the empirical analysis we have seen that Benford's law is quite good for two out of three characteristic claims development factors, namely

- the cumulative claims payments $C_{i,j}$ for $i = 0, \ldots, I$ and $j = 0, \ldots, J$ with $i + j \leq I$ and
- the incremental claims payments $X_{i,j}$ for $i = 0, \ldots, I$ and $j = 0, \ldots, J$ with $i + j \leqslant I$.

Of course, this analysis is done with development trapezoids containing "all lines of business" which are middle to long tailed. Thereby, the results only hold true (in an empirical sense) for this kind of triangles/trapezoids. For other data, e.g. for short tail lines of business, the same analysis has to be done a second time.

Due to the fact that Benford's law holds true for some characteristic factors, this result can be used to check a given development triangle/trapezoid against plausibility, outliers, fraud, etc. Of course, if an actuary detects inconsistency in a given dataset, he has to do some further research to determine the exact problem in the dataset. In this case, Benford's law can be seen as a first tool (among others) to automatically detect problems in a dataset.

References

1. Arnold, T. A. & Emerson, J. W. (2011): Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions. The R Journal, 3(2): 34–39 (cit. on p. 8).
2. Benford, F. (1938): The Law of Anomalous Numbers. Proceedings of the American Philosophical Society, 78(4): 551–572 (cit. on pp. 1, 5).
3. Berger, A. & Hill, T. P. (2011): A basic theory of Benford's Law. Probability Surveys, 8: 1–126 (cit. on p. 6).
4. Diaconis, P. (1977): The Distribution of Leading Digits and Uniform Distribution Mod 1. The Annals of Probability, 5(1): 72–81 (cit. on p. 6).
5. Diekmann, A. & Jann, B. (2010): Benford's Law and Fraud Detection: Facts and Legends. German Economic Review, 11(3): 397–401 (cit. on p. 2).
6. Durtschi, C.; Hillison, W. & Pacini, C. (2004): The Effective Use of Benford's Law to Assist in Detecting

Fraud in Accounting Data. Journal of Forensic Accounting, 5(1): 17–34 (cit. on p. 2).
7. England, P. D. & Verrall, R. J. (2001): A flexible framework for stochastic claims reserving. Proceedings of the Casualty Actuarial Society. Vol. 88. 1: 1–38 (cit. On p. 2).14
8. England, P. D. & Verrall, R. J. (1999): Analytic and bootstrap estimates of prediction errors in claims reserving. Insurance: Mathematics and Economics, 25(3): 281–293 (cit. on p. 2).
9. Gleser, L. J. (1985): Exact Power of Goodness-of-Fit Tests of Kolmogorov Type for Discontinuous Distributions. Journal of the American Statistical Association, 80(392): 954–958 (cit. on p. 8).
10. Govindarajulu, Z. (2007): Nonparametric Inference. World Scientific (cit. on p. 2).
11. Heberle, J.; Huergo, L. & Merz, M. (2010): Bootstrapping the Chain-Ladder-Method of Several Correlated Run-Off Portfolios. Zeitschrift für die gesamte Versicherungswissenschaft, 98(5): 541–564 (cit. on p. 2).
12. Mack, T. (1993): Distribution-free calculation of the standard error of chain ladder reserve estimates. ASTIN Bulletin, 23(2): 213–225 (cit. on p. 2).
13. Merz, M. & Wüthrich, M. V. (2007): Prediction error of the expected claims development result in the chain ladder method. Bulletin of Swiss Association of Actuaries,1: 117–137 (cit. on p. 2).
14. Pinheiro, P. J. R.; Andrade e Silva, J. M. & de Lourdes Centeno, M. (2003): Bootstrap Methodology in Claim Reserving. Journal of Risk and Insurance, 70(4): 701–714 (cit. on p. 2).
15. R Development Core Team (2014): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (cit. on p. 8).