GRAS Newsletter 58

May 2013

OpRisks by the Number

Summary: Quantitative ORM is usually restricted to Loss data, KRIs and Capital calculations. There is another kind of quantitative analysis OpRisk managers can fruitfully explore. Using a statistical property of many financial data sets, it is possible to detected if a data set contains 'unusual' entries. This analysis provides an invaluable tool to better understand data series, with special applications in fraud detection, already well known in the anti-fraud community. OpRisk managers would do well to also get into number crunching, not only for fraud management but also to better understand business processes and the ORM data itself.

Dear reader,

As an OpRisk manager, you may not have heard of Benford's Law and yet it is a very useful tool for ORM that can be used to identify irregularities in banking data. The law is a statistical curiosity which is widely known to forensic accountants as a way to detect fabricated data. It involves examining the frequency of the most leading digits (the 'leftmost' digits) in a data set, which often follow a distinct pattern. Benford's Law[1] is one such dominant pattern, which has been found to be valid for so-called 'natural' data sets. Examples of natural data sets are population sizes, contract sizes or number of stocks traded[2]. An astonishing number of (financial) data sets are natural data sets and thus should conform to Benford's law. This opens up the possibility to test observed data series against the expected distribution, thus identifying anomalies. Figure 1 shows the distribution of the most significant digit according to Benford's law, which we will call **B**. What is surprising to many people is that nearly half the numbers (47.7% to be exact) have '**1**' or '**2**' as the leading digit. We will show how we might use this and other properties of **B** in a simple ORM example.

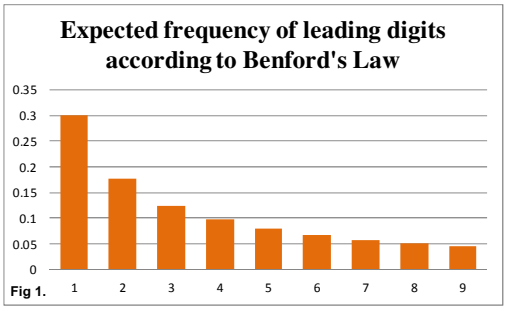


Fig 1.

An example using Loss data

The collection of OpRisk loss data is fraught with many difficulties. Some losses are covered up on purpose, not all losses are correctly identified, and, even with the best intentions, it is not always clear what the gross loss amount even is. The reason for that is that an OpRisk loss is not just an accounting entry, but emanates from ill-defined events and their consequences. Nevertheless, all banks are expected to monitor and report their loss data, which makes loss data an good topic for research. The loss amounts typically span several orders of magnitude and there are plenty of data entries which makes this data series a 'natural' data set which should follow **B**.

[1] Data sets satisfy Benford's law if the probability of the most significant digits follow $P(d) \approx \mathrm{Log}_{10}\{1+(1/d)\}$ where $d \in \{1,2,..\}$. A vast majority of 'natural' numbers creep up in finance and forensic accountants have used this data successfully on many occasions. The thorough treatment can be found in Mark Nigrini (2012), *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. A more gentle introduction is his well-known article *I've Got Your Number* which can be found at: http://www.journalofaccountancy.com/issues/1999/may/nigrini

[2] Such sets exclude assigned numbers (such as ZIP codes or ISBN numbers), bounded series (such as the number of passengers on a Boeing 747) and numbers that act as labels (such as 1 = Very Good, 2 = Good, ···, 5 = Very Bad).

OpRisks by the Number

To get started, we take the gross losses as reported between 2007 – 2011 from four banks, all converted to EUR. In this data set, none of the banks use a threshold. Figure 2 shows the expected frequency of the leading digits as per **B**, as well as the actual observed values. Visually, the fit looks remarkably good. We do notice a slight difference in the leading digit '**4**', but a simple goodness of fit test (here we use χ^2 for the time being) shows the similarity of the two distributions is (just) significant at the 5% level.

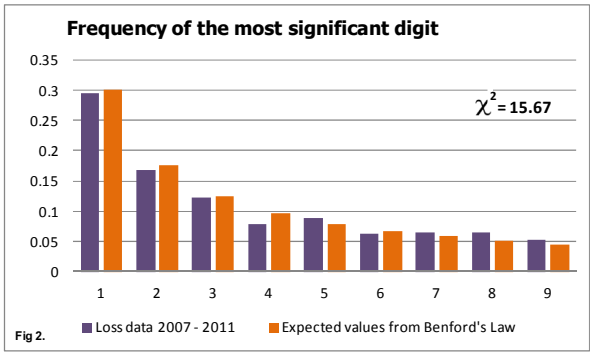


Fig 2.

Further analysis

Testing the leading digit is rarely sufficient to decide on the normalcy of a data set. The distribution of the *two* leading digits, however, is much more telling. That analysis does indeed suggest anomalies. When we examine figure 3 below, showing the distribution of the two leading

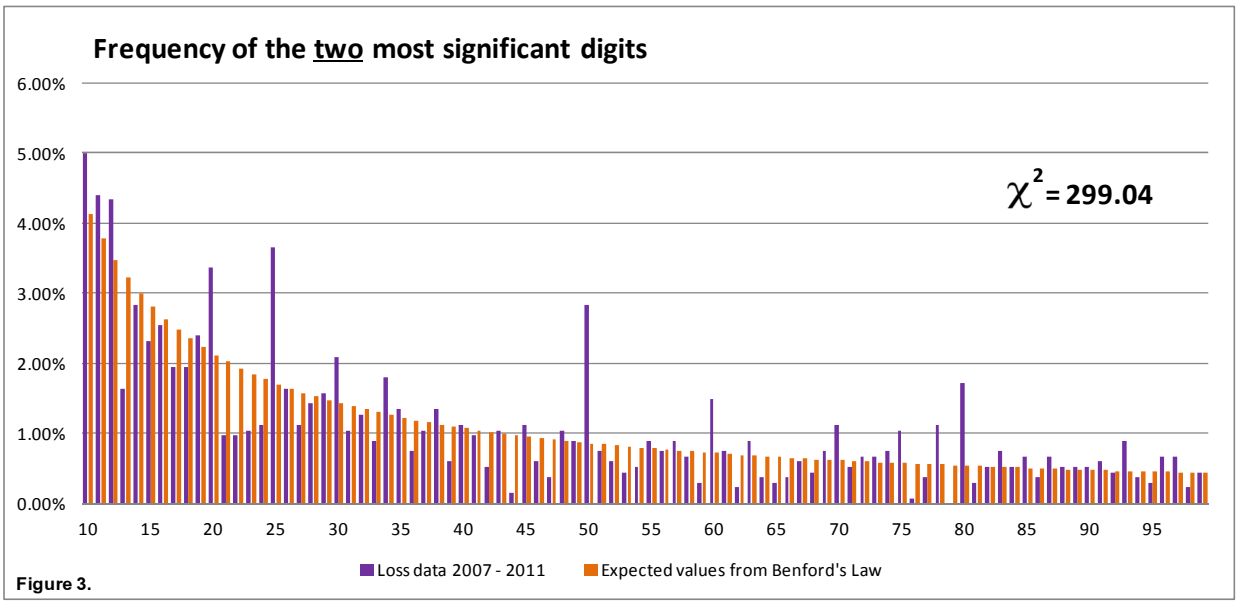
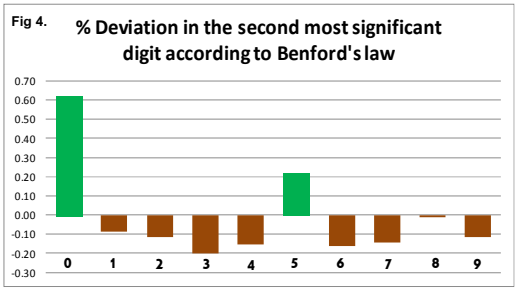


Figure 3.

digits {10…99}, the anomalies are clear. We notice very large spikes around round multiples of 10 as well as significantly reduced values surrounding them and some other outliers. Especially the multiples of 10 are too much of a coincidence. As an example, we would expect to find the leading digits '**50**' with a frequency of 0.86%. Instead, we find a frequency of 2.84% a significant deviation. We also note that observations starting with '**21**', '**22**', '**23**', or '**24**', occur 30% less than their expected frequency. The χ^2 of 299.04 also indicates that the H_0-hypothesis that the observed distribution conforms to **B** must be rejected at the 5% level.

OpRisks by the Number

Delving deeper: the *second-most* significant digit



Digging still further, we examine the frequency of the second-most significant digit in Figure 4. Now the anomaly is starting at us. Of the digits in second place, the number '**0**' is overrepresented by 60%, and the number '**5**' is overrepresented by 22%. What is even more significant is that these two digits are the only ones that are overrepresented. All other digits are underrepresented in the sample which indicates that the data is not at all clean. A data set such as this which should satisfy **B** quite closely but still fails needs to be examined further. As a preliminary explanation, could it be that people reporting losses are rounding up and down to, say 25···· if the true value is e.g. 24····? Note that that the anomaly in itself does not automatically imply fraud, but it does call for further examination to understand the nature and origin of the deviant pattern.

Where can we use this?

This property of natural occurring numbers can be used to spot anomalies in financial data series. As an example, Table 1 shows the kind of data that should follow a **B** distribution. The reason for that is that people who make up numbers typically do one of three things: they either repeat valid transactions (believing they will look good, ignoring that there may be too many of them), they target a specific range of numbers (such as submitting travel invoices that are just under the daily limit) or they tend to distribute their digits fairly uniformly, believing that will arouse less suspicion. The first two types are relatively easy to spot visually, but the latter type of distortion can easily be identified statistically yet is very hard for fraudsters to avoid.

Accounts payable transactions
Credit card transactions
POS transactions
Purchasing orders
Loan applications
End of day balances
Journal entries
Inventory prices
Customer refunds

Table 1. Bank 'natural' data series

Applying Benford's Law to truncated data sets

Note that not all made-up numbers are necessarily fraudulent and that not all data series that violate Benford's law are problematic. Bounded data, which we often see in financial data(e.g. a minimum cash balance of 20 USD or maximum overdrafts of 1,500 USD) do not conform to Benford's law, regardless of the number of observations.

When we are faced with limits on the data set, as we often are in ORM, we can still use the same principle with slight modifications, using the characteristic digit distribution of our own data set. The appendix to this newsletter provides an example of this.

Conclusion

Benford's law opens a whole new field of analysis and research for OpRisk Managers. It can help to identify fraud, to spot unusual transactions, discover errors in systems, and, in general, can be used to augment the KRI and Loss data analysis. ORM does not have a wealth of data. All the more reason to employ every technique available on the available data, including Benford's Law.

Appendix: The Loss Data Law

Example using bounded data

As an example of the use of Benford's law for series that follow a different pattern, we will use the example of loss data with a threshold. Loss data consortia, such as ORX, use a fixed cut-off threshold for loss reporting to ensure comparability and to reduce the burden on data collection and cleansing. A common threshold is 20,000 EUR, which we will now apply to our original dataset.

The Loss Data Law with a Threshold

To create a benchmark, we start by segmenting our data into two sets, applying a threshold of 20,000 EUR to both. We first data set comprises the loss data from three banks and will be used to create a benchmark. This we will test against Benford's **B**. The result of that exercise is shown in figure 5. The first digit distribution does not conform to **B**, notably for '**1**' and '**2**' and the χ^2 is way off. But it is with the two most significant digits in figure 6 below that this series, or at least the χ^2 associated with it, explodes completely.

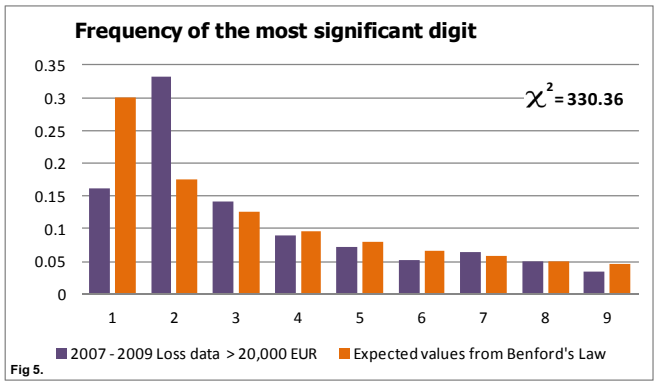


Fig 5.

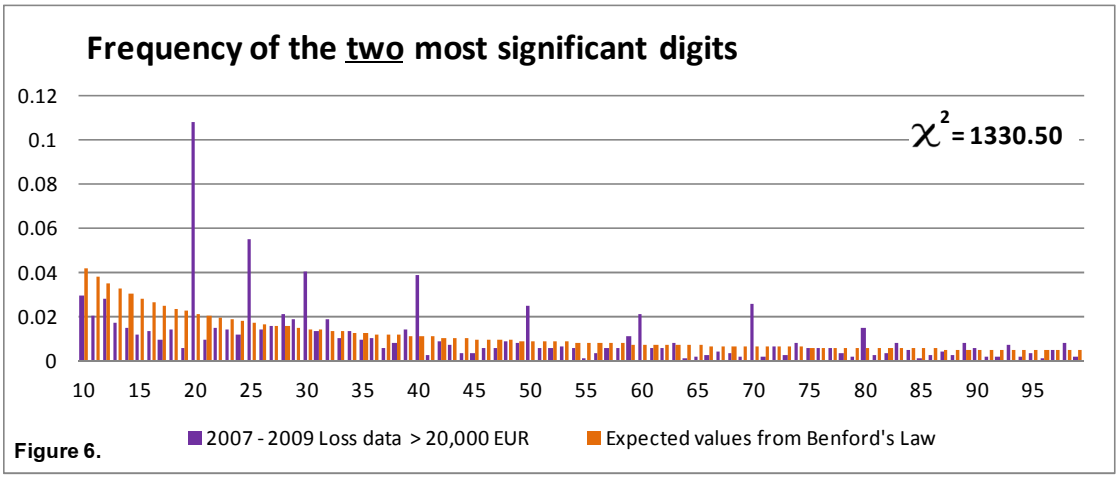


Figure 6.

This is a case where a **B** distribution is clearly unusable. But we can still use the central tenet behind Benford's law, namely that the frequency distribution of the leading digits follows a distinct pattern, much like a fingerprint. What we can do is capture that fingerprint and use that to test observed values the expected values.

The solution is fairly simple. If we wish to test the loss data for a particular bank, we should compare it not to **B** but we should build our own Loss Data Law incorporating the threshold. We will name that truncated distribution **T**, using again the data of three banks.

Appendix: The Loss Data Law

When we test the data of the fourth bank to the **T** distribution, we get new graphs as shown in figure 7 and 8. This looks much less out of sync.

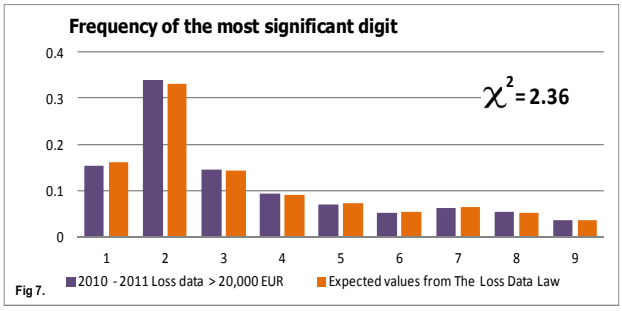


Fig 7.

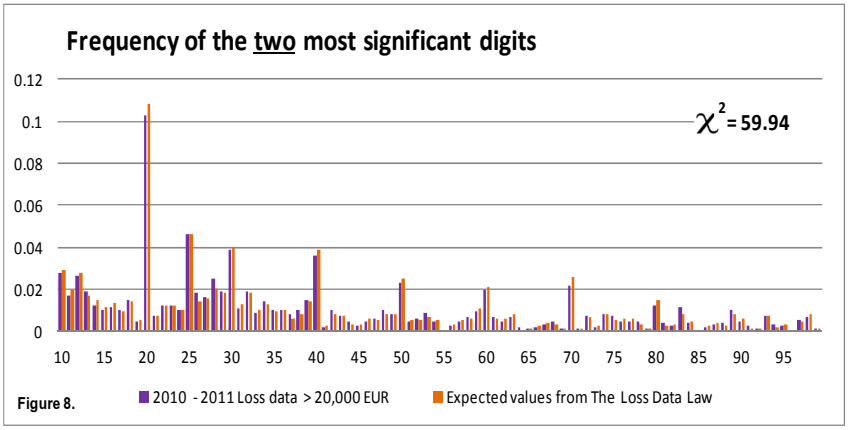


Figure 8.

Our test bank (in purple) conforms to the Truncated Loss data law **T** (orange) and the correspondence is significant at the 5% level. We still see the effect of the threshold. Also, the test of the two most significant digits, which is considered to be a much stronger test, shows an acceptable value for χ^2. Visually, we do see some discrepancies in **'20'**, **'40'**, **'70'** and **'80'**, but the goodness of fit test is easily satisfied at the 5% level. Using this information, it becomes possible to spot outliers even in some non-natural data sets.

Testing for conformity

To properly test for conformity to a distribution, we cannot only rely on visual inspection but we will need a robust, statistical test. We have so far used a χ^2-test, which is fine for relatively small numbers of 1,000 to 25,000 observations, such as we have here. If we have a larger dataset, this statistic becomes too stringent and needs to be replaced with a test that is independent of the sample size.

One such recommended statistic is the Mean Absolute Deviation, which is defined as:

$$\sum_{i=1}^{N} \frac{\left| ObservedValue_i - ExpectedValue_i \right|}{N} \qquad (1)$$

This simple test is suitable even for very large datasets, which are very common in finance. Some p-values have been developed for this statistic which makes it extremely useful for transaction data and client records. It will be beneficial for ORM managers who wish to expand their grasp on business processes.