Benford Distribution in Science

Fabio Gambarara & Oliver Nagy

July 17, 2004
Preface

This work was done at the ETH Zürich in the summer semester 2004 and is related to the the Mensch, Technik, Umwelt (MTU) lectures held at the electrical engineering department.

Authors:
- Oliver Nagy (nagyo@ee.ethz.ch)
- Fabio Gambarara (gafabio@ee.ethz.ch)

Tutor: Andreas Diekmann (diekmann@soz.gess.ethz.ch)

Semester: 6th

Department: D-ITET

Expected Hours: ~ 60

Credits: 6
Contents

1 Introduction 4

2 History of the Benford Distribution 6

3 Properties of the Benford Distribution 8
   3.1 Distribution Density 9
   3.2 More Digits 10
      3.2.1 Second Digit 12
   3.3 Scale Invariance 12
   3.4 Examples 15
      3.4.1 Exponential Growth 15
      3.4.2 Fibonacci Sequence 18

4 Real World Data and Simulation 20
   4.1 Pharmacy 21
   4.2 Photonic Crystals 27
   4.3 Micro Turbines 32
   4.4 Files on Computers 37

5 The Survey 40
   5.1 Online Questionnaire 40
      5.1.1 First Page 40
      5.1.2 Second Page 41
      5.1.3 What Results Were Expected 41
      5.1.4 Questionnaire Procedure 41
   5.2 Analysing the Numbers 42
      5.2.1 Fulfilment Criteria 42
      5.2.2 General Analysis 43
      5.2.3 Analysis with Division in Groups 44
   5.3 Analysis of the Digit Distribution 47
      5.3.1 Faking a Graph 47
      5.3.2 Invent Amount of Chemical Solutions 51
      5.3.3 Invent Numbers for the Parameters of a Machine 53
   5.4 Evaluation of the Dice Question 55
      5.4.1 General Probability of Numbers 55
      5.4.2 Sequences of Same Numbers (SoSN) 55
   5.5 What We Could Have Done Better 57
A Appendix

A.1 The Original Questionnaire ........................................ 58
A.2 Principles the People Used ......................................... 60
Chapter 1

Introduction

Science is performed by humans and therefore lacks some typical human behaviour. One of them is (scientific) misconduct and data fraud as for example in the case of the physician Jan Hendrik Schön [1] or in the medical cancer research by Herrmann [2]. As a consequence, the science community started a discussion on how to avoid such behaviour in the future. One already widely applied approach is to install rules for appropriate scientific behaviour and procedures to follow in suspicious cases, as like for example at ETH [3] or plenty of German universities [5].

Clear rules how to deal with such situations are important, but when done consistently and consequently, data substitution and manipulation can still occur. Therefore, other tools are needed and currently evaluated to give clues on manipulated data. One such tool is the statistical behaviour of the presented numbers.

This work focuses on the statistical distribution of single digits of numbers. An awesome property of naturally arising numbers is that they are more likely to start with a lower digit (e.g. 1) rather than with a higher one, (e.g. 8 or 9). Even more exciting is the fact that in some cases a deterministic probability for each digit can be identified. The so called Benford’s law describes this behaviour through a distribution called Benford distribution. This law fits very often when looking at naturally generated numbers, for example lake dimensions, death rates, sport results, financial data and many others. Benford’s law can therefore be used to analyse data sets to hint out possible data manipulation.

Unfortunately, it is not that easy. Whereas the Benford distribution can be found very often, it still matters how the data was generated. If there are artificial upper and lower bounds for example, the resulting numbers are not necessarily Benford distributed anymore, although they are not faked. Another typical scientific example is when measuring data that is expected to be around a certain level (e.g. 5V), then the first digit can definitively not be Benford distributed.

Despite that Benford’s law can not always be applied, it might be still a powerful tool to reveal data fraud if it can be. In finance, this method can be very useful as pointed out in [4]. If such fraud analysis is feasible for scientific results is not yet clear.

This report focuses therefore on the analysis of various scientific data sets using mainly Benford’s law. It starts with a brief history (chapter 2) of the
Benford distribution, followed by some mathematical properties that are either relevant for our work or just interesting to know about it (chapter 3). Chapter 4 shows the digit distributions of various data files taken from various institutes at ETH. Finally the results of a simple survey asking students to fake numbers will be presented in chapter 5. The aim hereby is to find out whether data invented by humans shows statistical properties that could be useful in fraud detection as well.
Chapter 2

History of the Benford Distribution

As introduced in the last chapter, the law describing the probability of the first digit(s) is referred to as Benford’s Law. This report almost only deals with this law, so its history is briefly introduced here.

The American astronom Simon Newcomb discovered an ‘odd’ behaviour when looking at his logarithm book. Within this book the logarithms of plenty of numbers were given, but the first pages in the book (those with numbers starting with the digit ‘1’) were worn much more than the last ones. It seems that he was either more interested in numbers starting with a low digit, or there are simply more given numbers starting out with 1’s, than with any other digit. As the first hypotheses is very unlikely, Newcomb started to analyse various data sets and empirically derived what is nowadays known as Benford’s law.

Frank Benford, a physician at General Electric, discovered the same behaviour in the year 1938. He studied more than 20,000 different data sets, mixing all kinds of data together. Included in the set were areas of rivers, baseball statistics, numbers in magazines and so forth [6]. Benford, like Newcomb before, empirically derived a rule for the probability of each digit. Benford himself did not know about the work of Newcomb (and neither did many others), so the distribution was named after him and is since referred to as Benford Distribution or Benford’s Law. Another name often seen in the literature is The power of One due to the high probability of the number one.

For a very long time, no mathematical or just plausible argument could explain why such a law should exist. In 1961 the mathematician Roger Pinkham came up with the statement that when such a law exists for whatever numbers generated by the universe, then it must be scale invariant. This conclusion just follows from the fact that Benford’s law seems to hold for many completely different and independent quantities like the number of atoms in an object and the stock market. Especially the latter example is very instructive, because the digit distribution fore casted by the Benford distribution holds for whatever currency or metric system in general on the planet (and would hold for any other currency on any other planet as well).

Pinkham could show that Benford’s law is not only scale invariant, but it is also the only possible distribution for achieving scale invariance. This also
implies that any statistical phenomenon out there which is unaffected by a scale change must be Benford distributed.

Still, there is one major drawback: Benford’s law obviously neither fits on assigned numbers like bank account numbers nor does it work on purely random values. In 1992, Mark Nigrini wrote his PhD thesis studying the possible use of Benford’s law in economics. It turned out that the whole stock market, or just almost any quantity in economics that is based on natural observations, follows that law. Based on that discovery, Nigrini successfully applied Benford’s law to detect tax frauds, and created the phrase digital analysis.

A program developed by him was used in the Brooklyn District Attorney to detect tax evasion, and it was successful in a test run on already proofed cases. As an interesting side note, the test was also applied to the tax return of former US president Bill Clinton, but no fraud hints were found [7].

Nigrini also pointed out that the law is not universal. If the probability behind a process is known to be uniformly distributed, like in the lottery, then there is no way to gain a statistical advantage, because the thrown out digits are really uniformly distributed. Also biased samples or lower and upper limits are problem.

The question when Benford’s law applies and when not remained until 1996. It was then, when the mathematician Theodore Hill of Georgia Institute of Technology in Atlanta gave the answer. In 1996 he published a paper explaining the origin of Benford’s law. He found out that when taking a random number of samples from random distributions, the digit distribution converges towards Benford’s law.

From this point of view it suddenly makes sense why Benford’s law is found so often. Every process in nature depends on various parameters. Each of those parameters are random with a certain distribution (not necessarily known). Therefore any process that is observed, is a combination of plenty of other random processes. This in turn full fills the condition of taking a random number of samples from random distributions. In such a sense, the Benford distribution can be called the distribution of distributions. More details can be found at [9].

Another possible application of Benford’s law is to validate the quality of simulations. The weather data for instance perfectly matches the Benford distribution. On simple test for weather simulations could therefore be to analyse the digit distribution of the first digit. If it does not match the Benford distribution, then the simulation obviously does not generate realistic results.

Yet another field where Benford’s law holds is for an exponential growth. Take for instance 1000€ and increase this amount by 10% each year due to interests. It will take significantly longer to cross 2000€ than it will need to get to 3000€, which still takes longer than getting to 4000€ in turn and so on.
Chapter 3

Properties of the Benford Distribution

So far it was pointed out that the probability of the first digit being 1 is higher than any other digit being the first, without stating the probability explicitly.

The probability, that the first digit $D$ equals $d$ in the decimal system is given by the following law

$$P[D = d] := \log_{10} \left( 1 + \frac{1}{d} \right) \quad \{d \in \mathbb{N} : 1 \leq d \leq 9\} \quad (3.1)$$

The probability of each digit according to equation (3.1) is summarised in table (3.1). A figure showing the distribution graphically is given in figure 3.1 According to equation (3.1), the probability of the first digit being 1 is not 1/9 as would be intuitively expected (zero cannot be the first digit).

<table>
<thead>
<tr>
<th>Digit</th>
<th>Probability in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.10</td>
</tr>
<tr>
<td>2</td>
<td>17.61</td>
</tr>
<tr>
<td>3</td>
<td>12.49</td>
</tr>
<tr>
<td>4</td>
<td>09.69</td>
</tr>
<tr>
<td>5</td>
<td>07.92</td>
</tr>
<tr>
<td>6</td>
<td>06.69</td>
</tr>
<tr>
<td>7</td>
<td>05.80</td>
</tr>
<tr>
<td>8</td>
<td>05.12</td>
</tr>
<tr>
<td>9</td>
<td>04.58</td>
</tr>
</tbody>
</table>

Table 3.1: Probability of the Digits 1 to 9 according to Benford’s Law

From table (3.1) it is clearly seen, that the digit 1 has a very unexpected high occurrence probability of about 30%. Comparing to the usually expected probability of $P[D = 1] = 1/9 \approx 11.11\%$, this is nearly three times as high.

A quite funny representation of Benford’s law is given on [8] and looks the following way:

```
1---------2---------3---------4---------5---------6---------7---------8---------9
```
3.1 Distribution Density

When claiming that equation (3.1) is a discrete density function, it must integrate to 1. That this is in fact the case, is shown below.

\[
\sum_{d=1}^{9} P[D = d] = \sum_{d=1}^{9} \log_{10} \left( 1 + \frac{1}{d} \right) \\
= \sum_{d=1}^{9} \log_{10} \left( \frac{1 + d}{d} \right) \\
= \log_{10} \left( \prod_{d=1}^{9} \frac{1 + d}{d} \right) \\
= \log_{10} \left( \frac{(1 + 9)!}{9!} \right) \\
= \log_{10}(10) = 1 \quad (3.2)
\]

Furthermore, the Benford distribution is base invariant. For any base \( B \) the density function is given by equation (3.3).

\[
P[D = d] = \log_B \left( \frac{1 + d}{d} \right) \quad \forall \ d \in [1, B - 1] \quad (3.3)
\]

Figure 3.2 shows the probability of the first digit for various bases. An interesting case here is the binary case (\( B = 2 \)) that can trap one, because the
probability of the first digit being 1 becomes 100%. In a binary system however, there are only the two digits, 0 and 1, and 0 can never be the first one. Therefore, the first digit must be 1.

If (3.3) is still a valid density function for any given base $B$, then it must integrate to 1 as well. Proofing this is similar to (3.2), but shows the general case. Therefore, the analysis shown below holds for any base $B$ with $\{B \in \mathbb{N} : B \geq 2\}$.

$$\sum_{d=1}^{B-1} P[D = d] = \sum_{d=1}^{B-1} \log_B \left( \frac{1 + d}{d} \right)$$

$$= \log_B \left( \prod_{d=1}^{B-1} \frac{1 + d}{d} \right)$$

$$= \log_B \left( \frac{(1 + B - 1)!}{(B - 1)!} \right)$$

$$= \log_B \left( \frac{(B)!}{(B - 1)!} \right)$$

$$= \log_B(B) = 1$$

(3.4)

3.2 More Digits

The law also extends to more digits. When interested in the probability of the first two digits being for example 13, then simply set $d = 13$ and use equation
(3.1) for the decimal system. As a simple but imperfect proof, it is shown here that the probability of $d$ being the first digit equals the sum of the probabilities of all first two digits starting with $d$. For example, the probability of the digit 1 must be equal to the sum of the probabilities of the first two digits being 10 to 19.

$$\sum_{k=0}^{9} P[D = 10d + k] = \sum_{k=0}^{9} \log_{10} \left[ \frac{10d + k + 1}{10d + k} \right]$$

$$= \log_{10} \left[ \prod_{k=0}^{9} \frac{10d + k + 1}{10d + k} \right]$$

$$= \log_{10} \left[ \frac{(10d + 1 + 9)! (10d - 1)!}{(10d + 9)! (10d)!} \right]$$

$$= \log_{10} \left[ \frac{(10d - 1)!}{(10d)!} \right]$$

$$= \log_{10} \left[ \frac{10d + 10}{10d} \right]$$

$$= \log_{10} \left[ \frac{d + 1}{d} \right]$$

(3.5)

Figure 3.3 shows the probabilities of the first two digits in the decimal system. The $x$-axis here shows the first two digits, not the number itself. So at position $x = 24$, the probability of the first two digits being 2 and 4, and not the probability of the number 24 itself is shown.
Finally, another trap should be pointed out when looking at bases greater than 10, for instance the hexadecimal system (base 16). Calculating the probability for each digit is still simple, but one might be fooled thinking that numbers greater than 9 (e.g. 10 to 15 in the hexadecimal case) are represented by two digits. This is definitely not the case. In the Arabic system we are simply running out of digits for such a case. Computer scientists start using alphanumerical ‘digits’, for instance an $A$ for the number 10, $B$ for 11 and so on. When therefore asking for the probability of the first digit being 12, it is asked for the first digit being $C$, and this is still only a single digit, no matter how many Arabic digits would be necessary to represent it.

3.2.1 Second Digit

When analysing the data later on, also the distribution of the second digit will be needed. The idea to get to it is quite simple: consider one is interested in the second digit being $d_2 = 3$, then the probability is simply the sum of the probabilities of the first two digits being 13, 23, \ldots, 93. It is important to note here that the second digit can be 0.

This sounds very simple but bringing it to formula is a nasty task. The probability for the second digit being $d_2$ is

$$P_2[D_2 = d_2] = \sum_{k=0}^{9} \log_{10} \left( \frac{10k + d_2 + 1}{10k + d_2} \right)$$  \hspace{1cm} (3.6)

Extending this to the case of an arbitrary digit position $z \geq 2$, the formula extends to

$$P_{z}[D_z = d_z] = \sum_{k=1}^{10^{(z-1)}} \log_{10} \left( \frac{10k + d_z + 1}{10k + d_z} \right)$$  \hspace{1cm} (3.7)

Setting $z = 1$ does not reduce the last formula to the first digit formula (3.1) due to the first digit cannot be a zero, whereas the proceeding ones can be.

3.3 Scale Invariance

As has been pointed out several times by now, the Benford distribution is scale invariant. Here the derivation of this probably most important property is given. As a bonus, the base invariance property will come along for free.

Roger Pinkham stated that if such an overall universal law exists, it must be scale invariant to any arbitrary but constant nonzero scaling factor $\alpha$. Formally speaking, the probability density function $p(x)$ must full fill the following property

$$p(x) \equiv f(\alpha) \cdot p(\alpha x) \hspace{1cm} \{x \in \mathbb{R} : x \geq 1\} \hspace{1cm} \{\alpha \in \mathbb{R} : \alpha \neq 0\}$$  \hspace{1cm} (3.8)

Note that $x$ is not a discrete variable here like used for digits. This derivation covers a more general approach which will become very valuable later on to determine the digit probabilities.
If \( p(x) \) should be a probability density function, then it must integrate to 1, e.g:

\[
\int_{-\infty}^{+\infty} p(x) \, dx = 1 \quad (3.9)
\]

According to (3.8), the same must hold for the scaled version.

\[
\int_{-\infty}^{+\infty} p(\alpha x) \, dx = \int_{-\infty}^{+\infty} p(z) \, \frac{dz}{\alpha} = \frac{1}{\alpha} \quad (3.10)
\]

where in the last formula the substitution \( z := \alpha x \) was used. To fulfil the equality in (3.8), \( f(\alpha) \) must be

\[
f(\alpha) = \alpha \quad (3.11)
\]

Substituting \( f(\alpha) \) in equation (3.8) then yields

\[
\frac{p(x)}{\alpha} = p(\alpha x) \quad (3.12)
\]

To find the function \( p(x) \), first the two sides of the equation are derivated with respect to \( \alpha \), where the substitution \( z := \alpha x \) is used again here:

\[
\frac{d}{d\alpha} \left( \frac{p(x)}{\alpha} \right) = \left( \frac{\partial}{\partial z} p(z) \right) \frac{dz}{d\alpha} = p(x) \frac{-1}{\alpha^2} \left( \frac{\partial}{\partial z} p(z) \right) x \quad (3.13)
\]

On the right hand side on the first line, the chain rule for derivations was used. Now a trick from [10] is used by setting \( \alpha = 1 \). Equation (3.8) must be valid for any \( \{ \alpha \in \mathbb{R} \setminus 0 \} \), so choosing \( \alpha = 1 \) is definitely valid, but not the only possible choice here.

Substituting \( \alpha \), this implies \( z = \alpha x = x \).

\[
p(x) \frac{-1}{1^2} = \left( \frac{\partial}{\partial x} p(x) \right) x = -p(x) = x \frac{\partial}{\partial x} p(x) \quad (3.14)
\]

This now however is a simple differential equation that is solved using the separation technique

\[
-p(x) = x \frac{\partial p(x)}{\partial x}
\]

\[
-\frac{\partial x}{x} = \frac{\partial p(x)}{p(x)}
\]

\[
-\int \frac{\partial x}{x} = \int \frac{\partial p(x)}{p(x)}
\]

\[
-\ln(x) + c_0 = \ln(p(x))
\]

\[
x^{-1} e^{c_0} = p(x)
\]

\[
\frac{\kappa}{x} = p(x) \quad (3.15)
\]
where another substitution \( \kappa := e^c \) was used in the last line. The result is surprisingly simple, but not as pleasing as it seems in the first place, because (3.15) is not a valid density function. To be, it must integrate to one, but the integral
\[
\int_{-\infty}^{+\infty} \frac{\kappa}{x} \, dx
\] (3.16)
diverges, and no value of \( \kappa \) could ever change that.

On the other hand, the result is not as bad as it might seem at the second look, because from the application point of view, the following constraints arise:

- A digit itself cannot have a sign, so negative numbers are dropped
- \( x \) can not be smaller than 1, as there is no lower first digit, which changes the lower limit to 1 instead of \(-\infty\)
- Only finite number bases are considered, so the upper limit is not \(+\infty\) but a finite number

With that restrictions that do not affect the properties we are interested in, a usable density function can be constructed from (3.15). Depending on the used number base, the value of \( \kappa \) has to be determined. Calculating \( \kappa \) for a given base \( B \) is easy. Say \( B = 10 \), e.g. the decimal system, we have to integrate (3.15) over all possible numbers in the set \([1, 10)\) and adjust \( \kappa \) that it is exactly one. In a general case, \( \kappa \) is determined by:
\[
\int_1^B p(x) \, dx = \int_1^B \frac{\kappa}{x} \, dx = \kappa \cdot \ln x |_1^B = \kappa \cdot \ln(B)
\] (3.17)

To let (3.17) integrate to one, the choice for \( \kappa \) is obviously
\[
\kappa \cdot \ln(B) \overset{!}{=} 1 \quad \Rightarrow \quad \kappa = \frac{1}{\ln(B)}
\] (3.18)

So the probability density function \( p_B \) for a given base \( B \) is
\[
p_B(x) = \frac{1}{\ln(B)} \cdot \frac{1}{x}
\] (3.19)

With a useful probability density function at hand, the probability of the first digit being \( d \) can now be calculated.

\[
P[D = d] = P[d \leq x < d + 1] = \int_d^{d+1} p(x) \, dx = \frac{1}{\ln(B)} \int_d^{d+1} \frac{1}{x} \, dx = \frac{\ln(d+1)}{\ln(B)} - \frac{\ln(d)}{\ln(B)} = \log_B \left( \frac{d+1}{d} \right)
\] (3.20)
Voila, here it is: the Benford distribution. The fact that it can be applied to any number base arises from the need to create a valid density function \( p(x) \), so this is automatically fulfilled.

Equation (3.20) can also be used to calculate the probability of any two or more digits being first. Consider to calculate the probability that the first two digits are 2 and 3, simply integrate \( p(x) \) over the interval \([2.3, 2.4)\). We therefore specify \( d = 23 \) (no decimal point here), and change the formula a bit to get down to the interval of \([2.3, 2.4)\) by dividing \( d \) and \((d+1)\) by ten each.

\[
P[D = d] = P\left[ \frac{d}{10} \leq x \leq \frac{d+1}{10} \right] = \int_{d/10}^{(d+1)/10} p(x) \, dx
\]

\[
= \frac{1}{\ln(B)} \int_{d/10}^{(d+1)/10} \frac{x}{x} \, dx
\]

\[
= \frac{\ln(\frac{d+1}{d})}{\ln(B)}
\]

\[
= \log_B \left( \frac{d+1}{d} \right)
\]

(3.21)

When using longer digit combination, for example four digits, than \( d \) and \((d+1)\) must be divided by a scaling factor of 100 or whatever to get back into the interval \([1, 10)\). From the equations (3.20) and (3.21) it is seen that the scaling factor always cancels in the end. So specifying the digit combination simply as a positive integer number and using that value for \( d \), will result in the wanted probability.

For instance, if interested in the probability of the first digits being the first five digits of \( \pi \) (3.1415) in the decimal system, just specify \( d = 31415 \) without the decimal point and use formula (3.20).

\[
P[D = 31415] = \log_{10} \left( \frac{31416}{31415} \right) \approx 13.824 \cdot 10^{-6}
\]

(3.22)

### 3.4 Examples

To show the application of the formulae and to give a proof that the Matlab scripts were implemented correctly, two examples are considered here.

#### 3.4.1 Exponential Growth

The first example shows the digit distribution of an exponential growth. Here the start value was 1,000, the interest rate was 0.001 (0.1%), and a total of 6912 steps were computed. The amount of steps is not arbitrary but was chosen to cover one full decade, so the start value will be 1,000, and the final value will be approximately 10,000. Figure 3.4 shows the distribution of the first digit, and figure 3.5 the distribution of the second one.
Both, the first and the second digit match the predicted distribution almost perfectly. The title in the figures also indicates the standard deviation from the theoretical values.

Figure 3.4: Exponential Growth (6912 values)
Properties of the Benford Distribution

Figure 3.5: Exponential Growth (6912 values)
3.4.2 Fibonacci Sequence

The second computer experiment was the Fibonacci sequence. Any element in this series is the sum of its two predecessors, and the sequence starts with two ones. In figure 3.6 and 3.7, the distribution of the first and second digit are plotted respectively. The used Fibonacci sequence had an overall length of 10 million values. Taking a look at the figures, the first digit again matches the Benford distribution very well, but the second digit is not perfectly aligned anymore.

Figure 3.6: Fibonacci Sequence (10,000,000 values)
Figure 3.7: Fibonacci Sequence (10,000,000 values)
Chapter 4

Real World Data and Simulation

In this chapter various data sets from three different institutes located at the ETH Zürich are presented. All data sets were either taken from numerical computer simulations or measurements.

Three of the experiments or simulations were not carried out by ourselves, nor were they created on purpose for us. The last data set (files on computer) however was gathered by ourselves. Also the kind of experiment was not specified by us - we simply asked a few people whether they could hand us some theoretical or practical data sets. The details of the experiments are hidden to us too, but this is rather an advantage than not. The reason is simply that without detailed knowledge, we picked all data elements that did not have any obvious pattern in it (like numbering, time stamps, deterministic parameter changes, etc) and analysed the whole set for the different experiments. Thinking of a possible practical applications of the Benford rule, the data analysis is most likely not carried out by the authors who did the experiment, or at least claimed to do so, but this is the job of reviewers who became suspicious. Those people then also only see the tables with the values and have to filter out what seems to be measured and what are deterministic data sets.

After evaluating the data sets, some were discarded because the Benford distribution will definitely not hold for them. Such sets either showed pretty constant values (for instance 5 Volts) or were simply noise. Analysing noise may yield interesting information as well but is definitely beyond the scope of this work.

Another thing to mention is that some of the data files were declared to be confidential, so neither performance graphs, detailed description or the like could be expected nor will be provided when questioned. Such knowledge on the other hand is irrelevant for the analysis and interpretation of the data, so no disadvantage is experienced here.

In the end, four experiment domains were chosen

1. Pharmacy
2. Photonic Crystals
3. Micro Turbines
4. Files on Computers

4.1 Pharmacy

This experiment dealt with the concentration of dialysis cells. The solutions are man made but the numerical results were created automatically by a computer.

There were several states of the experiment documented and there are no upper limits. The lower limit of course is zero due negative weights or the like don’t exist. This constraint however should still be OK on first sight to be Benford distributed, like the area of lakes or rivers, which cannot be negative too.

In the figures 4.1 to 4.5 the occurrence of the first digit is shown. The continuous red line indicates the probability predicted by Benford’s law. In the title of the figure the standard deviation to the predicted values is given as well as the total amount of numbers used to create the plot. The standard deviation is normally used as a quality measurement tool but more or less fails here (and in the later sections), because it does not take the distribution trend into account.

The five figures show five related experiments performed by different people. Comparing it to the Benford distribution, it does not look even similar. Still however, there seems to be a preference for the first digit, except in figure 4.2.

The following five figures show the same data set, but the distribution of the second digit is examined there (figures 4.6 to 4.10). In some cases the trend slightly follows the Benford distribution, but in general, not even a suitable rule of thumb could be derived here.
Figure 4.1: Pharmacy: First Absorption Measurement (102 Values)

Figure 4.2: Pharmacy: Second Absorption Measurement (306 Values)
Figure 4.3: Pharmacy: Third Absorption Measurement (583 Values)

Figure 4.4: Pharmacy: Fourth Absorption Measurement (75 Values)
Figure 4.5: Pharmacy: Fifth Absorption Measurement (96 Values)

Figure 4.6: Pharmacy: First Absorption Measurement (102 Values)
Figure 4.7: Pharmacy: Second Absorption Measurement (306 Values)

Figure 4.8: Pharmacy: Third Absorption Measurement (583 Values)
Figure 4.9: Pharmacy: Fourth Absorption Measurement (75 Values)

Figure 4.10: Pharmacy: Fifth Absorption Measurement (96 Values)
4.2 Photonic Crystals

Here a selection of simulations and measurements around Photonic Crystals is analysed. There were plenty of data sets available, but for convenience, not every single one but only a selection is shown here.

The selection includes the first, third and fifth data set, the distribution of the others look basically the same. In figures 4.11 to 4.16 the first and second digit of the corresponding data set is plotted. For the first digit there could maybe a tendency towards the Benford distribution be identified. Especially the first digit being preferably 1 is seen clearly here. For the second digit, Benford’s law seems not to be applicable.

Finally, all results were analysed together, and as a big surprise the digits of the whole sets are more or less perfectly Benford distributed - even the second one. This is astonishing but could eventually be a result of mixing different distributions together, which in turn must converge to the Benford distribution with an increasing number of samples.

Figure 4.11: Photonic Crystals: First Measurement (100 Values)
Figure 4.12: Photonic Crystals: First Measurement (100 Values)

Figure 4.13: Photonic Crystals: Third Measurement (100 Values)
Figure 4.14: Photonic Crystals: Third Measurement (100 Values)

Figure 4.15: Photonic Crystals: Fifth Measurement (100 Values)
Figure 4.16: Photonic Crystals: Fifth Measurement (100 Values)

Figure 4.17: Photonic Crystals: All Measurements together (69098 values)
Figure 4.18: Photonic Crystals: All Measurements together (69098 values)
4.3 Micro Turbines

This data was taken from a term project which dealt with micro turbines. The goal was to develop and test some of the electric components like the parameters of the inductors and then test the performance of the turbine. The data values cover various parameters like the voltage, the gap or the rotational speed.

The distribution of the first digit is shown in figures 4.19 to 4.22. It seems that again the nature of the measurement here complies with the Benford rule. Except for the fourth measurement (figure 4.22), at least the predicted distribution trend is met quite well. Even for the second digit (figure 4.23 to 4.26) the law is about to hold, although not as strictly as before anymore.

![Figure 4.19: Micro Turbine: First Measurement (438 Values)](image-url)
Figure 4.20: Micro Turbine: Second Measurement (720 Values)

Figure 4.21: Micro Turbine: Third Measurement (150 Values)
Figure 4.22: Micro Turbine: Fourth Measurement (99 Values)

Figure 4.23: Micro Turbine: First Measurement (438 Values)
Figure 4.24: Micro Turbine: Second Measurement (720 Values)

Figure 4.25: Micro Turbine: Third Measurement (150 Values)
Figure 4.26: Micro Turbine: Fourth Measurement (99 Values)
4.4 Files on Computers

Another experiment that is a bit amusing is to count the number of entries in any directory on a computer. Different directories there are used for different purposes, and the number of files in each depends heavily on that. The conditions therefore could meet the one for the Benford law. The test run on one of the authors home computers (Linux) and on the computing cluster at the electro technique department (Solaris).

Looking at the figures, one can notice that there is really a Benford trend for at least the first digit. It is better on the home computer (figure 4.27), but still there is a lack of values starting with the digit 1. On the whole cluster at the department (figure 4.29), the results are not quite as good. One explanation here might be that due to our user permissions, plenty of directories could not be accessed (but still 10268), so maybe this is one possible reason. Another reason pointed out by the system administrators is that there is quite some artificial structure in the setup that can never be compared to ones personal computer.

The second digit then is far apart from the Benford distribution, especially being 0 is outstanding, both at the home computer (figure 4.28) and on the whole cluster (4.30).

![Figure 4.27: Computer Files: Home Computer (17325 values)](image-url)
Figure 4.28: Computer Files: Home Computer (17325 values)

Figure 4.29: Computer Files: Electro Technique Department (10268 values)
Figure 4.30: Computer Files: Electro Technique Department (10268 values)
Chapter 5

The Survey

This chapter deals with the survey done to see if people automatically make up numbers that obey to Benford’s law. First, the questionnaire is presented, then the evaluation is shown and finally the results are given. In the appendix to this report the original questionnaire and the list of used principles declared by the participants can be found.

5.1 Online Questionnaire

5.1.1 First Page

One goal of the questionnaire was to study whether or not artificially faked numbers are Benford distributed, so in its first page people were asked to construct numbers. This page contained four questions: in the first three, the participants had to take on the role of a scientist and write numbers which they found to be realistic.

1. think like a physician and put realistic measured values to approximate a given graph.
2. provide realistic amounts of substances for a chemical experiment
3. take on the role of an engineer and specify characteristic numbers of a machine like the number of coil winding of the engines or resistor values
4. write down an arbitrary number, and think up a realistic sequence of six dice throws

The discussion of the results of the fourth question is postponed to section 5.4.

The first three questions were used explicitly to see if the Benford distribution had been used by the participants. At question 2 and 3 the first digit of the number given by the participants could directly be analysed. The data given in question 1 would never accord to Benford’s law if handled this way, because the graph we gave to approximate had values only between 60 and 73 and the people mostly gave numbers which started with 6 or 7. This question intended to find out whether the difference or the ratio to the original values accord to Benford’s law.
Question four had nothing to do with the Benford distribution at all, but it was interesting to see if the given dice-number sequences is in the correct ratio compared to real dice throws.

5.1.2 Second Page

In the second page of the questionnaire the participants were requested to give some information about themselves. These information were

1. if they believed to have written realistic numbers
2. if they had ever heard about the Benford rule and if yes . . .
3. . . .if they answered the questions on the first page by applying the Benford rule
4. if they had used other principles to fill in the first page
5. in which department and . . .
6. . . .in which semester they study

The time needed to fill out the survey was recorded too.

With the information on page two of our survey we wanted to compare the first four questions to each other. This is achieved by grouping the people according to what they declared on the second page. We could for example analyse if people who believed in given realistic numbers had more Benford distributed data sets than the others. Another possibility is to check whether students in higher semesters applied more to the Benford distribution than those in the first semester.

5.1.3 What Results Were Expected

We expected students in higher semesters to give more Benford distributed numbers than the others, because they had handled more with numbers and could have adopted the Benford distribution passively over time. We were especially interested in whether students from departments related to a specific question make up other numbers than those who didn’t know the topic at all. We also hoped to find out differences between those who knew and applied Benford’s law and those who don’t. Any other interesting result was of course welcomed too.

5.1.4 Questionnaire Procedure

The questionnaire was made as an Internet page. On Thursday the 24\textsuperscript{th} of June 2004 at nine o’clock in the morning an email with a link to our page was sent to all students of ETH Zürich. On Wednesday the 30\textsuperscript{th} of June 2004 all the given numbers were taken out of the database and processed locally on the computer. Until that moment 734 students had filled in the questionnaire, most of them at the first day of the survey. Regarding that the mail was sent to 7972 students, and 19 mails were rejected due to delivery errors, a return ratio of 9.2\% is not that high.
The Survey

Sadly not all sent data were usable because people wrote words where numbers were asked or just didn’t understand the questions correctly. The usable sets of numbers were therefore reduced to 711. From now on, only those 711 people are referred to as the participants.

5.2 Analysing the Numbers

A Matlab program was written to analyse the data from the questionnaire.

5.2.1 Fulfilment Criteria

The core of the program had to find out which sets of first digits accorded to Benford’s law. Because the exact distribution of the digits is quite unlikely to be achieved and a reasonable approximation could also be interpreted as a good result, different fulfilment criteria were defined. The criteria are summarised below, where $d_1$ denotes the amount of the first digit being '1', $d_2$ the amount of the first digit being '2' and so on:

$\hat{K}1$: \[d_1 + d_2 + d_3 + d_4 \geq d_5 + d_6 + d_7 + d_8 + d_9\]

$\hat{K}2$: \[d_1 + d_2 \geq d_3 + d_4 \text{ and } d_3 + d_4 \geq d_5 + d_6 \text{ and so on} \ldots\]

$\hat{K}3$: the occurrence of the first digit doesn’t exceed the real occurrence given by the Benford law by a factor of 2

$\hat{K}4$: the occurrence of the first digit doesn’t exceed the real occurrence given by the Benford law by a factor of 1.4

$\hat{K}0$: none of $K1$ to $K4$ are fulfilled. The set of first digits has nothing to do with Benford’s law.

The task of the written program was to determine which of these criteria were met for a given set of numbers. $K1$ is of course the easiest criterion to be fulfilled but shows clearly that people prefer numbers that start with lower digits. $K4$ sounds strange but is a very good way to find out sets of first digits that represent quite strictly the true Benford distribution.

As an example, figure 5.1 shows a graph which belongs to the second question (chemist). In the upper part of the graph each group of bars equals to a group of participants. The first group thinks to not have given realistic values. The second group thinks to have given realistic values just sometimes and the third group thinks to have given realistic values most of the time. The fourth group is the one of the people who did not declare (-ND-) whether they thought to have written realistic values and therefore couldn’t be assigned to one of the three main groups. The last group is the main average (-MA-). Every participant whatever he declared on page two counts also for the main average.

In the lower part of the graph you can see how many participants belong to each group: for example 490 participants (roughly 70% of all) thought to provide at least sometimes realistic values.

Sticking to this group and looking to the upper part of the graph, the amount of participants fulfilling the various criterions is given. 76% of these participants fulfill criterion $K1$ and 26% fulfill criterion $K2$. The first bar $K0$ of this group shows that approximately 23% don’t fulfill any criteria from $K1$ to $K4$. 

42
Grouping according to personal judgement for question 2 (chemist)

Figure 5.1: Analysis of the second question with grouping according to whether people think to have written realistic values

5.2.2 General Analysis

Unfortunately, our interpretations aren’t as significant as they would have been if more students had participated.

To make a general analysis of each question it is sufficient to look at the main average (-MA-) block of each figure, which is always the very right one.

Question One

The average fulfilment of criterion K1 to K4 is higher in question 1 than in the others, even if just lightly. This applies most if analysing not the difference but the ratio of the value given by the participant to the original function value. This is either due to the mathematical construction of the ratio that slightly aids the data set to automatically fulfil the criteria or because the people managed to write values that accord more to Benford’s distribution. The main average for question 1 using the differences is seen on the right in figure 5.2.

Another approach for evaluating question 1 is given in section 5.3 which shows the digit distribution directly.

Question Two

The last block of bars in picture 5.1 shows that about 74% of the participants prefer to numbers starting with lower digits (criterion 1). This points clearly in the direction of Benford’s theory.
Grouping according to personal judgement for question 1 (physician) as ratio

![Bar chart showing percentage of people fulfilling each criterion.]

---

**Figure 5.2:** Analysis of the first question with grouping according to whether people think to have written realistic values

**Question Three**

Looking at figure 5.3, even 81% of the people fulfilled criterion K1. The fulfil quota for criterion K2 is also quite high too. On contrary, K3 and K4 are less fulfilled, and nobody fulfilled K4 for question three.

**Conclusion**

Most of the people write numbers that start with low digits when they were asked to write realistic numbers. This points definitely in the direction of Benford’s law, even if the people did not succeed in providing a close match for all digits.

**5.2.3 Analysis with Division in Groups**

**Grouping by Self Judgement**

The biggest difference in fulfilment criteria comes when they are split up between those who thought to fill in realistic values and those who don’t. Figure 5.1 shows this relations.

For the answers to the first (figure 5.2) and third (figure 5.3) question the division in groups doesn’t show much differences between them. In question 1 people often have worse covering of the Benford criteria if they believed to have realistic values.
Figure 5.3: Analysis of the third question with grouping according to whether people think to have written realistic values

**Grouping by Knowledge About the Benford Theory.**

In the second question on page two of the questionnaire people could tell if they have ever heard about the Benford law. Those who already had heard about it had more Benford distributed sets of numbers than the others at the first and third question. In figure 5.4 the analysis for the third question is depicted.

**Grouping in Regard of the Application of the Benford Theory.**

Because just five persons (less than 1% of all participants) declared to have written numbers using the Benford law, it is impossible to speak of a trend and it is therefore difficult to make significant interpretations. On almost all questions except the third their values were less Benford distributed than those of the other people that know the theory of Benford but did not construct the numbers according to it. For the first question maybe they did not know that the difference or the ratio will be looked at.

**Grouping in Regard of the Use of Other Principles**

If people used other principles, it didn’t make a big difference for the fulfilment of the criteria K1 to K4. With this question mainly mathematical principles were meant, but the participants often told their own one, which was unexpected but nevertheless perfectly fine too.
Grouping in Regard of the Student’s Department

At the first question (difference), the students of the 'Enterprise and Production Science' department fulfilled the Benford distribution best. The physicians that actually were addressed by that question had fulfilment quotas like the average.

At question 2, which was targeted mostly to the chemistry and related students, those students provided values that were less Benford distributed than most of the others.

At question 3 electrical engineering, mechanical engineering and computer science students wrote values that were as Benford distributed as the main average.

Grouping by Semester

The grouping of participants by the same semester does not show who has constructed numbers that obey better to Benford’s law.

Grouping by Time Needed to Fill Out the First Page

Grouping in two groups of same size between those who needed more time to fill out the first page than the others yields that both groups fulfil the Benford criteria in the same way.

Figure 5.4: Analysis of the third question with grouping according to whether people know the Benford distribution
5.3 Analysis of the Digit Distribution

5.3.1 Faking a Graph

The first question in the survey dealt with manufacturing results. For this purpose a function graph was plotted and the people were asked to approximate the values of the function.

The hypotheses here was that artificially created numbers are not Benford distributed, although at least a preference for low digits could be expected. This due to the participants probably expected the differences to the real value not being too big.

To analyse the numbers, all answers were merged together and no grouping whatsoever was performed. Here this is considered valid because no one should have a preference for this question, unlike for example the second question which addresses chemistry and related students.

Looking at the results for the first digit (figure 5.5), this hypothesis is definitely wrong. Not only the lower digits are preferred being first, but the whole distribution looks pretty much Benford like.

For the second digit the match is worse than for the first digit (figure 5.6), but still the first digit being ‘1’ occurs most.

To give an idea whether the fabrication was carried out well, white Gaussian noise with same variance and mean than that of the faked data was analysed. As always in surveys, there are outliers, so calculating the mean and variance of the faked differences is worth not much, and figure 5.7 shows why. There were very few people who most likely inserted some funny numbers, which dominate the mean and the variance. No one said that such values are not allowed, but for the comparison to noise with identical parameters they must be removed. To get rid of them, simply the twenty smallest and biggest values were dropped. For the rest, a mean value (\(\mu = 0.4714\)) and a variance (\(\sigma^2 = 249.23\)) could be identified. Those two parameters were then in turn used to generate the noise. The digit analysis of the noise is shown in figure 5.8 for the first digit, and in 5.9 for the second digit.

Especially the first digit does not show any significant differences, whereas the second digit seems to be more smooth for the noise than for the fabricated numbers. Anyway, generalising this result is dangerous, because no one said that the noise must be white, or Gaussian or even both. It could be anything, but white Gaussian noise shows up quite often and having a look at it seemed worthy to us.
Figure 5.5: First Digit of the differences (6868 values)

Figure 5.6: Second Digit of the differences (6868 values)
Figure 5.7: Differences of all users together

Figure 5.8: First Digit of White Gaussian Noise with Mean $\mu = 0.4714$ and Variance $\sigma^2 = 249.23$
Figure 5.9: Second Digit of White Gaussian Noise with Mean $\mu = 0.4714$ and Variance $\sigma^2 = 249.23$
5.3.2 Invent Amount of Chemical Solutions

The second question of the survey treated faked numbers for chemical experiments. Each student should think up reasonable amounts of solutions used for an experiment. In the figures 5.10 to 5.12 the distribution of the first, second and third digit are shown.

Comparing the distribution to the one predicted by Benford, it seems that especially for the first digit the distribution is met accordingly. The second and third digit on the other hand are not even close to the Benford distribution, but the digit 0 is outstanding. A simple reason could be that the numbers were just not long enough, so there might not be a second or third digit at all, which are then implicitly assumed being zero.

However, what could be noted when looking at the figures is the preference for the digit 5.

![First Digit: Question 2 (Chemistry)](chart.png)

Figure 5.10: First Digit: Question 2 (Chemistry)
Figure 5.11: Second Digit: Question 2 (Chemistry)

Figure 5.12: Third Digit: Question 2 (Chemistry)
5.3.3 Invent Numbers for the Parameters of a Machine

The participants were asked to provide some machine characteristics that sound reasonable to them. The bottom line is pretty much the same as in the second question - the first digit is again pretty much Benford distributed, the second and third are not.

Here again, the digit 5 is used unusually often, but the digit 9 on the contrary is used too seldom for the first digit (figure 5.13). For the second (figure 5.14) and third (figure 5.15) digit, a value of 0 again dominates the distribution. The reason here again is probably that the people just did not provide more digits. Like for the first digit, a 5 is used more frequently here.

Figure 5.13: First Digit: Question 3 (Engineer)
Figure 5.14: Second Digit: Question 3 (Engineer)

Figure 5.15: Third Digit: Question 3 (Engineer)
5.4 Evaluation of the Dice Question

In the fourth question of our questionnaire the participants had to invent a realistic sequence of six dice throws. We can analyse them under two aspects: the general probability of a number to be thrown and the amount of sequences of the same number.

5.4.1 General Probability of Numbers

When throwing a dice there is one possibility of six to get a specific number. Throwing a die six times, it is very unlikely that each number from one to six shows up exactly once in a realistic experiment. Having an union of a lot of such sequences, it can only be realistic if all numbers of 1 to 6 show up roughly equally often (if it is an unbiased die of course). Table 5.1 shows the occurrence of each number that was provided by the participants. The average amount is 695, so number 2 and 3 were written too often. It is impressive to see how the amounts differ only a little from the average amount and that number 2 and 3 have the same amounts.

<table>
<thead>
<tr>
<th>Number</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>689</td>
</tr>
<tr>
<td>2</td>
<td>733</td>
</tr>
<tr>
<td>3</td>
<td>733</td>
</tr>
<tr>
<td>4</td>
<td>665</td>
</tr>
<tr>
<td>5</td>
<td>650</td>
</tr>
<tr>
<td>6</td>
<td>660</td>
</tr>
</tbody>
</table>

Table 5.1: Real Occurrence of Each Number in the Die Experiment

5.4.2 Sequences of Same Numbers (SoSN)

If we throw twice the same number, we have a sequence of same numbers (SoSN) of length two. To start such a sequence we have to throw first a number that is different from the last one thrown, which happens with a probability of $\frac{5}{6}$. If we want a SoSN with length 1, then the next number has again to be different, which happens with a probability of $\frac{5}{6}$. So the probability to get an isolated number is $\frac{5}{6} \cdot \frac{5}{6}$. A SoSN of length 2 happens with a probability of $\frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6}$, where the new factor $\frac{1}{6}$ is the probability that the second number is the same than the first one. We can conclude that a SoSN of length $n$ happens with a probability

$$P(\text{SoSN}|n) = \frac{5}{6} \left(\frac{1}{6}\right)^{n-1} \frac{5}{6}$$  \hspace{1cm} (5.1)

A SoSN for a given number (SoSNgN) happens six times less than a generalised SoSN for any number. The following list shows the probability for SoSNgN of length one to six:
In the questionnaire the participants couldn’t write in more than six numbers, so they didn’t have the possibility to write a SoSN longer than six. If we count all the SoSNgN given by the participants and divide them through the total number of SoSN they wrote, we get table 5.2 that shows the SoSNgN occurrence proportions. Comparing this to the real SoSNgN probability given before yields:

1. SoSN with length \( n = 1 \) occur just a little too much
2. SoSN with length \( n = 2 \) occur about one third less
3. SoSN with length \( n = 3 \) occur by far too less
4. SoSN with length \( n = 4 \) occur too much for the numbers 1 and 6

Obviously, number 2 is really popular for SoSN of length two but does not occur in any SoSN longer than this. The number that had the highest occurrences of SoSN was number 6. Maybe this is due to 6 is often a good throw in table games and the participants realise and remember such sequences of 6 more than sequences of other numbers.

As a result, we can conclude that sequences of same numbers occurred by far too seldom.
5.5 What We Could Have Done Better

One thing we could have done better is to ameliorate the rate of people who sent usable data back. With usable data we mean just numbers. Many people wrote also words or letters in the field of the first page of the questionnaire, which made the set of answers impossible to be processed for the analysis. A question that was not understood correctly was the one about about the amount of basic blocks in the chip at the engineer question. Maybe many people don’t know what this really means because they tried to write a kind of serial number of this chip (e.g. sdfg23fs8d32, 280BP-12T-STX, XAP23, or numbers with more points in it). If thinking of a chip, such codes sound realistic, even more than number that would answer the question correctly. To not loose all this people for the evaluation, we took the first digit of their answer to be the value we would analyse if it was a number and not a letter. We also could have done a hidden counter on the first site to see how many of the people who visited the site have gone away without completing the questionnaire.
Willkommen zu unserer Umfrage!

In den unten vorgestellten Situationen kann man aus irgend einem Grund die Originalwerte nicht brauchen und entscheidet sich zu unwissenschaftlichem Vorgehen und erfindet die Daten selbst. Bitte auch dort ausfüllen, wo ihr keine Ahnung haben könntet, was realistisch ist.

Wir wollen hier allerdings nicht den Eindruck erwecken, dass Fälschen in der Wissenschaft akzeptabel ist!

Frage 1
Du bist ein Physiker.
Im Traum kam dir die Formel in den Sinn, nach der du schon lange gesucht hast. Erfinde realistisch aussehende Messwerte, um deren Kurve zu bestätigen.

<table>
<thead>
<tr>
<th>Vorheriger Wert</th>
<th>Erfundener Messwert</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y(10) = 62.2920 )</td>
<td>( y(20) = 70.0507 )</td>
</tr>
<tr>
<td>( y(30) = 72.8022 )</td>
<td>( y(40) = 69.8953 )</td>
</tr>
<tr>
<td>( y(50) = 68.9631 )</td>
<td>( y(60) = 70.5677 )</td>
</tr>
<tr>
<td>( y(70) = 70.8529 )</td>
<td>( y(80) = 69.8737 )</td>
</tr>
<tr>
<td>( y(90) = 69.7113 )</td>
<td>( y(100) = 70.1915 )</td>
</tr>
</tbody>
</table>

Frage 2
Du bist ein Chemiker.
Für ein durchgeführtes Experiment hast du 10 Substanzen benutzt. Erfinde plausible Mengenangaben.

<table>
<thead>
<tr>
<th>Substanz Nr. 1</th>
<th>mg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substanz Nr. 2</td>
<td>mg</td>
</tr>
<tr>
<td>Substanz Nr. 3</td>
<td>mg</td>
</tr>
<tr>
<td>Substanz Nr. 4</td>
<td>mg</td>
</tr>
<tr>
<td>Substanz Nr. 5</td>
<td>mg</td>
</tr>
<tr>
<td>Substanz Nr. 6</td>
<td>mg</td>
</tr>
<tr>
<td>Substanz Nr. 7</td>
<td>mg</td>
</tr>
<tr>
<td>Substanz Nr. 8</td>
<td>mg</td>
</tr>
<tr>
<td>Substanz Nr. 9</td>
<td>mg</td>
</tr>
</tbody>
</table>

Figure A.1: First page of the questionnaire, part one
Substanz Nr. 10: mg

Frage 3
Du bist ein Ingenieur.
Du hast eine neue Maschine konstruiert. Erfinde einige Kennzahlen die für dich realistisch klingen.

Anzahl Windungen Motor 1: __________________________
Anzahl Windungen Motor 2: __________________________
Anzahl Windungen Motor 3: __________________________
Anzahl Windungen Motor 4: __________________________
Widerstand Nr. 1 [Ohm]: __________________________
Widerstand Nr. 2 [Ohm]: __________________________
Widerstand Nr. 3 [Ohm]: __________________________
Anzahl Basic-Logic-Blöcke des Prozessors: __________________________
Zeilen Code für das Betriebssystem deiner Maschine: __________________________

Frage 4
Jetzt bist du einfach du selbst.

• Schreibe die erste Zahl ein, die dir einfällt: __________________________
• Stell dir vor, du hast sechs mal gewürfelt. Schreibe hier die (erfindenen) Zahlen auf:
  Würfelwurf Nr.: 1 2 3 4 5 6
  Erhaltene Zahl: __________________________

Ok, jetzt nur noch Daten abschicken: Daten senden

Figure A.2: First page of the questionnaire, part two

Danke für das Ausfüllen.
Jetzt nur noch ein paar Fragen, die wir für das Evaluieren brauchen:

• Hast du es deiner Meinung geschafft, realistische Messwerte einzusetzen?
  • Ja
  • Manchmal
  • Nein

• Hast du schon einmal von der Benford’schen Regel (Benford’s law) oder "The Power of One"-Regel gehört?
  • Ja
  • Nein

• Wenn Ja, hast du die Werte der ersten drei Fragen (Physiker, Chemiker, Ingenieur) bewusst nach der Benford-Regel konstruiert?
  • Ja
  • Nein

• Hast du sonst noch ein anderes spezielles Prinzip angewendet?
  • Ja, nämlich __________________________
  • Nein

Bei welchem Departement studierst/arbeitest du? __________________________
In welchem Jahr studierst du? __________________________

• Möchtest du an der Pralinen-Verlosung teilnehmen?
  • Ja
  • Nein

• Willst du über das Ergebnis der Studie informiert werden?
  • Ja
  • Nein

Falls du mindestens eine der letzten zwei Fragen mit "Ja" beantwortet hast, benötigen wir noch deine Email Adresse.
Email: __________________________

Daten abschicken: senden

Figure A.3: Second page of the questionnaire
A.2 Principles the People Used

As said before, less than 1% of the participants used Benford’s law to fill out the questionnaire. On the second page we asked them if they used some other kind of principle and this are the answers we got:

- möglichst zufällig verteilt, aber nicht unplausibel
- Runden
- unauffälligkeit
- augenschliessen, und erste zahl dich ’visuell’ sehe wahlen
- mein Gefühl
- nicht zu verschiedene Zahlen
- keine Extremwerte, die aber realistischer wären, da bei Messungen vielfach unvorhergesehene Faktoren die Messreihe beeinflussen und somit nicht schön geschwungene Kurven ergeben
- Die Zahlen der Nachkommastellen einfach zufällig Permutiert, so bleibt die Statistik der einzelnen Ziffern erhalten
- schätzen :)
- zufall
- gleiche grössenordnung
- immer etwas mehr und etwas weniger
- Lust und Laune
- zufällige Abweichungen von Realwerten
- Werte der E Reihen bei Widerständen verwenden
- teilweise habe ich darauf geachtet, dass ich nicht zahlenreihen verwendet habe (zb. 4.234)
- common sense
- zufällsverteilung
- nicht zu viele ”runden” Zahlen, aber ein paar.
- random numbers
- schnell&einfach
- konstante pos. Abweichung, wenig 8,9 einsetzen
- auf ”schöne Werte” geachtet (Runden)
- einfach öppis...
- Wahrscheinlichkeit mehrmals dieselbe Zahl bei Würfel
• Normalverteilte Zufallszahlen

• möglichst wenig runde Zahlen (d.h. nicht mit 0 oder 5 endend) und auch keine Zahlenreihen (234...)

• Schwächsinn

• runden

• Lieblingszahlen

• mal höher, mal tiefer

• bei 1 eine schwingung mit harmonischer dämpfung

• nie den exakten Wert

• grenzwert bei 1.

• bei dem Motor habe ich Windungen proprotional zu den Widerständen gesetzt

• Für die 2. Kommastellen blind auf die Tasten gedrückt

• Runde, Mittel

• Zufälliges herumtippen auf dem Ziffernblock

• nicht so antworten, wie alle anderen...bsp:würfel:nicht alle zahlen würfeln!

• meinen internen zufallsgenerator

• an logischen mengen beim chemiker orientiert, die werte der wunschkurve so verändert, dass sie (gefittet) wahrscheinlich herauskommt...

• ”Humaner Zufallsgenerator” :-)

• nicht zu genaue Werte wählen, Messwerte sind immer etwas ungenau

• hau in die tasten - prinzip

• Nicht immer unterschiedliche Zahlen, sondern auch Wiederholungen einbauen! Abweichungen nicht immer gleich wählen

• Douglas Adamses 42!

• Messwerte innerhalb 5

• Wiederholungen nicht vermeiden

• wahlloses auf die Tasten hauen.

• meines

• Zufallsprinzip im sinnvollen Rahmen

• Zufallsprinzip

• Erfahrungswerte in der Chemie
Möglichst assortiert die Zahlen des Zahlenblocks drücken

Zufallsprinzip

Wiederholungen

kalkulierter zufall

öfters mal gleiche Zahlen nacheinander

habe ich eine Laptop-Tastatur und darum alle Zahlen hintereinander angeordnet. Somit ist es unwahrscheinlicher, dass ich z.B. 1 und 0 nacheinander schreibe, weil sie so weit auseinanderliegen.

blind getippt

möglichst KEIN Prinzip

Zufall

Erfahrung, runde Zahlen


Zufall

Zufall & Uniformverteilung

versucht nicht krampfhaft muster zu vermeiden (Enigma ... kein muster ist ein muster)

ich versuche sowohl ungerade wie gerade Endziffern zu haben. Ich suche ein systemisches Messresultat zu habe, welches gewisse Messsprünge macht (tt teilweise angewendet)

ich habe versucht, auch nullen und fünf en zu schreiben

Ruden

Erfahrungswerte (Windungen, Widerstände usw.)

Berücksichtigung der Art der Auswirkungen von Fehlerquellen (ungenaue Messgeräte, Bereiche, in denen es Abweichungen vom idealisierten, mathematischen Modell gibt)

Die Zahlen der Nachkommastellen einfach zufällig Permutiert, so bleibt die Statistik der einzelnen Ziffern erhalten

kalkulierter zufall

ich versuche sowohl ungerade wie gerade Endziffern zu haben. Ich suche ein systemisches Messresultat zu habe, welches gewisse Messsprünge macht (tt teilweise angewendet)
List of Tables

3.1 Probability of the Digits 1 to 9 according to Benford’s Law . . . 8
5.1 Real Occurrence of Each Number in the Die Experiment . . . . 55
5.2 Proportion of SoSNgN given by the participants . . . . . . . 56
List of Figures

3.1 Digit Probabilities according to Benford’s Law in the Decimal System .......................... 9
3.2 Probabilities of First Digit in various Systems ......................................................... 10
3.3 Probabilities of First Two Digits in Decimal System ................................................. 11
3.4 Exponential Growth (6912 values) ........................................................................... 16
3.5 Exponential Growth (6912 values) ........................................................................... 17
3.6 Fibonacci Sequence (10,000,000 values) ................................................................. 18
3.7 Fibonacci Sequence (10,000,000 values) ................................................................. 19

4.1 Pharmacy: First Absorption Measurement (102 Values) ........................................... 22
4.2 Pharmacy: Second Absorption Measurement (306 Values) ....................................... 22
4.3 Pharmacy: Third Absorption Measurement (583 Values) ......................................... 23
4.4 Pharmacy: Fourth Absorption Measurement (75 Values) .......................................... 23
4.5 Pharmacy: Fifth Absorption Measurement (96 Values) ............................................. 24
4.6 Pharmacy: First Absorption Measurement (102 Values) ........................................... 24
4.7 Pharmacy: Second Absorption Measurement (306 Values) ....................................... 25
4.8 Pharmacy: Third Absorption Measurement (583 Values) ......................................... 25
4.9 Pharmacy: Fourth Absorption Measurement (75 Values) .......................................... 26
4.10 Pharmacy: Fifth Absorption Measurement (96 Values) ............................................ 26
4.11 Photonic Crystals: First Measurement (100 Values) ................................................... 27
4.12 Photonic Crystals: First Measurement (100 Values) ................................................... 28
4.13 Photonic Crystals: Third Measurement (100 Values) ................................................... 28
4.14 Photonic Crystals: Third Measurement (100 Values) ................................................... 29
4.15 Photonic Crystals: Fifth Measurement (100 Values) ................................................... 29
4.16 Photonic Crystals: Fifth Measurement (100 Values) ................................................... 30
4.17 Photonic Crystals: All Measurements together (69098 values) .................................. 30
4.18 Photonic Crystals: All Measurements together (69098 values) .................................. 31
4.19 Micro Turbine: First Measurement (438 Values) ....................................................... 32
4.20 Micro Turbine: Second Measurement (720 Values) .................................................... 33
4.21 Micro Turbine: Third Measurement (150 Values) ...................................................... 33
4.22 Micro Turbine: Fourth Measurement (99 Values) ....................................................... 34
4.23 Micro Turbine: First Measurement (438 Values) ....................................................... 34
4.24 Micro Turbine: Second Measurement (720 Values) .................................................... 35
4.25 Micro Turbine: Third Measurement (150 Values) ...................................................... 35
4.26 Micro Turbine: Fourth Measurement (99 Values) ....................................................... 36
4.27 Computer Files: Home Computer (17325 values) .................................................... 37
4.28 Computer Files: Home Computer (17325 values) .................................................... 38
4.29 Computer Files: Electro Technique Department (10268 values) ............................. 38
<table>
<thead>
<tr>
<th>Figure No.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.30</td>
<td>Computer Files: Electro Technique Department (10268 values)</td>
<td>39</td>
</tr>
<tr>
<td>5.1</td>
<td>Analysis of the second question with grouping according to whether people think to have written realistic values</td>
<td>43</td>
</tr>
<tr>
<td>5.2</td>
<td>Analysis of the first question with grouping according to whether people think to have written realistic values</td>
<td>44</td>
</tr>
<tr>
<td>5.3</td>
<td>Analysis of the third question with grouping according to whether people think to have written realistic values</td>
<td>45</td>
</tr>
<tr>
<td>5.4</td>
<td>Analysis of the third question with grouping according to whether people know the Benford distribution</td>
<td>46</td>
</tr>
<tr>
<td>5.5</td>
<td>First Digit of the differences (6868 values)</td>
<td>48</td>
</tr>
<tr>
<td>5.6</td>
<td>Second Digit of the differences (6868 values)</td>
<td>48</td>
</tr>
<tr>
<td>5.7</td>
<td>Differences of all users together</td>
<td>49</td>
</tr>
<tr>
<td>5.8</td>
<td>First Digit of White Gaussian Noise with Mean $\mu = 0.4714$ and Variance $\sigma^2 = 249.23$</td>
<td>49</td>
</tr>
<tr>
<td>5.9</td>
<td>Second Digit of White Gaussian Noise with Mean $\mu = 0.4714$ and Variance $\sigma^2 = 249.23$</td>
<td>50</td>
</tr>
<tr>
<td>5.10</td>
<td>First Digit: Question 2 (Chemistry)</td>
<td>51</td>
</tr>
<tr>
<td>5.11</td>
<td>Second Digit: Question 2 (Chemistry)</td>
<td>52</td>
</tr>
<tr>
<td>5.12</td>
<td>Third Digit: Question 2 (Chemistry)</td>
<td>52</td>
</tr>
<tr>
<td>5.13</td>
<td>First Digit: Question 3 (Engineer)</td>
<td>53</td>
</tr>
<tr>
<td>5.14</td>
<td>Second Digit: Question 3 (Engineer)</td>
<td>54</td>
</tr>
<tr>
<td>5.15</td>
<td>Third Digit: Question 3 (Engineer)</td>
<td>54</td>
</tr>
<tr>
<td>A.1</td>
<td>First page of the questionnaire, part one</td>
<td>58</td>
</tr>
<tr>
<td>A.2</td>
<td>First page of the questionnaire, part two</td>
<td>59</td>
</tr>
<tr>
<td>A.3</td>
<td>Second page of the questionnaire</td>
<td>59</td>
</tr>
</tbody>
</table>
Bibliography


[6] [http://www.intuitor.com/statistics/Benford%27s%20Law.html]

[7] [http://www.rexswain.com/benford.html]


[9] [http://www.fortunecity.com/emachines/e11/86/one.html]

[10] [http://mathworld.wolfram.com/BenfordsLaw.html]