

FRAUDE DETECTEREN MET BEHULP VAN DE WET VAN BENFORD

EEN SIMULATIESTUDIE



AZIZ MOHAMMADI

MEI 2008

FRAUDE DETECTEREN MET BEHULP VAN DE WET VAN BENFORD

EEN SIMULATIESTUDIE

AUTEUR

Aziz Mohammadi
mohammadiaziz@gmail.com

BEGELEIDER

Dr. Marianne Jonker

FOTO VOORPAGINA

Frank Benford
<http://www.nigrini.com/images/Benford2.jpg>

**Vrije Universiteit Amsterdam
Faculteit Exacte Wetenschappen
Bedrijfswiskunde en Informatica
De Boelelaan 1081a
1081 HV Amsterdam**

MEI 2008

VOORWOORD

Het BWI-werkstuk is één van de laatste drempels die genomen moet worden op de weg die leidt naar het Master diploma van de opleiding Bedrijfswiskunde en Informatica aan de Vrije Universiteit in Amsterdam. Het doel van dit werkstuk is om de resultaten van een onderzoek, dat zelfstandig is uitgevoerd, op een heldere manier te beschrijven. Het onderwerp van het werkstuk mag hierbij zelf gekozen worden. Wel dient in het onderzoek, naast de wiskunde en informatica aspecten, ook het bedrijfsgerichte aspect van de studie terug te komen.

Het vinden van een geschikt onderwerp voor het BWI-werkstuk was een ware speurtocht die mij langs vele docenten heeft geleid. De reden hiervoor was dat ik op zoek was naar een interessant onderwerp waaraan ik een geruime tijd gemotiveerd kon werken. De onderwerpen die door de meeste docenten werden aangereikt voldeden echter niet aan deze eis totdat ik, op aanraden van Fetsje Bijma, terecht kwam bij Marianne Jonker. Marianne Jonker liet mij in een gesprek kennis maken met de wet van Benford en zijn toepassingen op het gebied van fraudedetectie. Alhoewel ik nooit van deze wet had gehoord, was mijn interesse voor dit onderwerp gelijk gewekt. Ik zag voldoende perspectief om een onderzoek uit te kunnen voeren dat tot betekenisvolle resultaten zou kunnen leiden. Ik wist op dat moment dat ik “mijn” onderwerp had gevonden.

Het uitvoeren van het onderzoek liep erg soepel. In samenwerking met Marianne Jonker heb ik een duidelijke onderzoeksvraag geformuleerd, die in het onderzoek stap voor stap is uitgewerkt. Hierin heeft Marianne Jonker mij uitstekend begeleid. Dankzij haar enthousiasme heb ik altijd met veel plezier gewerkt aan mijn BWI-werkstuk. Ik wil haar dan ook bij deze bedanken voor deze uitstekende begeleiding.

Het harde werken heeft uiteindelijk geresulteerd in het BWI-werkstuk dat nu voor u ligt. In dit werkstuk wordt antwoord gegeven op de vraag wat de wet van Benford inhoudt en wordt uitgebreid ingegaan op de wiskundige eigenschappen van deze wet. Daarnaast wordt besproken hoe deze wet gebruikt kan worden om fraude te detecteren in boekhoudingen van bedrijven. Het vroegtijdig ontdekken van deze fraude kan bedrijven namelijk veel geld besparen. Met behulp van diverse simulaties wordt inzicht gegeven in het aantal gefraudeerde bedragen dat nodig is om een geval van fraude met een bepaalde zekerheid aan het licht te stellen.

Aziz Mohammadi

Amsterdam, mei 2008

SAMENVATTING

De wet van Benford is vernoemd naar de natuurkundige Frank Benford, die in het jaar 1938 een wetenschappelijk artikel publiceerde waarin hij, op basis van empirisch bewijs, een wet beschreef die stelt dat de verdeling van de begincijfers van getallen die deel uitmaken van een dataset logaritmisch verdeeld zijn en geen uniforme verdeling volgen, zoals men intuïtief zou verwachten. Deze wet stelt dat de kans dat een willekeurig getal in een dataset begint met het cijfer $d \in \{1, 2, \dots, 9\}$ gelijk is aan $P(\text{begincijfer} = d) = \log_{10}(1 + 1/d)$. Volgens deze wet komen getallen met een laag begincijfer (1, 2 of 3) vaker voor dan getallen die beginnen met een hoog begincijfer. Zo is de kans dat een willekeurig getal uit een dataset begint met het cijfer 1 ongeveer gelijk aan 0.30 terwijl de kans op een 9 als begincijfer ongeveer gelijk is aan 0.046.

Helaas voldoen niet alle datasets aan de wet van Benford. Zo geldt bijvoorbeeld dat getallen die gegenereerd zijn aan de hand van een random proces, zoals lotterijgetallen of dobbelsteenworpen, niet de verdeling volgen die de wet van Benford voorschrijft. De wiskundige Theodore Hill heeft aangetoond dat een dataset aan de wet van Benford voldoet indien de verdeling van de getallen waaruit deze dataset is samengesteld een combinatie is van andere verdelingen. Om deze reden voldoet accountingdata (bij benadering) aan de wet van Benford. Een simpel voorbeeld is het bedrag aan openstaande vorderingen dat berekend wordt door het aantal verkochte producten, dat een bepaalde verdeling volgt, te vermenigvuldigen met de prijs per product, welke afkomstig is uit een andere verdeling.

De wet van Benford kent vele toepassingen in de praktijk. De belangrijkste toepassing van deze wet bevindt zich op het gebied van fraudedetectie in financiële overzichten van bedrijven, zoals boekhoudingen. Dé persoon die een grote rol heeft gespeeld in het populariseren van het gebruik van deze wet als fraudedetectietechniek is Mark Nigrini. Hij paste als eerste onderzoeker de wet van Benford toe op accountingdata en heeft in de afgelopen twaalf jaar verschillende artikelen gepubliceerd op dit gebied. In deze artikelen worden meerdere testen beschreven, die gebaseerd zijn op de wet van Benford, waarmee fraude in financiële overzichten ontdekt kan worden. Wat wij echter missen, in deze en andere artikelen, is inzicht in het minimale aantal gemanipuleerde bedragen dat nodig is om met een bepaalde zekerheid een geval van fraude te ontdekken wanneer gebruik wordt gemaakt van een bepaalde (statistische) toets.

Het doel van dit werkstuk is dan ook, om aan de hand van een uitgebreide simulatiestudie, inzicht te verschaffen in het aantal gemanipuleerde bedragen dat nodig is om een geval van fraude aan het licht te stellen indien men gebruik maakt van Pearson's chikwadraat-toets. De onderzoeksvraag die centraal staat in dit werkstuk is: hoeveel gefraudeerde bedragen n dienen er minstens toegevoegd te worden aan een boekhouding, die voldoet aan de wet van Benford en welke bestaat uit N bedragen, om een fraudeur met een kans van tenminste p te pakken. Hierbij wordt aangenomen dat de fraudeur de begincijfers van de n bedragen uniform verspreidt over de cijfers $1, 2, \dots, 9$.

De bovenstaande onderzoeksvragen hebben wij beantwoord door middel van simulaties. Deze simulaties zijn uitgevoerd in het statistische pakket R. Hiervoor is een functie geschreven die de nulhypothese toetst die stelt dat de verdeling van de begincijfers van bedragen in een boekhouding van een bedrijf voldoen aan de wet van Benford. Om de kans op het verwerpen van de nulhypothese zo nauwkeurig mogelijk te schatten wordt deze nulhypothese meerdere keren getoetst waarbij de begincijfers van de bedragen N en n telkens opnieuw worden getrokken uit de bijbehorende verdelingen. Het aantal keren dat de nulhypothese wordt verworpen, als fractie van het totale aantal keer dat de nulhypothese getoetst wordt, vormt dan een goede benadering van deze kans. Door de simulaties voor verschillende waarden van n en N uit te voeren, kan een duidelijk inzicht verkregen worden in het aantal bedragen dat nodig is om de fraude met een bepaalde kans te ontdekken.

De belangrijkste resultaten die zijn voortgekomen uit de uitgevoerde simulaties zijn in de onderstaande drie tabellen weergegeven. In deze tabellen is aangegeven hoeveel gefraudeerde of gemanipuleerde bedragen aan een boekhouding toegevoegd dienen te worden zodanig dat de eerder genoemde nulhypothese met een gegeven zekerheid verworpen kan worden bij een vooraf gedefinieerde onbetrouwbaarheidsdrempel¹.

De uitgevoerde simulaties hebben aangetoond dat het aantal gemanipuleerde bedragen, als fractie van de gehele boekhouding, dat nodig is om de fraudeur met een bepaalde zekerheid in de kraag te vatten kleiner wordt naarmate de initiële boekhouding groter wordt. De significante afwijking ten opzichte van de wet van Benford werd hierbij voornamelijk veroorzaakt door een tekort aan bedragen die beginnen met het cijfer één en aan een overschot van bedragen die beginnen met het cijfer zeven, acht of negen, respectievelijk. Daarnaast hebben de simulaties uitgewezen dat des te groter de onbetrouwbaarheidsdrempel is, des te minder gefraudeerde bedragen er nodig zijn om de nulhypothese, met een gegeven kans, te verwerpen. Op basis van gevonden resultaten zijn wij tot de conclusie gekomen dat de wet van Benford een eenvoudige en handige techniek is die gebruikt kan worden voor het detecteren van structurele fraude.

Tabel I: Het aantal gefraudeerde bedragen dat toegevoegd dient te worden aan de boekhouding, waarvan de grootte in de eerste kolom is weergegeven, zodanig dat de nulhypothese met de gegeven kans verworpen kan worden bij een onbetrouwbaarheidsdrempel van 0.25.

Grootte boekhouding	P(H ₀ wordt verworpen)≈0,75	P(H ₀ wordt verworpen)≈0,95	P(H ₀ wordt verworpen)≈0,99
1.000	140	220	260
10.000	425	625	750
100.000	1.350	2.000	2.350
1.000.000	4.200	6.200	7.400

Tabel II: Het aantal gefraudeerde bedragen dat toegevoegd dient te worden aan de boekhouding, waarvan de grootte in de eerste kolom is weergegeven, zodanig dat de nulhypothese met de gegeven kans verworpen kan worden bij een onbetrouwbaarheidsdrempel van 0.05.

Grootte boekhouding	P(H ₀ wordt verworpen)≈0,75	P(H ₀ wordt verworpen)≈0,95	P(H ₀ wordt verworpen)≈0,99
1.000	200	280	320
10.000	600	775	925
100.000	1.850	2.400	2.800
1.000.000	5.800	7.600	8.800

Tabel III: Het aantal gefraudeerde bedragen dat toegevoegd dient te worden aan de boekhouding, waarvan de grootte in de eerste kolom is weergegeven, zodanig dat de nulhypothese met de gegeven kans verworpen kan worden bij een onbetrouwbaarheidsdrempel van 0.01.

Grootte boekhouding	P(H ₀ wordt verworpen)≈0,75	P(H ₀ wordt verworpen)≈0,95	P(H ₀ wordt verworpen)≈0,99
1.000	240	320	360
10.000	725	900	1.025
100.000	2.200	2.750	3.100
1.000.000	7.000	8.600	9.800

¹ In deze tabellen is met andere woorden het onderscheidend vermogen van de gebruikte toets weergegeven als functie van de onbetrouwbaarheidsdrempel, de grootte van de boekhouding (N) en het aantal toegevoegde gefraudeerde bedragen (n). De bedragen in deze tabellen zijn afgerond op vijftallen of tientallen.

SUMMARY

The law of Benford is named after the physicist Frank Benford who published a scientific paper in the year 1938 in which he derived an empirical law that states that the distribution of the first digits of numbers in a particular dataset are not uniformly distributed, as one might expect, but rather follow a logarithmic distribution. According to this law, the probability that a number in a dataset begins with the digit $d \in \{1, 2, \dots, 9\}$ is given by $P(\text{first digit} = d) = \log_{10}(1 + 1/d)$. This implies that numbers with a low first digit (1, 2 or 3) occur more often in datasets compared to numbers with a high first digit. Using this formula, the probability that the first digit of a certain number in a dataset is one is about 0.3 while the probability that the first digit is a nine is approximately equal to 0.046.

However, not every dataset conforms to Benford's law. For example, numbers that are generated through some random process like lottery numbers do not obey to the law of Benford. Theodore Hill, a well known mathematician who performed pioneering research on this topic, proved that the numbers in a dataset comply to Benford's law if the distribution of these numbers is the result of combining different other distributions. This explains why accounting data conforms to Benford's law since they are composed of samples from many distributions. A simple example is an account receivable which is the number of items sold that comes from one distribution multiplied by the price per item, which comes from another distribution.

Benford's law has several applications in practice. The most important one is the detection of fraud in accounting data. The person who has played a major role in popularising the use of this law as a fraud detection tool is Mark Nigrini. He was the first researcher to apply Benford's law extensively to accounting numbers with the goal to detect fraud. In the last twelve years he has published many scientific papers on this topic in which he described numerous tests that can be used to detect fraud in accounting data and which are based on Benford's law. However, what is lacking in these, and other articles related to this topic, is a clear insight in the amount of manipulated numbers that is needed to detect a certain case of fraud when a particular test is applied.

Therefore, the goal of this thesis is to gain insight into the amount of fraudulent numbers that is needed to detect a fraud case when one applies Pearson's χ^2 goodness-of-fit test. This is done by means of an extensive simulation study. The research question to which this thesis is mainly devoted can be formulated as follows: how many fraudulent numbers n do we need to add to an account, which conforms to Benford's law and exists of N numbers, in order to detect a case of fraud with a probability of at least p (for example $p = 0.95$). Hereby it is assumed that the person who committed the fraud spreads the first digits of the n fraudulent numbers evenly across the digits 1, 2, ..., 9.

By means of simulations, attempts were made to answer the research questions mentioned above. These simulations were carried out in the statistical package R. To this end, a function which tests the null hypothesis stating that the distribution of the first digits in an account conform to Benford's law, was programmed. We tested this hypothesis a predefined number of times, each time selecting the first digits of the numbers N and n randomly from the matching distributions, in order to obtain a good estimate of the probability that the null hypothesis will be rejected. An approximation of the desired probability was now given by the number of that times that this null hypothesis was rejected as fraction of the total number of times that this hypothesis was tested. A good insight into the amount of numbers that is needed to detect a fraud case with a certain predefined probability was gained by varying the numbers n and N .

The most important results, which were obtained from the performed simulations, are summarized in the three tables below. These tables show the number of fraudulent numbers that has to be added to an existing account in order to reject the afore mentioned null hypothesis with a certain probability given a predefined confidence level¹. Interestingly, the performed simulations demonstrated that the proportion of fraudulent or manipulated numbers in the total account that is needed to detect the fraud with a certain probability decreases as the size of the initial account increases. Hereby it was observed that a significant deviation from Benford's law was mainly due to a lack of numbers that begin with a one and by a surplus of numbers beginning with a seven, eight or nine, respectively. In addition, the simulations revealed that a high significance level will lead to a lower number of rejections of the null hypothesis compared to a low confidence level. Based on these findings it can be concluded that Benford's law is an easy and useful tool for detecting structural fraud.

Table I: The number of fraudulent numbers that has to be added to an account, with a predefined size, in order to reject the null hypothesis with the mentioned probability, given a confidence level of 0.25.

Size of the accounting	P(H ₀ will be rejected)≈0,75	P(H ₀ will be rejected)≈0,95	P(H ₀ will be rejected)≈0,99
1.000	140	220	260
10.000	425	625	750
100.000	1.350	2.000	2.350
1.000.000	4.200	6.200	7.400

Table II: The number of fraudulent numbers that has to be added to an account, with a predefined size, in order to reject the null hypothesis with the mentioned probability, given a confidence level of 0.05.

Size of the accounting	P(H ₀ will be rejected)≈0,75	P(H ₀ will be rejected)≈0,95	P(H ₀ will be rejected)≈0,99
1.000	200	280	320
10.000	600	775	925
100.000	1.850	2.400	2.800
1.000.000	5.800	7.600	8.800

Table III: The number of fraudulent numbers that has to be added to an account, with a predefined size, in order to reject the null hypothesis with the mentioned probability, given a confidence level of 0.01.

Size of the accounting	P(H ₀ will be rejected)≈0,75	P(H ₀ will be rejected)≈0,95	P(H ₀ will be rejected)≈0,99
1.000	240	320	360
10.000	725	900	1.025
100.000	2.200	2.750	3.100
1.000.000	7.000	8.600	9.800

¹ Stated differently: these tables display the power of the test that is used in the simulations as a function of the confidence level, the size of the accounting (N) and the amount of added fraudulent numbers (n). The numbers in these tables are rounded to the nearest 5 or 10.

INHOUDSOPGAVE

Voorwoord	i
Samenvatting	iii
Summary	v
1 Inleiding	1
2 De wet van Benford	3
2.1 De oorsprong van de wet van Benford	3
2.2 De wiskunde achter de wet van Benford	4
2.3 Datasets die voldoen aan de wet van Benford	7
2.4 De wet van Benford toegepast op een dataset	7
3 Praktische toepassingen van de wet van Benford	11
3.1 Fraude detectie aan de hand van de wet van Benford.....	11
3.2 Overige toepassingen van de wet van Benford.....	16
4 Simulatie	17
4.1 Maximum likelihood-schatter van de multinomiale verdeling.....	17
4.2 De chikwadraat-toets en de likelihood-ratiotoets	20
4.3 Resultaten van de simulaties.....	22
4.4 Samenvatting en conclusies	35
5 Conclusie en discussie	39
Literatuurlijst	41
A Geprogrammeerde functies in R	43
A.1 Test van de begincijfers	43
A.2 Test van de tweede cijfers.....	46
A.3 Test van de eerste twee cijfers	49
A.4 Functies die gebruikt zijn voor het uitvoeren van de simulaties.....	52

HOOFDSTUK 1

INLEIDING

In het jaar 1938 publiceerde Frank Benford, destijds werkzaam als natuurkundige bij General Electric in Amerika, een wetenschappelijk artikel waarin hij, op basis van empirisch bewijs, een wet afleidde die stelt dat de verdeling van de begincijfers van getallen in een dataset logaritmisch verdeeld zijn en geen uniforme verdeling volgen, zoals men intuïtief zou verwachten [2]. Deze wetmatigheid was al eerder ontdekt door de astronoom en wiskundige Simon Newcomb, die dezelfde bevindingen publiceerde in een wetenschappelijk artikel in 1881 [15]. Dit artikel werd echter niet opgemerkt door andere wetenschappers, in tegenstelling tot het artikel van Benford dat veel aandacht verwierf. Aangezien iedereen het artikel van Newcomb reeds was vergeten, kwam de bovenstaande wet bekend te staan als de wet van Benford.

De wet van Benford impliceert dat getallen met een laag begincijfer vaker voorkomen in datasets dan getallen die beginnen met een hoog begincijfer. Deze wet stelt dat de kans dat een willekeurig getal in een dataset begint met het cijfer $d \in \{1, 2, \dots, 9\}$, gelijk is aan $P(\text{begincijfer} = d) = \log_{10}(1 + 1/d)$. Volgens de wet van Benford is de kans dat een willekeurig getal uit een dataset begint met het cijfer 1 ongeveer gelijk aan 0.30 en is de kans op een 9 als begincijfer ongeveer gelijk aan 0.046. Ter vergelijking: er geldt dat $P(\text{begincijfer} = d) = 1/9 \approx 0.11 \forall d = 1, 2, \dots, 9$, indien de begincijfers uniform verdeeld zouden zijn. In figuur 1.1 is ter illustratie de kansverdeling van de begincijfers, zoals deze volgt uit de wet van Benford, weergegeven.

Eén van de belangrijkste toepassingen van de wet van Benford in de praktijk bevindt zich op het gebied van fraudedetectie. Zo wordt deze wet tegenwoordig steeds meer gebruikt voor het opsporen van fraude en onjuistheden in accountingdata van bedrijven. In de praktijk komt het vaak voor dat een werknemer die fraude pleegt dit probeert te maskeren door verzonnen of gemanipuleerde bedragen zodanig te kiezen dat de begincijfers ongeveer even vaak voorkomen. Indien de werknemer veelvuldig bedragen verzint, zal dit er uiteindelijk voor zorgen dat de verdeling van de begincijfers in de boekhouding verschilt van de verdeling die door de wet van Benford wordt voorspeld. Een accountant zal wantrouwend zijn indien hij bijvoorbeeld vaststelt dat 10% van de bedragen in de boekhouding begint met een 9 terwijl dit percentage volgens Benford's wet in de buurt moet liggen van 4.6%. De accountant zal in dit geval de boekhouding zeker nader onderzoeken om vast te stellen wat de oorzaak is van deze afwijking. Hierdoor bestaat er een grote kans dat de fraude aan het licht komt.

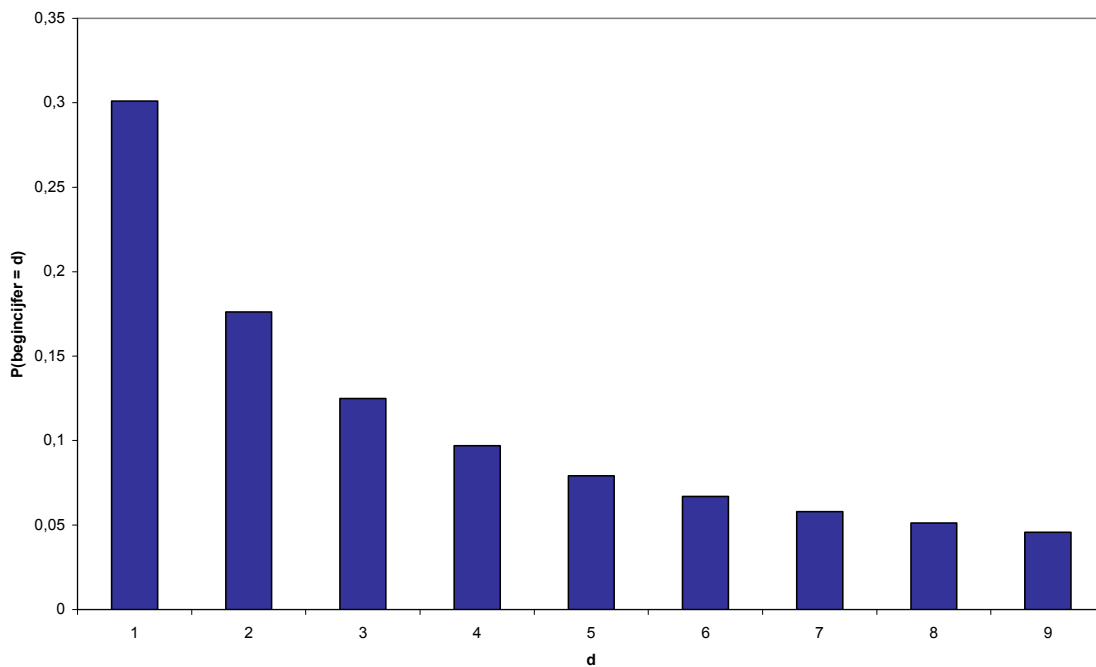
Nigrini was één van de eerste onderzoekers die de wet van Benford succesvol toepaste op accountingdata en kan op dit gebied gezien worden als een echte pionier. Hij heeft meerdere wetenschappelijke artikelen geschreven waarin hij verschillende testen beschrijft die gebaseerd zijn op de wet van Benford. Een groot deel van deze testen heeft hij zelf ontwikkeld en kunnen gebruikt worden om fraude in financiële overzichten, zoals boekhoudingen van bedrijven, mee op te sporen [zie 8, 16, 17 en 18]. Wat wij echter missen in deze, en andere wetenschappelijke artikelen op dit gebied, is een duidelijk inzicht in het aantal gemanipuleerde bedragen dat nodig is om een bepaald geval van fraude te ontdekken. Het doel van dit werkstuk is dan ook om dit inzicht, aan de hand van een uitgebreide simulatiestudie, te verschaffen. Dit inzicht kan een belangrijke rol spelen in het bepalen of het gebruik van een test die gebaseerd is op de wet van Benford in bepaalde situaties zinvol is.

De onderzoeksvraag waaraan dit werkstuk grotendeels is gewijd kan als volgt worden geformuleerd: stel dat de begincijfers van een boekhouding van een bepaald bedrijf of bedrijfs onderdeel, welke bestaat uit N bedragen, voldoen aan de wet van Benford. Stel nu dat een kwaadwillend persoon binnen het bedrijf snel rijk wil worden. Hij verzint een aantal van n bedragen die hij door middel van een fictief bedrijf overmaakt op zijn eigen privérekening. Wij nemen aan dat deze persoon niet op de

hoogte is van de wet van Benford. De begincijfers van de bedragen die hij verzint verdeelt hij uniform over de cijfers 1,2,...,9. De gefraudeerde bedragen komen uiteindelijk terecht in de boekhouding.

De vraag is nu hoeveel gefraudeerde bedragen n er aan een gegeven boekhouding, die bestaat uit N bedragen, toegevoegd dienen te worden voordat de fraudeur met een kans van tenminste p (bijvoorbeeld $p=0.95$) gepakt kan worden. Deze onderzoeksvraag hebben wij beantwoord door middel van simulaties die gebaseerd zijn op de chi-kwadraattoets.

Het vervolg van dit werkstuk is als volgt ingedeeld: in het volgende hoofdstuk wordt de wet van Benford nader bestudeerd. Hierbij komen de wiskundige eigenschappen van deze wet uitgebreid aan de orde. Daarnaast wordt in dit hoofdstuk nagegaan of een zelf verzamelde dataset voldoet aan de wet van Benford. Daaropvolgend worden in het derde hoofdstuk de veelvuldige toepassingen van de wet van Benford in de praktijk beschreven. Hierbij wordt uitgebreid stilgestaan bij het gebruik van de wet van Benford als fraudedetectietechniek. In hoofdstuk vier worden vervolgens de resultaten van de uitgevoerde simulaties beschreven. De opzet van de simulaties en de gebruikte statistische toetsen worden hierbij uitgebreid besproken. In het laatste hoofdstuk komen de conclusies aan bod. Ten slotte kan men in de bijlagen de broncode vinden van diverse functies die geprogrammeerd zijn in het statistische pakket R. Deze functies zijn onder andere gebruikt voor het uitvoeren van de simulaties.



Figuur 1.1: Histogram van de kansverdeling van de begincijfers volgens de wet van Benford.

HOOFDSTUK 2

DE WET VAN BENFORD

Stel dat wij een willekeurige editie van het gratis dagblad Spits pakken en wij verzamelen alle getallen die daarin gepubliceerd zijn. Hoeveel procent van deze getallen zal beginnen met het cijfer 1? Verrassend genoeg is het antwoord ruim 30%. Dit percentage wordt voorspeld door de wet van Benford. In dit hoofdstuk zullen wij de wet van Benford nader bestuderen. Als eerst zullen wij de oorsprong van de wet van Benford beschrijven en zullen wij dieper ingaan op de wiskunde achter deze wet. Daarnaast zal aangegeven worden welke datasets wel en welke niet voldoen aan de wet van Benford. In het laatste gedeelte van dit hoofdstuk zal de wet van Benford toegepast worden op een dataset die wij zelf samengesteld hebben.

2.1 DE OORSPRONG VAN DE WET VAN BENFORD

In 1881 ontdekte de astronoom en wiskundige, Simon Newcomb, dat de eerste pagina's van de boeken met logaritmetabellen, die zich bevonden in de bibliotheek, veel vuiler en meer versleten waren dan de laatste pagina's [9, 15]. Op de eerste pagina's van deze boeken kon men de logaritmen vinden van getallen met lage begincijfers en op de laatste pagina's die van getallen met hoge begincijfers. Newcomb leidde op basis van zijn observaties af dat deze boeken vaker gebruikt werden om logaritmen op te zoeken van getallen die beginnen met een 1 dan getallen die beginnen met een 2, een 3, enzovoort. Hij concludeerde daarom dat er meer getallen bestaan die met het cijfer 1 beginnen dan getallen die beginnen met een hoger cijfer. In zijn artikel toonde Newcomb op basis van een heuristiek aan, dat de kans dat een willekeurig getal uit een dataset met het cijfer d begint, gelijk is aan:

$$P(\text{begincijfer} = d) = \log_{10}(1 + 1/d), \quad d = 1, 2, \dots, 9 \quad (2.1)$$

Volgens deze formule is de kans dat een willekeurig getal uit een dataset begint met het cijfer 1 ongeveer gelijk aan 0.30 en is de kans op een 9 als begincijfer ongeveer gelijk aan 0.046. Dit is opmerkelijk aangezien men eerder geneigd is om aan te nemen dat alle begincijfers even vaak voorkomen. Newcomb toonde daarnaast ook aan dat de kans dat het tweede cijfer van een willekeurig getal gelijk is aan d , onafhankelijk van het begincijfer, gelijk is aan:

$$P(\text{tweede cijfer} = d) = \sum_{m=1}^9 \log_{10}(1 + (10m + d)^{-1}), \quad d = 0, 1, \dots, 9 \quad (2.2)$$

Hill geeft in [10] een definitie van de algemene formule die betrekking heeft op alle positieve getallen: voor alle positieve integers k , alle $d_i \in \{1, 2, \dots, 9\}$ en alle $d_j \in \{0, 1, \dots, 9\}$, $j = 2, \dots, k$, geldt dat voor alle getallen die uit ten minste k cijfers bestaan:

$$P(D_1 = d_1, \dots, D_k = d_k) = \log_{10} \left[1 + \left(\sum_{i=1}^k d_i \times 10^{k-i} \right)^{-1} \right]. \quad (2.3)$$

De formule in (2.3) geeft dus de kans weer op het getal $(d_1 \times 10^{k-1} + d_2 \times 10^{k-2} + \dots + d_{k-1} \times 10 + d_k)$.

In tabel 2.1 zijn de kansen $P(D_i = d_i)$ voor $i = 1, \dots, 4$ weergegeven. Op basis van deze tabel kan men afleiden dat de verdeling van het tweede cijfer minder scheef is dan de verdeling van het eerste cijfer.

Daarnaast blijkt uit deze tabel dat de verdeling van het cijfer op een bepaalde positie in een getal convergeert naar de uniforme verdeling naarmate deze positie steeds verder naar rechts verschuift. Newcomb gaf geen theoretische verklaring voor het fenomeen dat hij beschreef. Zijn artikel werd niet door het grote publiek opgemerkt. Zevenenvijftig jaar later, in het jaar 1938, ontdekte de natuurkundige Frank Benford die werkzaam was bij General Electric, onafhankelijk van het artikel van Newcomb, precies hetzelfde patroon in de logaritmeboeken [2]. In zijn artikel dat verscheen in de *Proceedings of the American Philosophical Society* kwam hij tot dezelfde conclusies als Newcomb. Benford leverde echter, in tegenstelling tot Newcomb, wel empirisch bewijs voor het fenomeen dat hij beschreef. Hiervoor verzamelde en onderzocht Benford 20.229 observaties afkomstig uit diverse bronnen. Zo bestond een deel van de data van Benford uit de lengte van diverse rivieren, honkbal scores, het atomische gewicht van diverse elementen en getallen die afkomstig zijn uit artikelen die gepubliceerd zijn in het tijdschrift *Reader's Digest*. De observaties waren in totaal verspreid over 20 verschillende datasets. Voor elke afzonderlijke dataset bepaalde hij de fractie getallen waarvan het begincijfer gelijk was aan $1, 2, \dots, 9$. Alhoewel sommige datasets afzonderlijk niet aan (2.1) voldeden, voldeden de gemiddelde fracties berekend over alle 20 datasets daarentegen uitzonderlijk goed aan deze wet. Benford maakte hiermee aannemelijk dat een dataset die samengesteld is uit verschillende datasets (welke mogelijk afzonderlijk afwijken van (2.1)) aan formule (2.1) voldoet.

Diaconis en Freedman hadden wel enige kritiek op de werkwijze van Benford [7]. In het artikel dat verscheen in de *Journal of the American Statistical Association* leverde zij overtuigend bewijs dat aantoonde dat Benford de afrondfouten manipuleerde om zo de verzamelde data beter aan formule (2.1) te laten voldoen. Hill toont in [10] echter aan dat zelfs de onbewerkte data uitzonderlijk goed aan (2.1) voldoet.

Het artikel van Benford kreeg veel aandacht mede doordat het in hetzelfde nummer verscheen als het kritische artikel van Hans Bethe over de verspreiding van elektronen [3]. Aangezien iedereen het artikel van Newcomb vergeten was, kwam formule (2.1) bekend te staan als de wet van Benford.

Cijfer	Positie in het getal			
	1 ^e positie	2 ^e positie	3 ^e positie	4 ^e positie
0		0,11968	0,10178	0,10018
1	0,30103	0,11389	0,10138	0,10014
2	0,17609	0,10882	0,10097	0,10010
3	0,12494	0,10433	0,10057	0,10006
4	0,09691	0,10031	0,10018	0,10002
5	0,07918	0,09668	0,09979	0,09998
6	0,06695	0,09337	0,09940	0,09994
7	0,05799	0,09035	0,09902	0,09990
8	0,05115	0,08757	0,09864	0,09986
9	0,04576	0,08500	0,09827	0,09982

Tabel 2.1: De kansen op een bepaald cijfer op de eerste vier posities van een willekeurig getal volgens de wet van Benford.

2.2 DE WISKUNDE ACHTER DE WET VAN BENFORD

Een afleiding van formule (2.1) kan gevonden worden in [23]. Elk positief getal x kan in de volgende vorm geschreven worden: $x = m \times 10^k$ met $1 \leq m < 10$ en $k \in \mathbb{Z}$. Hierbij stelt m de mantisse van $x = m \times 10^k$ voor. Het begincijfer van x is nu gelijk aan het begincijfer van de mantisse m . Laat D het

begincijfer van een willekeurig getal $X = M \times 10^K$ zijn en veronderstel nu dat de logaritme van de mantisse, $Y := \log_{10} M$, uniform verdeeld is op het interval $[0,1]$.

Nu geldt dat

$$\begin{aligned} P(D = d) &= P(d \leq M < d+1) \\ &= P(\log_{10}(d) \leq Y < \log_{10}(d+1)) \\ &= \log_{10}(d+1) - \log_{10}(d) \\ &= \log_{10}\left(\frac{d+1}{d}\right) \\ &= \log_{10}(1 + 1/d), \text{ voor } d = 1, 2, \dots, 9. \end{aligned}$$

Het is niet zo dat alleen datasets waarvan de logaritme van de mantissen uniform verdeeld is op $[0,1]$ voldoen aan de wet van Benford. Ook datasets waarvan de logaritmen van de mantissen niet gezien kunnen worden als realisaties uit de standaard homogene verdeling kunnen aan de wet van Benford voldoen.

Stel dat een dataset aan de wet van Benford voldoet. Men zou dan intuïtief verwachten dat de cijfers in een willekeurig getal uit deze dataset onafhankelijk van elkaar zijn. Een simpel voorbeeld laat echter zien dat dit niet het geval is. De kans dat het tweede cijfer D_2 gelijk is aan d_2 , gegeven het feit dat het eerste cijfer D_1 gelijk is aan d_1 , kan op de volgende manier afgeleid worden:

$$P(D_2 = d_2 | D_1 = d_1) = \frac{P(D_1 = d_1 \wedge D_2 = d_2)}{P(D_1 = d_1)} = \frac{\log_{10}(1 + 1/(10 \times d_1 + d_2))}{\log_{10}(1 + 1/d_1)}, \quad (2.4)$$

waarbij gebruik is gemaakt van (2.1) en (2.2). Volgens (2.4) is de voorwaardelijke kans dat het tweede cijfer gelijk is aan twee, gegeven het feit dat het begincijfer een 1 is, gelijk aan: $P(D_2 = 2 | D_1 = 1) = \log_{10}(13/12) / \log_{10}(2) \approx 0.11548$. Uit (2.2) volgt echter dat de kans op een 2 als tweede cijfer gelijk is aan 0.10882, onafhankelijk van het begincijfer. Aangezien deze twee kansen niet gelijk zijn, volgt hieruit dat de cijfers in een getal niet onafhankelijk zijn van elkaar.

De afgelopen zestig jaar hebben wiskundige, statistici, natuurkundigen, en zelfs amateurs getracht de wet van Benford (wiskundig) te bewijzen, maar men stuitte hierbij echter op twee grote struikelblokken [10, 12]. Het eerste, relatief simpele probleem, is dat het niet duidelijk was welke datasets wel aan de wet van Benford voldeden en welke niet. Er was geen statistische test voorhanden waarmee dit vastgesteld kon worden. De onderzoekers gooiden het echter over een andere boeg. Zij waren niet zozeer geïnteresseerd in de vraag waarom een bepaalde dataset wel of niet aan de wet van Benford voldoet, maar zij probeerden daarentegen te bewijzen dat de wet van Benford een ingebouwde karakteristiek is van ons getallensysteem. De onderzoekers trachtten te bewijzen dat alle getallen in ons getallensysteem voldoen aan de wet van Benford en wilden aan de hand hiervan een verklaring geven voor het empirische bewijs dat Benford verzamelde. Geen van de onderzoekers slaagde er echter in om een sluitend bewijs te geven.

Pinkham [19] slaagde er wel in om een belangrijke stelling te bewijzen die zorgde voor een nieuwe invalshoek in het onderzoek naar de wet van Benford. Pinkham redeneerde hierbij als volgt [12, 14, 17 en 19]: stel dat de cijfers van de lengte van alle rivieren over de gehele wereld voldoen aan de wet van Benford. Volgens Pinkham zou het dan niet mogen uitmaken of deze lengte wordt uitgedrukt in mijlen of in kilometers, in beide gevallen zouden de begincijfers aan de wet van Benford moeten voldoen. Met andere woorden: indien een verzameling van getallen voldoet aan de wet van Benford en deze verzameling wordt vermenigvuldigd met een constante groter dan nul, dan voldoet de nieuwe

verzameling getallen nog steeds aan de wet van Benford. Deze stelling wordt de stelling van schaal-invariantie genoemd.

Wiskundig kan de stelling van schaal-invariantie als volgt geformuleerd worden: laat A_p de verzameling van getallen zijn waarvan het begincijfer kleiner of gelijk is aan p , dan zou moeten gelden:

$$P(x \in A_p) = P(x \in kA_p), \quad \forall k > 0. \quad (2.5)$$

Pinkham bewees dat, onder de aanname dat de verdelingsfunctie van de onderliggende data continu is, de volgende relatie geldig is: als vergelijking (2.5) geldt, dan geldt ook:

$$P(x \in A_p) = \log_{10}(p+1), \quad \text{voor } p = 1, 2, \dots, 9, \quad (2.6)$$

wat op het hetzelfde neerkomt als de wet van Benford. Pinkham bewees daarnaast dat alleen de wet van Benford schaal-invariant is.

Alhoewel men er in geslaagd was om aan te tonen dat de wet van Benford schaal-invariant is, was er nog een groot struikelblok. De wiskundige bewijzen waren niet zo rigoureuus wat de huidige standaarden uit de kansrekening betreft. De huidige theorie uit de kansrekening gaat uit van het “countable additivity” axioma. Dit axioma stelt dat indien een willekeurig geheel getal wordt gekozen en $P(1)$ is de kans dat het getal 1 wordt gekozen, $P(50)$ de kans dat het getal 50 gekozen wordt enzovoort, dan moet gelden dat:

$$P(1) + P(2) + P(3) + \dots = 1. \quad (2.7)$$

De bewijzen die waren gepubliceerd voor het jaar 1995 voldeden niet aan dit axioma. Het was Theodore Hill die in 1995 voor het eerst bewijs leverde dat wel aan (2.7) voldeed [10, 12]. Hill borduurde voort op het idee van Pinkham en introduceerde de grondgetal-invariantie van de wet van Benford. De grondgetal-invariantie houdt in dat de wet van Benford geldt ongeacht het grondgetal dat gebruikt wordt voor de logaritme. Hill ontdekte dat het bovenstaande probleem geëlimineerd wordt indien gekeken wordt naar equivalente verzamelingen van getallen met hetzelfde begincijfer in plaats van naar afzonderlijke getallen. Dit stelde Hill in staat om te bewijzen dat de wet van Benford de enige kansverdeling is die zowel schaal- als grondgetal-invariant is.

Alhoewel men nu iets meer wist over de eigenschappen van de wet van Benford, kon men nog steeds niet verklaren waarom sommige datasets wel aan de wet van Benford voldeden en waarom andere datasets dat niet deden. Het was wederom Hill [10] die hier een verklaring voor gaf. Zijn verklaring is gebaseerd op het feit dat de verdeling van getallen die afkomstig zijn uit verzamelingen die voldoen aan de wet van Benford tweede generatie verdelingen zijn. Tweede generatie verdelingen zijn verdelingen die opgebouwd zijn uit combinaties van andere verdelingen. Hill bewees dat indien oneindig veel verschillende verdelingen random geselecteerd worden en er willekeurig steekproeven uit deze verdelingen geselecteerd en samengevoegd worden, de verdeling van de begincijfers van de getallen in de samengestelde steekproef zal convergeren naar de verdeling van Benford ongeacht of de geselecteerde verdelingen afzonderlijk aan de verdeling van Benford voldoen.

De sleutel ligt dus in het combineren van data uit verschillende bronnen. Dit verklaart ook gelijk waarom het gemiddelde dat Benford in zijn artikel had berekend voldeed aan de wet van Benford. Dit gemiddelde was, zoals eerder aangegeven, samengesteld uit 20 verschillende bronnen. Het belangrijkste punt van de stelling van Hill is dat er meerdere steekproefprocedures zijn die leiden naar de wet van Benford, wat verklaart waarom zoveel verschillende datasets voldoen aan de wet van Benford.

Een intuïtieve verklaring voor de wet van Benford die makkelijk te volgen is, kan gevonden worden indien men bijvoorbeeld een onderneming beschouwt met een marktwaarde van 1 miljoen euro [8]. Stel dat deze marktwaarde met 5% groeit per jaar. Het duurt dan ruim 14 jaar voordat de marktwaarde met 100% gegroeid is (tot 2 miljoen). Gedurende deze 14 jaar is het begincijfer van de marktwaarde steeds gelijk aan 1. Indien de marktwaarde eenmaal gelijk is aan 2 miljoen is een stijging van 50% voldoende om de marktwaarde te laten stijgen naar 3 miljoen en het begincijfer 2 te laten veranderen in een 3. Dit duurt ruim 8 jaar. Wanneer de marktwaarde gelijk is aan 9 miljoen is een groei van 11.1% voldoende om het begincijfer van een 9 te laten veranderen in een 1 (en de marktwaarde dus te laten groeien van 9 naar 10 miljoen). Hiervoor zijn iets meer dan 2 jaar nodig. Het begincijfer van de marktwaarde zal nu weer gelijk zijn aan 1 totdat de marktwaarde weer groeit met 100% (wat weer ruim 14 jaar duurt). Hieruit blijkt dus dat het begincijfer van de marktwaarde veel vaker gelijk zal zijn aan 1 dan alle andere begincijfers. Een zelfde verklaring kan gevonden worden in [17], waar in plaats van de marktwaarde van een onderneming de populatie van een stad wordt beschouwd.

2.3 DATASETS DIE VOLDOEN AAN DE WET VAN BENFORD

De afgelopen jaren zijn veel datasets onderzocht met als doel om vast te stellen of zij aan de wet van Benford voldoen. Zo is bijvoorbeeld aannemelijk gemaakt dat de lengte van rivieren, de populaties van steden, het inkomen van huishoudens, de bedragen op energierekeningen, de marktwaarde en het inkomen van bedrijven die genoteerd zijn aan een index allemaal (bij benadering) voldoen aan de wet van Benford [4]. Daarnaast is bewezen dat Fibonacci en Lucas getallen aan de wet van Benford voldoen. Ook de twintig meest gebruikte constanten in de natuurkunde voldoen bij benadering aan de wet van Benford. Dit is maar een kleine greep uit de grote verzameling van datasets die voldoen aan de wet van Benford.

Natuurlijk voldoen niet alle datasets aan de wet van Benford. Zo geldt bijvoorbeeld dat getallen die gegenereerd zijn aan de hand van een random proces (zoals loterijgetallen of dobbelsteenworpen) niet voldoen. Ook mogen de getallen niet verzonden of toegekend zijn door derden. Telefoonnummers, rekeningnummers en postcodes voldoen om deze reden niet.

Drake en Nigrini [8] geven een lijst met drie criteria waaraan een dataset moet voldoen wil deze in aanmerking komen om aan de wet van Benford te voldoen. Deze criteria zijn als volgt gedefinieerd:

1. De data moet hetzelfde fenomeen beschrijven, dit betekent dat alle bedragen in een dataset bijvoorbeeld in dezelfde valuta weergegeven dienen te worden (euro's of dollars).
2. De getallen in de dataset mogen geen ingebouwd minimum of maximum hebben.
3. De getallen in een dataset mogen niet toegekend zijn (door derden).

Daarnaast geldt dat de wet van Benford het best toegepast kan worden op grote datasets die minstens 10.000 observaties bevatten.

2.4 DE WET VAN BENFORD TOEGEPAST OP EEN DATASET

De wet van Benford klinkt, wanneer men er voor het eerst mee geconfronteerd wordt, ongeloofwaardig. Om aan te tonen dat deze wet geen klinkklare onzin is, hebben wij een dataset samengesteld met daarin de inwoneraantallen van alle landen in de wereld². In tabel 2.2 zijn enkele beschrijvende statistieken van deze dataset weergegeven. De dataset bestaat, zoals blijkt uit tabel 2.2, uit 236 inwoneraantallen. Het land met de kleinste populatie is Pitcairn, dat slechts 50 inwoners telt. China is met ruim 1.3 miljard inwoners het land met de grootste populatie in de gehele wereld, op de

² Bron: http://en.wikipedia.org/wiki/List_of_countries_by_population

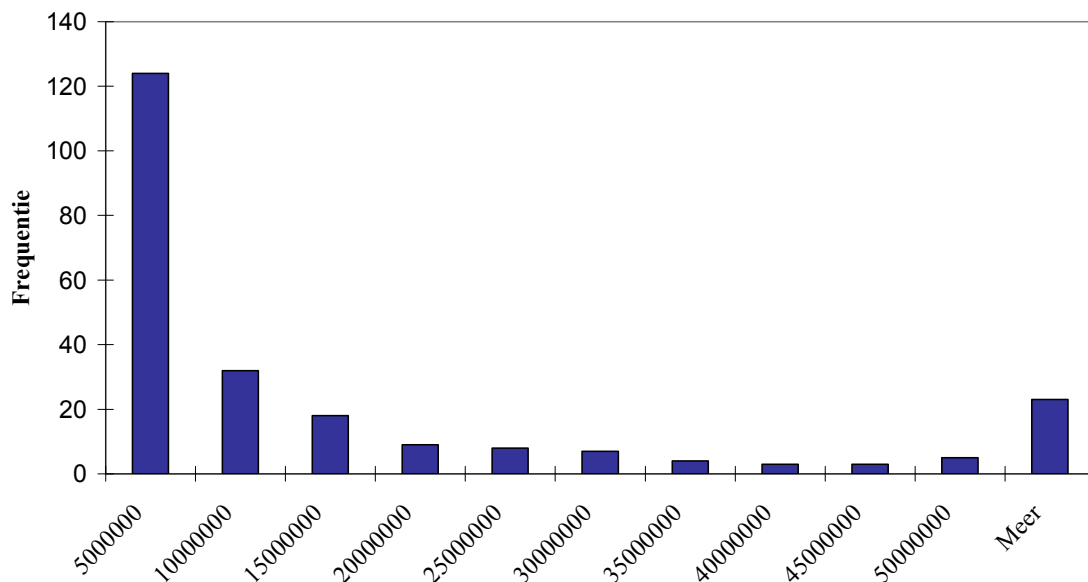
voet gevolgd door India met een populatiegrootte van ruim 1.1 miljard. Nederland staat met zijn 16.5 miljoen inwoners op de 61^e plaats.

Tabel 2.2: Enkele numerieke grootheden voor de samengestelde dataset.

n	236
Minimum	50
Maximum	1,323,040,000
Gemiddelde	28,132,192
Mediaan	4,431,500
Standaardafwijking	117,578,516
Scheefheid	9,38

In figuur 2.1 is een histogram weergegeven van de inwoneraantallen. Uit dit histogram blijkt dat de verdeling van de inwoneraantallen rechtsscheef is, wat bevestigd wordt door tabel 2.2 waaruit blijkt dat de mediaan veel kleiner is dan het gemiddelde van de inwoneraantallen. De reden hiervoor is dat de meeste landen een populatiegrootte hebben die niet groter is dan 5 miljoen.

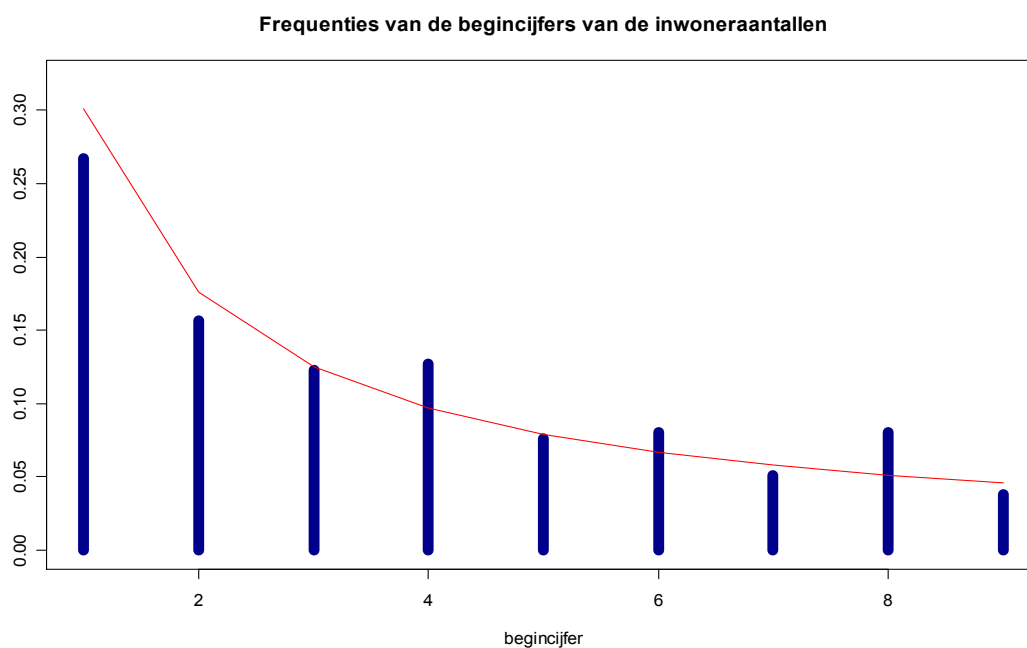
Histogram van de inwoneraantallen van alle landen in de wereld



Figuur 2.1: Histogram van de populatiegroottes van alle landen in de wereld.

Wallace [25] suggereert in zijn artikel dat de kans groot is dat een verzameling getallen aan de wet van Benford voldoet indien het gemiddelde van de verzameling getallen groter is dan de mediaan en de scheefheid positief is. Hij beweert dat des te groter het verschil tussen het gemiddelde en de mediaan is, des te beter de verzameling getallen aan de wet van Benford zal voldoen. De reden hiervoor is dat datasets die aan de wet van Benford voldoen een voorkeur hebben voor kleine getallen. Aangezien onze dataset aan de bovenstaande criteria voldoet, verwachten wij dat deze dataset aan de wet van Benford zal voldoen.

In figuur 2.2 is een histogram afgebeeld van de frequenties van de begincijfers van de inwoneraantallen samen met de frequenties die voorspeld worden door de wet van Benford³. Op basis van deze figuur kunnen wij concluderen dat de inwoneraantallen op het eerste oog voldoen aan de wet van Benford. Er is wel een redelijk grote afwijking waarneembaar voor het begincijfer 8. Wij hebben echter geen statistische test voorhanden waarmee wij kunnen vaststellen of een dataset aan de wet van Benford voldoet en waarmee wij kunnen beoordelen hoe sterk het bewijs is. Deze test zullen wij in hoofdstuk 4 beschrijven.



Figuur 2.2: Histogram van de frequenties van de begincijfers 1 tot en met 9 van de dataset met de inwoneraantallen van alle landen in de wereld. De rode lijn geeft de verwachte frequenties op basis van de wet van Benford weer.

³ Dit histogram is gecreëerd aan de hand van een functie die geprogrammeerd is in het statistische pakket R. De broncode van deze functie kan men in bijlage A.1 terugvinden.

HOOFDSTUK 3

PRAKTISCHE TOEPASSINGEN VAN DE WET VAN BENFORD

De wet van Benford werd na zijn ontdekking gezien als een wiskundig fenomeen waarvan men aanvankelijk dacht dat het geen meerwaarde zou hebben in de praktijk. Dit is niet waar gebleken. De afgelopen jaren kent de wet van Benford juist steeds meer toepassingen in de praktijk. In dit hoofdstuk zullen wij deze toepassingen nader beschrijven. In de eerste paragraaf zullen wij beschrijven hoe de wet van Benford wordt gebruikt door accountants om fraude te ontdekken in boekhoudingen. In de daaropvolgende paragraaf zullen wij andere interessante applicaties van de wet van Benford beschrijven.

3.1 FRAUDE DETECTIE AAN DE HAND VAN DE WET VAN BENFORD

Varian, een econoom, suggereerde als eerste dat de wet van Benford gebruikt kon worden als een test om de geldigheid van wetenschappelijke data te testen [24]. Het duurde echter tot het eind van de jaren tachtig voordat dit idee door accountants werd opgepikt. Carslaw paste als eerste de wet van Benford toe in een accountingcontext [5]. Hij gebruikte deze wet om aan te tonen dat managers van bedrijven in Nieuw Zeeland de netto winst opzettelijk omhoog afronden omdat zij hier baat bij hadden. Carslaw suggereerde dat indien er sprake was van opzettelijk omhoog afronden dit tot uiting zou moeten komen in een overschot aan getallen waarvan het tweede cijfer gelijk is aan 0 en een tekort aan getallen met als tweede cijfer een 9. Een netto winstcijfer van €290.000 zou bijvoorbeeld afgerond worden op €300.000. Alhoewel het procentuele verschil tussen deze twee getallen erg klein is, lijkt het bedrag van €300.000 gevoelsmatig toch veel groter. Carslaw toonde aan dat de verdeling van de tweede cijfers van de netto winstcijfers van bedrijven in Nieuw Zeeland niet aan de wet van Benford voldeden. Hij toonde aan dat er daadwerkelijk sprake was van een tekort aan negens en een overschot aan nullen als tweede cijfer. Thomas ontdekte een soortgelijk patroon in de netto winstcijfers van Amerikaanse bedrijven [22].

Nigrini was de eerste onderzoeker die de wet van Benford toepaste als fraudedetectietechniek en kan op dit gebied gezien worden als een echte pionier. Hij gebruikte de wet om een methode te ontwikkelen waarmee nagegaan kan worden of belastingbetalers fraude hebben gepleegd bij het invullen van het belastingaangifteformulier [16]. Daarnaast toonde hij aan dat belastingbetalers met een laag inkomen eerder geneigd waren om verzonnen getallen op te geven in hun belastingaangifte.

De reden waarom de wet van Benford kan helpen bij het opsporen van fraude is dat het aannemelijk is dat accountingdata (bij benadering) voldoet aan de wet van Benford. Accountinggetallen zijn meestal een combinatie van verschillende andere getallen, die alle een verschillende verdeling hebben. Zo geldt bijvoorbeeld dat het bedrag aan debiteuren gelijk is aan het aantal items dat men verkocht heeft vermenigvuldigd met de prijs per item.

Nigrini en anderen [8, 9, 16, 17 en 18] hebben verschillende artikelen geschreven omtrent het ontdekken van fraude in accountingdata op basis van de wet van Benford. In deze artikelen wordt gedetailleerd aangegeven welke stappen ondernomen moeten worden om een goede analyse van de data uit te voeren. Zo moet men volgens Drake en Nigrini [8] eerst nagaan of aan een aantal randvoorwaarden wordt voldaan voordat men overgaat tot het analyseren van de begincijfers van de accountingdata op afwijkingen van de wet van Benford. Zo dient als eerst nagegaan te worden of de data een gehele tijdsperiode omspannen zoals een maand, een kwartaal of een fiscaal jaar. Daarnaast dient de data per bedrijfs onderdeel bestudeerd te worden. De reden hiervoor is dat afwijkingen in de

data verloren kunnen gaan indien data die afkomstig is van verschillende ongerelateerde bedrijfsonderdelen wordt samengevoegd. De derde voorwaarde is dat de data op een zo'n specifiek mogelijk niveau wordt geanalyseerd. De achterliggende reden hiervan is dat gemanipuleerde bedragen mogelijk niet opgemerkt worden indien bepaalde bedragen gesommeerd worden. Zo kan het bijvoorbeeld voorkomen dat de som van de reiskostenvergoeding en een algemene onkostenvergoeding wel voldoet aan de wet van Benford, waardoor gemanipuleerde bedragen die zich in één van de afzonderlijke vergoedingen bevinden mogelijk onopgemerkt blijven. De kans op het ontdekken van deze gemanipuleerde bedragen is veel groter indien beide vergoedingen apart onderzocht zouden worden.

Als laatst dient men bedragen die kleiner zijn dan €10 (of \$10) en negatieve getallen uit de data te verwijderen voordat men deze gaat analyseren. De bedragen die kleiner zijn dan €10 worden verwijderd zodat men verzekerd is van het feit dat alle bedragen een eerste en tweede cijfer hebben. Hiermee wordt dus voorkomen dat gehele getallen zoals 1,2,...,9, die geen tweede cijfer hebben, in de data voorkomen. Deze cijfers lenen zich namelijk niet voor een test van de tweede cijfers die in het vervolg van deze paragraaf beschreven zal worden. Een andere reden is dat in sommige computerprogramma's gehele getallen worden voorzien van een extra nul. Een getal als 5 wordt dan bijvoorbeeld weergegeven als 5.0. Dit zorgt dus voor een overschot aan nullen als tweede cijfer wat een negatieve invloed heeft op de resultaten van de testen die hierna beschreven zullen worden. Negatieve getallen worden meestal apart onderzocht omdat deze onderhevig zijn aan andere type fouten.

Drake en Nigrini stellen drie verschillende testen voor om een goed beeld van de data te krijgen en de kans op het ontdekken van fraude te vergroten. In deze testen wordt gebruik gemaakt van de gemiddelde absolute afwijking als grootte om vast te stellen of een willekeurige verzameling cijfers aan de wet van Benford voldoet. De gemiddelde absolute afwijking wordt in het geval van de 9 begincijfers gevonden door de absolute waarde van de afwijkingen van de gevonden proporties ten op zichte van de voorspelde proporties (volgens de wet van Benford) te sommeren en deze som te delen door 9:

$$\text{Gemiddelde absolute afwijking} = \frac{1}{9} \sum_{d=1}^9 |\text{gevonden proportie cijfer } d - \text{verwachte proportie}| \quad (3.1)$$

Wij zullen de drie verschillende testen hierna kort beschrijven.

Richtlijnen voor de test van de begincijfers

Als eerst dient onderzocht te worden of de begincijfers van de getallen die afkomstig zijn uit de verzamelde accountingdata aan de wet van Benford voldoen. Deze test geeft een eerste indruk en kan gebruikt worden om vast te stellen of de data afwijkingen bevat die nader onderzoek vereisen. Uitschieters in de vorm van grote pieken in het histogram met de gevonden proporties (zie bijvoorbeeld figuur 2.2) zijn een indicatie hiervan. Drake en Nigrini hebben daarnaast, op basis van hun ervaring, de volgende richtlijnen vastgesteld met betrekking tot de gemiddelde absolute afwijking:

- Er sprake van sterke conformiteit aan de wet van Benford indien de gemiddelde absolute afwijking zich bevindt in het interval [0.000, 0.004).
- Er sprake van accepteerbare conformiteit indien de gemiddelde absolute afwijking zich bevindt in het interval [0.004, 0.008).
- Er sprake van marginaal accepteerbare conformiteit indien de gemiddelde absolute afwijking zich bevindt in het interval [0.008, 0.012).
- De begincijfers voldoen niet aan de wet van Benford indien de gemiddelde absolute afwijking groter is dan 0.012.

Daarnaast geven zij aan dat verwacht mag worden dat de gevonden proporties in grote datasets met zeer diverse data, die een grote tijdsspanne beslaan, dichtbij de verwachte proporties zullen liggen die voorspeld worden door de wet van Benford.

Richtlijnen voor de test van de tweede cijfers

Nadat de begincijfers van de verzameling getallen zijn onderzocht dient men ook de tweede cijfers van deze getallen te onderzoeken. Deze test is ook bedoeld om te bepalen of de data afwijkingen bevat en om vast te stellen waar deze afwijkingen zich precies bevinden. Drake en Nigrini merken op dat de bedragen die vermeld worden op facturen in Amerika en Canada meer nullen en vijven als tweede cijfers bevatten dan men zou verwachten op basis van de wet van Benford. De reden hiervoor is dat mensen geneigd zijn om factuurbedragen af te ronden op gehele cijfers wat tot uiting komt in een overschot aan de cijfers 0 en 5 als tweede cijfer. Voorbeelden hiervan zijn de getallen €199.99 en €149.99 die mogelijk worden afgerond tot €200 en €150, respectievelijk. Dit geldt mogelijk ook voor factuurbedragen in Nederland alhoewel wij hier niks over hebben kunnen vinden in de literatuur.

Drake en Nigrini bieden de volgende richtlijnen in het geval van de tweede cijfers:

- Er sprake van sterke conformiteit aan de wet van Benford indien de gemiddelde absolute afwijking zich bevindt in het interval $[0.000, 0.008)$.
- Er sprake van accepteerbare conformiteit indien de gemiddelde absolute afwijking zich bevindt in het interval $[0.008, 0.012)$.
- Er sprake van marginaal accepteerbare conformiteit indien de gemiddelde absolute afwijking zich bevindt in het interval $[0.012, 0.016)$.
- De begincijfers voldoen niet aan de wet van Benford indien de gemiddelde absolute afwijking groter is dan 0.016.

Wederom geldt dat verwacht mag worden dat de gevonden proporties in grote datasets met zeer diverse data, die een grote tijdsspanne beslaan, dichtbij de verwachte proporties zullen liggen die voorspeld worden door de wet van Benford.

Richtlijnen voor de test van de eerste twee cijfers

Er zijn 90 verschillende combinaties mogelijk van de eerste twee cijfers, namelijk 10, 11, ..., 98, 99. Het doel van deze test is om die combinaties te identificeren waarvan de waargenomen proporties veel groter zijn dan de verwachte proporties. Dit zal resulteren in grote pieken in het histogram waarin de proporties van alle negentig combinaties zijn uitgezet. De getallen die deze combinaties als eerste cijfers hebben dienen dan nader onderzocht te worden door de accountant.

De accountant dient ook te onderzoeken of er pieken zichtbaar zijn bij getallen die psychologische barrières vormen. Voorbeelden van deze getallen zijn 500, 1.000, 10.000 en 100.000. Zichtbare pieken bij cijfercombinaties zoals 48, 49, 98 of 99 zijn interessant omdat deze pieken er bijvoorbeeld op kunnen wijzen dat managers getallen opzettelijk afronden zodat zij net onder de psychologische barrières vallen. De managers denken namelijk dat de kans groot is dat getallen die boven deze barrières liggen onderzocht worden door accountants.

Drake en Nigrini stellen ook voor om de cijfercombinaties te onderzoeken die dichtbij interne bedrijfslimieten liggen. Een manager in een bedrijf kan bijvoorbeeld een uitgavenlimiet hebben van €4.000. Het is dan verstandig om cijfercombinaties zoals 38 en 39 te onderzoeken. Volgens Drake en Nigrini duiden grote pieken bij deze cijfercombinaties erop dat de manager mogelijk abnormale uitgaven doet die hij opsplits in meerdere delen die net onder zijn uitgavenlimiet liggen.

De volgende richtlijnen worden door Drake en Nigrini, in het geval van de eerste twee cijfers, gehanteerd:

- Er sprake van sterke conformiteit aan de wet van Benford indien de gemiddelde absolute afwijking zich bevindt in het interval $[0.0000, 0.0006)$.
- Er sprake van accepteerbare conformiteit indien de gemiddelde absolute afwijking zich bevindt in het interval $[0.0006, 0.0012)$.

- Er sprake van marginaal accepteerbare conformiteit indien de gemiddelde absolute afwijking zich bevindt in het interval $[0.0012, 0.0018]$.
- De begincijfers voldoen niet aan de wet van Benford indien de gemiddelde absolute afwijking groter is dan 0.0018.

Men moet bij het gebruik van alle bovenstaande testen bedacht zijn op het feit dat non-conformiteit niet hoeft te betekenen dat er sprake is van fraude. De afwijkingen kunnen namelijk ook het gevolg zijn van typefouten of kunnen het gevolg zijn van de manier waarop een onderneming zaken doet. Zo is het algemeen bekend dat winkels (zoals supermarkten) een voorkeur hebben voor bepaalde onafgeronde getallen. Zo verkoopt een product beter als het €9.99 kost in plaats van €10 euro. Men moet dus eerst zeker weten dat er sprake is van fraude voordat men maatregelen neemt. Daarnaast is de wet van Benford het meest geschikt voor het opsporen van structurele fraude. Een groot bedrag dat eenmalig naar een privé-rekening is gestuurd zal niet zo gauw met de wet van Benford gedetecteerd worden. Het is daarom van belang om ook andere opsporingsmethoden te gebruiken.

De bovenstaande drie testen hebben één groot nadeel: het is namelijk niet meteen duidelijk welke gevonden proporties significant afwijken van de verwachte proporties. Zo voorspelt de wet van Benford dat ongeveer 30.103% van de begincijfers zal beginnen met een 1. Het geobserveerde percentage zal nooit precies gelijk zijn aan dit percentage puur vanwege het toeval. De vraag is nu op welk punt de afwijkingen statistisch significant zijn.

Nigrini gebruikt een Z-statistiek om te bepalen of een geobserveerde proportie van een bepaald cijfer afwijkt van de verwachte proportie volgens de wet van Benford [16]. Laat X_i het aantal getallen zijn dat begint met cijfer $i=1,2,\dots,9$. X_i is binomiaal verdeeld met parameters m en p_i . De Z-statistiek toetst nu de volgende hypothese:

$$\begin{aligned} H_0 : p_i &= \log_{10}(1+1/i) \\ H_1 : p_i &\neq \log_{10}(1+1/i), \end{aligned} \quad (3.2)$$

en is als volgt gedefinieerd:

$$z = \frac{(|p_o - p_e| - 1/(2m))}{s}, \quad (3.3)$$

hierbij stelt $p_o = (X_i / m) = \hat{p}_i$ de geobserveerde proportie van een bepaald cijfer (of cijfercombinatie) in de dataset voor, p_e stelt de verwachte proportie van dit cijfer volgens de wet van Benford voor, s is de standaarddeviatie van $(p_o - p_e)$ en m is gelijk aan het aantal waarnemingen waaruit de dataset bestaat. De term $1/(2m)$ is een continuïteitscorrectie die wordt toegepast indien $1/(2m) < |p_o - p_e|$. De standaarddeviatie s voor elk cijfer 1 tot en met 9 wordt als volgt berekend:

$$s = \sqrt{\text{Var}(p_o - p_e)} = \sqrt{\text{Var}(p_o)} = \sqrt{\frac{mp_i(1-p_i)}{m^2}} = \sqrt{\frac{p_i(1-p_i)}{m}} = \sqrt{\frac{p_e(1-p_e)}{m}}, \quad (3.4)$$

waarbij de laatste stap volgt uit het feit dat onder de nulhypothese geldt dat $p_i = p_e$.

De Z-statistiek is onder de nulhypothese bij benadering standaard normaal verdeeld. Hieruit volgt dat een gevonden waarde van 1.96 voor de Z-statistiek overeenkomt met een overschrijdingskans van 0.05 en een Z-statistiek die gelijk is aan 1.64 overeenkomt met een overschrijdingskans van 0.10.

In [9] wordt een interessant probleem aangekaart: indien wij een dataset met veel waarnemingen onderzoeken is procentueel een kleiner aantal getallen nodig om een statistisch significante afwijking te creëren dan in een kleine dataset met weinig observaties. Dit kan aan de hand van een simpel voorbeeld aangetoond worden. Stel wij hebben de beschikking over twee datasets, één dataset met 1.000 waarnemingen en één dataset met 10.000 getallen. Wij willen nu weten hoeveel gefraudeerde getallen, die beginnen met een 1, er minstens nodig zijn zodat er sprake is van een statistisch significante afwijking bij een betrouwbaarheid van 90%. Dit aantal kan, in het geval van de dataset met 10.000 observaties, als volgt berekend worden:

$$\frac{\frac{|p_o - p_e|}{\sqrt{\frac{p_e(1-p_e)}{m}}}}{\sqrt{\frac{p_e(1-p_e)}{m}}} \geq 1.64 \Rightarrow \frac{|mp_o - mp_e|}{\sqrt{mp_e(1-p_e)}} \geq 1.64 \Rightarrow \frac{|M - 10000 \cdot \log_{10}(2)|}{\sqrt{10000 \cdot \log_{10}(2)(1 - \log_{10}(2))}} \geq 1.64 \Rightarrow$$

$$|M - 10000 \cdot \log_{10}(2)| \geq 75.228.$$

Er is dus een verschil van minstens 76 waarnemingen nodig voordat er sprake is van een statistisch significante afwijking. Dit komt overeen met een percentage van 0.76% van het totale aantal waarnemingen. Men kan op dezelfde manier berekenen dat in het geval van de dataset die 1.000 waarnemingen bevat er 24 gefraudeerde getallen nodig zijn, wat overeenkomt met 2.4% van het totaal aantal waarnemingen. De reden hiervoor is dat in een grote dataset meer data voorhanden is op basis waarvan de onbekende kansen p_i nauwkeuriger geschat kunnen worden en dus ook beter onderscheiden kunnen worden van de kansen die de wet van Benford voorspelt. Hierdoor zijn er procentueel minder waarnemingen nodig om een statistisch significante afwijking te creëren.

In het algemeen geldt dat hoe lager de betrouwbaarheid des te minder gefraudeerde getallen er nodig zijn voordat de fraude statistisch significant is en hoe hoger de betrouwbaarheid des te meer gefraudeerde getallen benodigd zijn. Er is hier echter sprake van een tweestrijd. Een lage betrouwbaarheid zal in het algemeen resulteren in veel vals-positieve resultaten waarbij de test aangeeft dat er fraude is gepleegd terwijl dit helemaal niet het geval is. Een hoge betrouwbaarheid zal er daarentegen voor zorgen dat de test in minder gevallen zal aangeven dat er sprake is van fraude.

Men moet er ook rekening mee houden dat de kans op een fout van de eerste soort toeneemt wanneer de bovenstaande toets voor alle begincijfers tegelijk wordt uitgevoerd [1]. Een fout van de eerste soort houdt in dat wij de nulhypothese onterecht verwerpen. Een accountant die de toets op deze manier gebruikt zal vaker een vals-positief resultaat tegenkomen dan wanneer hij bijvoorbeeld de chi-kwadraat toets gebruikt die in paragraaf 4.3 beschreven wordt. Het grootste nadeel hiervan is dat de toets zijn impact verliest als iedere keer blijkt dat de nulhypothese onterecht wordt verworpen. De accountant moet namelijk iedere keer uitzoeken wat de reden is dat de nulhypothese verworpen wordt. Als hij er dan steeds achter komt dat de nulhypothese onterecht verworpen is, dan zal hij in het vervolg eerder geneigd zijn om te denken dat het om een vals alarm gaat, zonder dit nader te onderzoeken.

Het bovenstaande probleem kan verholpen worden door toepassing van de Bonferroni correctie [23]. Deze correctie houdt in dat iedere individuele toets bij onbetrouwbaarheid α/M uitgevoerd wordt om een overall onbetrouwbaarheid van α te verkrijgen bij het tegelijk uitvoeren van M toetsen. Een nadeel van deze correctie is dat de overall onbetrouwbaarheid vaak veel kleiner is dan de gewenste α . Een uitgebreid bewijs hiervan kan in hoofdstuk 4 van [23] gevonden worden.

3.2 OVERIGE TOEPASSINGEN VAN DE WET VAN BENFORD

Prognoses

De wet van Benford kan gebruikt worden bij het maken van prognoses zoals bevolkingsprognoses of het maken van voorspellingen omtrent de prijs van aandelen. Het idee hierachter is als volgt: indien de huidige data voldoet aan de wet van Benford of het is aannemelijk dat de data aan deze wet voldoet dan moeten de prognoses of voorspellingen ook aan de wet van Benford voldoen. Anders dient dit model misschien vervangen te worden door een ander model dat wel aan deze eis voldoet. Dit noemt men ook wel de “Benford-in, Benford out” test. Deze test is dus alleen een extra hulpmiddel, maar kan zelf niet gebruikt worden bij het maken van de voorspellingen.

Ontwerp van computers

De wet van Benford kan ook gebruikt worden bij het ontwerpen van computers. De meeste onderzoekers gaan ervan uit dat de computers in de toekomst berekeningen zullen uitvoeren waarvan de uitkomsten afkomstig zullen zijn uit verschillende willekeurige verdelingen. Dit zou betekenen dat de uitkomsten niet uniform verdeeld zullen zijn, maar eerder de verdeling zullen volgen die de wet van Benford voorschrijft. Als dit inderdaad het geval is dan kan men computers ontwerpen die hier slim gebruik van maken. Als het cijfer 1 veel vaker voorkomt dan het cijfer 9, zoals de wet van Benford voorspelt, dan is het misschien mogelijk om de computers te ontwerpen die deze informatie gebruiken om de hoeveelheid geheugenruimte die nodig is om data op te slaan te minimaliseren. De Duitse wiskundige Peter Schatte heeft, gebruik makend van dit principe, reeds berekend dat computers die werken op basis van de logaritme met 8 als grondgetal de benodigde hoeveelheid geheugen voor dataopslag het meest minimaliseren. Onderzoekers zijn nu bezig om uit te zoeken hoe logaritmische computers gebruikt kunnen worden bij het versnellen van grote berekeningen.

Onderzoeksdata

In [21] wordt beschreven hoe de wet van Benford gebruikt is bij het ontdekken van verzonden antwoorden in huis aan huis enquêtes in Amerika. De Consumer Expenditure Survey is een onderzoek dat elk jaar in Amerika gehouden wordt om uit te vinden hoe huishoudens hun geld uitgeven. Enquêteurs gaan hierbij op bezoek bij verschillende huishoudens om de enquête af te nemen. Hoe meer enquêtes zij afnemen des te meer zij betaald krijgen. Sommige enquêteurs komen in de verleiding om zelf de enquêtes in te vullen om op deze manier zoveel mogelijk geld te verdienen. Zij verzinnen de antwoorden op alle vragen inclusief de bijbehorende bedragen. Dit kan de resultaten van het onderzoek ernstig schaden. De wet van Benford kan gebruikt worden om oplichters te identificeren en de enquêtes die zij hebben afgenomen te verwijderen uit de verzameling.

Ministerie van Financiën

Nigrini beschrijft in [17] hoe het Nederlandse ministerie van financiën in 1995 de begincijfers van een selectie belastingteruggaven heeft onderzocht op basis van de wet van Benford. Het doel van dit onderzoek was om vast te stellen of de belastingteruggaven afwijkingen bevatten en om de bruikbaarheid van deze methode vast te stellen. Een steekproef van 30.000 belastingteruggaven afkomstig uit het jaar 1992 werd hiervoor onderzocht. De eerste en tweede cijfers van deze teruggaven bleken zeer goed aan de wet van Benford te voldoen.

Klantenrelatie

Er is door verschillende onderzoekers gesuggereerd dat de duur van een relatie tussen een klant en een leverancier een Benford verdeling heeft. Als dit inderdaad het geval is dan heeft dit mogelijk implicaties voor de manier waarop de relaties met klanten onderhouden worden. Het ligt dan meer voor de hand om meer tijd en moeite te steken in het onderhouden van relaties met de huidige loyale klanten dan deze tijd en moeite te investeren in het werven van nieuwe klanten. Het is namelijk zeer goed mogelijk dat de relatie met de loyale klant langer stand houdt dan de relatie die wordt aangegaan met de nieuwe klant.

HOOFDSTUK 4

SIMULATIE

In hoofdstuk 2 hebben wij beschreven wat de wet van Benford inhoudt waarna wij in hoofdstuk 3 uitvoerig hebben besproken hoe deze wet gebruikt kan worden door accountants bij het opsporen van fraude in financiële boekhoudingen. In dit hoofdstuk zullen wij antwoord geven op de onderzoeksvraag, die beschreven is in de inleiding en betrekking heeft op het ontdekken van fraude in boekhoudingen. Deze fraude kan door middel van een statistische toets ontdekt worden. In paragraaf 4.2 zullen wij twee statistische toetsen beschrijven die in dit specifieke geval gebruikt kunnen worden: de chikwadraat-toets en de likelihood-ratiotoets. De likelihood-ratiotoets maakt gebruik van de maximum likelihood-schatter die wij om deze reden in de eerste paragraaf zullen afleiden. Wij hebben er uiteindelijk voor gekozen om de onderzoeksvraag door middel van simulaties, die gebaseerd zijn op de chikwadraat-toets, te beantwoorden. In paragraaf 4.3 wordt de opzet van deze simulaties besproken alsmede de resultaten die uit deze simulaties zijn voortgekomen. Tenslotte zullen in de laatste paragraaf de conclusies worden beschreven.

4.1 MAXIMUM LIKELIHOOD-SCHATTER VAN DE MULTINOMIALE VERDELING

In de geformuleerde onderzoeksvraag hebben wij de beschikking over een boekhouding die bestaat uit N bedragen. Deze bedragen kunnen gezien worden als een steekproef uit de multinomiale verdeling met parameters N en kansen p_1, p_2, \dots, p_9 , waarvoor geldt dat $p_1 + p_2 + \dots + p_9 = 1$. De kansen p_i geven aan wat de kans is dat een bepaald getal zal beginnen met het cijfer $i = 1, 2, \dots, 9$.

De maximum likelihood-schatter kan gebruikt worden om de kansen p_i te schatten aan de hand van de N waarnemingen waaruit de boekhouding bestaat. De N waarnemingen dienen hiervoor geschreven te worden als een vector $X = (X_1, X_2, \dots, X_9)$, waarvoor geldt dat X_i het aantal getallen aangeeft dat begincijfer i heeft en waarvoor geldt dat $X_1 + X_2 + \dots + X_9 = N$. De kansen p_i zijn onbekend en gaan wij om deze reden schatten met de maximum likelihood-schatter waarbij aangenomen wordt dat de parameter N bekend is. De te schatten parameter is dan $\theta = (p_1, p_2, \dots, p_8) \in \Theta$. p_9 ligt vast zodra p_1, p_2, \dots, p_8 bekend zijn, er geldt namelijk dat $p_9 = 1 - p_1 - p_2 - \dots - p_8$. Θ is de natuurlijke parameterverzameling $\left\{ p \in [0, 1]^8, \sum_{i=1}^8 p_i \leq 1 \right\}$, ook wel eenheidssimplex genoemd.

De likelihood-functie is in dit geval gedefinieerd als:

$$\theta \mapsto \binom{N}{X_1 \dots X_8} \left(\prod_{i=1}^8 p_i^{X_i} \right) (1 - p_1 - p_2 - \dots - p_8)^{X_9} \quad (4.1)$$

De maximum likelihood-schatter is die schatter $\hat{\theta} \in \Theta$ die de likelihood-functie maximaliseert. Om het rekenwerk te beperken is het handiger om de log likelihood te beschouwen:

$$\theta \mapsto \log \binom{N}{X_1 \dots X_8} + \sum_{i=1}^8 X_i \log p_i + X_9 \log (1 - p_1 - p_2 - \dots - p_8) \quad (4.2)$$

Nu geldt dat indien de likelihood-functie differentieerbaar is in $\theta \in \Theta$ en zijn maximum in een inwendig punt van θ aanneemt dat [zie 23]:

$$\frac{\partial}{\partial \theta_j} \log L(\theta; X) \Big|_{\theta=\hat{\theta}} = 0, \quad j = 1, 2, \dots, 8, \quad (4.3)$$

waarbij L de likelihood-functie voorstelt. Het oplossen van dit stelsel van aannemelijkheidsvergelijkingen leidt tot de maximum likelihood-schatter. De acht aannemelijkheidsvergelijkingen zijn in ons geval gelijk aan:

$$\frac{X_i}{\hat{p}_i} - \frac{X_9}{1 - \hat{p}_1 - \hat{p}_2 - \dots - \hat{p}_8} = 0, \quad \text{voor } i = 1, \dots, 8 \quad (4.4)$$

Deze vergelijkingen kunnen herschreven worden als:

$$\hat{p}_i = X_i \frac{\hat{p}_9}{X_9}, \quad \text{voor } i = 1, \dots, 8 \quad (4.5)$$

Sommatie van alle acht vergelijkingen geeft:

$$\sum_{i=1}^8 \hat{p}_i = \frac{\hat{p}_9}{X_9} \sum_{i=1}^8 X_i \Rightarrow 1 - \hat{p}_9 = \frac{\hat{p}_9}{X_9} (N - X_9) \Rightarrow 1 - \hat{p}_9 = \frac{N\hat{p}_9}{X_9} - \hat{p}_9 \Rightarrow \hat{p}_9 = \frac{X_9}{N} \quad (4.6)$$

Wij vinden dus uitgaande van $\theta = (p_1, p_2, \dots, p_8)$ de schatter voor p_9 in plaats van de schatters voor p_1, \dots, p_8 . Deze schatters kunnen echter makkelijk afgeleid worden uit vergelijking (4.5). Uit (4.6) volgt namelijk dat $(\hat{p}_9 / X_9) = (1/N)$. Dit resultaat substitueren in (4.5) geeft:

$$\hat{p}_i = \frac{X_i}{N}, \quad \text{voor } i = 1, \dots, 8 \quad (4.7)$$

Wij dienen nu te bewijzen dat de schatter $\hat{\theta} = \left(\frac{X_1}{N}, \frac{X_2}{N}, \dots, \frac{X_8}{N} \right)$ ook daadwerkelijk het punt is waar de likelihood-functie maximaal is. Omdat wij hier te maken hebben met een functie van meerdere variabelen dient dit te geschieden aan de hand van de Hessiaan matrix van de log likelihood-functie in het punt $\hat{\theta}$. Deze matrix is in ons geval gedefinieerd als:

$$H \log L(\hat{\theta}, X) = \begin{bmatrix} \left. \frac{\partial^2 \log L(\theta, X)}{\partial \theta_1^2} \right|_{\theta=\hat{\theta}} & \left. \frac{\partial^2 \log L(\theta, X)}{\partial \theta_1 \partial \theta_2} \right|_{\theta=\hat{\theta}} & \dots & \left. \frac{\partial^2 \log L(\theta, X)}{\partial \theta_1 \partial \theta_8} \right|_{\theta=\hat{\theta}} \\ \left. \frac{\partial^2 \log L(\theta, X)}{\partial \theta_2 \partial \theta_1} \right|_{\theta=\hat{\theta}} & \left. \frac{\partial^2 \log L(\theta, X)}{\partial \theta_2^2} \right|_{\theta=\hat{\theta}} & \dots & \left. \frac{\partial^2 \log L(\theta, X)}{\partial \theta_2 \partial \theta_8} \right|_{\theta=\hat{\theta}} \\ \vdots & \vdots & \ddots & \vdots \\ \left. \frac{\partial^2 \log L(\theta, X)}{\partial \theta_8 \partial \theta_1} \right|_{\theta=\hat{\theta}} & \left. \frac{\partial^2 \log L(\theta, X)}{\partial \theta_8 \partial \theta_2} \right|_{\theta=\hat{\theta}} & \dots & \left. \frac{\partial^2 \log L(\theta, X)}{\partial \theta_8^2} \right|_{\theta=\hat{\theta}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{-X_1}{\left(\frac{X_1}{N}\right)^2} - \frac{X_9}{\left(1 - \frac{X_1}{N} - \dots - \frac{X_8}{N}\right)^2} & -\frac{X_9}{\left(1 - \frac{X_1}{N} - \dots - \frac{X_8}{N}\right)^2} & \dots & -\frac{X_9}{\left(1 - \frac{X_1}{N} - \dots - \frac{X_8}{N}\right)^2} \\ -\frac{X_9}{\left(1 - \frac{X_1}{N} - \dots - \frac{X_8}{N}\right)^2} & \frac{-X_2}{\left(\frac{X_2}{N}\right)^2} - \frac{X_9}{\left(1 - \frac{X_1}{N} - \dots - \frac{X_8}{N}\right)^2} & \dots & -\frac{X_9}{\left(1 - \frac{X_1}{N} - \dots - \frac{X_8}{N}\right)^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{X_9}{\left(1 - \frac{X_1}{N} - \dots - \frac{X_8}{N}\right)^2} & -\frac{X_9}{\left(1 - \frac{X_1}{N} - \dots - \frac{X_8}{N}\right)^2} & \dots & \frac{-X_8}{\left(\frac{X_8}{N}\right)^2} - \frac{X_9}{\left(1 - \frac{X_1}{N} - \dots - \frac{X_8}{N}\right)^2} \end{bmatrix} \quad (4.8)$$

Volgens de ‘‘Multivariable Second Derivative Test’’ die beschreven wordt in het boek Calculus van Edwards en Penney⁴, geldt dat $\log L(\hat{\theta}, X)$ een globaal maximum is indien de eigenwaarden van de matrix $H \log L(\hat{\theta}, X)$ alle negatief zijn en de determinant $|H \log L(\hat{\theta}, X)| \neq 0$.

Alhoewel de 8×8 Hessiaan matrix een eenvoudige structuur heeft, kost het veel tijd en moeite om de eigenwaarden van deze matrix handmatig te berekenen. Dit is eenvoudiger in het geval van een multinomiale verdeling met drie parameters p_1, p_2 en p_3 . In dit specifieke geval geldt dat de aannemelijkheidsvergelijkingen als volgt zijn gedefinieerd:

$$\frac{X_i}{\hat{p}_i} - \frac{X_3}{1 - \hat{p}_1 - \hat{p}_2} = 0, \text{ voor } i = 1, 2 \quad (4.9)$$

Hieruit volgt dat de schatter gelijk is aan $\hat{\theta} = \left(\frac{X_1}{N}, \frac{X_2}{N}\right)$. Een bewijs dat aantoont dat deze schatter de likelihood-functie maximaliseert zullen wij nu afleiden. De Hessiaan matrix van de log likelihood-functie in het punt $\hat{\theta}$ is gelijk aan:

$$H \log L(\hat{\theta}, X) = \begin{bmatrix} -\frac{N^2}{X_1} - \frac{N^2}{X_3} & -\frac{N^2}{X_3} \\ -\frac{N^2}{X_3} & -\frac{N^2}{X_2} - \frac{N^2}{X_3} \end{bmatrix} \quad (4.10)$$

Het is duidelijk dat de determinant van deze matrix ongelijk is aan 0. Hiermee is dus voldaan aan de voorwaarde van de ‘‘Multivariable Second Derivative Test’’. De eigenwaarden van deze matrix zijn de oplossingen van de karakteristieke vergelijking:

$$|H \log L(\hat{\theta}, X) - \lambda I| = \begin{vmatrix} -\frac{N^2}{X_1} - \frac{N^2}{X_3} - \lambda & -\frac{N^2}{X_3} \\ -\frac{N^2}{X_3} & -\frac{N^2}{X_2} - \frac{N^2}{X_3} - \lambda \end{vmatrix} = 0 \quad (4.11)$$

⁴ Bron: Edwards, C.H. en D. E. Penney (2002), *Calculus, Early Transcendentals, Matrix Version*. Prentice Hall, 6^e editie, blz. 977-978.

Uitwerken van deze vergelijking leidt tot:

$$(X_1X_2X_3)\lambda^2 + (X_2X_3N^2 + X_1X_3N^2 + 2X_1X_2N^2)\lambda + N^5 = 0 \quad (4.12)$$

Beide kanten delen door N^2 geeft:

$$\left(\frac{X_1X_2X_3}{N^2}\right)\lambda^2 + (X_2X_3 + X_1X_3 + 2X_1X_2)\lambda + N^3 = 0 \quad (4.13)$$

De eigenwaarden worden nu gevonden door toepassing van de abc-formule:

$$\lambda_1, \lambda_2 = \frac{-(X_2X_3 + X_1X_3 + 2X_1X_2) \pm \sqrt{(X_2X_3 + X_1X_3 + 2X_1X_2)^2 - 4X_1X_2X_3N}}{2X_1X_2X_3 / N^2} \quad (4.14)$$

Wij dienen nu aan te tonen dat deze eigenwaarden negatief zijn. Als eerst merken wij op dat $X_1, X_2, X_3 > 0$ omdat wij hebben aangenomen dat de maximum likelihood-schatter een inwendig punt van het eenheidssimplex is. De factor $(2X_1X_2X_3 / N^2)$ kunnen wij buiten beschouwing laten omdat deze groter is dan 0. Er geldt dat $\sqrt{(X_2X_3 + X_1X_3 + 2X_1X_2)^2 - 4X_1X_2X_3N} > 0$ is, waaruit wij kunnen afleiden dat $-(X_2X_3 + X_1X_3 + 2X_1X_2) - \sqrt{(X_2X_3 + X_1X_3 + 2X_1X_2)^2 - 4X_1X_2X_3N} < 0$. Hiermee is dus aangetoond dat λ_2 negatief is.

Men kan ook aantonen dat λ_1 negatief is. Wij weten namelijk dat $4X_1X_2X_3N > 0$. Hieruit volgt dat $(X_2X_3 + X_1X_3 + 2X_1X_2) > \sqrt{(X_2X_3 + X_1X_3 + 2X_1X_2)^2 - 4X_1X_2X_3N}$. Op basis hiervan kunnen wij afleiden dat $-(X_2X_3 + X_1X_3 + 2X_1X_2) < -\sqrt{(X_2X_3 + X_1X_3 + 2X_1X_2)^2 - 4X_1X_2X_3N}$. Hieruit volgt uiteindelijk dat $-(X_2X_3 + X_1X_3 + 2X_1X_2) + \sqrt{(X_2X_3 + X_1X_3 + 2X_1X_2)^2 - 4X_1X_2X_3N} < 0$. Hiermee is aangetoond dat λ_1 ook negatief is.

Wij concluderen dus dat de schatter $\hat{\theta} = \left(\frac{X_1}{N}, \frac{X_2}{N}\right)$ de likelihood-functie inderdaad maximaliseert. Op

basis van deze conclusie gaan wij er vanuit dat de maximum likelihood-schatter $\hat{\theta} = \left(\frac{X_1}{N}, \frac{X_2}{N}, \dots, \frac{X_8}{N}\right)$

ook de likelihood-functie maximaliseert in het oorspronkelijke geval, zonder een wiskundig bewijs hiervoor te leveren.

4.2 DE CHIKWADRAAT-TOETS EN DE LIKELIHOOD-RATIOTOETS

In de onderzoeksvraag zijn wij erop uit om een significante afwijking te ontdekken van de gevonden proporties van de begincijfers in de boekhouding ten opzichte van de verwachte proporties die gespecificeerd worden door de wet van Benford. Laat X_i het aantal getallen zijn met begincijfer $i = 1, 2, \dots, 9$. De vector $X = (X_1, X_2, \dots, X_9)$ heeft, zoals wij reeds in paragraaf 4.1 hebben aangegeven, een multinomiale verdeling met parameters N en $p = (p_1, p_2, \dots, p_9)$. Laat $p_0 = (p_{01}, p_{02}, \dots, p_{09})$ de vector met kansen zijn die gevonden worden met behulp van de wet van Benford. Wij willen nu de volgende hypothesen toetsen [zie 6]:

$$\begin{aligned} H_0 : p = p_0 &\equiv \forall p_i : (p_i = p_{0i}) \\ H_1 : p \neq p_0 &\equiv \exists p_i : (p_i \neq p_{0i}) \end{aligned} \quad (4.15)$$

De nulhypothese geeft aan dat alle p_i en p_{0i} gelijk zijn en de alternatieve hypothese geeft aan dat er minstens een p_i bestaat waarvoor geldt dat $p_i \neq p_{0i}$.

Een veel gebruikte toets waarmee de bovenstaande hypothese getoetst kan worden is Pearson's chikwadraat-toets. De toetsingsgrootte van deze toets is gedefinieerd als:

$$X^2 = \sum_{i=1}^9 \frac{(X_i - Np_{0i})^2}{Np_{0i}} \quad (4.16)$$

Deze toetsingsgrootte meet de afwijking tussen het geobserveerde aantal waarnemingen en het verwachte aantal waarnemingen onder H_0 op een genormaliseerde manier. De nulhypothese wordt voor grote waarden van de toetsingsgrootte verworpen. Onder de nulhypothese heeft X^2 bij benadering een chikwadraat-verdeling met 8 vrijheidsgraden. Deze benadering is alleen betrouwbaar voor niet te kleine waarden van N . In de praktijk betekent dit dat het verwachte aantal waarnemingen, Np_{0i} voor $j = 1, 2, \dots, 9$, minstens gelijk dient te zijn aan 5.

Een andere toets die veel gebruikt wordt om de hypothesen in (4.15) te toetsen is de log likelihood-ratiotoets. Laat $X = (X_1, \dots, X_9)$ wederom een steekproef zijn uit de multinomiale kansverdeling met parameters N en $p = (p_1, p_2, \dots, p_9)$. In de voorgaande paragraaf hebben wij reeds aangegeven dat de likelihood-functie gegeven wordt door:

$$p \mapsto \binom{N}{X_1 \dots X_9} p_1^{X_1} \dots p_9^{X_9}. \quad (4.17)$$

Daarnaast hebben wij in de vorige paragraaf aangetoond dat de maximum likelihood-schatter voor p relatief tot het eenheidssimplex gelijk is aan $\hat{p} = X/N$. De log likelihood-ratiostatistiek voor het toetsen van H_0 zoals weergegeven in (4.15) is gelijk aan:

$$2 \log \lambda_n(X) = 2 \log \frac{\binom{N}{X_1 \dots X_9} \prod_{i=1}^9 (X_i/N)^{X_i}}{\binom{N}{X_1 \dots X_9} \prod_{i=1}^9 p_{0i}^{X_i}} = 2 \sum_{i=1}^9 X_i \log \frac{X_i}{Np_{0i}}. \quad (4.18)$$

De log likelihood-ratiostatistiek in (4.18) is gebaseerd op de maximum likelihood-schatter van de multinomiale verdeling met parameters N en $p = (p_1, p_2, \dots, p_9)$. Deze schatter kan men terugvinden in de teller van de breuk die is weergegeven in (4.18) en hebben wij om deze reden reeds in paragraaf 4.1 afgeleid.

Alhoewel wij te maken hebben met een enkelvoudige nulhypothese $P_0 = \{p_0\}$ heeft de log likelihood-ratiostatistiek in (4.18) onder de nulhypothese een gecompliceerde verdeling. In de praktijk gebruikt men om deze reden een benadering van deze verdeling bij het uitvoeren van de toets. Aangezien X kan worden opgevat als de som van N onderling onafhankelijke multinomiale grootheden met parameters p en $N=1$, en de kansdichtheid van X en de simultane dichtheid van deze steekproef

(welke het product van de marginale dichtheden is) proportioneel zijn, is stelling 4.18 in [23] toepasbaar. Op basis van deze stelling leiden wij af dat $2\log \lambda_n(X)$ asymptotisch χ_8^2 verdeeld is. Het is deze benadering die in de praktijk wordt toegepast.

De likelihood-ratiotoets is asymptotisch equivalent aan Pearson's chikwadraat-toets. Gebruikmakend van de bekende Taylor-benadering $x \log \left(\frac{x}{x_0} \right) \approx (x - x_0) + \frac{1}{2} \frac{(x - x_0)^2}{x_0}$ vinden wij voor grote N :

$$2\log \lambda_n(X) = 2 \sum_{i=1}^9 X_i \log \frac{X_i}{Np_{0i}} \approx 2 \sum_{i=1}^9 (X_i - Np_{0i}) + \sum_{i=1}^9 \frac{(X_i - Np_{0i})^2}{Np_{0i}} \approx \sum_{i=1}^9 \frac{(X_i - Np_{0i})^2}{Np_{0i}}. \quad (4.19)$$

De laatste stap volgt uit het feit dat $\sum_{i=1}^9 (X_i - Np_{0i}) = \sum_{i=1}^9 X_i - N \sum_{i=1}^9 p_{0i} = N - N = 0$.

Wij geven de voorkeur aan de toetsingsgrootheid in (4.16) die toebehoort aan de chikwadraat-toets. De reden hiervoor is dat deze toetsingsgrootheid makkelijker te interpreteren is dan de toetsingsgrootheid in (4.18) die toebehoort aan de likelihood-ratiotoets. Bij het gebruik van de toetsingsgrootheid in (4.16) geldt dat de nulhypothese in (4.15) bij onbetrouwbaarheidsdrempel α_0 wordt verworpen indien

$$X^2 \geq \chi_{8, \alpha_0}^2, \quad (4.20)$$

met χ_{8, α_0}^2 het bovenste α_0 -kwantiel van de χ^2 -verdeling met 8 vrijheidsgraden. Indien vastgesteld wordt dat de nulhypothese daadwerkelijk wordt verworpen, dan kan men vervolgens de Z-statistiek (zie formule 3.2) toepassen om te bepalen voor welke begincijfers het geobserveerde aantal waarnemingen significant afwijkt van het verwachte aantal waarnemingen.

4.3 RESULTATEN VAN DE SIMULATIES

Opzet van de simulaties

Wij hebben een functie geschreven die de simulaties uitvoert in het statistische pakket R⁵. Met behulp van deze simulaties kan antwoord gegeven worden op onze onderzoeksvraag: hoeveel gefraudeerde bedragen n dienen er minstens toegevoegd te worden aan een boekhouding, die voldoet aan de wet van Benford en welke bestaat uit N bedragen, om een fraudeur met een kans van tenminste p te pakken⁶.

De opgestelde functie maakt gebruik van vier invoerparameters: N , n , *onbetrouwbaarheidsdrempel* en *aantalSimulaties*. De parameter N stelt hierbij het aantal initiële bedragen in de boekhouding voor, voordat er gefraudeerde bedragen aan toegevoegd worden. De begincijfers van deze getallen worden op willekeurige wijze getrokken uit de Benford verdeling, welke gespecificeerd wordt door (2.1). De parameter n stelt het aantal gefraudeerde bedragen voor die aan de boekhouding worden toegevoegd. De begincijfers van deze getallen worden willekeurig getrokken uit de uniforme verdeling:

$$p(i) = 1/9, \text{ voor } i = 1, 2, \dots, 9. \quad (4.21)$$

De betreffende functie toetst de nulhypothese in (4.15) gebruikmakend van de toetsingsgrootheid in (4.16) die toebehoort aan de Pearson's chikwadraat-toets. De nulhypothese wordt, zoals reeds in de voorgaande paragraaf is aangegeven, verworpen bij onbetrouwbaarheidsdrempel

⁵ De broncode van deze functie is terug te vinden in bijlage A.4.

⁶ Zie hoofdstuk 1 voor een gedetailleerde beschrijving van deze onderzoeksvraag.

α_0 indien $X^2 \geq \chi_{8, \alpha_0}^2$, met χ_{8, α_0}^2 het bovenste α_0 -kwantiel van de χ^2 -verdeling met 8 vrijheidsgraden. De onbetrouwbaarheidsdrempel α_0 wordt in de functie aangegeven met de parameter *onbetrouwbaarheidsdrempel*. Bij het toetsen van de gespecificeerde nulhypothese zijn er twee conclusies mogelijk: verwerp H_0 en accepteer H_1 of verwerp H_0 niet. De eerste conclusie is een sterke conclusie en geeft aan dat de verdeling van de begincijfers significant afwijkt van de verdeling van Benford. De tweede conclusie is eigenlijk geen conclusie en geeft aan dat er meer informatie nodig is om tot een conclusie te komen.

De functie die wij geschreven hebben schat voor de gegeven N , n en onbetrouwbaarheidsdrempel α_0 de kans dat de nulhypothese wordt verworpen. Om deze kans te schatten wordt de nulhypothese meerdere keren getoetst waarbij de begincijfers van de bedragen N en n telkens opnieuw worden getrokken uit de bijbehorende verdelingen. Het aantal keren dat de nulhypothese wordt verworpen, als fractie van het totaal aantal keer dat de nulhypothese getest is, vormt dan een benadering van de kans dat de nulhypothese wordt verworpen. Het aantal keren dat de nulhypothese bij gegeven N , n en α_0 getoetst dient te worden kan in de functie met de parameter *aantalSimulaties* aangegeven worden.

In de uitgevoerde simulaties wordt de opgestelde functie voor verschillende n , bij vaste N , *onbetrouwbaarheidsdrempel* en *aantalSimulaties*, aangeropen. De resulterende kansen worden in een grafiek uitgezet als functie van n . Deze grafiek geeft een beeld van het onderscheidend vermogen van de toets. Het onderscheidend vermogen geeft hierbij aan hoe groot de kans is dat de nulhypothese wordt verworpen. Daarnaast wordt voor vier opeenvolgende waarden van n , die vooraf gespecificeerd zijn, een histogram afgebeeld van de frequenties van de begincijfers van de gehele boekhouding samen met de frequenties die voorspeld worden door de wet van Benford. De afgebeelde frequenties zijn een gemiddelde dat berekend is over het totaal aantal simulaties. De histogrammen verschaffen ons een goed beeld van de verschuiving in de verdeling van de begincijfers naarmate het aantal gefraudeerde bedragen in de boekhouding toeneemt. Daarnaast krijgen wij aan de hand van deze histogrammen een goed beeld van de grootte van de afwijkingen per begincijfer.

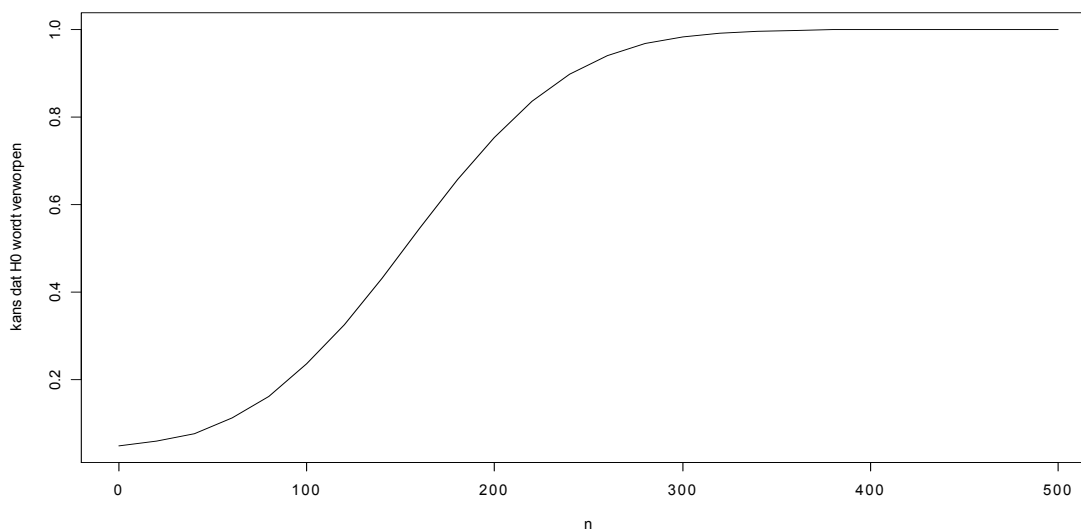
Resultaten

$N = 1.000$, *onbetrouwbaarheidsdrempel* = 0.05 en *aantal Simulaties* = 100.000

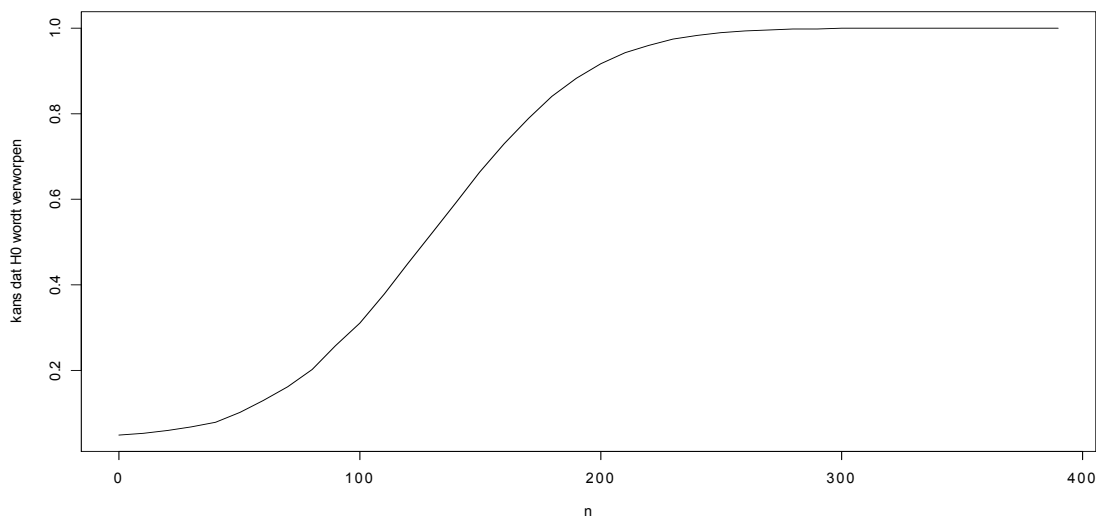
In de grafiek in figuur 4.1 is de kans weergegeven dat de nulhypothese in (4.15) wordt verworpen als functie van n , waarbij de overige parameters als volgt gedefinieerd zijn: $N = 1.000$, *onbetrouwbaarheidsdrempel* = 0.05 en *aantalSimulaties* = 100.000. Het eerste wat opvalt aan deze grafiek is dat er een kans bestaat dat de nulhypothese onterecht wordt verworpen ook als men geen gefraudeerde getallen toevoegt aan de boekhouding (dus $n = 0$). Deze kans is, om precies te zijn, gelijk aan 0.0497. De onbetrouwbaarheidsdrempel geeft in dit geval aan dat deze kans hoogstens 0.05 mag zijn. In dit specifieke geval wordt dus aan deze eis voldaan.

Op basis van figuur 4.1 kunnen wij nu een antwoord formuleren op onze onderzoeksvraag. Uit deze grafiek blijkt dat de nulhypothese met een kans van minstens 0.95 wordt verworpen indien het aantal gefraudeerde getallen n groter of gelijk is aan 280. Dit komt met ongeveer 21.88% ($= 280 / (1000 + 280)$) van de totale boekhouding overeen. Dit is vrij veel, maar kan intuïtief als volgt verklaard worden: de begincijfers van de gefraudeerde getallen worden uniform verspreid waardoor het langer duurt voordat de genormaliseerde som van de afwijkingen per begincijfer groot genoeg is om statistisch significant te zijn. De fraudeur zou eerder tegen de lamp lopen indien hij een voorkeur zou hebben voor bepaalde begincijfers en daardoor alleen deze begincijfers zou gebruiken. Dit kan met behulp van een simpel voorbeeld aangetoond worden. Stel dat de fraudeur om de één of andere

reden alleen bedragen verzint waarvan de begincijfers uniform verspreid zijn over de cijfers 1 t/m 5. Deze fraudeur maakt dus geen gebruik van de begincijfers 6 t/m 9. In figuur 4.2 is nu de kans weergegeven dat de nulhypothese in dit specifieke geval wordt verworpen, als functie van n . Uit deze figuur blijkt nu dat slechts 215 gefraudeerde bedragen benodigd zijn om de nulhypothese met een kans van 0.95 te verwerpen.

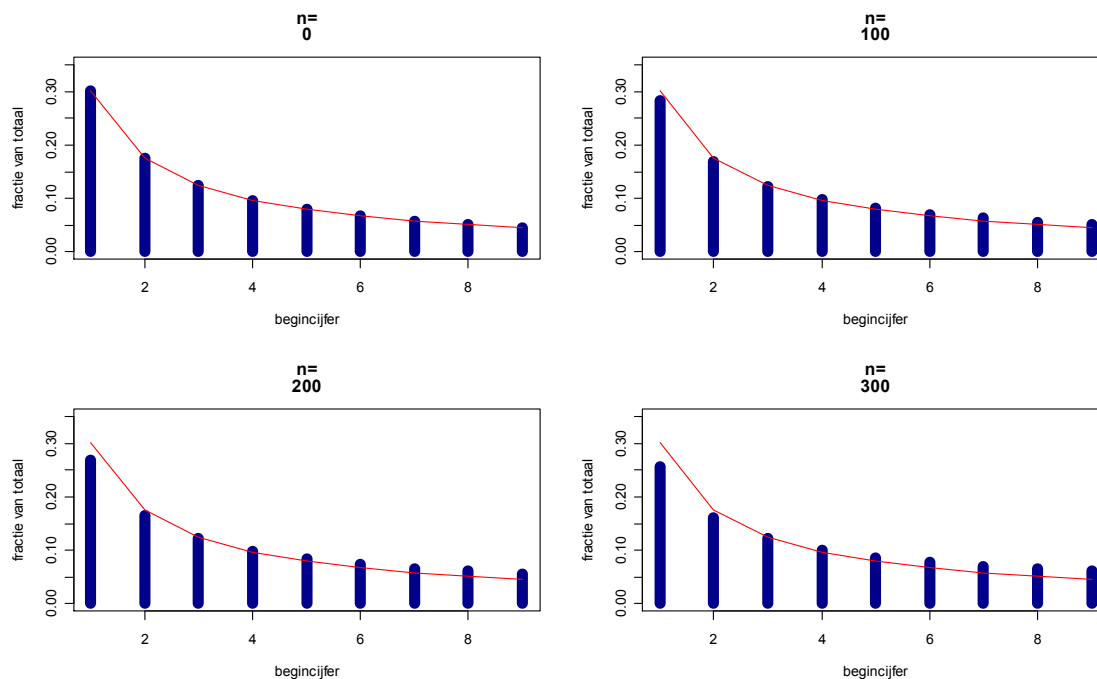


Figuur 4.1: De kans dat de nulhypothese wordt verworpen als functie van de toegevoegde gefraudeerde getallen (n).



Figuur 4.2: De kans dat de nulhypothese wordt verworpen als functie van de toegevoegde gefraudeerde getallen (n), indien de begincijfers van de gefraudeerde getallen uniform worden verspreid over de cijfers 1 t/m 5.

In figuur 4.3 zijn vier histogrammen weergegeven met de gemiddelde frequenties van de begincijfers van de gehele boekhouding samen met de frequenties die voorspeld worden door de wet van Benford voor $n = 0, 100, 200$ en 300 , respectievelijk. Uit deze histogrammen blijkt dat de afwijking van de begincijfers vooral te danken is aan een overschot van bedragen die beginnen met een 6, 7, 8 of 9 en aan een tekort van bedragen die beginnen met het cijfer 1.



Figuur 4.3: Histogrammen van de frequenties van de begincijfers 1 tot en met 9 van de bedragen in de gehele boekhouding (dus inclusief de 1000 initiële bedragen) in het geval van $n = 0$, $n = 100$, $n = 200$ en $n = 300$, respectievelijk. De rode lijn geeft de verwachte frequenties op basis van de wet van Benford weer.

De nulhypothese wordt in het geval van $n = 300$ met een kans van 0.9839 verworpen bij een betrouwbaarheid van 95%. Het is nu interessant om in dit specifieke geval na te gaan welke begincijfers een significante afwijking vertonen. Hiervoor kan men gebruik maken van de Z-statistiek van Nigrini zoals die beschreven is in paragraaf 3.1. Het is daarnaast ook interessant om de test van de begincijfers, die eveneens in deze paragraaf beschreven wordt, toe te passen op de gemiddelde frequenties zoals die zijn weergegeven in figuur 4.2 in het geval van $n = 300$. Wij hebben daarom een functie geschreven die de test van de begincijfers uitvoert in combinatie met de negen toetsen die gebaseerd zijn op de Z-statistiek⁷. In tabel 4.1 zijn de resultaten weergegeven.

Op basis van de overschrijdingskansen die vermeld zijn in deze tabel kunnen wij concluderen dat de geobserveerde frequenties van de begincijfers 1, 8 en 9 een statistisch significante afwijking vertonen op significantieniveau 0.05 ten opzichte van de verwachte frequenties. Volgens de fracties in deze tabel zijn er minder bedragen die met het cijfer 1 beginnen dan verwacht. Er is daarentegen juist een overschot aan bedragen die beginnen met de cijfers 8 en 9.

Uit de bovenstaande analyse blijkt dus dat de afwijkingen zich voornamelijk voordoen aan de uiteinden van de verdeling van de begincijfers van de bedragen in de boekhouding. Een logische verklaring hiervoor kan gevonden worden indien men de verdeling van Benford (wat een logaritmische verdeling is) vergelijkt met de uniforme verdeling. Deze twee verdelingen wijken in de staarten het meeste van elkaar af. Zo is volgens de Benford verdeling de kans op bedragen die beginnen met de cijfers 1 of 2 veel groter dan de kans die de uniforme verdeling toekent aan deze gebeurtenissen, en is de kans op bedragen die beginnen met een 8 of 9 juist kleiner dan de kansen die de uniforme verdeling voorschrijft. Wanneer men nu gefraudeerde bedragen waarvan het eerste cijfer uniform verdeeld is, toevoegt aan de aanwezige bedragen waarvan de begincijfers de Benford verdeling volgen, zal dit er uiteindelijk toe leiden dat men op de lange duur te weinig gefraudeerde bedragen toevoegt die met een 1 of 2 beginnen en zal men daarentegen teveel gefraudeerde bedragen

⁷ De broncode van deze functie is te vinden in bijlage A.1.

toevoegen die met een 8 of 9 beginnen. Dit verklaart waarom de significante afwijkingen zich juist in de staarten van de waargenomen verdeling van de begincijfers bevinden.

Uit tabel 4.1 blijkt daarnaast dat de gemiddelde absolute afwijking gelijk is aan 0.0138. Aangezien deze waarde groter is dan 0.012 voldoen de begincijfers volgens de richtlijnen van Drake en Nigrini niet aan de wet van Benford. Deze conclusie komt dus overeen met de conclusie die men verkrijgt op basis van de chikwadraat-toets.

Tabel 4.1: Het resultaat van de test van de begincijfers en de resultaten van de negen toetsen die gebaseerd zijn op de Z-statistiek toegepast op de gemiddelde frequenties in het geval van $n = 300$.

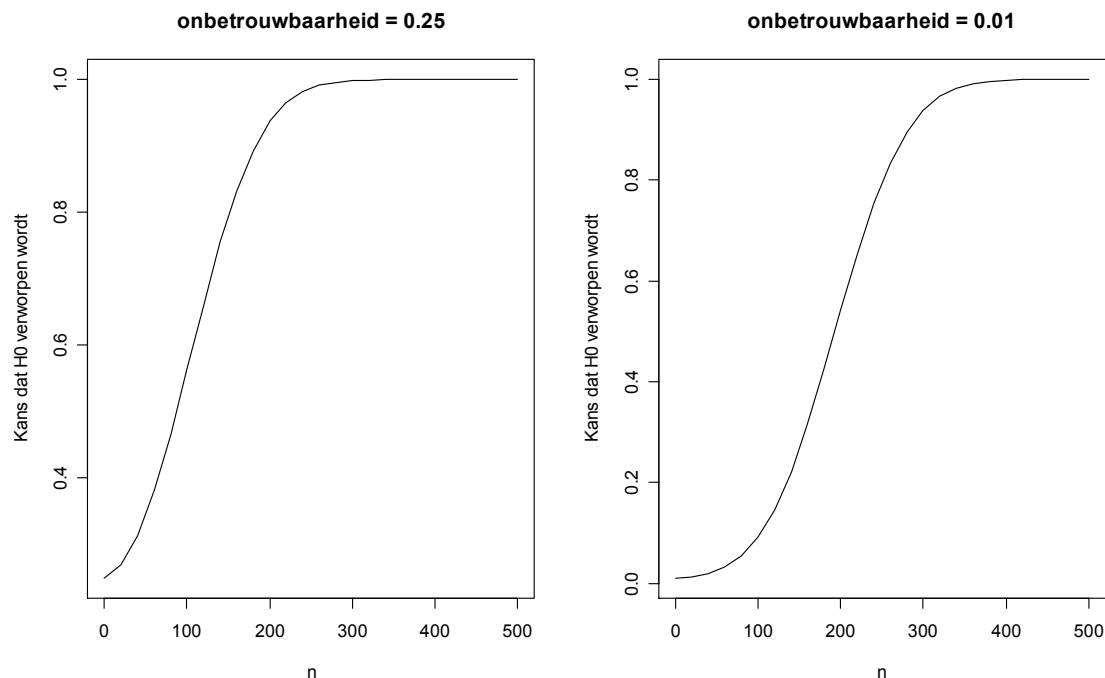
Begincijfer	Aantal	Fractie	Benford's wet	Verschil	Z-statistiek	p-waarde
1	334	0,2569	0,3010	-0,0441	3,4367	0,0003
2	210	0,1615	0,1761	-0,0146	1,3412	0,0899
3	158	0,1215	0,1249	-0,0034	0,3288	0,3711
4	130	0,1000	0,0969	0,0031	0,3297	0,3708
5	113	0,0869	0,0792	0,0077	0,9824	0,1630
6	100	0,0769	0,0669	0,0100	1,3837	0,0832
7	91	0,0700	0,0580	0,0120	1,7931	0,0365
8	85	0,0654	0,0512	0,0142	2,2663	0,0117
9	79	0,0608	0,0458	0,0150	2,5239	0,0058

Gemiddelde absolute afwijking
0,0138

De invloed van de onbetrouwbaarheidsdrempel op de kans waarmee de nulhypothese wordt verworpen kan worden afgeleid uit figuur 4.4⁸. De linkergrafiek in deze figuur geeft deze kans weer als functie van n , bij een onbetrouwbaarheidsdrempel van 0.25. In de rechtergrafiek is hetzelfde weergegeven, maar nu is de onbetrouwbaarheidsdrempel gelijk aan 0.01. Wanneer wij deze twee grafieken vergelijken met de grafiek in figuur 4.1 dan concluderen wij dat een hogere betrouwbaarheid ervoor zorgt dat er meer gefraudeerde bedragen nodig zijn om de nulhypothese met een bepaalde kans te verwerpen. Zo blijkt uit de linkergrafiek in de onderstaande figuur dat er slechts 210 bedragen nodig zijn om de nulhypothese met een kans van minstens 0.95 te verwerpen bij een onbetrouwbaarheidsdrempel van 0.25. Uit de rechtergrafiek blijkt dat er ongeveer 310 gefraudeerde bedragen nodig zijn om dezelfde nulhypothese met dezelfde kans te kunnen verwerpen bij een onbetrouwbaarheidsdrempel van 0.01. Dit betekent dus dat er 100 extra gefraudeerde bedragen nodig zijn, wat overeenkomt met een toename van ongeveer 48%.

Een grote onbetrouwbaarheid zorgt er weliswaar voor dat er minder gefraudeerde bedragen nodig zijn om een statistisch significante afwijking te ontdekken, maar zorgt er aan de andere kant ook voor dat de nulhypothese vaak onterecht wordt verworpen (dit komt overeen met een fout van de eerste soort). Dit is een ongewenst effect waar men bij een kleine onbetrouwbaarheid geen last van heeft. Het nadeel van een kleine onbetrouwbaarheid is echter dat de nulhypothese vaak niet wordt verworpen terwijl deze nulhypothese wel incorrect is (dit komt overeen met een fout van de tweede soort). De reden hiervoor is dat er veel gefraudeerde bedragen nodig zijn om een statistisch significante afwijking te creëren. In de praktijk hanteert men vaak een onbetrouwbaarheidsdrempel van 0.05.

⁸ Men moet bij het vergelijken van de twee grafieken in deze figuur rekening houden met feit dat de indeling van de y-as in de linkergrafiek verschilt van de indeling in de rechtergrafiek. Zo begint de y-as in de rechtergrafiek bij het punt 0 terwijl dit niet het geval is in de linkergrafiek. Dit geldt tevens voor de grafieken in de figuren 4.7, 4.10 en 4.13, respectievelijk.



Figuur 4.4: Het onderscheidend vermogen van de gehanteerde toets bij een onbetrouwbaarheidsdrempel van 0.01 en 0.25.

$N = 10.000$, onbetrouwbaarheidsdrempel = 0.05 en aantal Simulaties = 100.000

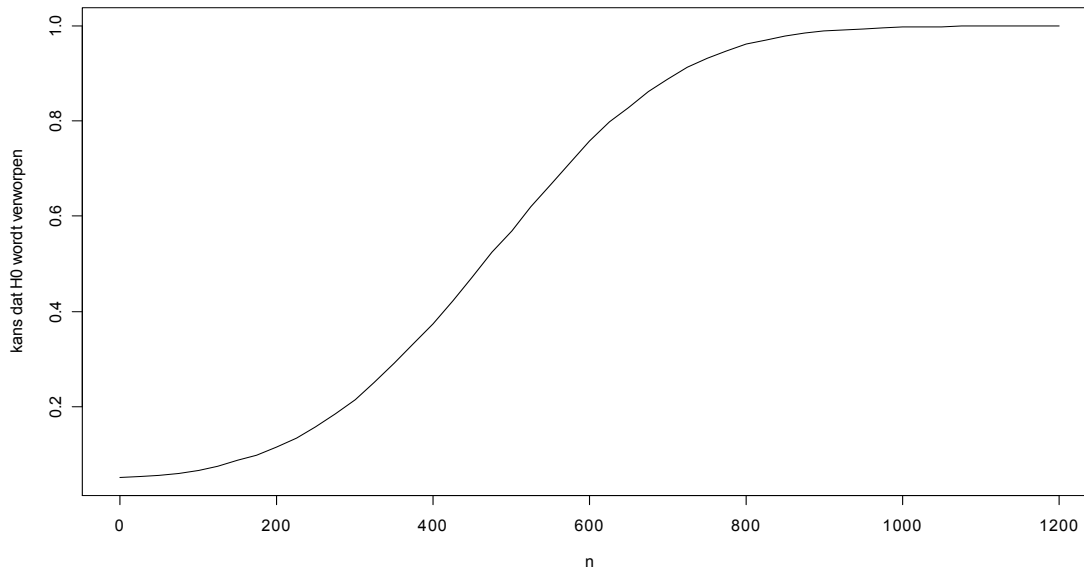
In de grafiek in figuur 4.5 is het onderscheidend vermogen van de chikwadraat-toets weergegeven als functie van n , bij een onbetrouwbaarheidsdrempel van 0.05. Het initiële aantal bedragen N waaruit de boekhouding bestaat is in dit geval gelijk aan 10.000 en het aantal simulaties is wederom gelijk aan 100.000. Aan de hand van deze grafiek maken wij op dat er ongeveer 800 gefraudeerde bedragen nodig zijn om de nulhypothese in (4.15) met een kans van tenminste 0.95 te verwerpen. Dit komt ongeveer overeen met 7.41% ($= 800/10800$) van het totaal aantal bedragen in de boekhouding. In het voorgaande geval, waarin N gelijk is aan 1.000, is dit percentage gelijk aan 21.88%. Hieruit leiden wij af dat, naarmate het aantal initiële bedragen groter wordt, er procentueel minder gefraudeerde bedragen nodig zijn om de fraude met een bepaalde kans te detecteren. Daarnaast blijkt uit figuur 4.5 dat de fraude met bijna 100% zekerheid wordt gedetecteerd indien er meer dan 1.000 gefraudeerde bedragen aan de boekhouding zijn toegevoegd.

Wanneer wij de eerder beschreven resultaten bestuderen kunnen wij reeds opmaken dat de chikwadraat-toets die gebaseerd is op de wet van Benford alleen gebruikt kan worden om structurele fraude mee te detecteren. Een persoon die eens in de zoveel tijd een gefraudeerde factuur indient zal met deze methode niet opgespoord kunnen worden, ook al gaat het hierbij om een zeer groot bedrag. Wij raden dan ook aan om naast deze toets ook andere methoden te gebruiken waarmee de eventuele fraude gedetecteerd kan worden.

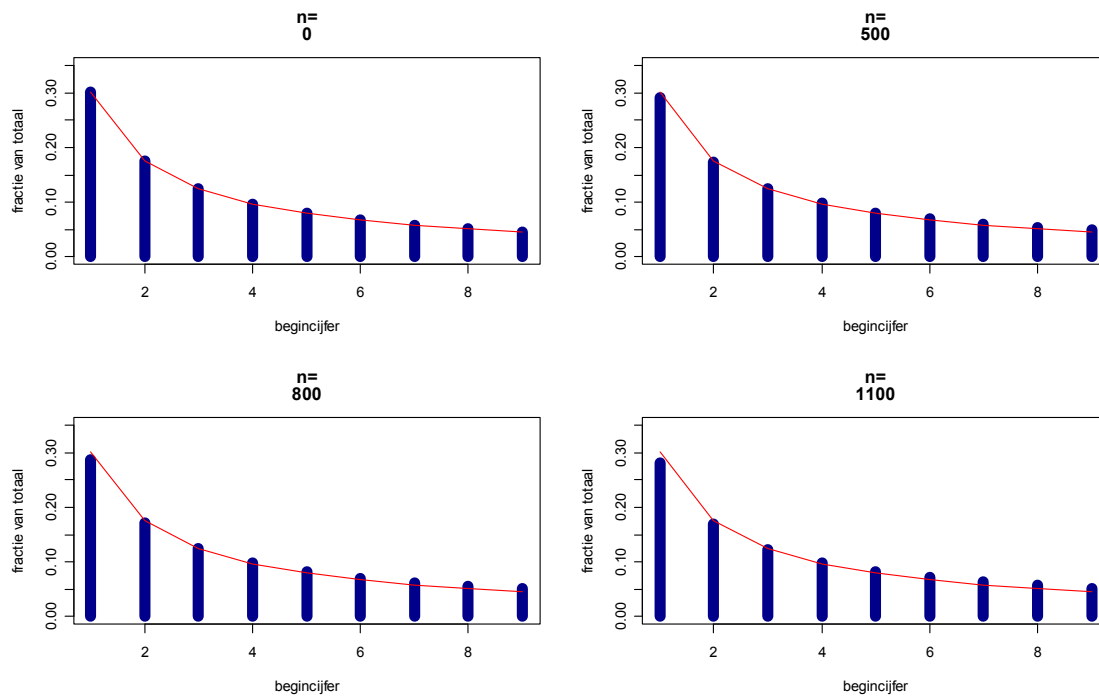
De reden waarom er procentueel steeds minder gefraudeerde bedragen nodig zijn om de fraude met een bepaalde kans te detecteren is dat naarmate de boekhouding groter wordt er meer informatie beschikbaar komt waardoor de parameter $\hat{\theta} = \left(\frac{X_1}{N}, \frac{X_2}{N}, \dots, \frac{X_8}{N} \right)$ nauwkeuriger geschat kan worden.

Hierdoor zijn wij beter in staat om de geschatte parameter $\hat{\theta}$ te onderscheiden van de parameter $\theta = (p_1, p_2, \dots, p_8)$ in de nulhypothese. Het onderscheidend vermogen voor een bepaalde parameter

onder de alternatieve hypothese neemt dus toe naarmate de boekhouding groter wordt. Als gevolg hiervan wordt de nulhypothese sneller verworpen en zijn er in een grote boekhouding relatief minder gefraudeerde bedragen nodig om hetzelfde onderscheidend vermogen te krijgen als in een kleinere boekhouding.



Figuur 4.5: De kans dat de nulhypothese wordt verworpen als functie van de toegevoegde gefraudeerde getallen (n).



Figuur 4.6: Histogrammen van de frequenties van de begincijfers 1 tot en met 9 van de bedragen in de gehele boekhouding in het geval van $n = 0$, $n = 500$, $n = 800$ en $n = 1.100$, respectievelijk. De rode lijn geeft de verwachte frequenties op basis van de wet van Benford weer.

In figuur 4.6 zijn vier histogrammen weergegeven met de gemiddelde frequenties van de begincijfers van de gehele boekhouding samen met de frequenties die voorspeld worden door de wet van Benford voor $n = 0, 500, 800$ en 1.100 , respectievelijk. Deze histogrammen geven een goed overzicht van de verschuivingen in de verdeling van de begincijfers als gevolg van de toegevoegde gefraudeerde bedragen. Wanneer men deze histogrammen zeer grondig bestudeert dan blijkt dat de afwijking van de begincijfers vooral veroorzaakt wordt door een overschot aan bedragen die beginnen met een 6, 8 of 9 en aan een tekort van bedragen die beginnen met het cijfer 1. De afwijkingen bevinden zich dus wederom aan de uiteinden van de verdeling van de begincijfers. Hierbij dient opgemerkt te worden dat de afwijkingen in dit geval niet zo duidelijk zichtbaar zijn als in het voorgaande geval.

Uit de hierboven beschreven simulaties blijkt dat de nulhypothese met een kans van 0.99966 wordt verworpen indien er 1.100 gefraudeerde bedragen worden toegevoegd aan de boekhouding, bij een onbetrouwbaarheid van 0.05. Om vast te stellen welke begincijfers in dit geval een significante afwijking vertonen kan wederom de Z-statistiek gebruikt worden. In tabel 4.2 is het resultaat weergegeven van de toepassing van de Z-statistiek op de gemiddelde verdeling van de begincijfers, zoals deze is weergegeven in figuur 4.6 voor $n = 1.100$. Uit deze tabel blijkt dat alle waargenomen fracties van de begincijfers 1, 6, 7, 8, en 9 een statistische afwijking vertonen ten opzichte van de verwachte frequenties in geval van een significantieniveau ter grootte van 0.05. Er is hierbij sprake van een tekort aan bedragen die beginnen met een 1 en een overschot aan getallen die beginnen met een 7, 8 of 9.

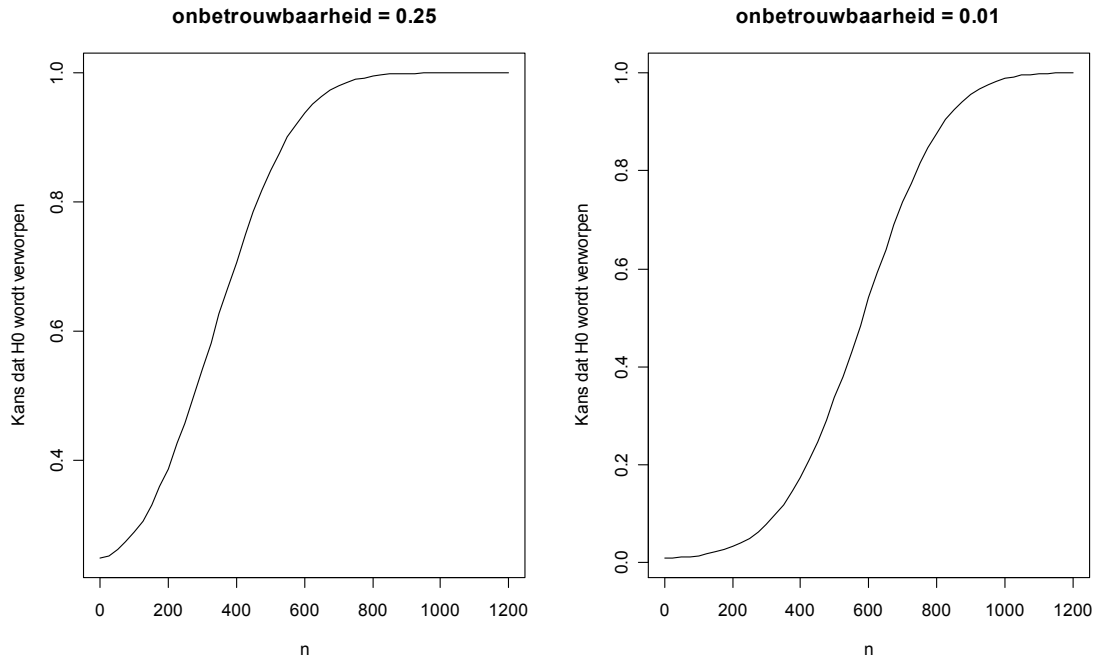
In tabel 4.2 is ook de gemiddelde absolute afwijking weergegeven, die in dit geval gelijk is aan 0.0059. Volgens de richtlijnen van Drake en Nigrini betekent dit dat er sprake is van accepteerbare conformiteit van de begincijfers aan de wet van Benford. Deze conclusie komt niet overeen met de conclusie die wij zouden trekken indien wij de chikwadraat-toets zouden gebruiken. Een belangrijke reden hiervoor is dat er geen wiskundige afleiding aan de richtlijnen van Drake en Nigrini ten grondslag ligt. Deze richtlijnen zijn puur opgesteld op basis van de ervaring van deze twee personen en zijn onafhankelijk van de grootte van de boekhouding. Hierdoor zijn deze richtlijnen niet in iedere situatie toepasbaar en dient men hier niet blind op te vertrouwen. Het is in onze ogen beter om te allen tijde een statistische toets te gebruiken om vast te stellen of de begincijfers een significante afwijking vertonen ten opzichte van de wet van Benford.

Tabel 4.2: Het resultaat van de test van de begincijfers en de resultaten van de negen toetsen die gebaseerd zijn op de Z-statistiek toegepast op de gemiddelde frequenties in het geval van $n = 1.100$.

Begincijfer	Aantal	Fractie	Benford's wet	Vershil	Z-statistiek	p-waarde
1	3132	0,2822	0,3010	-0,0189	4,3233	7,70E-06
2	1883	0,1696	0,1761	-0,0065	1,7721	0,0380
3	1372	0,1236	0,1249	-0,0013	0,4111	0,3405
4	1091	0,0983	0,0969	0,0014	0,4748	0,3175
5	914	0,0823	0,0792	0,0032	1,2158	0,1120
6	792	0,0714	0,0669	0,0044	1,8377	0,0331
7	702	0,0632	0,0580	0,0053	2,3468	0,0095
8	634	0,0571	0,0512	0,0060	2,8309	0,0023
9	580	0,0523	0,0458	0,0065	3,2520	0,0006

Gemiddelde absolute
afwijking

0,0059

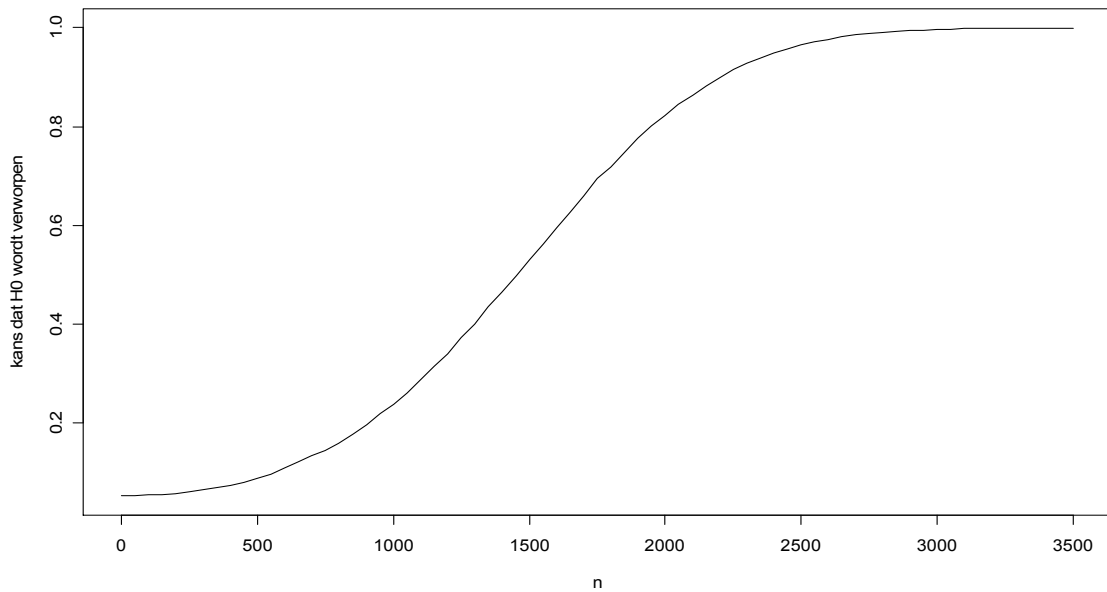


Figuur 4.7: Het onderscheidend vermogen van de gehanteerde toets bij een onbetrouwbaarheidsdrempel van 0.01 en 0.25.

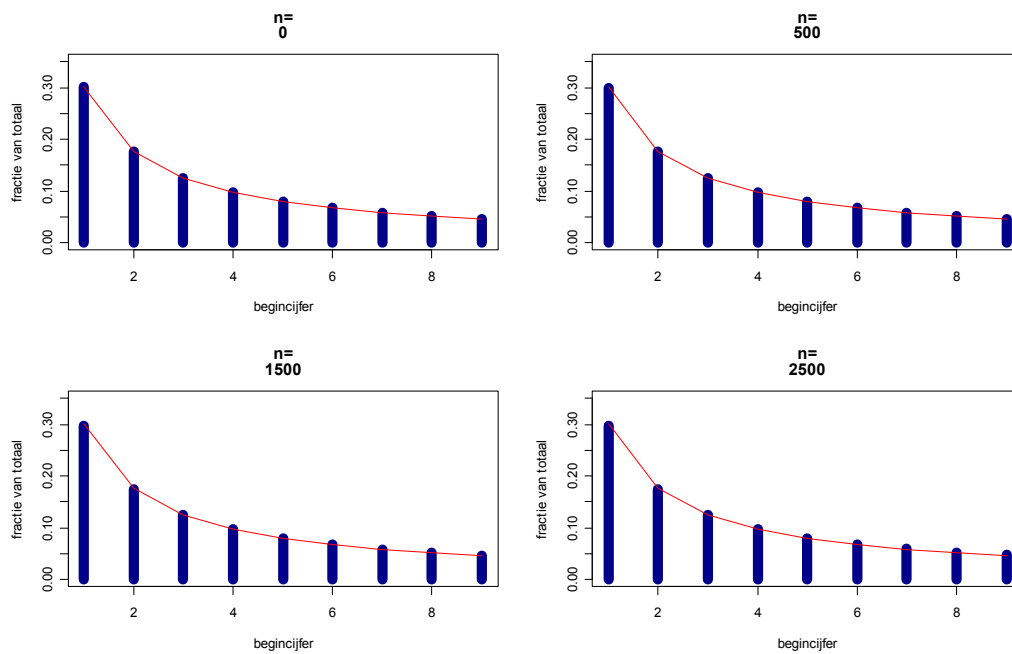
De invloed van de onbetrouwbaarheid op de kans waarmee de getoetste nulhypothese wordt verworpen kan afgeleid worden uit figuur 4.7. Het valt ons op dat de grafiek in de linkerfiguur, waarin de onbetrouwbaarheid gelijk is aan 0.25, veel steiler is dan de grafiek in de rechterfiguur. Dit betekent dat er minder gefraudeerde bedragen nodig zijn om de fraude met een bepaalde vooraf vastgestelde kans te detecteren indien de onbetrouwbaarheid gelijk is aan 0.25. Zo blijkt uit onze simulaties dat er in dit geval ongeveer 625 gefraudeerde bedragen nodig zijn om de nulhypothese met een kans van minstens 0.95 te verworpen tegenover 900 bedragen in het geval dat de onbetrouwbaarheid gelijk is aan 0.01. Dit komt overeen met een verschil van 44% in het benodigd aantal gefraudeerde bedragen. Hier staat wel tegenover dat een groot gedeelte van de ontdekte fraudes vals alarm zal zijn in het geval dat de onbetrouwbaarheid gelijk is aan 25%.

$N = 100.000$, $onbetrouwbaarheidsdrempel = 0.05$ en $aantal\ Simulaties = 100.000$

Over het algemeen zullen boekhoudingen van grote bedrijven doorgaans uit meer dan 1.000 of 10.000 getallen bestaan. In figuur 4.8 is daarom het onderscheidend vermogen van de chikwadraat-toets weergegeven voor een initiële boekhouding die bestaat uit 100.000 bedragen. Wederom geldt dat de onbetrouwbaarheidsdrempel gelijk is aan 0.05 en dat het aantal simulaties, waarop de weergegeven kansen zijn gebaseerd, gelijk is aan 100.000. Uit de grafiek maken wij op dat de nulhypothese met een kans van tenminste 0.95 wordt verworpen indien het aantal gefraudeerde bedragen minstens 2.400 bedraagt. Dit komt overeen met 2.34% ($= 2400/102400$) van het totaal aantal bedragen waaruit de boekhouding bestaat. Om de nulhypothese met een kans van 0.99 te verworpen zijn ongeveer 2.800 bedragen nodig, wat overeenkomt met 2.72% van de gehele boekhouding.



Figuur 4.8: De kans dat de nulhypothese wordt verworpen als functie van de toegevoegde gefraudeerde getallen (n).



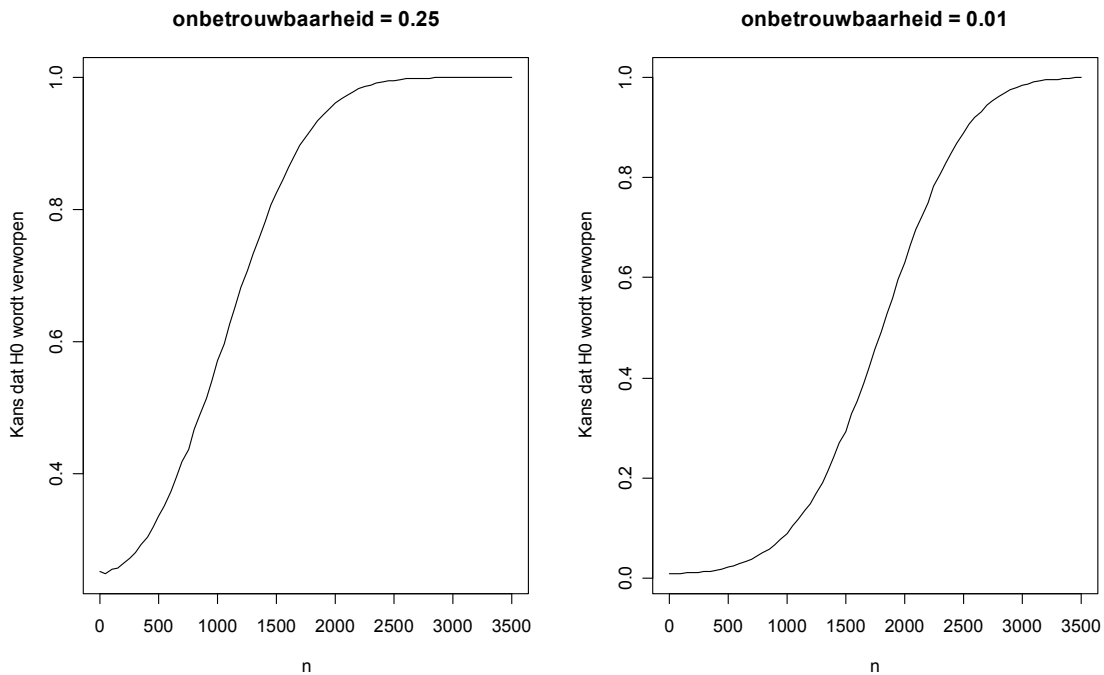
Figuur 4.9: Histogrammen van de frequenties van de begincijfers 1 tot en met 9 van de bedragen in de gehele boekhouding in het geval van $n = 0$, $n = 500$, $n = 1.500$ en $n = 2.500$, respectievelijk. De rode lijn geeft de verwachte frequenties op basis van de wet van Benford weer.

Uit de histogrammen in figuur 4.9 kunnen wij afleiden dat de significante afwijking voornamelijk wordt veroorzaakt door een overschot aan getallen die beginnen met het cijfer 7, 8 en 9. Omdat het hier om kleine afwijkingen gaat, is het moeilijk om op basis van deze histogrammen, met het blote oog aan te geven of deze afwijkingen significant zijn. Hiervoor gebruiken wij dan ook de Z-statistiek. Deze Z-statistiek hebben wij toegepast op de waargenomen frequenties zoals deze zijn weergegeven in het onderstaande histogram voor $n = 2.500$. De resultaten hiervan zijn weergegeven in tabel 4.3. Op basis van deze tabel leiden wij af dat er sprake is van een afwijking op significantieniveau 0.05 in het geval van de begincijfers 1, 8 en 9. In het eerste geval is er sprake van een tekort en in de overige gevallen is er sprake van een overschot aan bedragen die beginnen met deze cijfers.

Tabel 4.3: Het resultaat van de test van de begincijfers en de resultaten van de negen toetsen die gebaseerd zijn op de Z-statistiek toegepast op de gemiddelde frequenties in het geval van $n = 2.500$.

Begincijfer	Aantal	Fractie	Benford's wet	Vershil	Z-statistiek	p-waarde
1	30381	0,2964	0,3010	-0,0046	4,3233	0,0006
2	17886	0,1745	0,1761	-0,0016	1,7721	0,0909
3	12772	0,1246	0,1249	-0,0003	0,4111	0,3750
4	9969	0,0973	0,0969	0,0003	0,4748	0,3550
5	8196	0,0800	0,0792	0,0008	1,2158	0,1791
6	6972	0,0680	0,0669	0,0011	1,8377	0,0857
7	6077	0,0593	0,0580	0,0013	2,3468	0,0385
8	5393	0,0526	0,0512	0,0015	2,8309	0,0171
9	4854	0,0474	0,0458	0,0016	3,2520	0,0073

Gemiddelde absolute afwijking
0,00146



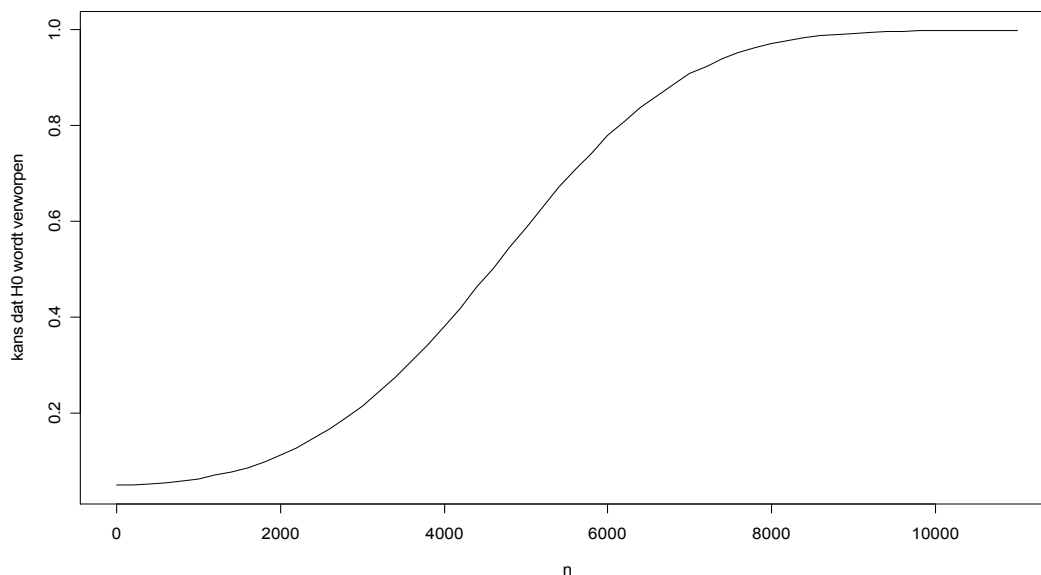
Figuur 4.10: Het onderscheidend vermogen van de gehanteerde toets bij een onbetrouwbaarheidsdrempel van 0.01 en 0.25.

In tabel 4.3 is daarnaast de gemiddelde absolute afwijking weergegeven, die in dit specifieke geval gelijk is aan 0.00146. Volgens de richtlijnen van Nigrini betekent dit dat de begincijfers erg goed voldoen aan de wet van Benford. Dit is wederom een foutieve conclusie. Volgens de uitgevoerde simulaties wordt de nulhypothese in het geval dat $n = 2.500$ met een kans van 0.9661 verworpen op basis van de chikwadraat-toets. Dit bewijst eens te meer dat de richtlijnen van de test van de begincijfers, die vastgesteld zijn door Drake en Nigrini, niet al te betrouwbaar zijn.

Ten slotte hebben wij in figuur 4.10 het onderscheidend vermogen van dezelfde toets weergegeven, maar nu in het geval van de onbetrouwbaarheidsdrempels 0.25 en 0.01. Op basis van deze figuur kan men afleiden dat in het eerste geval ongeveer 2.000 gefraudeerde bedragen nodig zijn om de nulhypothese met een kans van tenminste 0.95 te verwerpen. Bij een onbetrouwbaarheid van 0.01 zijn hier ongeveer 2.750 gefraudeerde bedragen voor nodig. Dit komt overeen met een verschil van 37.5%.

$N = 1.000.000$, $onbetrouwbaarheidsdrempel = 0.05$ en $aantal\ Simulaties = 100.000$

In de laatste simulaties die wij zullen beschrijven bestaat de initiële boekhouding uit $N = 1$ miljoen bedragen. Het onderscheidend vermogen van de gehanteerde statistische toets is voor dit specifieke geval weergegeven in figuur 4.11, waarbij het aantal simulaties wederom gelijk is aan 100.000. Uit deze figuur blijkt dat de nulhypothese in (4.15) met een kans van minstens 0.95 wordt verworpen indien het aantal gemanipuleerde bedragen, dat aan de bestaande boekhouding wordt toegevoegd, ongeveer gelijk is aan 7.600. Dit komt met slecht 0.75% ($= 7600 / (1007600)$) van de totale boekhouding overeen. Om dezelfde nulhypothese met een zekerheid van 99% te verwerpen zijn ongeveer 8.800 bedragen nodig, overeenkomend met 0.87% van de totale boekhouding. Procentueel gezien zijn er dus erg weinig gefraudeerde bedragen nodig om de fraude te ontdekken. De reden hiervoor hebben wij al uitgebreid beschreven in het geval van $N=10.000$. Echter, in dit geval is de initiële boekhouding groter waardoor er procentueel nog minder gefraudeerde bedragen nodig zijn om de nulhypothese te verwerpen.

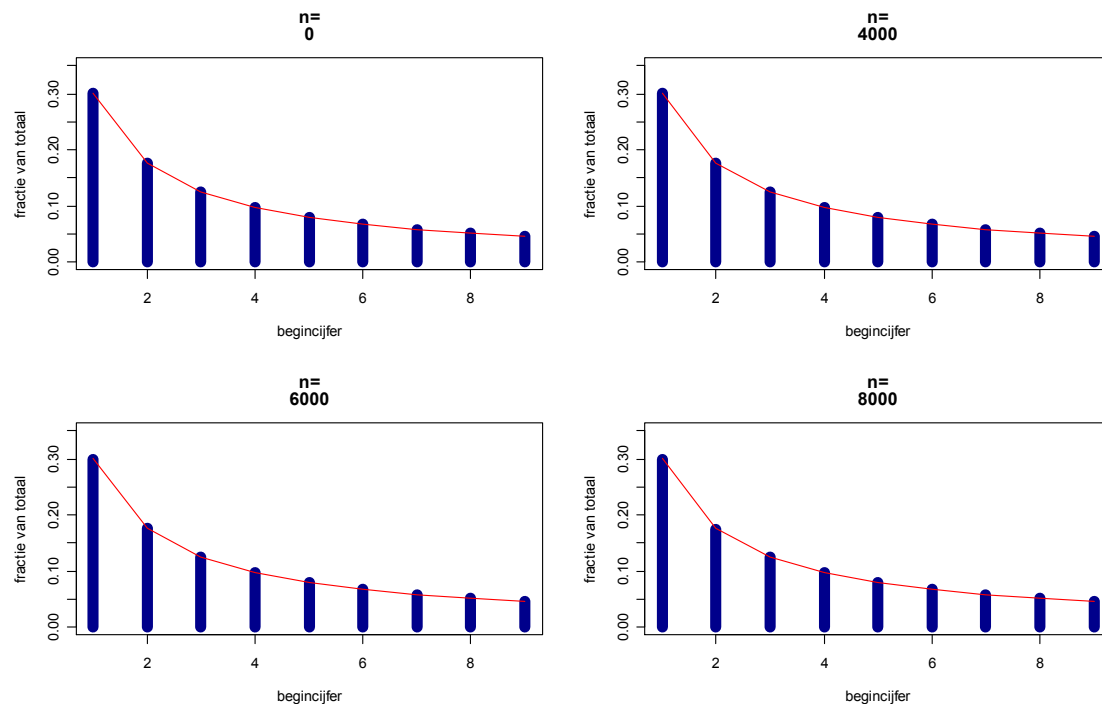


Figuur 4.11: De kans dat de nulhypothese wordt verworpen als functie van de toegevoegde gefraudeerde getallen (n).

In figuur 4.12 zijn vier histogrammen afgebeeld met de gemiddelde frequenties van de begincijfers van de gehele boekhouding in combinatie met de frequenties die voorspeld worden door de wet van Benford voor $n = 0, 4.000, 6.000$ en 8.000 , respectievelijk. Het is in dit geval niet meer mogelijk om op basis van deze histogrammen af te leiden welke begincijfers de oorzaak zijn van de gevonden afwijking. Wij zijn dus genoodzaakt om de Z-statistiek toe te passen. Wij hebben deze Z-statistiek toegepast op de waargenomen gemiddelde frequenties zoals deze zijn weergegeven in het onderstaande histogram voor $n = 8.000$. De resultaten zijn weergegeven in tabel 4.4. Deze tabel maakt duidelijk dat er een significante afwijking bestaat, op niveau 0.05, in het geval van de begincijfers 1, 8 en 9, respectievelijk. Wederom geldt dat in het eerste geval sprake is van een tekort en in de overige gevallen van een overschot aan bedragen die beginnen met de aangegeven cijfers.

De uitkomst van de test van de begincijfers, die gebaseerd is op de gemiddelde absolute afwijking, blijkt wederom niet betrouwbaar te zijn. Uit tabel 4.4 blijkt dat de gemiddelde absolute afwijking in dit geval gelijk is aan 0.00047, wat volgens de richtlijnen van Drake en Nigrini betekent dat de verdeling van de begincijfers zeer goed overeenkomt met de verdeling die de wet van Benford voorspelt. Volgens de uitgevoerde simulaties wordt de nulhypothese echter met een kans van 0.97150 verworpen. Dit bevestigt dat de richtlijnen, die zijn opgesteld door Drake en Nigrini, niet juist zijn.

Het onderscheidend vermogen van de gehanteerde toets is in het geval van de onbetrouwbaarheidsdrempels 0.25 en 0.01 weergegeven in figuur 4.13. Aan de hand van deze figuur leiden wij af dat er ongeveer 6.200 gefraudeerde bedragen aan de boekhouding toegevoegd dienen te worden om de fraude met een kans van minstens 0.95 aan het licht te stellen. Dit komt met ongeveer 0.62% van de gehele boekhouding overeen. In het geval dat de onbetrouwbaarheidsdrempel gelijk is aan 0.01 zijn hier ongeveer 8.600 gefraudeerde bedragen voor nodig, wat overeenkomt met 0.85% van de gehele boekhouding.

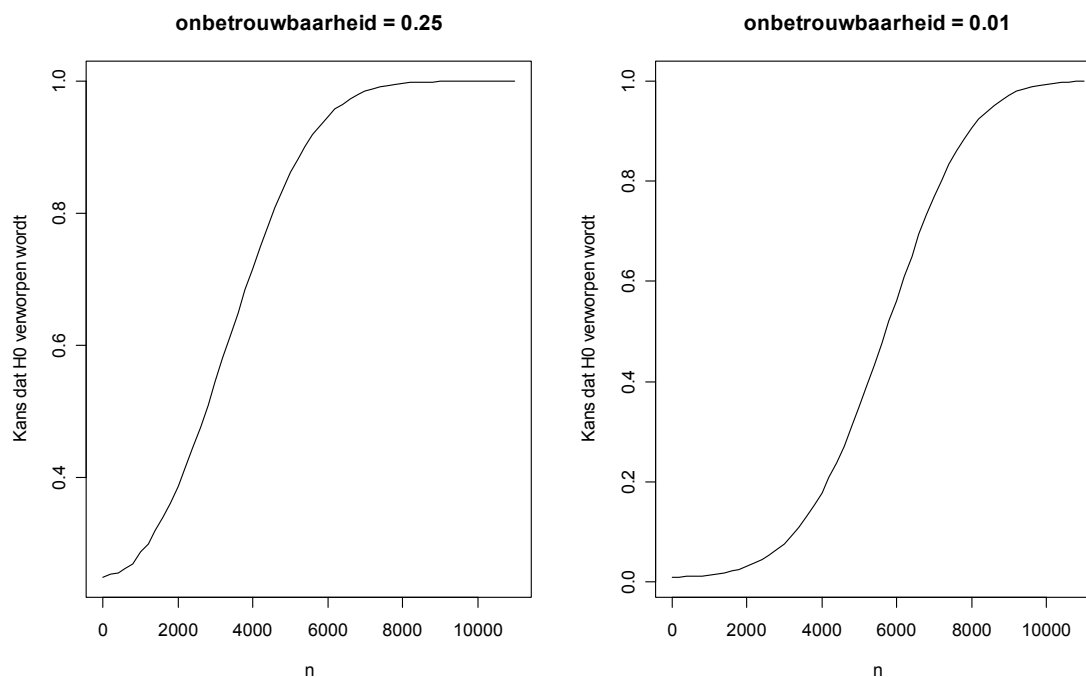


Figuur 4.12: Histogrammen van de frequenties van de begincijfers 1 tot en met 9 van de gehele boekhouding in het geval van $n = 0, n = 4.000, n = 6.000$ en $n = 8.000$, respectievelijk. De rode lijn geeft de verwachte frequenties op basis van de wet van Benford weer.

Tabel 4.4: Het resultaat van de test van de begincijfers en de resultaten van de negen toetsen die gebaseerd zijn op de Z-statistiek toegepast op de gemiddelde frequenties in het geval van $n = 8.000$.

Begincijfer	Aantal	Fractie	Benford's wet	Vershil	Z-statistiek	p-waarde
1	301919	0,2995	0,3010	-0,00151	3,2978	0,00049
2	176978	0,1756	0,1761	-0,00052	1,3637	0,0863
3	125829	0,1248	0,1249	-0,00011	0,3276	0,3716
4	97799	0,0970	0,0969	0,00011	0,3811	0,3515
5	80071	0,0794	0,0792	0,00025	0,9436	0,1727
6	67836	0,0673	0,0669	0,00035	1,4073	0,0797
7	58881	0,0584	0,0580	0,00042	1,8095	0,0352
8	52041	0,0516	0,0512	0,00047	2,1645	0,0152
9	46646	0,0463	0,0458	0,00052	2,4880	0,0064

Gemiddelde absolute
afwijking
0,00047



Figuur 4.13: Het onderscheidend vermogen van de gehanteerde toets bij onbetrouwbaarheidsdrempels 0.25 en 0.01, respectievelijk.

4.4 SAMENVATTING EN CONCLUSIES

Samenvatting

In tabel 4.5, 4.6 en 4.7 zijn de resultaten van de simulaties op een overzichtelijke manier samengevat. In deze tabellen is aangegeven hoeveel gefraudeerde of gemanipuleerde bedragen aan een boekhouding toegevoegd dienen te worden zodanig dat de nulhypothese in (4.15) met een gegeven kans verworpen kan worden bij een vooraf gedefinieerde onbetrouwbaarheidsdrempel. De bedragen zijn hierbij afgerond op vijftallen of tientallen.

Tabel 4.5: Het aantal gefraudeerde bedragen dat toegevoegd dient te worden aan de boekhouding, waarvan de grootte in de eerste kolom is weergegeven, zodanig dat de nulhypothese in (4.15) met de gegeven kans verworpen kan worden bij een onbetrouwbaarheidsdrempel van 0.25.

Grootte boekhouding	$P(H_0 \text{ wordt verworpen}) \approx 0,75$	$P(H_0 \text{ wordt verworpen}) \approx 0,95$	$P(H_0 \text{ wordt verworpen}) \approx 0,99$
1.000	140	220	260
10.000	425	625	750
100.000	1.350	2.000	2.350
1.000.000	4.200	6.200	7.400

Tabel 4.6: Het aantal gefraudeerde bedragen dat toegevoegd dient te worden aan de boekhouding, waarvan de grootte in de eerste kolom is weergegeven, zodanig dat de nulhypothese in (4.15) met de gegeven kans verworpen kan worden bij een onbetrouwbaarheidsdrempel van 0.05.

Grootte boekhouding	$P(H_0 \text{ wordt verworpen}) \approx 0,75$	$P(H_0 \text{ wordt verworpen}) \approx 0,95$	$P(H_0 \text{ wordt verworpen}) \approx 0,99$
1.000	200	280	320
10.000	600	775	925
100.000	1.850	2.400	2.800
1.000.000	5.800	7.600	8.800

Tabel 4.7: Het aantal gefraudeerde bedragen dat toegevoegd dient te worden aan de boekhouding, waarvan de grootte in de eerste kolom is weergegeven, zodanig dat de nulhypothese in (4.15) met de gegeven kans verworpen kan worden bij een onbetrouwbaarheidsdrempel van 0.01.

Grootte boekhouding	$P(H_0 \text{ wordt verworpen}) \approx 0,75$	$P(H_0 \text{ wordt verworpen}) \approx 0,95$	$P(H_0 \text{ wordt verworpen}) \approx 0,99$
1.000	240	320	360
10.000	725	900	1.025
100.000	2.200	2.750	3.100
1.000.000	7.000	8.600	9.800

Conclusies

De uitgevoerde simulaties hebben ten eerste aangetoond dat het aantal gefraudeerde of gemanipuleerde bedragen, als percentage van de gehele boekhouding, dat nodig is om de fraudeur met een bepaalde zekerheid in de kraag te vatten kleiner wordt naarmate de initiële boekhouding groter wordt. De reden hiervoor is dat in grote boekhoudingen de onbekende kansen nauwkeuriger geschat kunnen worden waardoor deze beter te onderscheiden zijn van de gespecificeerde kansen in de nulhypothese (het onderscheidend vermogen neemt dus toe). Hierdoor kunnen kleine afwijkingen in een grote boekhouding sneller ontdekt worden. Uiteindelijk zorgt dit ervoor dat er in een grote boekhouding relatief minder gefraudeerde bedragen nodig zijn om hetzelfde onderscheidend vermogen te verkrijgen als in een kleinere boekhouding.

Uit de analyse van de simulaties is daarnaast naar voren gekomen dat de significante afwijking ten opzichte van de wet van Benford voornamelijk wordt veroorzaakt door een tekort aan bedragen die beginnen met het cijfer één en aan een overschot van bedragen die beginnen met het cijfer acht of negen, respectievelijk. Dit is te wijten aan de verschillen in de staarten van de verdeling van de begincijfers van de bedragen die reeds in de boekhouding aanwezig zijn (Benford verdeling) en de verdeling van de begincijfers van de gefraudeerde bedragen die aan deze boekhouding worden toegevoegd (uniforme verdeling).

De simulaties hebben tevens aangetoond dat des te groter de onbetrouwbaarheidsdrempel is, des te minder gefraudeerde bedragen er nodig zijn om de nulhypothese, met een gegeven kans, te verwerpen. Op basis van de gevonden resultaten zijn wij tot de conclusie gekomen dat de wet van Benford alleen gebruikt kan worden in die gevallen dat er sprake is van structurele fraude.

Men dient wel rekening te houden met het feit dat de chi-kwadraattoets zoals wij die in de simulaties hebben gebruikt alleen aangeeft of de verdeling van de begincijfers significant afwijkt van de verdeling die voorspeld wordt door de wet van Benford. Deze toets geeft daarentegen niet aan welke

begincijfers verantwoordelijk zijn voor de waargenomen afwijking. Hiervoor kan men de Z -statistiek gebruiken zoals deze is beschreven in het artikel van Drake en Nigrini [8]. In dit artikel wordt ook de test van de begincijfers beschreven. Tijdens het uitvoeren van de simulaties is gebleken dat de uitkomsten van deze test niet betrouwbaar zijn. De richtlijnen die in dit artikel worden aangegeven en welke gebaseerd zijn op de ervaring van de auteurs, stroken niet met de uitkomsten van de chi-kwadraattoets. Wij raden het gebruik van deze toets dan ook af. Het is in onze ogen te allen tijde beter om gebruik te maken van de chi-kwadraattoets of de likelihood-ratio toets.

HOOFDSTUK 5

CONCLUSIE EN DISCUSSIE

De wet van Benford kan, in combinatie met andere fraudedetectietechnieken, een handig hulpmiddel zijn bij het opsporen van fraude en onjuistheden in financiële overzichten, zoals de boekhouding van een bedrijf. Aan het gebruik van deze wet als fraudedetectietechniek kleven zowel voor als nadelen. De voordelen zijn dat de wet eenvoudig is, en makkelijk toepasbaar is in de praktijk. Zo is een functie die een analyse uitvoert op basis van de wet van Benford in een mum van tijd geprogrammeerd. Deze wet verschaft daarnaast op een snelle en eenvoudige manier inzicht in het feit of de begincijfers van de bedragen in een boekhouding voldoen aan de wet van Benford. De uitkomst van deze test kan dan als startpunt dienen voor een uitgebreide analyse.

Een fraudedetectietechniek die gebaseerd is op de wet van Benford geeft alleen aan of de verdeling van de begincijfers van de bedragen waaruit de boekhouding is samengesteld een significante afwijking vertonen ten op zichte van de verdeling die de wet van Benford voorspelt. Het nadeel hiervan is dat deze wet niet aangeeft welke bedragen verdacht zijn. Om te achterhalen welke bedragen mogelijk gemanipuleerd zijn dienen andere fraudedetectietechnieken gehanteerd te worden. Een ander nadeel van deze wet is dat een fraudeur alleen gepakt kan worden indien hij structureel bedragen manipuleert. Indien deze persoon één keer een groot bedrag fraudeert, zal dit niet met de wet van Benford ontdekt worden. Daarnaast wordt er ook geen rekening gehouden met de grootte van de bedragen in de boekhouding omdat alleen de begincijfers van deze bedragen worden beschouwd. Deze wet maakt daardoor geen onderscheid tussen een bedrag ter grootte van €1 miljoen en een bedrag ter grootte van €10 miljoen omdat zij hetzelfde begincijfer hebben. Dit onderscheid is in sommige gevallen van cruciaal belang.

Een groot gedeelte van de bovenstaande voor en nadelen zijn aan het licht gekomen tijdens het uitvoeren van de simulaties. Deze simulaties hebben ons een duidelijk inzicht verschaft in het benodigde aantal gefraudeerde bedragen, dat aan een bestaande boekhouding toegevoegd dient te worden, zodanig dat een geval van fraude met een bepaalde zekerheid ontdekt kan worden. Hierbij zijn echter wel enkele aannamen gemaakt. Zo hebben wij aangenomen dat de fraudeur niet op de hoogte is van de wet van Benford en de begincijfers van de bedragen die hij verzint uniform verspreidt over alle mogelijke begincijfers. Wij hadden ook een andere verdeling voor de begincijfers kunnen kiezen. Wij hebben dit echter achterwege gelaten omdat wij ervan uit zijn gegaan dat een fraudeur de kans dat hij gepakt wordt wil minimaliseren door zo min mogelijk op te vallen. Een uniforme verdeling van de begincijfers is in onze ogen dan de meest voor de hand liggende strategie.

Een ander punt, dat betrekking heeft op de simulaties, is dat wij zijn uitgegaan van een initiële boekhouding met een vaste grootte waaraan wij de gefraudeerde bedragen hebben toegevoegd. Men had hier ook anders te werk kunnen gaan door aan te nemen dat een bepaald percentage van de initiële boekhouding bestaat uit gefraudeerde bedragen. In de simulaties kan dit percentage dan steeds gevarieerd worden om het aantal gemanipuleerde bedragen vast te stellen dat nodig is om de fraude op te sporen. Hierbij worden er dus geen nieuwe bedragen aan de boekhouding toegevoegd. De implementatie van deze methode is echter ingewikkelder dan de implementatie van de methode die wij gehanteerd hebben. Dit is dan ook de voornaamste reden waarom wij niet voor deze aanpak hebben gekozen.

Al met al kunnen wij concluderen dat de bekendheid van de wet van Benford steeds meer toeneemt. Deze wet wordt door steeds meer accountants gebruikt en wij verwachten dan ook dat deze wet in de toekomst steeds meer toegepast zal worden in de praktijk. Een nadeel is dat met het bekend worden van de wet van Benford ook fraudeurs op de hoogte raken van deze wet, met alle gevolgen van dien.

LITERATUURLIJST

- [1] Applying Digital Analysis using Benford's Law to Detect Fraud: A Practice Note about the Dangers of Type I and Type II Errors.
(<http://aaahq.org/audit/midyear/04midyear/papers/Dangers%20of%20Applying%20Benford's%20Law%20-%20Paper%20%20December%202003.doc>)
- [2] Benford, F. (1938), The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 4(78), blz. 551-572.
- [3] Brooks, D. (2002), Naked eye quantum mechanics: practical applications of Benford's law for integer quantities. *Frequencies: The Journal of Size Law Applications, special paper #1*.
- [4] Busta, B. en R. Weinberg (1998), Using Benford's law and neural networks as a review procedure. *Managerial Auditing Journal*, 13(6), blz. 356-366.
- [5] Carslaw, C. A. P. N. (1988), Anomalies in income numbers: Evidence of goal oriented behaviour. *The Accounting Review*. LXIII(2), blz. 321-327.
- [6] Cressie, N. en T. R. C. Read (1984), Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B (Methodological)*, 3(46), blz. 440-464.
- [7] Diaconis, P. en D. Freedman (1979), On rounding percentages. *Journal of the American Statistical Association*, 74, blz. 359-364.
- [8] Drake, P., en M. Nigrini (2000), Computer assisted analytical procedures using Benford's Law. *Journal of Accounting Education*, 18, blz. 127-146.
- [9] Durtschi, C., W. Hillison en C. Pacini (2004), The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. *Journal of Forensic Accounting*, blz. 17-34.
- [10] Hill, T.P. (1995), A statistical derivation of the significant-digit law. *Statistical Science*, 4(86), blz. 354-363.
- [11] Hill, T.P. (1996), The difficulty of faking data.
(<http://www.math.gatech.edu/~hill/publications/cv.dir/faking.pdf>)
- [12] Hill, T.P. (1998), The first digit phenomenon. *American Scientist*, (86), blz. 358-363.
- [13] Ley, E. (1996), On the peculiar distribution of the U.S. stock indexes' digits. *The American Statistician*, 4(50), blz. 311-313.
- [14] Long, A. First Significant Digit Law and Fraud Detection in Factoring Data.
(<http://www-users.cs.york.ac.uk/~lillian/Benford.pdf>)
- [15] Newcomb, S. (1881), Note of the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, (4), blz. 39-40.
- [16] Nigrini, M. (1996), A taxpayer compliance application of Benford's law. *Journal of the American Taxation Association*, 1(18), blz. 72-91.

- [17] Nigrini, M., en L. Mittermaier (1997), The use of Benford's Law as an Aid in Analytical Procedures. *Auditing: A Journal of Practice & theory*, 2(16), blz. 52-67.
- [18] Nigrini, M. (1999), Adding value with digital analysis. *The internal auditor*, 1(56), blz. 21- 23.
- [19] Pinkham, R.S. (1961), On the distribution of first significant digits. *Annals of Mathematical Statistics*, 32, blz. 1223-1230.
- [20] Sandron, F. (2002), Do Populations conform to the Law of Anomalous Numbers. *Population (English Edition)*, 4(57), blz. 755-761.
- [21] Swanson, D., M.J. Cho en J. Eltinge (2003), Detecting possibly fraudulent or error-prone survey data using Benford's law. *Joint Statistical Meetings*, blz. 4172-4177.
- [22] Thomas, J. K. (1989), Unusual patterns in reported earnings. *The accounting Review*, LXIV(4), blz. 773-787.
- [23] Vaart, A.W. van der en anderen (2008), *Algemene statistiek*. Collegedictaat, Vrije Universiteit, Amsterdam.
- [24] Varian, H.R. (1972), Benford's law. *The American Statistician*, 26, blz. 65-66.
- [25] Wallace, W. A. (2002), Assessing the quality of data used for benchmarking and decision-making. *The Journal of Government Financial Management*, 51(3), blz. 16-22.

BIJLAGE A

GEPROGRAMEERDE FUNCTIES IN R

A.1 TEST VAN DE BEGINCIJFERS

```
fdBenford<-function(boekhouding){

# Deze functie voert de test van de begincijfers uit op een meegegeven
# BOEKHOUDING. Deze test is beschreven in het artikel van Drake en Nigrini
# [Drake, P., en M. Nigrini (2000), Computer assisted analytical procedures
# using Benford's Law. Journal of Accounting Education, 18, blz. 127-146].
# Daarnaast voert deze functie de chikwadraat-toets uit en past hij de
# Z-statistiek toe op alle begincijfers van de bedragen waaruit de
# boekhouding is samengesteld. Voor een gedetailleerde beschrijving van de
# werking van deze toetsen wordt verwezen naar hoofdstuk 3 en 4 van dit
# werkstuk. Deze functie retourneert uiteindelijk een tabel waarin de
# uitkomsten van deze toetsen overzichtelijk zijn weergegeven.

    somEen    = 0;
    somTwee   = 0;
    somDrie   = 0;
    somVier   = 0;
    somVijf   = 0;
    somZes    = 0;
    somZeven  = 0;
    somAcht   = 0;
    somNegen  = 0;

    N = length(boekhouding);

    # De tabel waarin de resultaten worden opgenomen wordt hier
    # gecreëerd.
    tabel<-
    matrix(0,9,6,dimnames=list(c("1","2","3","4","5","6","7","8","9"),
    c("Aantal","Fractie","Benford's Law","Verschil","Z-statistiek",
    "p-waarde")));

    for(j in 1:9){
        tabel[j,3] = log10(1 + (1/j));
    }

    # Hier wordt het begincijfer van elk bedrag in de boekhouding
    # bepaald.
    for (i in 1:N) {
        while(boekhouding[i] >= 10){
            boekhouding[i] = boekhouding[i]/10;
        }
        boekhouding[i] = floor(boekhouding[i]);

        if (boekhouding[i]==1){
            somEen = somEen + 1;
        }
        else{
            if (boekhouding[i]==2){
                somTwee = somTwee + 1;
            }
        }
    }
}
```

```

    }
    else{
    if (boekhouding[i]==3){
        somDrie = somDrie + 1;
    }
    else{
    if (boekhouding[i]==4){
        somVier = somVier + 1;
    }
    else{
    if (boekhouding[i]==5){
        somVijf = somVijf + 1;
    }
    else{
    if (boekhouding[i]==6){
        somZes = somZes + 1;
    }
    else{
    if (boekhouding[i]==7){
        somZeven = somZeven + 1;
    }
    else{
    if (boekhouding[i]==8){
        somAcht = somAcht + 1;
    }
    else{
        somNegen = somNegen + 1;
    }}}}}}}}}
}

tabel[1,1] = somEen;
tabel[2,1] = somTwee;
tabel[3,1] = somDrie;
tabel[4,1] = somVier;
tabel[5,1] = somVijf;
tabel[6,1] = somZes;
tabel[7,1] = somZeven;
tabel[8,1] = somAcht;
tabel[9,1] = somNegen;

tabel[1,2] = somEen/N;
tabel[2,2] = somTwee/N;
tabel[3,2] = somDrie/N;
tabel[4,2] = somVier/N;
tabel[5,2] = somVijf/N;
tabel[6,2] = somZes/N;
tabel[7,2] = somZeven/N;
tabel[8,2] = somAcht/N;
tabel[9,2] = somNegen/N;

# Het verschil tussen de waargenomen proporties en de proporties die
# worden voorspeld volgens de wet van Benford wordt hier bepaald.
for(k in 1:9){
    tabel[k,4] = tabel[k,2] - tabel[k,3];
}

# De chikwadraat-toetsingsgrootheid wordt hier berekend samen met de
# bijbehorende overschrijdingskans.
chiKwadraat = (((somEen-log10(2)*N)^2)/(log10(2)*N))+(((somTwee-
```



```

log10(3/2)*N)^2)/(log10(3/2)*N))+(((somDrie-
log10(4/3)*N)^2)/(log10(4/3)*N))+(((somVier-
log10(5/4)*N)^2)/(log10(5/4)*N))+(((somVijf-
log10(6/5)*N)^2)/(log10(6/5)*N))+(((somZes-
log10(7/6)*N)^2)/(log10(7/6)*N))+(((somZeven-
log10(8/7)*N)^2)/(log10(8/7)*N))+(((somAcht-
log10(9/8)*N)^2)/(log10(9/8)*N))+(((somNegen-
log10(10/9)*N)^2)/(log10(10/9)*N));

pValue = 1-pchisq(chiKwadraat,8);

# Hier wordt de gemiddelde absolute afwijking berekend.
absoluteAfwijkingEen = abs((somEen/N)-log10(2));
absoluteAfwijkingTwee = abs((somTwee/N)-log10(3/2));
absoluteAfwijkingDrie = abs((somDrie/N)-log10(4/3));
absoluteAfwijkingVier = abs((somVier/N)-log10(5/4));
absoluteAfwijkingVijf = abs((somVijf/N)-log10(6/5));
absoluteAfwijkingZes = abs((somZes/N)-log10(7/6));
absoluteAfwijkingZeven = abs((somZeven/N)-log10(8/7));
absoluteAfwijkingAcht = abs((somAcht/N)-log10(9/8));
absoluteAfwijkingNegen = abs((somNegen/N)-log10(10/9));

gemiddeldeAbsoluteAfwijking =
(absoluteAfwijkingEen+absoluteAfwijkingTwee+absoluteAfwijkingDrie+
absoluteAfwijkingVier+absoluteAfwijkingVijf+absoluteAfwijkingZes+
absoluteAfwijkingZeven+absoluteAfwijkingAcht+absoluteAfwijkingNegen)/
9;

vector<-
c(absoluteAfwijkingEen,absoluteAfwijkingTwee,absoluteAfwijkingDrie,
absoluteAfwijkingVier,absoluteAfwijkingVijf,absoluteAfwijkingZes,
absoluteAfwijkingZeven,absoluteAfwijkingAcht,absoluteAfwijkingNegen);

# Hier wordt de Z-statistiek toegepast.
for(i in 1:9){
  if(1/(2*N) < vector[i]){
    tabel[i,5] = (vector[i] - 1/(2*N))/
      (sqrt((log10(1 + (1/i))*(1-log10(1 + (1/i))))/N));
  }
  else{
    tabel[i,5] = (vector[i])/
      (sqrt((log10(1 + (1/i))*(1-log10(1 + (1/i))))/N));
  }

  tabel[i,6] = 1 - pnorm(tabel[i,5],0,1);
}

# De eerder gecreëerde tabel wordt afgerond en geretourneerd.
w <- list(tabel, chiKwadraat, pValue, gemiddeldeAbsoluteAfwijking);
names(w) <- c("Tabel", "X^2", "p", "Gemiddelde Absolute Afwijking");

w
}

```

A.2 TEST VAN DE TWEDE CIJFERS

```

sdBenford<-function(boekhouding){

# Deze functie voert de test van de tweede cijfers uit op een
# meegegeven BOEKHOUDING. Deze test is beschreven in het artikel van Drake
# en Nigrini [Drake, P., en M. Nigrini (2000), Computer assisted analytical
# procedures using Benford's Law. Journal of Accounting Education, 18, blz.
# 127-146]. Daarnaast voert deze functie de chikwadraat-toets uit en past
# hij de Z-statistiek toe op alle tweede cijfers van de bedragen waaruit de
# boekhouding is samengesteld. Voor een gedetailleerde beschrijving van de
# werking van deze toetsen wordt verwezen naar hoofdstuk 3 en 4 van dit
# werkstuk. Deze functie retourneert uiteindelijk een tabel waarin de
# uitkomsten van deze toetsen overzichtelijk zijn weergegeven.

    somNul    = 0;
    somEen    = 0;
    somTwee   = 0;
    somDrie   = 0;
    somVier   = 0;
    somVijf   = 0;
    somZes    = 0;
    somZeven  = 0;
    somAcht   = 0;
    somNegen  = 0;

    N = length(boekhouding);

# De tabel waarin de resultaten worden opgenomen wordt hier
# gecreëerd.
tabel<-
matrix(0,10,6,dimnames=list(c("0","1","2","3","4","5","6","7","8","9"
),c("Aantal","Fractie","Benford's Law","Verschil","Z-statistiek",
"p-waarde")));

for(j in 0:9){
    eersteTweeCijfers = 10 + j;
    for(k in 1:9){
        tabel[j+1,3] = tabel[j+1,3] +
            log10(1 +(1/eersteTweeCijfers));
        eersteTweeCijfers = eersteTweeCijfers + 10;
    }
}

for (i in 1:N) {
    while(boekhouding[i] >= 100){
        boekhouding[i] = boekhouding[i]/10;
    }
    boekhouding[i] = (floor(boekhouding[i]))%%10;

    if (boekhouding[i]==0){
        somNul = somNul + 1;
    }
    else{
        if (boekhouding[i]==1){
            somEen = somEen + 1;
        }
    }
    else{
        if (boekhouding[i]==2){

```

```

        somTwee = somTwee + 1;
    }
else{
    if (boekhouding[i]==3){
        somDrie = somDrie + 1;
    }
else{
    if (boekhouding[i]==4){
        somVier = somVier + 1;
    }
else{
    if (boekhouding[i]==5){
        somVijf = somVijf + 1;
    }
else{
    if (boekhouding[i]==6){
        somZes = somZes + 1;
    }
else{
    if (boekhouding[i]==7){
        somZeven = somZeven + 1;
    }
else{
    if (boekhouding[i]==8){
        somAcht = somAcht + 1;
    }
else{
    somNegen = somNegen + 1;
    }}}}]]]]}}
}

tabel[1,1] = somNul;
tabel[2,1] = somEen;
tabel[3,1] = somTwee;
tabel[4,1] = somDrie;
tabel[5,1] = somVier;
tabel[6,1] = somVijf;
tabel[7,1] = somZes;
tabel[8,1] = somZeven;
tabel[9,1] = somAcht;
tabel[10,1] = somNegen;

tabel[1,2] = somNul/N;
tabel[2,2] = somEen/N;
tabel[3,2] = somTwee/N;
tabel[4,2] = somDrie/N;
tabel[5,2] = somVier/N;
tabel[6,2] = somVijf/N;
tabel[7,2] = somZes/N;
tabel[8,2] = somZeven/N;
tabel[9,2] = somAcht/N;
tabel[10,2] = somNegen/N;

# Het verschil tussen de waargenomen proporties en de proporties die
# worden voorspeld volgens de wet van Benford worden hier bepaald.
for(k in 1:10){
    tabel[k,4] = tabel[k,2] - tabel[k,3];
}

# De chikwadraat-toetsingsgrootheid wordt hier berekend samen met de

```

```

# bijbehorende overschrijdingskans.
chiKwadraat = (((somNul-tabel[1,3]*N)^2)/(tabel[1,3]*N))+(((somEen-
  tabel[2,3]*N)^2)/(tabel[2,3]*N))+(((somTwee-
  tabel[3,3]*N)^2)/(tabel[3,3]*N))+(((somDrie-
  tabel[4,3]*N)^2)/(tabel[4,3]*N))+(((somVier-
  tabel[5,3]*N)^2)/(tabel[5,3]*N))+(((somVijf-
  tabel[6,3]*N)^2)/(tabel[6,3]*N))+(((somZes-
  tabel[7,3]*N)^2)/(tabel[7,3]*N))+(((somZeven-
  tabel[8,3]*N)^2)/(tabel[8,3]*N))+(((somAcht-
  tabel[9,3]*N)^2)/(tabel[9,3]*N))+(((somNegen-
  tabel[10,3]*N)^2)/(tabel[10,3]*N));

pValue = 1-pchisq(chiKwadraat,9);

# Hier wordt de gemiddelde absolute afwijking berekend.
absoluteAfwijkingNul = abs((somNul/N)-tabel[1,3]);
absoluteAfwijkingEen = abs((somEen/N)-tabel[2,3]);
absoluteAfwijkingTwee = abs((somTwee/N)-tabel[3,3]);
absoluteAfwijkingDrie = abs((somDrie/N)-tabel[4,3]);
absoluteAfwijkingVier = abs((somVier/N)-tabel[5,3]);
absoluteAfwijkingVijf = abs((somVijf/N)-tabel[6,3]);
absoluteAfwijkingZes = abs((somZes/N)-tabel[7,3]);
absoluteAfwijkingZeven = abs((somZeven/N)-tabel[8,3]);
absoluteAfwijkingAcht = abs((somAcht/N)-tabel[9,3]);
absoluteAfwijkingNegen = abs((somNegen/N)-tabel[10,3]);

gemiddeldeAbsoluteAfwijking =(absoluteAfwijkingNul +
absoluteAfwijkingEen+absoluteAfwijkingTwee+absoluteAfwijkingDrie
+absoluteAfwijkingVier+absoluteAfwijkingVijf+absoluteAfwijkingZes+
absoluteAfwijkingZeven+absoluteAfwijkingAcht+absoluteAfwijkingNegen)/
10;

vector<-
c(absoluteAfwijkingNul,absoluteAfwijkingEen,absoluteAfwijkingTwee,
absoluteAfwijkingDrie,absoluteAfwijkingVier,absoluteAfwijkingVijf,
absoluteAfwijkingZes,absoluteAfwijkingZeven,absoluteAfwijkingAcht,
absoluteAfwijkingNegen);

# Hier wordt de Z-statistiek toegepast.
for(i in 1:10){
  if(1/(2*N) < vector[i]){
    tabel[i,5] = (vector[i] - 1/(2*N))/
      (sqrt((tabel[i,3]*(1-tabel[i,3]))/N));
  }
  else{
    tabel[i,5] = (vector[i])/
      (sqrt((tabel[i,3]*(1-tabel[i,3]))/N));
  }

  tabel[i,6] = 1 - pnorm(tabel[i,5],0,1);
}

# De gecreëerde tabel wordt afgerond en geretourneerd.
w <- list(tabel, chiKwadraat, pValue, gemiddeldeAbsoluteAfwijking);

names(w) <- c("Tabel", "X^2", "p", "Gemiddelde Absolute Afwijking");

w
}

```

A.3 TEST VAN DE EERSTE TWEE CIJFERS

```

fsdBenford<-function(boekhouding){

# Deze functie voert de test van de eerste twee cijfers uit op een
# meegegeven BOEKHOUDING. Deze test is beschreven in het artikel van Drake
# en Nigrini [Drake, P., en M. Nigrini (2000), Computer assisted analytical
# procedures using Benford's Law. Journal of Accounting Education, 18, blz.
# 127-146]. Daarnaast past deze functie de Z-statistiek toe op alle
# mogelijke combinaties van eerste twee cijfers. Voor een gedetailleerde
# beschrijving van de werking van deze toetsen wordt verwezen naar
# hoofdstuk 3 en 4 van dit werkstuk. Deze functie retourneert uiteindelijk
# een tabel waarin de uitkomsten van deze toetsen overzichtelijk zijn
# weergegeven.

  begincijferEen   = array(0,10);
  begincijferTwee  = array(0,10);
  begincijferDrie  = array(0,10);
  begincijferVier  = array(0,10);
  begincijferVijf  = array(0,10);
  begincijferZes   = array(0,10);
  begincijferZeven = array(0,10);
  begincijferAcht  = array(0,10);
  begincijferNegen = array(0,10);

  N = length(boekhouding);

  # De tabel waarin de resultaten worden opgenomen wordt hier
  # gecreëerd.
  tabel<-
  matrix(0,90,6,dimnames=list(c(10:99),c("Aantal","Fractie","Benford's
  Law","Verschil","Z-statistiek","p-waarde")));

  for(j in 10:99){
    tabel[j-9,3] = log10(1 + (1/j));
  }

  for (i in 1:N) {
    while(boekhouding[i] >= 100){
      boekhouding[i] = boekhouding[i]/10;
    }
    boekhouding[i] = floor(boekhouding[i]);

    if (boekhouding[i]>=10 && boekhouding[i]<=19){
      for(k in 10:19){
        if(boekhouding[i]==k){
          begincijferEen[k-9] = begincijferEen[k-9] + 1;
        }
      }
    }
    else{
      if (boekhouding[i]>=20 && boekhouding[i]<=29){
        for(k in 20:29){
          if(boekhouding[i]==k){
            begincijferTwee[k-19] =
            begincijferTwee[k-19] + 1;
          }
        }
      }
    }
  }
}

```

```
else{
  if (boekhouding[i]>=30 && boekhouding[i]<=39){
    for(k in 30:39){
      if(boekhouding[i]==k){
        begincijferDrie[k-29] =
          begincijferDrie[k-29] + 1;
      }
    }
  }
  else{
    if (boekhouding[i]>=40 && boekhouding[i]<=49){
      for(k in 40:49){
        if(boekhouding[i]==k){
          begincijferVier[k-39] =
            begincijferVier[k-39] + 1;
        }
      }
    }
  }
  else{
    if (boekhouding[i]>=50 && boekhouding[i]<=59){
      for(k in 50:59){
        if(boekhouding[i]==k){
          begincijferVijf[k-49] =
            begincijferVijf[k-49] + 1;
        }
      }
    }
  }
  else{
    if (boekhouding[i]>=60 && boekhouding[i]<=69){
      for(k in 60:69){
        if(boekhouding[i]==k){
          begincijferZes[k-59] =
            begincijferZes[k-59] + 1;
        }
      }
    }
  }
  else{
    if (boekhouding[i]>=70 && boekhouding[i]<=79){
      for(k in 70:79){
        if(boekhouding[i]==k){
          begincijferZeven[k-69] =
            begincijferZeven[k-69] + 1;
        }
      }
    }
  }
  else{
    if (boekhouding[i]>=80 && boekhouding[i]<=89){
      for(k in 80:89){
        if(boekhouding[i]==k){
          begincijferAcht[k-79] =
            begincijferAcht[k-79] + 1;
        }
      }
    }
  }
  else{
    if (boekhouding[i]>=90 && boekhouding[i]<=99){
      for(k in 90:99){
        if(boekhouding[i]==k){
          begincijferNegen[k-89] =
            begincijferNegen[k-89] + 1;
        }
      }
    }
  }
}
```

```

    }
  }
}

for(i in 1:10){
  tabel[i,1] = begincijferEen[i];
  tabel[i,2] = begincijferEen[i]/N;
}

for(i in 1:10){
  tabel[i+10,1] = begincijferTwee[i];
  tabel[i+10,2] = begincijferTwee[i]/N;
}

for(i in 1:10){
  tabel[i+20,1] = begincijferDrie[i];
  tabel[i+20,2] = begincijferDrie[i]/N;
}

for(i in 1:10){
  tabel[i+30,1] = begincijferVier[i];
  tabel[i+30,2] = begincijferVier[i]/N;
}

for(i in 1:10){
  tabel[i+40,1] = begincijferVijf[i];
  tabel[i+40,2] = begincijferVijf[i]/N;
}

for(i in 1:10){
  tabel[i+50,1] = begincijferZes[i];
  tabel[i+50,2] = begincijferZes[i]/N;
}

for(i in 1:10){
  tabel[i+60,1] = begincijferZeven[i];
  tabel[i+60,2] = begincijferZeven[i]/N;
}

for(i in 1:10){
  tabel[i+70,1] = begincijferAcht[i];
  tabel[i+70,2] = begincijferAcht[i]/N;
}

for(i in 1:10){
  tabel[i+80,1] = begincijferNegen[i];
  tabel[i+80,2] = begincijferNegen[i]/N;
}

for(k in 10:99){
  tabel[k-9,4] = tabel[k-9,2] - tabel[k-9,3];
}

# Hier wordt de gemiddelde absolute afwijking berekend.
somAbsoluteAfwijkingen = 0;
for(i in 1:90){
  somAbsoluteAfwijkingen = somAbsoluteAfwijkingen +
    abs(tabel[i,2]-tabel[i,3]);
}

```

```

gemiddeldeAbsoluteAfwijking = somAbsoluteAfwijkingen/90;

# Hier wordt de Z-statistiek toegepast.
for(i in 1:90){
  if(1/(2*N) < abs(tabel[i,2]-tabel[i,3])){
    tabel[i,5] = (abs(tabel[i,2]-tabel[i,3]) - 1/(2*N))/
      (sqrt((tabel[i,3]*(1-tabel[i,3]))/N));
  }
  else{
    tabel[i,5] = (abs(tabel[i,2]-tabel[i,3]))/
      (sqrt((tabel[i,3]*(1-tabel[i,3]))/N));
  }

  tabel[i,6] = 1 - pnorm(tabel[i,5],0,1);
}

# De eerder gecreëerde tabel wordt afgerond en geretourneerd.
w <- list(tabel, gemiddeldeAbsoluteAfwijking);

names(w) <- c("Tabel", "Gemiddelde Absolute Afwijking");

w
}

```

A.4 FUNCTIES DIE GEBRUIKT ZIJN VOOR HET UITVOEREN VAN DE SIMULATIES

```

pWaarde<-function(N,n,aantalSimulaties,onbetrouwbaarheidsdrempel){

# Deze functie retourneert de kans dat de nulhypothese in (4.16)
# wordt verworpen voor gegeven N, n, aantalSimulaties en
# onbetrouwbaarheidsdrempel. Deze functie wordt uitgebreid beschreven
# in paragraaf 4.3 van dit werkstuk onder de kop "opzet van de
# simulaties".

totaleSomEen = 0;
totaleSomTwee = 0;
totaleSomDrie = 0;
totaleSomVier = 0;
totaleSomVijf = 0;
totaleSomZes = 0;
totaleSomZeven = 0;
totaleSomAcht = 0;
totaleSomNegen = 0;

aantalKeerVerworpen = 0;

# De vier verschillende waarde van n, waarvoor een histogram met de
# gemiddelde frequenties gemaakt dient te worden, kunnen hier
# gespecificeerd worden.
nEersteHistogram = 500;
nTweedeHistogram = 1500;
nDerdeHistogram = 2500;
nVierdeHistogram = 3000;

for(j in 1:aantalSimulaties){
# De begincijfers van de boekhouding ter grootte van N worden
# random geselecteerd uit de Benford verdeling.
begincijfersBoekhouding = rmultinom(1, size = N,
prob=c(log10(2),log10(3/2),log10(4/3),log10(5/4),log10(6/5),
log10(7/6),log10(8/7),log10(9/8),log10(10/9)));

```



```

somEen   = begincijfersBoekhouding[1];
somTwee  = begincijfersBoekhouding[2];
somDrie  = begincijfersBoekhouding[3];
somVier  = begincijfersBoekhouding[4];
somVijf  = begincijfersBoekhouding[5];
somZes   = begincijfersBoekhouding[6];
somZeven = begincijfersBoekhouding[7];
somAcht  = begincijfersBoekhouding[8];
somNegen = begincijfersBoekhouding[9];

# De begincijfers van de gefraudeerde cijfers ter grootte van n
# worden willekeurig uit de uniforme verdeling geselecteerd.
gefraudeerdeBeginCijfers = rmultinom(1, size = n,
prob=c(1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9));

somEen   = somEen   + gefraudeerdeBeginCijfers[1];
somTwee  = somTwee  + gefraudeerdeBeginCijfers[2];
somDrie  = somDrie  + gefraudeerdeBeginCijfers[3];
somVier  = somVier  + gefraudeerdeBeginCijfers[4];
somVijf  = somVijf  + gefraudeerdeBeginCijfers[5];
somZes   = somZes   + gefraudeerdeBeginCijfers[6];
somZeven = somZeven + gefraudeerdeBeginCijfers[7];
somAcht  = somAcht  + gefraudeerdeBeginCijfers[8];
somNegen = somNegen + gefraudeerdeBeginCijfers[9];

# Hier wordt de chikwadraat-toetsingsgrootte berekend samen met de
# bijbehorende overschrijdingskans.
chiKwadraat = (((somEen*log10(2)*(N+n))^2)/(log10(2)*(N+n)))+
(((somTwee*log10(3/2)*(N+n))^2)/(log10(3/2)*(N+n)))+
(((somDrie*log10(4/3)*(N+n))^2)/(log10(4/3)*(N+n)))+
(((somVier*log10(5/4)*(N+n))^2)/(log10(5/4)*(N+n)))+
(((somVijf*log10(6/5)*(N+n))^2)/(log10(6/5)*(N+n)))+
(((somZes*log10(7/6)*(N+n))^2)/(log10(7/6)*(N+n)))+
(((somZeven*log10(8/7)*(N+n))^2)/(log10(8/7)*(N+n)))+
(((somAcht*log10(9/8)*(N+n))^2)/(log10(9/8)*(N+n)))+
(((somNegen*log10(10/9)*(N+n))^2)/(log10(10/9)*(N+n)));

overschrijdingskans = 1-pchisq(chiKwadraat,8);

if(overschrijdingskans <= onbetrouwbaarheidsdrempel){
  aantalKeerVerworpen = aantalKeerVerworpen + 1;
}

if( n == nEersteHistogram || n == nTweedeHistogram ||
n == nDerdeHistogram || n == nVierdeHistogram){
  totaleSomEen   = totaleSomEen   + somEen;
  totaleSomTwee  = totaleSomTwee  + somTwee;
  totaleSomDrie  = totaleSomDrie  + somDrie;
  totaleSomVier  = totaleSomVier  + somVier;
  totaleSomVijf  = totaleSomVijf  + somVijf;
  totaleSomZes   = totaleSomZes   + somZes;
  totaleSomZeven = totaleSomZeven + somZeven;
  totaleSomAcht  = totaleSomAcht  + somAcht;
  totaleSomNegen = totaleSomNegen + somNegen;
}
}

# Hier worden de vier histogrammen gecreëerd en weergegeven.
if( n == nEersteHistogram || n == nTweedeHistogram ||
n == nDerdeHistogram || n == nVierdeHistogram){

```

```

gemiddeldeVerdeling<-
c(totaleSomEen,totaleSomTwee,totaleSomDrie,totaleSomVier,
totaleSomVijf,totaleSomZes,totaleSomZeven,totaleSomAcht,
totaleSomNegen)/(aantalSimulaties*(N+n));

benfordVerdeling<-
c(log10(2),log10(3/2),log10(4/3),log10(5/4),log10(6/5),log10(7/
6),log10(8/7),log10(9/8),log10(10/9));

maxY = max(gemiddeldeVerdeling,benfordVerdeling) + 0.05;

plot(gemiddeldeVerdeling,type="h",col="black",lwd=10,
ylim=c(0,maxY), main=c("n=",n), xlab="begincijfer",
ylab="fractie van totaal")

lines(benfordVerdeling,col="red");
}

# Hier wordt de kans dat de nulhypothese wordt verworpen berekend en
# geretourneerd.
kansVerwerpenNulhypothese = aantalKeerVerworpen / aantalSimulaties;

kansVerwerpenNulhypothese
}

grafiekOnderscheidendVermogen<-function(N,aantalSimulaties,
onbetrouwbaarheidsdrempel){

# De werking van deze functie is uitgebreid beschreven in paragraaf
# 4.3 van dit werkstuk onder de kop "opzet van de simulaties". Deze
# functie retourneert een grafiek van het onderscheidend vermogen van
# de toets bij een gegeven onbetrouwbaarheidsdrempel en vier
# histogrammen van de gemiddelde frequenties voor de gespecificeerde
# waarden van n. Hiervoor maakt deze functie gebruik van de functie
# pWaarde. Voorbeelden van grafieken die deze functie retourneert
# zijn te vinden in paragraaf 4.3 van dit werkstuk.

par(mfrow=c(2,2))
n = 0;
x<-c();
onderscheidendVermogen<-c();
stapgrootte = 50;

for (i in 1:70){
  print(i);
  x<-c(x,n);
  onderscheidendVermogen<-
  c(onderscheidendVermogen,pWaarde(N,n,aantalSimulaties,
  onbetrouwbaarheidsdrempel)[1]);
  n = n + stapgrootte;
}

x11()

figuur<-plot(x,onderscheidendVermogen,type="l",xlab="n",
ylab="kans dat H0 wordt verworpen");

w<-list(x,onderscheidendVermogen);

w
}

```