De Wet van Benford, 30% van alle getallen begint met een 1

1. Inleiding, probleemstelling

Een voorbeeld.
Als je een lijst maakt van de lengtes (in centimeters) van alle 16-jarigen in Nederland, dan kun je wel bedenken dat het cijfer 1 veruit het meest voorkomende eerste cijfer van die getallen is. Logisch toch? Want zoveel XXXL-ers van boven de 2 meter zijn er niet.
In de tekst noemen we het eerste cijfer vaak het "begincijfer".

Nog een voorbeeld.
Denk nu aan een lijst van inwoneraantallen van de gemeenten van Nederland. Dat loopt van 951 (Schiermonnikoog) tot 747093 (Amsterdam). Natuurlijk is deze lijst chaotisch en je mag verwachten dat de begincijfers van de aantallen ruwweg gelijk verdeeld zijn over. Logisch toch?

Helemaal niet, want dit blijkt: *in de lijst inwoneraantallen van de Nederlandse gemeenten begint ongeveer 30% van de aantallen met het cijfer 1*.
Om het nog gekker te maken: de oververtegenwoordiging van het cijfer 1 komt veel voor in de wereld. Vooral in getalbestanden die niet op een voor de hand liggende manier begrensd zijn zoals dat van de lichaamslengtes. Deze gekte heeft een naam: *de Wet van Benford*.

De Wet van Benford

Het fenomeen was al in de 19e eeuw ontdekt, maar Frank Benford publiceerde er in 1938 serieus over en gaf vele voorbeelden, waaruit bleek dat begincijfer 1 in veel databestanden ongeveer in 30% van de gevallen voorkwam.
Zijn wet levert de volgende tabel voor het voorkomen van begincijfers van getallen uit metingen:

cijfer:	1	2	3	4	5	6	7	8	9
als begincijfer (percentage):	30,1	17,6	12,5	9,7	7,9	6,7	5,8	5,1	4,6

De inwoneraantallen van de gemeenten zijn in goede benadering zo verdeeld.

Sterker nog!
Als je kijkt naar het totaal aantal vingers van de inwoners per gemeente, krijg je weer de verdeling volgens de Wet van Benford. Écht logisch: het begincijfer verandert dan namelijk niet. Als je kijkt naar het totaal aantal handen van de inwoners per gemeente, krijg je óók bij benadering de Benford-verdeling. Bizar, want de aantallen handen hebben andere begincijfers dan de aantallen personen. En als je kijkt naar de kwadraten van de aantallen inwoners? Jawel, Benford gaat alwéér op!

Probleemstelling

In deze opdracht ga je een redelijke verklaring zoeken voor dit bijzondere verschijnsel. Je gaat daarbij eerst bij een paar databestanden kijken of en in hoeverre de wet van Benford daar opgaat. Je gaat hetzelfde doen met een paar op de computer gemaakte getalbestanden. Daar blijkt de wet soms wel, soms niet op te gaan. Daarmee kom je mogelijk op het spoor van een verklaring, en ook van een voorspelling wanneer je de verdeling van Benford wel of niet kunt verwachten. Je leert mogelijk zelfs toepassingen kennen!

2. Ter oriëntatie: handmatig begincijferverdeling verkennen via alfabetisch rangschikken

Eerst een nader onderzoek gemeentes van Nederland.
Het bestand gemaantal.xls is een lijst van de Nederlandse gemeenten met hun inwoneraantallen in het jaar 2010. Aanvankelijk staat de lijst alfabetisch op plaatsnaam geselecteerd van Aa en Hunze tot Zwolle.

a. Open het bestand en sorteer het bestand nu op de inwoneraantallen in kolom B. Zo: –Selecteer alle kolommen. Kies de optie Data >> Sort. Geef in het keuzevenster de juiste kolom aan, in dit geval Column B.

b. Zoek het kleinste, middelste en grootste aantal op.
Is het middelste aantal ongeveer het gemiddelde van grootste en kleinste?

c. Sorteer het bestand nu zó, dat alle aantallen met begincijfer 1 boven komen, daarna die met begincijfer 2, etc. Daarvoor moet je de getallenkolom sorteren op de manier van *alfabetisch* rangschikken. Dat is in het bestand voorbereid in kolom C:
Sorteer dus op Column C. want in kolom C staan de getallen van kolom B zó vertaald, dat Excel ze leest als 'tekst' niet als 'getal'. Hoe dat gedaan is, zie je door cel C1 te selecteren; de formule in die cel schijnt bovenin. (Deze formule kun je later eventueel kopiëren voor andere onderzoekingen)

d. Maak met behulp van deze laatste sortering een tabel en staafdiagram (op papier) voor de verdeling van 1, 2, 3, 4, 5, 6, 7, 8 en 9 als begincijfer.

e. Vergelijk dit met de Benford-verdeling in de tabel op bladzijde 1.

3. Verkennen van bestanden en lijsten automatiseren

a. Open het bestand benfordtest.xls. Dit worksheet bevat meer hulpmiddelen voor het onderzoek. Je ziet hier direct de grafiek en de verdeling naar begincijfers van kolom B. Je kunt de werkelijke en relatieve frequenties aflezen in een tabel.

b. In dit Excelsheet kun je ook andere data creëren, invoeren via Copy/Paste of inlezen en direct informatie krijgen over de begincijferverdeling. Als voorbeeld zijn in kolom G van de inwoneraantallen de verschillen met het inwoneraantal van Amsterdam bepaald.
Verander nu de letter B in cel N13 in de letter G.
Je ziet dat de getallen uit kolom G naar de Q-kolom zijn gekopieerd en dat in kolom R de begincijfers zijn berekend. De tabel en grafiek worden uiteraard ook aangepast.

c. Een ander begincijfer komt nu in overgrote meerderheid voor. Hoe zou dat komen?

Hoe je in Excel voor elkaar krijgt dat je via één letter een ander kolom aanspreekt en hoe je in Excel telt hoe vaak een getal in kolom R voorkomt, hoef je niet te weten. Je zou het uit kunnen zoeken met de celformules in Q7 en M22 ... en de handleiding (Help!) van Excel.

Je wilt natuurlijk wel weten hoe het begincijfer van een getal kan worden uitgerekend. Want dat is leuk stukje wiskunde en het heeft niets met de werking van Excel te maken!
Pas op: Het begincijfer van een getal kun je natuurlijk direct zien, als je het getal voor je ziet. Maar heeft de computer niet veel aan als die van een getallenbestand alle begincijfers moet bepalen. Het begincijfer berekenen met een formule is dan nodig, en dat blijkt helemaal niet zo gemakkelijk!

4. Begincijfer van een getal berekenen met een formule

Hoe gaat dat? Stel je voor je hebt een willekeurig getal. In je achterhoofd denk je aan een voorbeeld, bijv. 56089.
Het basisidee voor dit voorbeeld is: deel het getal door een macht van 10, zodat je 5,6089 krijgt en bekijk daarna de gehele waarde.
De macht van 10 vind je met behulp van de ^{10}log en de gehele waarde vind je met behulp van INT (naar beneden afronden).
De log van 56089 is 4,748877… en de INT van 4,748877... is 4.
Als je 56089 deelt door 10^4 krijg je 5,6089. De INT van 5,6089 is 5. En dat is het begincijfer van 56089.

a. Zoek de INT-functie op je GRM. Controleer of de functie werkt zoals boven beschreven.

b. Bepaal met behulp van wat hierboven geschreven is het begincijfer van 243665.

c. Probeer een formule te maken die het begincijfer van een getal bepaalt. Als het je niet lukt, kun je kijken in het bestand formule.pdf.

5. De formule in Excel

a. Test met een zelfgekozen voorbeeldgetal tussen 1000 en 1000000000 dat de formule klopt.

b. Noteer bij een, eventueel ander, voorbeeldgetal alle tussenstappen van de berekening om te laten zien dat de formule inderdaad doet wat in het basisidee van de vorige paragraaf stond.

c. Open het bestand benfordtest.xls. De formule in R7 berekent het begincijfer van het getal in cel Q7. In deze cel staat de formule: INT(Q7/10^INT(LOG10(Q7))).
Ga na dat dit inderdaad dezelfde formule is, vertaald naar Excel.

6. Voortgezet onderzoek 1: schaalverandering

a. In het bestand rivieren.xls staan de lengtes in kilometers van een groot aantal rivieren. Onderzoek of deze rij getallen zich houdt aan de wet van Benford. Je kunt dat op de diverse manieren doen zoals je gezien hebt. Op de manier van paragraaf 2, maar handiger door overnemen van de data via Copy/Paste naar benfordtest.xls.
Kopiëren van data uit een bestand naar benfordtest.xls.
Selecteer in rivieren.xls de cellen die je wilt kopiëren via slepen en Ctrl^C.
Je kunt volstaan met de getallenkolom, maar de namen mogen ook mee.

Ga nu naar benfordtest.xls.
Zoals je kunt zien mag je de kolommen A t/m I gebruiken, maar ook de kolommen T en verder. De data moeten in regel 7 t/m 1006 terecht kommen. Klik dus bijvoorbeeld op cel C7 en tik Ctrl^V. Je data worden nu ingevoerd. In veld N13 geef je de te analyseren kolom aan.

b. De verdeling is weer niet gelijkmatig wat betreft begincijfers. Het zou heel raar zijn als dat een verschijnsel is dat van de grootte van de 'kilometer' afhangt. Maak een extra kolom getallen met de lengtes in 'mijlen'. Onderzoek die ook op verdeling naar begincijfer.

c. Vergelijkbaar onderzoek kun je doen met de lijst van landen en hun oppervlaktes in landopp.xls.

d. Of met de lijst van hoogtes van bergen op
http://nl.wikipedia.org/wiki/Lijst_van_bergen.
(Dit is een bestand dat niet luistert naar de wet van Benford.)

Voortgezet onderzoek 2: schaalveranderingen (zonder computer)

e. Kijk bij wijze van extreem geval eens naar de getallen 1, 2, 3, ... , 99.
Hoeveel van die getallen beginnen met het cijfer 1?
Hoe ziet de hele verdeling naar begincijfer er nu uit?

f. Nu kijken we naar het dubbele van die 99 getallen: 2, 4, 6, ... , 198.
Hoeveel van die getallen beginnen met het cijfer 1?
Hoe ziet de hele verdeling naar begincijfer er nu uit?

Voortgezet onderzoek 3: schaalveranderingen (met computer)

g. Voorbeeld onderzoek: begincijferverdeling bij bevolkingsgroei per gemeente. Keer terug naar het de verdeling van de inwoners over gemeenten.
Pas een groeifactor 1,8 toe. Zet daartoe in cel H7 de formule 1,8 * B7. Door slepen naar beneden kun je de formule de hele H-kolom door kopiëren. Alle H-getallen veranderen dan.
Vraag de Benford-verdeling van deze H-kolom.
Probeer het ook met ander factoren dan 1,8.

Voortgezet onderzoek 4: schaalveranderingen (getallen verdubbelen)

h. Zou een rij die de Benford-verdeling heeft, bij verdubbeling van alle getallen erin weer een nieuwe rij met Benford-verdeling geven?

7. Voortgezet onderzoek 5: random getallen

In het bestand random.xls is een kolom van 1000 getallen opgenomen; die is gemaakt met een randomgenerator op een computer. De getallen zijn behoorlijk gelijk verdeeld over de range 1-10000. Dat betekent dat in het interval 2000-2999 ongeveer 10% van de getallen moet liggen. Bij ideale gelijkverdeling, met veel meer getallen, is de fractie getallen in een interval steeds praktisch gelijk aan de verhouding van dat interval ten opzichte van de range. Er valt wat aan de random getallen te onderzoeken (met behulp van Excel).

a. Open random.xls en onderzoek de 10% claim voor intervallen van 1000 (bijvoorbeeld door sorteren).

b. Onderzoek ook of deze rij getallen zich houdt aan de wet van Benford.
(Lees daartoe de data weer in benfordtest.xls in.)

c. In de C-kolom van random.xls staat niet A1 zelf, maar 100*1,1^(A1/1000), of in een wat normalere formule: in plaats van x staat daar $100 \times 1{,}1^{x/1000}$.
Onderzoek die getallen ook op benford-gedrag.
Varieer de formule, door de constanten 100, 1000 en 1,1 aan te passen en kijk naar het benford-gedrag.

Voortgezet onderzoek 6: de natuurlijke getallen

d. Gebruik in een andere kolom (bijvoorbeeld H) de rij van natuurlijke getallen. Zet daartoe het getal 1 in H7 en daarna de formule = 1 + H7 in cel H8. Kopieer je cel H8 nu naar beneden met slepen, bijvoorbeeld tot cel H1005, dan heb je een eenvoudige rekenkundige rij gemaakt.
Doe het zelfde onderzoek als bij de random-rij van zoeven.
Ook met het transformeren van x naar $100 \times 1{,}1^{x/1000}$; gebruik een nieuwe kolom.

Voortgezet onderzoek 7 :een meetkundige rij

Meetkundige rijen, zijn rijen waarbij elke volgend getal een constante factor groter (of kleiner) is dan het vorige. De constante factor heet ook wel de 'reden' van de rij.
Voorbeeld: 2, 6, 18, 64, ...
De eerste term in deze rij is 2 en de 'reden' is hier 3.

Als een meetkundige rij beginterm a heeft en reden r, dan zijn dit de eerste termen: $a, a\cdot r, a\cdot r^2, a\cdot r^3, \ldots$. De n-de term is dan $a\cdot r^{n-1}$.

e. Denk aan 1000 termen van een meetkundige rij, waarvan de eerste term 1 is en de 1001-ste precies 10 is.
Bereken exact (met hulp van de logaritme).
–de constante factor die bij deze rij hoort
–hoeveel termen kleiner dan 2 zijn
–hoeveel termen kleiner dan 3 zijn,
Enzovoort.

f. Term 2001 van deze rij is exact gelijk aan 100.
Onderzoek ook hoe de eerste-cijfer-verdeling van de termen 1001 t/m 2000 is.

8. De echte Wet van Benford en de Hoofdvraag

Benford gaf stelde niet zijn wet op in de vorm van de tabel hierboven, maar in de vorm van een formule: De relatieve frequentie van begincijfer d in bestanden met 'afwijkende' getallen is: $\log\left(\frac{d+1}{d}\right)$

De hoofdvraag is:
Hoe kan die wet begrijpelijk gemaakt worden en kan aangegeven worden welk type getalbestanden aan de wet voldoen? Met name met behulp van de voorafgaande verkenningen.

In de verklaring zal zeker het verschil in gedrag van de gewone getallen en gewone random getallen vergeleken worden met de getallen die een meetkundige rij vormen en de getallen die ontstaan we als we de random-getallen als exponent gebruiken bij een vast getal dat dichtbij 1 ligt.

Afronding

Maak een *product* waarmee je je docent en medeleerlingen kunt laten zien wat je geleerd hebt van deze keuzeopdracht. Zo'n product kan een poster zijn waarop je de Wet van Benford uitlegt, of een uitwerking van een opgave. Maar iets anders mag ook.
Bedenk met elkaar een vraag die een medeleerling moet kunnen beantwoorden als hij/zij jullie product heeft bestudeerd. Welk(e) antwoord(en) zouden jullie op deze vraag willen krijgen?

9. Bronnen

Het is goed om bij dit onderzoek wat meer op te zoeken over Benford en zijn wet. Toepassingen van de wet vind je in de fraudebestrijding bij banken. In de volgende bronnen vind je daar mogelijk wat meer over.

Algemeen over de wet van Benford: via Google. Zoek gewoon op 'Benford'.

Aanbevelingswaardig zijn
http://nl.wikipedia.org/wiki/Wet_van_Benford (Nederlands, maar kijk ook naar de Engelse Wikipedia, die is vaak veel uitgebreider.)

http://www.inzichten.nl/wetenschap/weten_52.htm
Je vindt daar ook iets over toepassing in de fraudebestrijding.
Op dat gebied is ook nuttig: http://web.uvic.ca/econ/ewp0606.pdf

Databestanden op het web
Informatie over van alles in Nederland bij het Centraal Bureau voor de Statistiek: http://www.cbs.nl

Op deze site van de CIA vind je veel getalsmatige informatie over alle landen van de wereld:
https://www.cia.gov/library/publications/the-world-factbook/index.html

Enkele Nederlandse artikelen
Marleen de Wit en Aad Goddijn: Gemeenten verdelen cijfers oneerlijk. Nieuwe Wiskrant april 1993.
Simon van de Salm: Benfords logaritmische distributie van cijfers. Nieuwe Wiskrant december 2008.
In dat laatste artikel zal vooral het beeld van de ouderwetse rekenliniaal je kunnen helpen ....
Verder staat in beide artikelen ook behoorlijk wat verder gaande informatie.
Altijd goed om te zien dat die er is, zonder dat je verplicht ben die allemaal te begrijpen of te gebruiken!

---einde---