Data-analyse - een poging om de bomen in het bos te zien

In deze bijdrage probeer ik orde te scheppen in het landschap van methoden en technieken voor data-analyse bij de gegevensgerichte controle.

Paul van Batenburg

NBA Handreiking 1141 geeft een overzicht van voorwaarden voor het gebruik, en laat zien hoe data-analyse bijdraagt bij het vinden van controle-informatie in alle fasen van de controle. Wat ik wil toevoegen is een top-down-overzicht van methoden en technieken die bij de gegevensgerichte controle kunnen worden ingezet. Hierbij een poging tot classificatie van die methoden en technieken aan de hand van de norm die wordt gebruikt om gegevensgericht te controleren. Ik pretendeer niet dat dit overzicht volledig is, maar wel dat het inzicht geeft in de verschillende soorten methoden en technieken.

Waar is de norm?

Controleren is toetsen aan een norm. Laten we aannemen dat wij een bestand hebben dat te controleren gegevens Y en eventueel daarmee samenhangende gegevens X bevat. Controleren is het vergelijken van Y met een norm Z. Die norm is soms digitaal aanwezig, maar niet altijd. Dan kan je hem buiten de computer zoeken of in de computer proberen te maken.

Stel: de norm zit in een bestand!

Als die norm ook in een bestand staat, kunnen we met behulp van auditsoftware Y en Z integraal met elkaar vergelijken. Het zou vreemd zijn om dat met een steekproef te doen als het integraal kan. Auteurs die stellen dat data-analyse steekproeven overbodig maakt, dienen zich te realiseren dat dat alleen waar is als de norm in bestandvorm beschikbaar is.

Ik heb in een vorige column een steen in de vijver gegooid door te stellen dat als een accountant integraal kan controleren, hij of zij dat ook zou moeten doen. Die regel, misschien in afgezwakte vorm ('Als een accountant integraal kan controleren en dat niet doet, dient hij of zij aan te geven waarom'), zou in COS 500 niet misstaan.

Toch is dit makkelijker gezegd dan gedaan. Want, wie zegt dat dat bestand Z juist en volledig is? We zien dat Y en Z verschillen, maar weten we zeker dat Y moet worden gecorrigeerd en niet Z? Eigenlijk zijn we weer terug bij af, maar nu met Z als controleobject. Maar, misschien is het gemakkelijker om de bruikbaarheid van Z aan te tonen dan de juistheid van Y. Integrale controle lost het probleem op door het te verplaatsen.

Stel: de norm is niet digitaal beschikbaar!

Dan maar een steekproef?
Als de norm Z niet digitaal beschikbaar is, maar ouderwets op papier staat (of in een magazijnlocatie moet worden geteld) dan is een steekproef (*sample of documentary evidence*) misschien een oplossing. De gewenste onnauwkeurigheid (het verschil tussen de foutkans die de steekproef aangeeft en de werkelijkheid) en onbetrouwbaarheid (de kans dat dat verschil nog groter is) van de steekproef bepalen samen met de vermoedelijke fout de omvang, maar het werkelijk aantal fouten kan die onnauwkeurigheid veranderen. Er zijn toepassingen denkbaar waarbij steekproeven niet efficiënt zijn, met name als de vereiste onnauwkeurigheid erg laag is of als de vermoedelijke fout die norm nadert.

Of we verzinnen wel een norm?
Veel toepassingen van data-analyse in de controle vallen in deze categorie: de norm Z is niet, of niet op efficiënte wijze, bruikbaar en daarom maken we een norm Z'. Voorbeelden volgen, maar voor die tijd is het belangrijk om te beseffen dat als we Z' in plaats van Z als norm hanteren, er sprake kan zijn van ten onrechte goedkeuren of ten onrechte afkeuren zo lang Z' niet perfect met Z overeenkomt.

Normen op basis van Y zelf (Geauzomofo?)
De simpelste manier om zelf een norm op te stellen is het kijken naar extreme waarden: uitschieters, uitbijters, of gewoon de grootste en de kleinste uitkomsten. Alles buiten tweemaal de standaarddeviatie wordt vaak als extreem gezien, maar vereist wel een onderliggende Gausse-verdeling: *Box and Whiskers*-plots (alles buiten de mediaan plus of

min anderhalf keer de kwartielafstand) zijn wat complexer, maar zijn ook toepasbaar op niet-symmetrische verdelingen. Ik vind dat deze norm vaker te onpas dan te pas wordt gebruikt. Extreme waarden bestaan nu eenmaal. In een populatie van een "normaal" (Gauss) verdeelde grootheid zou het juist vreemd zijn als er geen vijf procent uitkomsten buiten tweemaal de standaarddeviatie zouden liggen.

Maar, belangrijker nog is om te beseffen dat plausibiliteit geen juistheid is. Iets dat niet als fout wordt herkend, is nog niet goed omdat we niet zeker weten dat we alle manieren van het herkennen van fouten hebben gebruikt.

Benford-analyse is een voorbeeld van het stellen dat gegevens aan een bepaalde kansverdeling (in dit geval een logistische) moeten voldoen. Alleen die gegevens die van die verdeling afwijken hoeven onderzocht te worden. Mij is geen (empirisch) bewijs bekend dat correcte datasets wel, en gefraudeerde datasets niet aan deze verdeling, of welke andere dan ook, voldoen, en inmiddels is de lijst van mogelijke toepassingen van Benford kleiner dan de lijst met toepassingen waar Benford niet bruikbaar is.

Eigenlijk kan men als data-analist zich hier uitleven op complexiteit: er zijn allerlei vormen van clustering mogelijk, waarbij de populatie wordt opgedeeld in groepen data die binnen elke groep sterk op elkaar lijken, maar tussen de groepen zo veel mogelijk verschillen. De vraag blijft echter of daarmee de onjuiste transacties boven tafel komen: een beetje fraudeur boekt onopvallende transacties, toch?

Belangrijk bij al deze manieren om uit de te controleren data zelf een norm af te leiden, is dat het een norm is om onjuiste gegevens te identificeren en niet om juiste gegevens te vinden. Ik heb dit in een eerdere column GEAUZOMOFO (geautomatiseerd zoeken naar mogelijke fouten) genoemd. In een bestand met geboortedata kun je wel aangeven welke data niet kunnen kloppen (31 februari bijvoorbeeld), maar een datum die niet op die lijst van onmogelijke data staat kan zo niet worden gecontroleerd.

Op grond van risicoanalyse bedenkt de accountant wat er fout kan gaan en hoe een onjuiste transactie eruitziet. De data-analist bouwt vervolgens een tool om die fout op te sporen. Maar wanneer heeft de accountant *alle* fouten gevonden? Die beperking van deze vorm van data-analyse wordt nog weleens over het hoofd gezien, maar COS 500 A52 stelt terecht dat het afwezig zijn van signalen voor onjuistheden nog niet betekent dat de uitkomst klopt.

Normen op basis van (de relatie tussen) Y en X, de overige data
Een geheel andere aanpak is om een norm te halen uit de relatie tussen Y en X, met de gegevens die horen bij dezelfde waarnemingen als de te controleren data. Ook daarvoor geldt weer dat moet worden vastgesteld dat X correct is. Op basis van hoe wij denken dat X invloed heeft op Y wordt een norm opgesteld voor de juiste waarde van Y. Marge-analyse is een eenvoudig voorbeeld daarvan, cijferanalyse kan met behulp van regressie of andere technieken om modellen te schatten uitgroeien tot een zeer complexe uitwerking.

Eigenlijk lijkt cijferanalyse wel heel veel op *machine learning*: de accountant traint een model op gecontroleerde data (X kan dus ook gecontroleerde waarden van Y buiten de huidige controle bevatten) en dat model is de norm voor de te controleren data. Belangrijke vraag bij het gebruiken van een model als norm is dat het model uit te leggen is in bedrijfseconomische termen en of de data X voldoende representatief zijn om als norm voor Y te dienen.

Of, anders gezegd: wanneer het model niet meer te begrijpen is, noemen we het een algoritme. Als we de hoop opgeven om het model dat Z uitspuugt te snappen, wordt het natuurlijk nog belangrijker om zeker te weten dat de data X waarop het model wordt getraind representatief zijn voor de controle op Y.

Deel dit artikel

Drs. Paul van Batenburg is zelfstandig adviseur die als statisticus met verstand van controleren de eenmanszaak en website steekproeven.eu voert.

GERELATEERD

STATISTICAL AUDITING (95) | *25 juli 2022*

De steekproefomvang ontmaskerd - een introductie

Steekproefomvangen berekenen doen we meestal met rekenbladen in Excel of statistische software. Hoewel velen van ons statistiek hebben gehad, is het berekenen van... →

DISCUSSIE | Opinie | *23 juni 2022*

Stille ramp in de Limpergzaal

Iets te laat liep ik op 16 juni de Limpergzaal bij de NBA in voor de Accounttech bijeenkomst. Het thema duurzaamheid stond prominent op de agenda. Ik was dan misschien... →

NIEUWS | *20 juni 2022*

EY gaat 1 miljard dollar investeren in assurance-technologie

EY is van plan in de komende vier jaar ruim één miljard dollar te investeren in een assurance-technologieplatform. Dat moet zorgen voor verdere integratie van bestaande... →

DISCUSSIE | Column | *17 mei 2022*

Drie zorgen over innovatie

Hoewel de grote woorden over Big Data, AI en ML weer wat voorbij lijken, is er op zijn minst een gezonde evolutie gaande in het accountantsvak. Toch is Arnout van... →

NIEUWS | *23 maart 2022*

MAB-themanummer over data-analyse in de controle

De MAB-uitgave 'De controle verklaard: data-analyse, toepassingen en uitdagingen in de accountantscontrole' is beschikbaar. Het is de vierde editie van de nieuwe... →
