Werking van een proces vaststellen met steekproeven*

Wout Schiphorst

15 december 2020

1 Inleiding

Wanneer auditors de werking van een IT- of bedrijfsproces vast willen stellen, nemen ze vaak een deelwaarneming. Op Internet zijn talloze voorbeelden te vinden van tabellen die aangeven hoeveel posten minimaal beoordeeld moeten worden, afhankelijk van hoevaak een proces in een jaar wordt uitgevoerd[1]. Een deelwaarneming is geen steekproef. Omdat met een deelwaarneming doorgaans minder posten worden gecontroleerd, is de zeggingskracht stukken lager dan wat met een steekproef mogelijk is. In dit artikel wordt toegelicht hoe een statistische steekproef zou kunnen worden gebruikt om een uitspraak te doen over de werking van een proces. Ingegaan wordt op de vraag wat een steekproef precies is, uit welke stappen die bestaat en welke alternatieven er zijn.

2 Wat zijn steekproeven?

Bij een steekproef wordt een beperkt aantal posten gecontroleerd. Het aantal moet groot genoeg zijn om een statistisch verantwoorde uitspraak over de hele populatie te kunnen doen. In een geldsteekproef wordt die populatie gevormd door de bedragen die in de jaarrekening worden genoemd. In een postensteekproef bestaat de populatie uit het aantal keren dat zich een gebeurtenis heeft voorgedaan, bijvoorbeeld hoe vaak een proces in een periode is uitgevoerd. In dit artikel gaat het uitsluitend over de postensteekproef. Gecontroleerd wordt dan of een beheersingsmaatregel in alle onderzochte gevallen effectief is geweest, bijvoorbeeld of de impact van een wijzigingsverzoek correct is bepaald.

[1] Zie bijvoorbeeld het Control Framework Horizontaal Toezicht Zorg, juli 2019, `https://www.horizontaaltoezichtzorg.nl/shared/content/uploads/2019/07/htz_control_framework_3.0_def.pdf`, geraadpleegd op 24 november 2020.

De populatie bestaat uit een aantal *elementen* waarvan de *kenmerken* worden bepaald. In het eerder genoemde voorbeeld zijn de elementen de verzameling wijzigingsverzoeken en is het bepalen van de impact een kenmerk. In een steekproef wordt een uitspraak gedaan of een kenmerk van een element goed of fout is en op basis hiervan statistisch getoetst of het aantal fouten in de populatie onder een vooraf vastgestelde grens ligt. Statistici spreken dan van een Bernoulli-experiment. Die uitspraak wordt met een bepaalde *betrouwbaarheid* gedaan. Omgekeerd aan de betrouwbaarheid is de kans dat het oordeel over de hele populatie verkeerd is. Gebruikelijk is een betrouwbaarheid van 95%. Dat betekent dat in hooguit 5% van de gevallen de auditor ten onrechte een goedkeurend oordeel geeft. Verder gaat het oordeel gepaard met een bepaalde *nauwkeurigheid*, oftewel de schatting van het aantal fouten in de populatie. Gebruikelijk is dat ten hoogste 1% fouten in de populatie acceptabel is. Accountants spreken in plaats van nauwkeurigheid over materialiteit.

Het uitvoeren van een steekproef is aan een aantal spelregels gebonden. Ten eerste moet duidelijk zijn welke elementen tot de populatie behoren en welke niet. Ten tweede mag geen misverstand bestaan hoe kenmerken gemeten worden en wanneer die als 'goed' worden bestempeld. Ten derde worden elementen willekeurig en onafhankelijk van de kenmerken geselecteerd om te controleren. Tot slot mag een element pas worden goedgekeurd als voldoende informatie is verzameld om te kunnen concluderen dat aan alle criteria is voldaan.

3 Zes stappen

3.1 Elementen en kenmerken bepalen

In de eerste stap wordt de totale populatie beschreven en afgebakend. Naast duidelijkheid over de elementen waaruit een populatie bestaat, speelt ook de beschouwingsperiode hierbij een rol. Een voorbeeld is alle geregistreerde incidenten in een ticketsysteem, die geclassificeerd zijn als IT-gerelateerd en die in het afgelopen jaar zijn aangemeld. Incidenten die bijvoorbeeld betrekking hebben op facilitaire zaken of die van vijf jaar geleden vallen er dan buiten. Voorwaarden zijn wel dat de registratie juist en volledig is en dat de opzet in de tijd stabiel is. Indien een procedure gedurende de beschouwingsperiode fundamenteel is herzien, dan moeten twee steekproeven worden getrokken: een voor de oude en een voor de nieuwe situatie. Als er sprake is van een diffuse situatie, kan geen steekproef worden getrokken.

Binnen een populatie kan sprake zijn van deelverzamelingen, bijvoorbeeld wijzigingsverzoeken met een lage, midden of hoge impact. Beheersingsmaatregelen kunnen afhankelijk zijn van de deelverzameling waartoe een element behoort. Zo kan een uitgebreide impactanalyse alleen verplicht zijn als de initiële impact hoog is geschat of indien er gevolgen zijn voor de informatiebeveiliging, terwijl die maatregel in andere gevallen niet van toepassing is. In voorkomende gevallen zijn bovendien meerdere niveaus in de populatie te onderscheiden. Geaccepteerde wijzigingsverzoeken kunnen in de realisatie bijvoorbeeld worden samengenomen tot een versie en in een keer worden getest en uitgerold. In die gevallen moet duidelijk zijn op welke niveaus getoetst wordt alvorens de populatie wordt

afgebakend en deelverzamelingen worden onderscheiden.

Naast de elementen moeten ook de kenmerken precies gedefinieerd worden, zodanig dat een element als zijnde 'goed' of 'fout' kan worden gekenmerkt. Een kenmerk kan bijvoorbeeld zijn dat de impact van een wijzigingsverzoek juist moet zijn bepaald. Dat wordt in de praktijk gebracht door de ingevulde impact op het formulier te toetsen aan de criteria in de procedure wijzigingsbeheer. Een element is alleen goed als een ingevulde impact aan alle criteria voldoet en fout als dat niet zo is. Bij ieder kenmerk moet de vraag worden gesteld of de nauwkeurigheid voldoende is. In het geval van een onvervangbare beheersingsmaatregel is iedere afwijking er een teveel en is het uitvoeren van een steekproef niet zinvol.

Het verdient de voorkeur om eerst de opzet te beoordelen en het bestaan vast te stellen aan de hand van een representatief voorbeeld dat de geauditeerde zelf aanlevert. Uitsluitend wanneer de opzet en het bestaan goed zijn, heeft het trekken van een steekproef zin, ook gezien de grote tijdsinvestering die met een dergelijk onderzoek naar de werking gemoeid is. Dat laatste pleit ervoor dat de gegevens over de elementen en kenmerken direct voorhanden moeten zijn en dat in een oogopslag duidelijk is of een kenmerk van een element voldoet of niet.

3.2 Gegevens opvragen

De elementen worden in een of meerdere systemen vastgelegd. Voor ieder te toetsen kenmerk wordt in kaart gebracht in welk systeem de registratie plaatsvindt en op welke locatie, zoals de naam van een bepaald veld. Per systeem wordt een totale lijst van de populatie opgevraagd of aangemaakt, waarbij de uitvraag gedocumenteerd wordt. In het geval van een geautomatiseerd systeem verdient het aanbeveling om de query vast te leggen, inclusief de datum en het tijdstip waarop die is afgevuurd. Dan vindt een globale analyse plaats om een eerste beeld van de data te krijgen: hoe groot is de totale populatie, zijn alle kenmerken aanwezig, komen alle deelverzamelingen terug en zijn er mogelijk meer deelverzamelingen dan eerst onderkend?

3.3 Steekproefomvang vaststellen

In het meest eenvoudige en meest voorkomende geval kan de minimale steekproefomvang met een gegeven betrouwbaarheid en nauwkeurigheid als volgt worden bepaald, namelijk als uit wordt gegaan van nul gevonden fouten in de steekproef om goed te mogen keuren. De minimale steekproefomvang (n) kan met onderstaande formule worden berekend:

$$nauwkeurigheid^n \leq (1 - betrouwbaarheid)$$

Als de gewenste nauwkeurigheid 0,99 is en de betrouwbaarheid 0,95, dan kan proefondervindelijk worden vastgesteld dat n groter dan of gelijk moet zijn aan 300. Immers, als het werkelijk aantal fouten in de populatie inderdaad maximaal 1% is, dan bedraagt de kans dat een willekeurig element goed is minimaal 99%. Als 300 posten op die manier worden gecontroleerd, is de kans lager dan 5% dat de auditor niet zou opmerken indien het werkelijke aantal fouten in de populatie groter is dan 1%.

		Betrouwbaarheid					
		95%	90%	86,4%	80%	63%	50%
Verwachte fouten	0	3,0	2,31	2,0	1,61	1,0	0,7
	1	4,75	3,89	3,5	3,0	2,14	1,68
	2	6,3	5,33	4,88	4,28	3,25	2,68
	3	7,76	6,69	6,18	5,52	4,35	3,68
	4	9,15	7,99	7,45	6,72	5,42	4,67
	5	10,51	9,27	8,69	7,91	6,49	5,67
	6	11,84	10,53	9,91	9,08	7,56	6,67

Tabel 1: R-factor

Lastiger wordt het als de auditor verwacht meer dan een fout in de steekproef aan te treffen. Om dan de minimale steekproefomvang te bepalen, wordt eerst de zogeheten R-factor in een tabel[2] opgezocht. Samen met de nauwkeurigheid wordt de omvang voor de steekproef bepaald. In tabel 1 zijn de R-waarden van enkele combinaties van het verwachte aantal fouten in de steekproef en de betrouwbaarheid weergegeven.

De minimale steekproefomvang wordt berekend door de R-factor te delen door 1 minus de nauwkeurigheid. Uit tabel 1 volgt dat bij een betrouwbaarheid van 95% de R-factor bij nul fouten 3 bedraagt. Bij een nauwkeurigheid van 99% wordt de R-factor gedeeld door 0,01, oftewel vermenigvuldigd met 100, zodat de minimale steekproefomvang 300 is.

3.4 Posten trekken

In de volgende stap worden de elementen willekeurig geselecteerd om daarvan de kenmerken vast te stellen. Bij het trekken van de posten moet zeker zijn gesteld dat iedere deelverzameling in voldoende mate terug komt. Bij voorkeur wordt daar vooraf rekening mee gehouden; er is dan sprake van een gestratificeerde steekproef. Iedere deelverzameling moet in beginsel proportioneel terugkomen in de steekproef. Stel dat 2% van de populatie tot een bepaalde deelverzameling behoort, dat dient hetzelfde percentage ook in de steekproef vertegenwoordigd te zijn. Er kan voor worden gekozen om bepaalde deelverzamelingen sterker te controleren, omdat daar een groter risico mee verbonden is, zoals de incidenten met een hoge impact. In dat geval wordt gesproken over een disproportionele gestratificeerde steekproef.

Bij een Bernoulli-experiment worden elementen teruggelegd nadat ze eenmaal zijn getrokken. In een audit willen we echter normaal gesproken niet een post dubbel controleren. In plaats van een volkomen willekeurige selectie kan beter een interval- of een celsteekproef worden uitgevoerd. In het geval van een intervalsteekproef wordt eerst de

[2]Met een spreadsheetprogramma kan de R-waarde voor iedere combinatie van het verwachte aantal fouten en de gewenste betrouwbaarheid zelf worden berekend. De benodigde functie is de inverse van de cumulatieve gammaverdeling, bijvoorbeeld in LibreOffice Calc: functie GAMMAINV. Op basis van de betrouwbaarheid, het aantal verwachte fouten en de constante 1 geeft deze functie de R-factor.

omvang van het interval bepaald door het aantal posten in de populatie (of deelverzameling) te delen door de steekproefomvang. Als de totale populatie uit 3000 elementen bestaat is en de steekproefomvang is bepaald op 300, dan is het interval 10 groot. Vervolgens wordt willekeurig een getal tussen 1 en de grootte van het interval gekozen, bijvoorbeeld het cijfer 6. De eerste getrokken post is dan de zesde. Voor de volgende wordt telkens een veelvoud van het interval erbij opgeteld, zodat posten 16, 26, 36 enzovoort worden geselecteerd. Een intervalsteekproef heeft als nadeel dat de uitkomsten mogelijk een verkeerd beeld geven als er patronen in de gegevens zitten. Om dat te ondervangen kan beter een celsteekproef worden gedaan. De populatie wordt in cellen ter grootte van een interval opgedeeld en binnen iedere cel wordt een willekeurig getal gekozen. Dan kan bijvoorbeeld de zesde post in de eerste cel worden geselecteerd en de negende in de tweede (= de negentiende post in de populatie).

De selectie geschiedt per deelverzameling met gebruikmaking van een getallengenerator, zoals die in een spreadsheetprogramma of op Internet is te vinden. Aangezien naderhand kan blijken dat een element niet onderzocht kan worden, moeten op voorhand extra elementen geselecteerd worden, bijvoorbeeld 10% bovenop het minimum. Dit kan zich voordoen indien een geselecteerde post achteraf niet van toepassing blijkt, zoals een wijzigingsverzoek dat is ingediend maar uiteindelijk niet is doorgezet.

3.5 Posten controleren

Nadat de posten geselecteerd zijn kan de controle daadwerkelijk beginnen. De controle kan het beste in tabelvorm worden vastgelegd met in de rijen de elementen, in de kolommen de kenmerken en in de cellen de uitspraak. De controle is ten einde als na afstemming met de geauditeerde alle cellen de waarden 'goed' of 'fout' hebben. In het geval van 'niet van toepassing' wordt de reden gedocumenteerd en aansluitend een nieuw element geselecteerd. Als aan het eind vragen onbeantwoord blijven, moet worden geconcludeerd dat er onvoldoende geschikte controle-informatie is.

Het trekken van een statistische steekproef kost veel tijd. In de praktijk worden minder elementen geselecteerd dan de minimum steekproefomvang, zoals dertig posten als een proces dagelijks of vaker wordt uitgevoerd. De redenering hierachter is dat de auditor gebruik maakt van voorkennis over de interne beheersing bij de controleklant. Indien de geauditeerde zijn risico's goed in de greep heeft, dan verkleint dat ook het risico voor de auditor dat hij een verkeerd oordeel geeft. In dergelijke gevallen wordt de minimumsteekproefomvang gehandhaafd, maar vindt reductie plaats door een aantal posten ongezien goed te keuren. Stel dat de steekproefomvang op 300 is bepaald en slechts 30 posten worden gecontroleerd, dan wordt dus aangenomen dat de resterende 270 posten goed zullen zijn.

Een andere strategie om minder tijd aan de steekproef te besteden is om naar mogelijkheden te zoeken om het onderzoek te automatiseren. Dat kan als alle gegevens digitaal en gestructureerd vastliggen. Controles kunnen geautomatiseerd worden door in een spreadsheet de inhoud van cellen met elkaar te vergelijken en vast te stellen dat ze verschillen om het vier-ogen-principe vast te stellen of door de inhoud van een cel te vergelijken tegen een stamgegeven, bijvoorbeeld de namen van functionarissen die mo-

gen goedkeuren. Indien de controle helemaal geautomatiseerd kan worden, is zelfs een integrale controle mogelijk.

3.6 Oordeel vormen

Per element wordt bepaald of alle kenmerken de waarde 'goed' hebben, in welk geval geconcludeerd wordt dat er geen afwijkingen zijn. Als er een kenmerk de waarde 'fout' heeft, dan is sprake van een afwijking. Voor de hele steekproef wordt het aantal afwijkingen geteld en vergeleken met het verwachte aantal fouten in de steekproef. Als het werkelijke aantal afwijkingen groter is dan de verwachting, luidt het oordeel over de steekproef na hoor en wederhoor in principe afkeurend. Bij een disproportionele gestratificeerde steekproef moet bij een afkeurend oordeel wel een nadere analyse plaatsvinden, omdat niet zonder meer kan worden gesteld dat de hele populatie meer fouten bevat dan is toegestaan. Een mogelijke oplossing is om het afkeurende oordeel te beperken tot bepaalde deelverzamelingen en om vast te stellen wat nu precies het risico is. Indien het een assurance-opdracht betreft, kan de geauditeerde in de gelegenheid worden gesteld om de afwijkingen te corrigeren en een nieuwe steekproef uit te voeren, opdat een goedkeurend oordeel wel mogelijk is. Ook bij een interne audit waarin een oordeel over de interne beheersing wordt gegeven bestaat deze optie. Het is dan wel zaak om hier in het rapport melding van te maken: er zijn meer afwijkingen aangetroffen dan de norm, maar herstelacties zijn in gang gezet. Het oordeel over de steekproef moet in samenhang worden gezien met dat over de andere auditwerkzaamheden, zoals over opzet en bestaan.

Indien tijdens de uitvoering van de steekproef blijkt dat het aantal fouten de verwachting overstijgt, doet zich de vraag voor of doorgaan zin heeft. Overleg met de opdrachtgever en de geauditeerde is dan nodig. Er zijn twee mogelijkheden: (1) de steekproef wordt voortijdig afgebroken met een afkeurend oordeel of (2) de steekproef wordt voortgezet om de omvang van de afwijkingen te kunnen schatten en de mogelijke oorzaken te analyseren. De laatste mogelijkheid heeft de voorkeur, hoewel met de eerste een besparing in tijd en geld kan worden behaald. Bij de analyse van de mogelijke oorzaken gaat het om de vraag of de afwijkingen zich voordoen in bepaalde elementen (rijen) of bepaalde kenmerken (kolommen). Het zou bijvoorbeeld kunnen dat een afwijking op een bepaalde afdeling of in een periode is terug te voeren (elementen) of dat een zekere beheersingsmaatregel niet heeft gewerkt en de andere wel (kenmerken). Het kan ook betekenen dat er hiaten in de opzet zijn.

4 Wel of geen steekproef?

Het uitvoeren van een statistische steekproef is geen sinecure en vraagt om een grote investering door de opdrachtgever. Een statistische steekproef is echter niet verplicht en er bestaan alternatieve technieken om de werking vast te stellen. In een niet-statistische steekproef of deelwaarneming zou een auditor met gebruik van voorkennis gericht posten kunnen selecteren als daar grotere risico's aan verbonden zijn. Daar moet dan wel rekening mee worden gehouden als op basis daarvan een uitspraak over de hele populatie wordt gedaan, omdat de representativiteit van de steekproef niet gewaarborgd is. Hoewel

reductie vaak wordt toegepast en niet-statistische steekproeven veelal worden uitgevoerd, moet bedacht worden dat een steekproefomvang van slechts dertig elementen ceteris paribus gepaard gaat met een betrouwbaarheid van 26% of een nauwkeurigheid van 90%. Dat betekent dat de kans dat de auditor een verkeerde conclusie trekt maar liefst 74% bedraagt of de populatie tot 10% fouten kan bevatten.

Naast steekproeven kunnen auditors ook andere technieken inzetten, zoals het leggen van verbanden tussen verschillende registraties, het vergelijken met vorige perioden of de controle tegen een plan of begroting. Dit is vergelijkbaar met cijferanalyses die accountants toepassen. In een IT-audit kan bijvoorbeeld gedacht worden aan het aantal personele mutaties volgens de personeelsafdeling en het aantal autorisatieaanvragen, waartussen een bepaald verband mag worden verwacht. Een ander voorbeeld is dat wijzigingsaanvragen in een ticketsysteem worden vastgelegd, terwijl ontwikkelaars andere systemen gebruiken om code te ontwikkelen, te testen en vrij te geven. Ook hiervoor geldt dat op totaalniveau verbanden moeten kunnen worden gelegd tussen de registraties. Binnen een registratie kan worden geanalyseerd welk type wijzigingen zich hebben voorgedaan en of een relatie kan worden gelegd met grote projecten in een organisatie. Tevens kan het aantal wijzigingen gerelateerd worden aan het aantal ontwikkelaars en worden beoordeeld of dat redelijk is en in lijn met vorige perioden. Uitzonderingen in de data kunnen worden verkend, bijvoorbeeld of het aantal incidenten over de tijd constant is of dat de wijzigingsverzoeken gelijkelijk over de uitvoerende medewerkers verdeeld zijn. In sommige gevallen is sprake van continue monitoring, zoals voor het meten van de beschikbaarheid van systemen. Het spreekt voor zich dat een steekproef dan achterwege kan blijven.

Dankwoord

Met dank aan Frank Nillesen die een eerdere versie van dit artikel heeft gelezen en van commentaar heeft voorzien.

Referenties

Kloosterman, H. en van Batenburg, P. (2018), *Essaybundel Statistical Auditing*, Publishing House Jacques de Swart.

De inhoud van dit artikel is gebaseerd op de NBA-cursus *Steekproeven: Audit efficiency met statistische steekproeven* door Hein Kloosterman en Ferry Geertman en de bijbehorende essaybundel zoals hierboven genoemd.