Les 13 - Data verwerken en klaarmaken voor rapportage (incl checks en weging) en verdieping op SPSS, evt R?

In dit hoofdstuk hebben we het over dataverwerking vanaf het moment dat er genoeg respondenten hebben deelgenomen aan het onderzoek (het veldwerk vol is). Een deel zal gaan over statistisch onderzoek. Ben je hier al bekend mee vanuit studie of eerder werk, dan kun je hier snel doorheen gaan. Vooraf zal ik aangeven dat het over een eventuele 'recap' gaat.

Analyseprogramma's

Programma's waarin gewerkt wordt bij onderzoeksbureaus voor dataverwerking zijn onder andere: SPSS, R en Python. Vanuit je studie of eerdere loopbaan ben je ongetwijfeld al minimaal met een van deze programma's bekend. Er zijn verschillen per programma in de mogelijkheden en bedrijven hebben verschillende voorkeuren.

Wat zijn de voor- en nadelen?

SPSS kan je zien als het meest gebruiksvriendelijke programma. De lay-out is zeer toegankelijk en met relatief weinig kennis kom je met SPSS het verst, maar omdat nieuwe toepassingen pas worden doorgevoerd wanneer deze door het programma zijn goedgekeurd, kunnen innovaties wat langer op zich laten wachten. R daarentegen werkt met zogenoemde pakketten. Deze pakketten kan men zelf ontwikkelen en zijn online te downloaden voor persoonlijk gebruik. Python is tevens open source en heeft uiteindelijk de meeste functionaliteiten. Hiervoor is wel extra kennis vereist en is lastiger snel onder de knie te krijgen dan bijvoorbeeld SPSS.

In deze module zullen we ons richten op SPSS, omdat je hier waarschijnlijk het meest mee te maken zult krijgen, tenzij je op een meer data science of machine learning-functie terecht zult komen.

Data-check/ kwaliteitscontrole

Zodra een onderzoek vol is, dat wil zeggen: de netto N is bereikt, kan je met de data aan de slag. Eerst wordt de data gecheckt op de volgende zaken:

- Zijn alle vragen beantwoord? In principe heb je tijdens de soft launch (zie eerdere module) natuurlijk al een check uitgevoerd, dus als het goed is zou je hier niks vreemds moeten vinden.
- Zijn de open vragen goed ingevuld? → non-antwoorden verwijderen.
- Controleren op straightlining → personen die overal hetzelfde invullen of overal 'weet niet' noemen, valt onder straightlining. Sommige bureaus hebben voor deze check speciale variabelen aangemaakt (flags). Check voor deze respondenten of hun antwoorden betrouwbaar zijn, zo niet, dan zul je deze respondenten moeten verwijderen.
- Check op invulduur → Voor een onderzoek staat altijd een x aantal minuten. Gemiddeld kan je uitgaan van 3 vragen per minuut. Als een respondent bijvoorbeeld 1 minuut doet over een vragenlijst met 30 vragen (in de meeste gevallen niet aan te raden), dan kan je ervan uitgaan dat de kwaliteit van de antwoorden niet goed is en kan je deze respondent beter verwijderen.

Omdat bovenstaande altijd in een bepaalde mate voorkomt in een dataset, doe je er verstandig aan altijd een aantal extra respondenten binnen te halen. Afhankelijk van de kwaliteit van het panel ligt dit op ongeveer 10 a 20 op de 1000. Op die manier is er ruimte om respondenten te verwijderen die het onderzoek niet serieus hebben ingevuld. Bij klantbestanden ligt het anders. Hierbij wordt geen totale N gebruikt, maar probeer je zoveel mogelijk klanten in je onderzoek te krijgen. Slechte invullers worden hier uiteraard wel nog gefilterd.

Wegen

Weeg je en zo ja waarop? Het wegen van data wordt gedaan om kleine afwijkingen in een dataset recht te trekken, zodat deze overeen komt met de werkelijke populatie waar je onderzoek naar doet.. (op onder andere sociaal-demografische kenmerken, op marktaandelen of wanneer je een zakelijke groep hebt bijvoorbeeld op bedrijfsgrootte of branche.) Kortom: het is dus van belang dat je de verhoudingen van bepaalde achtergrondkenmerken van je doelgroep helder zijn.

Gaat het om de totale populatie, dan kan de Gouden standaard je verder helpen. Dit is een jaarlijks geüpdatet bestand wat tot stand wordt gebracht door de MOA (Marktonderzoek Associatie) en het CBS (Centraal Bureau voor de Statistiek) waarin de achtergrondkenmerken zijn opgenomen van de Nederlandse populatie. (voorbeeld gouden standaard). Wanneer je niet in het bezit bent van een Gouden Standaard, dan kan je via het CBS ook veel vinden.

Bijvoorbeeld een bedrijf dat graszaad aan particulieren verkoopt, heeft een specifieke doelgroep. Personen in de grote steden hebben minder vaak een tuin, dus woonomgeving wijkt mogelijk af van het gemiddelde in Nederland. Heeft het bedrijf hier al eerder onderzoek naar gedaan, dan kan de opdrachtgever deze cijfers aanleveren.

Afhankelijk van het soort onderzoek 'kies' je een aantal variabelen die van belang zijn in je onderzoek. In een standaard onderzoek is dit meestal: geslacht, leeftijd, opleiding en soms regio. Zijn andere achtergrondkenmerken van belang voor het onderzoek, dan kun je deze ook meenemen in je weging. Denk bijvoorbeeld aan etniciteit, professie, sociale klasse of het wel/niet klant zijn van een bepaald merk. (Zoals al eerder in het hoofdstuk 'Steekproef trekken' langskwam, hier houd je dus vooraf al rekening mee en bepaal je niet pas wanneer alle data binnen is.) Let wel op: hoe meer variabelen je meeneemt, met hoe meer eventuele afwijkingen er tijdens de weging rekening gehouden moet worden. Probeer dit waar mogelijk dus te beperken.

2 opties van de gouden standaard:

1. Als je alleen leeftijd, geslacht en opleiding nodig hebt kan je werken met de standaard draaitabel. Je selecteert de groepen waarin je leeftijd wil indelen. Vervolgens update je de draaitabel en je ziet de percentages per groep. Hierin kan je kiezen of je per cel wil uitsplitsen of per randtotaal.

2. Heb je meer input nodig? Dan moet je gebruik maken van de aparte databestanden. In een overzicht kan je vinden welk databestand je nodig hebt. Dit is afhankelijk van de variabelen die je wilt meenemen.

Je kunt op verschillende manieren wegen: in marktonderzoek zal je het vaakst tegenkomen op celniveau of op randtotalen. Dat ziet er als volgt uit. (hier een tabel zodat de verschillen zichtbaar zijn op geslacht en provincie).

Bij celniveau wordt per vakje gewogen. Bijvoorbeeld: een vrouw, uit Noord-Holland. Bij randtotalen gaat het er alleen om dat het percentage vrouwen klopt ten opzichte van het percentage mannen en het percentage Noord-Holland ten opzichte van de overige provincies.

In de meeste gevallen wordt gebruik gemaakt van randtotalen. Celniveau is zo specifiek dat hiervoor een extra groot sample nodig is, omdat je heel specifiek kijkt naar een groep. (zie afbeelding 1: Gouden Standaard splits. De meest rechter kolom en onderste rij laten randtotalen zien. Grote gedeelte laat de celniveau percentages zien).

		OPLEIDING (GENOTEN)			TOTAAL
Geslacht	Leeftijd	laag (tm MAVO)	midden	hoog (vanaf HBO)	100,0%
Man	13-17	2,3%	1,3%	0,0%	3,6%
	18-34	0,9%	5,6%	6,0%	12,5%
	35-49	1,8%	4,7%	4,8%	11,3%
	50-64	2,7%	5,0%	4,4%	12,1%
	65 en ouder	3,3%	3,7%	3,0%	10,0%
Vrouw	13-17	2,0%	1,3%	0,0%	3,4%
	18-34	0,6%	4,9%	6,7%	12,2%
	35-49	1,7%	4,7%	4,9%	11,4%
	50-64	3,4%	5,2%	3,5%	12,1%
	65 en ouder	6,1%	3,4%	1,9%	11,4%
					0,0%
					0,0%
					0,0%
					0,0%
TOTAAL		24,8%	40,0%	35,2%	100,0%

Afbeelding 1: overzicht Gouden Standaard splits.

Weegvoorwaarden

Er zijn geen harde regels voor het wegen, maar bureaus (soms in overleg met de klant) houden vaak wel een maatstaf aan:

1. Maximale weegfactor die tijdens het wegen bereikt mag worden. De weegfactor is het maximale aantal keer dat een respondent meetelt in het onderzoek.

Er is niet een standaard voor de maximale weegfactor die door alle bureaus wordt gehanteerd. Persoonlijk zou ik niet tevreden zijn met een weegfactor hoger dan 2 omdat een respondent zijn mening anders te zwaar gaat meetellen. Andere bureaus hanteren 3. In elk geval belangrijk dat je dit afstemt met de klant wanneer deze relatief meer onderzoekservaring heeft. (zie afbeelding 2: weegfactor, kolom 'load').

2. Minimale weeg efficiency. De weeg efficiency is het percentage van je steekproef dat na je weging buiten de foutmarge valt. Een weegefficiency van onder de 70% zou ik persoonlijk niet goedkeuren. Tussen de 70%-80% is oke, maar liever heb je een efficiency van boven de 90%. (zie afbeelding 3: effective base)

Weighting report

Variable nr.	Variable label	Value nr.	Value label	Target	Sample	Weighted	Load	Fit
1	Geslacht	1	1	0,4940	0,4900	0,4940	1,0082	1,0000
		2	2	0,5060	0,5100	0,5060	0,9922	1,0000
2	Leeftijd	1	1	0,1350	0,1310	0,1350	1,0305	1,0000
		2	2	0,1540	0,1430	0,1540	1,0769	1,0000
		3	3	0,1440	0,1450	0,1440	0,9931	1,0000
		4	4	0,1770	0,1830	0,1770	0,9672	1,0000
		5	5	0,1640	0,1680	0,1640	0,9762	1,0000
		6	6	0,2260	0,2300	0,2260	0,9826	1,0000
3	Opleiding	1	1	0,2150	0,2070	0,2150	1,0386	1,0000
		2	2	0,4050	0,4060	0,4050	0,9975	1,0000
		3	3	0,3800	0,3870	0,3800	0,9819	1,0000
4	Regio	1	1	0,1600	0,1390	0,1600	1,1511	1,0000
		2	2	0,2900	0,3130	0,2900	0,9265	1,0000
		3	3	0,1000	0,1020	0,1000	0,9804	1,0000
		4	4	0,2100	0,2100	0,2100	1,0000	1,0000
		5	5	0,2400	0,2360	0,2400	1,0169	1,0000

Afbeelding 2: weegfactor

Summary

Sample size	1000	
Effective base	993 = 99,30%	Effective base = (sum of weight factors) squared / sum of the squared weight factors

Algorithm:

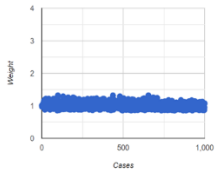
Sequence:	1;2;3;4
Iterations:	14
Optimal fit:	Yes
Max factor:	disabled
Population size:	disabled

Weights:

Min	0,86
Max	1,31
Average	1,00

Distribution:	Sample	Weighted
> 3,0	-	-
2,5 - 3,0	-	-
2,0 - 2,5	-	-
1,5 - 2,0	-	-
1,0 - 1,5	0,429	0,461
0,7 - 1,0	0,571	0,539
0,4 - 0,7	-	-
0,2 - 0,4	-	-
< 0,2	-	-

Scatterchart



Afbeelding 3: effective base

Het implementeren van de weging

Uit je huidige databestand selecteer je de variabelen waarop je gaat wegen (bijvoorbeeld geslacht en leeftijdsgroep. Wanneer je gaat wegen is het belangrijk dat je niet vergeet een koppelvariabele mee te nemen. Dit is een variabele die je meeneemt in het weegbestand zodat je later weet welk weeggetal bij welke respondent hoort. Je maakt een weegbestand aan met de koppelvariabelen en de achtergrondvariabelen waarop je gaat wegen. Er zijn verschillende programma's ontwikkeld waarmee je kunt wegen, maar globaal werken ze hetzelfde. Het programma voegt per respondent een weegfactor toe. (Dit onderwijl laten zien in een filmpje van SPSS).

Wanneer zet je de weging aan en wanneer niet? (N benoemen in rapport)

Wanneer je resultaten uitdraait zet je uiteraard de weging aan. Voor de N per vraag of over het totaal van het onderzoek wordt altijd de ongewogen N weergegeven. Daarnaast kan je in de rapportage een onderzoeksverantwoording opnemen waarin je de weging weergeeft. Laat dit afhangen van de klant, het kan namelijk ook juist verwarrend werken. (zie afbeelding 4: wegingsoverzicht).

Uitdraai weegvariabelen voor weging

Geslacht

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Man	490	49,0	49,0	49,0
	Vrouw	510	51,0	51,0	100,0
	Total	1000	100,0	100,0	

Leeftijd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	16-24 jaar	131	13,1	13,1	13,1
	25-34 jaar	143	14,3	14,3	27,4
	35-44 jaar	145	14,5	14,5	41,9
	45-54 jaar	183	18,3	18,3	60,2
	55-64 jaar	168	16,8	16,8	77,0
	65 jaar of ouder	230	23,0	23,0	100,0
	Total	1000	100,0	100,0	

Opleiding

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Laag	207	20,7	20,7	20,7
	Midden	406	40,6	40,6	61,3
	Hoog	387	38,7	38,7	100,0
	Total	1000	100,0	100,0	

Regio

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nielsen 1	139	13,9	13,9	13,9
	Nielsen 2	313	31,3	31,3	45,2
	Nielsen 3	102	10,2	10,2	55,4
	Nielsen 4	210	21,0	21,0	76,4
	Nielsen 5	236	23,6	23,6	100,0
	Total	1000	100,0	100,0	

Uitdraai weegvariabelen na weging

Geslacht

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Man	494	49,4	49,4	49,4
	Vrouw	506	50,6	50,6	100,0
	Total	1000	100,0	100,0	

Leeftijd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	16-24 jaar	135	13,5	13,5	13,5
	25-34 jaar	154	15,4	15,4	28,9
	35-44 jaar	144	14,4	14,4	43,3
	45-54 jaar	177	17,7	17,7	61,0
	55-64 jaar	164	16,4	16,4	77,4
	65 jaar of ouder	226	22,6	22,6	100,0
	Total	1000	100,0	100,0	

Opleiding

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Laag	215	21,5	21,5	21,5
	Midden	405	40,5	40,5	62,0
	Hoog	380	38,0	38,0	100,0
	Total	1000	100,0	100,0	

Regio

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nielsen 1	160	16,0	16,0	16,0
	Nielsen 2	290	29,0	29,0	45,0
	Nielsen 3	100	10,0	10,0	55,0
	Nielsen 4	210	21,0	21,0	76,0
	Nielsen 5	240	24,0	24,0	100,0
	Total	1000	100,0	100,0	

Afbeelding 4: Wegingsoverzicht

(recap)
Analyse en rapportage

Van tevoren is bepaald wat het doel is van een onderzoek. Je 'duikt' dus niet zomaar in de data, maar gaat gericht te werk op basis van hypotheses. In het marktonderzoek kan je denken aan: Heeft de televisie-inzet van de campagne een positief effect op de geholpen naamsbekendheid van merk x? Hebben vrouwen hebben meer interesse in dit product dan mannen? X% van de doelgroep staat open voor online boodschappen.

Afhankelijk van het soort onderzoek wordt via SPSS tabellen uitgedraaid zodat je per categorie kan zien of er verschillen zijn of via een dashboard wordt een bestand ingeladen dat op basis van de door jou aangegeven splitsingen tabellen en grafieken creëert. (in het volgende hoofdstuk zullen we verder ingaan op dashboarding).

Significante verschillen

Voor het berekenen van significanties hebben de meeste bureaus handige tools ontwikkeld of kun je deze gemakkelijk uitdraaien via SPSS. We gaan hierom verder niet in op de onderliggende formules, maar bespreken enkel de zaken waarmee je te maken krijgt tijdens het proces. Wellicht is (een deel) van deze informatie al bekend voor je. Je kunt hier in dat geval snel doorheen gaan.

Wat is significantie?

Significantie wordt in de statistiek gebruikt om aan te geven of een resultaat al dan niet berust op toeval. Draai altijd significanties uit over je resultaten via SPSS, maar focus je op de hypotheses die je vooraf hebt opgesteld. Overigens zijn niet uitsluitend significante resultaten van belang. Dus zijn resultaten niet significant, maar wel relevant, dan kun je deze alsnog rapporteren. Zo kan een trend over tijd bijvoorbeeld relevant zijn om te rapporteren of juist een verwachting die niet significant blijkt, juist daarom een goed inzicht geven dat deze verwachting niet overduidelijk terug te zien blijkt.

Betrouwbaarheidsinterval

Meestal wordt gewerkt met een 95% betrouwbaarheidsinterval. Dit interval wordt ook in de wetenschap het vaakst gebruikt. Dit wil zeggen dat wanneer je het onderzoek nogmaals uitvoert, en je zou 100 steekproeven trekken uit dezelfde populatie, je kan verwachten dat in 95 van de berekende intervallen het populatiegemiddelde zich hierin bevindt.

Eenzijdig of tweezijdig toetsen

Wanneer je een significantie wil berekenen, wordt je gevraagd of je eenzijdig of tweezijdig wil toetsen. Of je eenzijdig of tweezijdig toetst, hangt af van de hypothese die je hebt opgesteld. (hier voorbeeld van een normaalverdeling tonen en de standaard afwijking bij eenzijdig en tweezijdige toetsing). Weet je niet hoe je hypothese zal uitvallen, dan toets je tweezijdig. Het betrouwbaarheidsinterval kan in dat geval aan beide kanten van het verwachte gemiddelde vallen. Is de hypothese eenduidig, dan maak je gebruik van een eenzijdige toets. Je betrouwbaarheidsinterval valt dan aan een kant van je normaalverdeling.