



---

## **Richtlijnen voor output uit de microdataomgeving van het CBS**

Versie: januari 2026



---

## Inhoud

|                                                          |           |
|----------------------------------------------------------|-----------|
| <b>1. Inleiding en achtergrond</b>                       | <b>3</b>  |
| <b>2. Richtlijnen</b>                                    | <b>5</b>  |
| 2a. Richtlijnen met betrekking tot onthulling            | 5         |
| 2b. Richtlijnen met betrekking tot leesbaarheid          | 7         |
| 2c. Richtlijnen met betrekking tot technische vereisten  | 8         |
| <b>Bijlage 1: Voorbeelden beveiligingsrichtlijnen</b>    | <b>9</b>  |
| 1. Minimum aantal waarnemingen                           | 9         |
| 2. Modellen                                              | 11        |
| 3. Groepsonthulling                                      | 12        |
| 4. Dominantie                                            | 12        |
| 5. Aanvullend: Figuren                                   | 13        |
| <b>Bijlage 2: Standaardformulier beschrijving output</b> | <b>15</b> |
| <b>Bijlage 3: CBS-wet</b>                                | <b>16</b> |



## 1. Inleiding en achtergrond

Het is van groot belang dat data goed wordt beveiligd. Binnen het CBS wordt data zowel fysiek als juridisch beveiligd, ook voor gebruik van data binnen de CBS-microdataomgeving. Deze beveiliging wordt verzorgd via afscherming van de CBS-microdataomgeving, via contracten die CBS sluit met partijen die de CBS-microdataomgeving gebruiken en door statistische beveiliging, om te voorkomen dat individuele gegevens onthuld worden; dat wil zeggen herkenning van een eenheid én openbaarmaking van nadere gegevens over die eenheid. Deze richtlijnen hebben hoofdzakelijk betrekking op de statistische beveiliging, dus op het voorkomen van onthulling.

Al het gebruik van de microdataomgeving van het CBS wordt ingekaderd door enkele artikelen uit de 'Wet op het Centraal Bureau voor de Statistiek'.

### **CBS-wet (verkorte weergave)**

#### Artikel 37

De door het CBS verzamelde gegevens worden uitsluitend gebruikt voor statistische doeleinden en niet verstrekt aan anderen dan degenen die belast zijn met de uitvoering van de taak van het CBS.

Vervolgens worden de gegevens slechts zodanig openbaar gemaakt dat daaraan geen herkenbare gegevens over een afzonderlijk persoon, huishouden, onderneming of instelling kunnen worden ontleend.

#### Artikel 41

In afwijking van artikel 37 kan de directeur-generaal op verzoek, ten behoeve van statistisch of wetenschappelijk onderzoek, een verzameling van gegevens met betrekking tot het gebruik waarvan passende maatregelen zijn genomen om herkenning van afzonderlijke personen, huishoudens, ondernemingen of instellingen te voorkomen, verstrekken aan een dienst, organisatie of instelling, dan wel daartoe toegang verlenen.

#### Artikel 42

De directeur-generaal wiligt een verzoek als bedoeld in artikel 41 slechts in, indien de verzoeker naar het oordeel van de directeur-generaal voldoende maatregelen heeft getroffen om te voorkomen dat de verzameling van gegevens voor andere doeleinden dan statistisch of wetenschappelijk onderzoek wordt gebruikt.

*(zie bijlage 3 voor de volledige tekst)*

Naast deze wettelijke verplichting staat voorop dat de reputatie die de buitenwereld van het CBS heeft er één moet zijn van absolute betrouwbaarheid. De respons op enquêtes is ook afhankelijk van de belofte van het CBS dat individuele gegevens in veilige handen zijn. Wie wil er immers nog persoonlijke informatie verschaffen als er niet extreem zorgvuldig met deze data wordt omgegaan?

Enkele opmerkingen vooraf:

- Het voorkomen van onthulling valt onder gedeelde verantwoordelijkheid van het CBS en de onderzoeker die met de microdata<sup>1</sup> werkt. Zie hiervoor ook de geheimhoudingsverklaring.
- De outputcontroleurs beoordelen elke output afzonderlijk op onthulling. Of er in combinatie met eerder vrijgegeven output (ongeacht project of instelling) een risico op onthulling ontstaat, is de

<sup>1</sup> Microdata: data gebaseerd op enquêtes en registraties over individuele entiteiten, zoals personen, bedrijven en instellingen.



verantwoordelijkheid van de onderzoeker. Deze combinatie wordt niet gecontroleerd door de outputcontroleurs.

- De beveiligingsregels gelden voor datasets over zowel personen als bedrijven als andere instellingen, zoals scholen en ziekenhuizen.
- Voor een project worden alleen die bestanden beschikbaar gesteld die nodig zijn voor het beantwoorden van de onderzoeksvraag. De aangeboden output moet doelbinding hebben met het project en dus nauw gerelateerd zijn aan de onderzoeksvraag.
- Wanneer eigen bestanden worden geupload op de microdataomgeving van het CBS moet de output hiervan eveneens aan de outputrichtlijnen van het CBS voldoen.
- Naast de outputrichtlijnen bestaan er voor sommige CBS-microdatabestanden ook publicatierichtlijnen. Deze zijn terug te vinden in het documentatierapport van de desbetreffende bestanden. Bij de outputcontrole wordt niet gecontroleerd op publicatierichtlijnen. Het is de verantwoordelijkheid van de onderzoeker om te zorgen dat de output gepubliceerd wordt volgens de publicatierichtlijnen.

In de volgende hoofdstukken worden drie aspecten van de richtlijnen besproken. In hoofdstuk 2a worden de *beveiligingsrichtlijnen* besproken om te voorkomen dat er onthulling plaatsvindt. In hoofdstuk 2b wordt aandacht besteed aan de *leesbaarheid* van de output, zodat deze goed te controleren is door de outputcontroleurs. In hoofdstuk 2c worden de *technische* richtlijnen beschreven.



## 2. Richtlijnen

### 2a. Richtlijnen met betrekking tot onthulling

*Dit onderdeel van de richtlijnen is bedoeld om duidelijkheid te scheppen over welke output is toegestaan, waarbij gecontroleerd wordt op onthullingsrisico.*

*De gehanteerde normen zijn bewust aan de veilige kant gekozen, zodat een aantal veelvoorkomende beveiligingsrisico's bij voorbaat vrijwel volledig kunnen worden uitgesloten. Als minder strikte richtlijnen zouden worden toegepast, zou de onderzoeker bij elke output aanzienlijk meer achtergrondinformatie moeten aanleveren om een zorgvuldige controle mogelijk te maken.*

*Indien een onderzoeker output wil aanleveren die niet voldoet aan de richtlijnen, rust de bewijslast primair bij de onderzoeker om aan te tonen dat er desondanks geen sprake is van een onthullingsrisico. In dat geval moet vóór het aanbieden van de output expliciet toestemming worden gevraagd voor de desbetreffende uitzondering.*

Wij hanteren de volgende beveiligingsrichtlijnen:

0. **Geen data op individueel niveau:** Het is niet toegestaan om microdata als output aan te bieden. Denk hierbij aan:
  - individuele koppelsleutels die niet opgenomen mogen worden in de output, zoals RINPERSOON, RINOBJECTNUMMER, BEID, BRIN;
  - namen van bedrijven, scholen, instellingen, etc. die niet in de output mogen voorkomen;
  - aantallen op basis van één instelling, ook al anoniem, die niet zijn toegestaan.
1. **Minimum aantal waarnemingen:** Alle tabellen en soortgelijke output bevat **tenminste 10** eenheden (ongewogen) als grondslag voor elke cel of datapunt bevatten. Aantallen kleiner dan 10 moeten onderdrukt zijn, en niet terug te rekenen. Let op terugrekenmogelijkheden op basis van rij- en kolomtotalen, maar ook op basis van totalen op andere tabbladen, verschillende regionale indelingen en deels wel/ deels niet afgeronde aantallen. Hierbij moet ook rekening worden gehouden dat sommige totaalcellen groter zijn dan de som van de verschillende categorieën, omdat onderzoekers aantallen onder 10 weggelaten kunnen hebben.

Opmerkingen hierbij:

  - Wanneer data op zowel bedrijven als personen betrekking heeft, geldt dit minimum voor zowel het aantal bedrijven als het aantal personen.
  - Aantallen van 0 moeten op dezelfde manier onderdrukt worden als aantallen tussen de 1 en de 9. Veel statistische programma's laten categorieën zonder observaties weg bij een uitdraai, dus let hierop bij het maken van output. Een uitzondering hierop zijn nullen die bij definitie niet kunnen voorkomen in de gekozen selectie. Deze moeten in het outputformulier worden voorzien van een uitleg waarom het aantal op logische gronden 0 is en daardoor geen informatie vrijgeeft.
  - Indien de tabellen percentages of opgehoogde aantallen bevatten, moeten de *ongewogen aantallen* ook opgenomen worden om de controle van deze norm mogelijk te maken.
  - Een aantal kleiner dan 10 is bij *missing/onbekend* zonder verdere uitsplitsingen geen probleem, omdat onbekend niet herleidbaar is naar personen. Voor de categorie "overig" geldt wel het minimum aantal waarnemingen, omdat dit een samenvoeging is van (herleidbare) categorieën.
  - *Maxima en minima* worden niet vrijgegeven, aangezien zij normaal gesproken refereren aan één eenheid. Bovendien geven ze impliciet aan dat er binnen een bepaalde groep geen lagere of hogere waarden voorkomen. Uitzondering hierop zijn situaties waarin zulke uiterste waarden logisch vastliggen en geen onthullingsrisico

vormen, bijvoorbeeld het minimale/maximale loon van niet-werkenden, dat per definitie nul is

- Voor *percentielwaarden* geldt dat zij minimaal 10 datapunten van het minimum en maximum af moeten liggen. Dit houdt in dat de mediaan mag worden weergegeven bij een populatie vanaf 20 personen of instellingen, maar dat voor bijvoorbeeld data over de 10<sup>e</sup> percentielgroep pas mag worden weergegeven bij een populatie vanaf 100 personen of instellingen.

*Aanvullend voor modellen:*

2. **Modellen:** Alle gemodelleerde output moet **minstens 10** vrijheidsgraden overhouden, waarbij het aantal vrijheidsgraden gelijk is aan: het aantal waarnemingen minus het aantal variabelen en/of restricties. Tevens mogen modellen niet gebaseerd zijn op één eenheid; een tijdreeks gebaseerd op één bedrijf is dus niet toegestaan.
  - Dit geldt in het bijzonder voor regressiemodellen, hogere orde momenten (variantie, covariantie, kurtosis, skewness), samenvatting- en teststatistieken van schattingen ( $R^2$  en variaties hierop), geschatte variantie, informatiecriteria (bijv. AIC, BIC) of individuele en groepstesten en -statistieken (bijv. t, F, chi-kwadraat, Wald, Hausman)
  - Verder geldt dat residuen en grafieken/plots van residuen worden behandeld als zijnde informatie over individuen en daarom niet worden vrijgegeven.

*Aanvullend voor frequentietabellen<sup>2</sup>:*

3. **Groepsonthulling:** In alle frequentietabellen en soortgelijke output mag geen enkele cel meer dan **90% van het totaal** aantal eenheden in de rij of kolom bevatten. Op deze manier wordt voorkomen dat bepaalde variabelen in een tabel een herkenbare groep definiëren, waar andere variabelen in de tabel informatie onthullen die geldig is voor elk lid van de groep. Hoewel er geen individuele eenheid kan worden herkend, is de vertrouwelijkheid geschonden, omdat de informatie geldig is voor (vrijwel) elk lid van de groep en de groep als zodanig herkenbaar is.
  - Ook telt het als groepsonthulling wanneer af te leiden valt dat elke persoon in de groep eenzelfde kenmerk heeft, bijvoorbeeld omdat de gegeven standaarddeviatie 0 is van dat kenmerk. Ondanks dat er genoeg respondenten aanwezig zijn is daardoor informatie terug te leiden naar het individu.

*Aanvullend voor kwantitatieve tabellen<sup>3</sup>:*

4. **Dominantie:** Dominantie speelt een rol bij onder andere output geaggregeerd naar:
  - Bedrijfsinstellingen
  - Scholen
  - Ziekenhuizen
  - Woningcorporaties.

Bij aggregatie naar één van deze categorieën is het belangrijk dat er geen sprake is van dominantie. Dit betekent dat de grootste bijdrager aan een cel niet meer dan **50% van het celtotaal** mag vertegenwoordigen. De onderzoeker moet deze aantallen zelf aanleveren; doet hij/zij dit niet, mag dit worden opgevraagd.

- Dominantie kan ook een rol spelen wanneer deze termen niet direct worden gebruikt. Als er bijvoorbeeld geaggregeerd wordt naar SBI-indeling gaat het ook om bedrijven. Hetzelfde geldt bij tabellen van aantal vacatures met bedrijven onderliggend en voor bepaalde specialismen met ziekenhuizen onderliggend.

In bijlage 1 worden bovenstaande richtlijnen verder toegelicht per type output. Hier zijn ook voorbeelden opgenomen die deze richtlijnen verduidelijken.

---

<sup>2</sup> Frequentietabel: tabel waarin je kunt aflezen hoe vaak (frequentie) een bepaalde waarneming voorkomt. Het genoemde percentage is strikter dan het CBS zelf toepast voor publicaties omdat het onthullingsrisico afhankelijk is van wat er gepubliceerd wordt.

<sup>3</sup> Kwantitatieve tabel: een tabel waarbij de celwaarden de som zijn van een variabele over alle waarnemingen. Bijvoorbeeld loonsom en omzet, maar ook aantal werkzame personen en aantal vacatures (bij die laatste twee zijn dan de bedrijven de waargenomen eenheden).



## 2b. Richtlijnen met betrekking tot leesbaarheid

*Uitgangspunt voor onderstaande richtlijnen is dat de onderzoeker ervoor dient te zorgen dat de aangeleverde output in combinatie met het standaardformulier begrijpelijk is voor de outputcontroleurs.*

*De output wordt gecontroleerd door twee (objectieve) controleurs. Hierdoor wordt een onafhankelijke outputcontrole gewaarborgd met criteria gelijk voor alle onderzoekers. De controleurs zijn weliswaar ervaren, maar doorgaans niet inhoudelijk bekend met de projecten. Ook weten de controleurs vanwege de grote diversiteit aan onderwerpen die ze voorgeschiedt krijgen niet van alle datasets alle ins en outs. Dit betekent dat het van groot belang is dat de onderzoeker een heldere toelichting geeft op de analyses en de gepresenteerde variabelen. Hiermee heeft de onderzoeker zelf invloed op de snelheid en uitkomst van de outputcontrole.*

Wij hanteren de volgende richtlijnen:

1. Outputbestanden dienen vergezeld te gaan van een volledig en correct ingevuld **Standaardformulier Outputcontrole**. Dit formulier bevindt zich in de 'shared workspace' (H-schijf) of in 'metadata' (K-schijf) onder 'Utilities' en moet worden meegeleverd bij iedere output die ter controle wordt aangeboden.  
*Als voorbeeld is in bijlage 2 een volledig ingevuld outputformulier bijgesloten.*  
In het formulier moet het volgende worden beschreven:
  - of de output aan de richtlijnen voldoet en zo niet waarom er geen onthullingsrisico is;
  - wat het doel van de analyses is;
  - welke bronbestanden gebruikt zijn;
  - hoe het onderzoeksbestand is samengesteld (bijvoorbeeld door koppeling van verschillende bestanden of door subpopulaties te selecteren);
  - op welke populatie de output betrekking heeft;
  - waaruit de output bestaat, zoals tellingen, kruistabellen, modellen of analyses;
  - wat de relatie is tussen tabellen binnen een outputbestand; (indien van toepassing)
  - wat de relatie is tussen figuren en onderliggende modellen binnen een outputbestand waarop deze gebaseerd zijn (indien van toepassing);
  - wat de relatie is met eerder aangeleverde output (indien van toepassing).
2. Variabelen in de output dienen **gedocumenteerd** te zijn. Hiermee bedoelen we dat er een duidelijke omschrijving moet zijn van de betekenis van de variabele en zijn afzonderlijke waarden in het geval van discrete variabelen. Een mogelijkheid is het aanmaken van een apart document waarin deze gegevens worden vermeld. Dit document kan bij elke output worden meegeleverd en steeds aangevuld worden met nieuwe variabelen.
3. Elk onderdeel van de output moet voorafgegaan worden door een regel tekst waarin de **populatie of selectie** voor dat onderdeel wordt beschreven.
4. Bij *dominantie* en andere gevallen waarbij de informatie op controle op zichzelf al onthullend is, is het noodzakelijk om achterliggende tabellen mee te leveren die niet vrijgegeven kunnen worden. In dat geval dienen deze aangeboden te worden in een **apart zipbestand** met de naam **"GEENOUTPUT\_Achtergrondinformatie.zip"**. In deze documentatie moet per tabel duidelijk beschreven worden voor welk onderdeel van de output deze achterliggende tabel dient. Dit kan bijvoorbeeld door eenzelfde nummering aan te brengen in de output en de achtergrondinformatie. Dit zipbestand wordt uiteraard niet vrijgegeven.



## 2c. Richtlijnen met betrekking tot technische vereisten

*Naast beveiliging en leesbaarheid gelden er ook enkele technische vereisten voor de aan te leveren output. Deze richtlijnen zijn bedoeld om de controle efficiënt en betrouwbaar te laten verlopen, én om technische knelpunten te voorkomen bij het verwerken van bestanden vanuit de CBS-microdataomgeving.*

*Onderstaande richtlijnen hebben betrekking op onder andere de omvang, het bestandsformaat en de structuur van de outputbestanden. Door deze technische instructies zorgvuldig op te volgen, wordt de controle op output eenvoudiger en wordt vertraging in de verwerking voorkomen.*

1. Grootte van de output: Om de outputs controleerbaar te houden mag een output niet te groot zijn. Wij houden de volgende richtlijnen aan:
  - Ons uitgangspunt is dat een output binnen één uur gecontroleerd moet kunnen worden. De tarieven voor outputcontroles zijn hier ook op gebaseerd. Wanneer een output te groot is, zullen de kosten en doorlooptijd voor het controleren naar rato verhoogd worden. Hiervoor vragen wij dan wel eerst een kostenakkoord aan de onderzoeker.
  - Bied geen zipfiles in een zipfile of mappen in een zipfile aan.
2. Formaat: We hanteren de volgende voorkeursformaten:
  - Tabellen in: csv-, xlsx-, ods- of xml-format;
  - Scripts in: do-, R-, py-, sps- of txt-format;
  - Figuren in jpg-, png- of gif-format.
  - Meer informatie over soorten output en bijbehorende kosten staat in de [Dienstencatalogus](#).
  - Presentaties in .ppt- en .pptx-format (PowerPoint) en de databestanden met de extensie ".RData" staan we niet toe.
  - Grafieken zijn in principe toegestaan als output indien ze op veilige output zijn gebaseerd. Hierbij is het van belang om duidelijk uit te leggen welke informatie er in elke grafiek staat en op hoeveel waarnemingen deze is gebaseerd. Elk datapunt moet op minimaal 10 waarnemingen gebaseerd zijn; het is dus niet voldoende om alleen de totale populatieaantallen weer te geven.
3. Zichtbaarheid: Verberg geen output in de outputbestanden. Het is niet toegestaan om in Excel-bestanden kolommen, rijen of tabbladen te verbergen. Ook mogen draaitabellen niet ter controle worden aangeboden.
4. Overig:
  - De naam van de aangeboden output mag niet meer dan 100 posities bevatten, dit kan namelijk problemen geven met het overhalen van de output uit de CBS-microdataomgeving. Ook mag de naam van de aangeboden output niet beginnen met een datum.
  - De procedure van het aanbieden van output is te vinden op de webpagina [Aanbieden output of andere exportbestanden](#).
  - Zet de output in een zipfile in een van de bestaande exportmappen. Maak hier niet zelf een map aan, want dan wordt de output niet overgehaald van de CBS-microdataomgeving naar de CBS-omgeving.
  - Elk half uur wordt de output uit de map gehaald; dan is deze niet meer zichtbaar voor de onderzoeker. Wanneer de bestanden nog openstaan op dat moment, komen ze corrupt bij ons binnen en kunnen we ze niet openen. Daarom is het belangrijk om de output buiten de exportmap gereed te maken en daarna alleen nog naar de exportmap te kopiëren.



## Bijlage 1: Voorbeelden beveiligingsrichtlijnen

De volgende beveiligingsrichtlijnen worden in detail beschreven:

*Algemene richtlijnen:*

0. Geen microdata: Het is niet toegestaan om microdata als output aan te bieden.
1. Minimum aantal waarnemingen: Alle output bevat **tenminste 10** eenheden (ongewogen) als grondslag voor elke cel of datapunt.

*Aanvullend voor modellen:*

2. Modellen: Alle gemodelleerde output moet minstens 10 vrijheidsgraden hebben, waarbij het aantal vrijheidsgraden gelijk is aan: aantal waarnemingen  $\cdot$  aantal parameters  $\cdot$  andere modelrestricties. Verder geldt dat residuen en grafieken/plots van residuen niet vrijgegeven worden.

*Aanvullend voor frequentietabellen<sup>4</sup>:*

3. Groepsonthulling: In alle tabellen en soortgelijke output mag geen enkele cel meer dan 90% van het totaal aantal eenheden in de rij of kolom bevatten om groepsonthulling te voorkomen.

*Aanvullend voor kwantitatieve tabellen<sup>5</sup>:*

4. Dominantie: In alle tabellen en soortgelijke gegevens mag de grootste bijdrager aan een cel niet meer dan 50% van het celtotaal bijdragen.

De eerste genoemde richtlijn, geen microdata, is uiteraard van toepassing op elke output. Hieronder een tabel waarin per type output wordt aangegeven welke richtlijnen van toepassing zijn.

| Type statistiek                   | Type output                                                                | Beveiligingsrichtlijn(en) van toepassing |
|-----------------------------------|----------------------------------------------------------------------------|------------------------------------------|
| Beschrijvende statistieken        | Frequentietabellen                                                         | 1, 3                                     |
|                                   | Kwantitatieve tabellen en percentielen                                     | 1, 3, 4                                  |
|                                   | Maxima, minima                                                             | Niet vrijgeven                           |
|                                   | Modus, mediaan                                                             | 1, 3                                     |
|                                   | Gemiddelden, indices, verhoudingen, indicatoren                            | 1, 3, 4                                  |
|                                   | Concentratieverhoudingen                                                   | 1, 3, 4                                  |
|                                   | Verdelingsmaten (incl. variantie, covariantie, kurtosis, scheefheid)       | 1, 2                                     |
|                                   | Grafieken: grafische weergave van data                                     | 1, 2, 3, 4                               |
| Correlaties en Regressie-analyses | Lineaire regressiecoëfficiënten                                            | 1, 2                                     |
|                                   | Niet-lineaire regressiecoëfficiënten                                       | 1, 2                                     |
|                                   | Schattingsresiduen                                                         | Niet vrijgeven                           |
|                                   | Samenvattings- en teststatistieken van schattingen ( $R^2$ , $\chi^2$ enz) | 1, 2                                     |
|                                   | Correlatiecoëfficiënten                                                    | 1                                        |

### 1. Minimum aantal waarnemingen

Elke cel in een frequentie- of kruistabel moet minimaal 10 waarnemingen bevatten. In voorbeeld 1a is een kruistabel weergegeven met kleine aantallen waarin onthulling plaatsvindt. Na het zien van deze tabel weet je als inwoner van Zuiderwijk aan Zee dat het enige gezin met meer dan 20 kinderen in een lage inkomensklasse valt.

In dit geval kan je ervoor zorgen dat de tabel niet meer onthullend is door één van de volgende drie opties: de laatste 2 categorieën van het aantal kinderen samen te voegen (1b) of cellen te

<sup>4</sup> Frequentietabel: tabel waarin je kunt aflezen hoe vaak (frequentie) een bepaalde waarneming voorkomt.

<sup>5</sup> Kwantitatieve tabel: een tabel waarbij de celwaarden de som zijn van een variabele over alle waarnemingen. Bijvoorbeeld loonsom en omzet, maar ook aantal werkzame personen en aantal vacatures (bij die laatste twee zijn dan de bedrijven de waargenomen eenheden).



onderdrukken (1c) of door af te ronden op vijftallen (1d). Bij het onderdrukken van cellen moet je er tevens op letten dat de aantallen niet terug te rekenen zijn met informatie uit de andere cellen.

#### Voorbeeld 1a

Tabel: Huishoudens naar aantal kinderen en inkomensklasse in Zuiderwijk aan Zee.

|                 |        | Inkomensklasse |        |      |        |
|-----------------|--------|----------------|--------|------|--------|
|                 |        | Laag           | Midden | Hoog | Totaal |
| Aantal kinderen | 1-3    | 32             | 64     | 84   | 180    |
|                 | 4-10   | 28             | 48     | 28   | 104    |
|                 | 11-20  | 10             | 16     | 10   | 36     |
|                 | >20    | 1              | 0      | 0    | 1      |
|                 | Totaal | 71             | 128    | 122  | 321    |

#### Voorbeeld 1b

Tabel: Huishoudens naar aantal kinderen en inkomensklasse in Zuiderwijk aan Zee.

|                 |        | Inkomensklasse |        |      |        |
|-----------------|--------|----------------|--------|------|--------|
|                 |        | Laag           | Midden | Hoog | Totaal |
| Aantal kinderen | 1-3    | 32             | 64     | 84   | 180    |
|                 | 4-10   | 28             | 48     | 28   | 104    |
|                 | >10    | 11             | 16     | 10   | 37     |
|                 | Totaal | 71             | 128    | 122  | 321    |

#### Voorbeeld 1c

Tabel: Huishoudens naar inkomensklasse en aantal kinderen in Zuiderwijk aan Zee.

De cellen die onthullend waren zijn in deze tabel onderdrukt. Hierdoor wordt voor bijvoorbeeld het gezin met meer dan 20 kinderen geen extra informatie vrijgegeven.

|                 |        | Inkomensklasse |        |      |        |
|-----------------|--------|----------------|--------|------|--------|
|                 |        | Laag           | Midden | Hoog | Totaal |
| Aantal kinderen | 1-3    | 32             | 64     | 84   | 180    |
|                 | 4-10   | 28             | 48     | 28   | 104    |
|                 | 11-20  | x              | x      | x    | x      |
|                 | >20    | x              | x      | x    | x      |
|                 | Totaal | 71             | 128    | 122  | 321    |

#### Voorbeeld 1d

Tabel: Huishoudens naar aantal kinderen en inkomensklasse in Zuiderwijk aan Zee.

Wegens groepsonthulling zou dit óók niet mogen als inkomensklasse laag afgerond 5 was en midden/hoog 0. Toevallig dat het hier maar om één fictief gezin gaat.

|                 |        | Inkomensklasse |        |      |        |
|-----------------|--------|----------------|--------|------|--------|
|                 |        | Laag           | Midden | Hoog | Totaal |
| Aantal kinderen | 1-3    | 30             | 65     | 85   | 180    |
|                 | 4-10   | 30             | 50     | 30   | 105    |
|                 | 11-20  | 10             | 15     | 10   | 35     |
|                 | >20    | 0              | 0      | 0    | 0      |
|                 | Totaal | 70             | 130    | 120  | 320    |

## 2. Modellen

In onderstaande voorbeelden hebben we twee voorbeelden van modeluitkomsten gegeven: één met output vanuit Stata, en één vanuit R. Voor Stata controleren wij het aantal vrijheidsgraden dat bij “Residual” genoteerd staat (in het geval van voorbeeld 2a zijn dat er 72); niet de vrijheidsgraden die bij het model staat (die geeft alleen aan hoeveel parameters zijn opgenomen). Ook in R staat het te controleren aantal vrijheidsgraden bij de “Residuals” (in het geval van voorbeeld 2b zijn dat er 48).

De residuen zelf mogen niet in de output opgenomen worden, want deze geven informatie per waarneming. Deze kunnen worden opgevraagd als grafiek bij het uitvoeren van de regressie. Deze grafiek mag in de CBS-microdataomgeving bekeken worden, maar niet worden meegenomen als output. We zien in het voorbeeld dat R hier standaard informatie voor geeft, inclusief een minimum en maximum. Deze informatie moet dus weggehaald worden.

Bij complexe regressies is er niet altijd een aantal vrijheidsgraden aanwezig. Voor de onderliggende gegevens waarop het model zijn schattingen doet, blijft daarom vuistregel 1 altijd van kracht: een minimum aantal waarnemingen van 10. Ook moet bij weggevallende variabelen een uitleg bijgevoegd worden waarom deze zijn weggevallend; dit kan namelijk een gevolg zijn van een niet-voorkomende variabele binnenin de populatieselectie, waardoor je aantal van 0 zou kunnen afleiden.

### Voorbeeld 2a: Modeluitkomsten uit Stata

```
. sysuse auto
(1978 automobile data)
```

```
. reg price length
```

| Source   | SS        | df | MS         | Number of obs | = | 74     |
|----------|-----------|----|------------|---------------|---|--------|
| Model    | 118425867 | 1  | 118425867  | F(1, 72)      | = | 16.50  |
| Residual | 516639529 | 72 | 7175549.01 | Prob > F      | = | 0.0001 |
|          |           |    |            | R-squared     | = | 0.1865 |
|          |           |    |            | Adj R-squared | = | 0.1752 |
| Total    | 635065396 | 73 | 8699525.97 | Root MSE      | = | 2678.7 |

| price  | Coefficient | Std. err. | t     | P> t  | [95% conf. interval] |
|--------|-------------|-----------|-------|-------|----------------------|
| length | 57.20224    | 14.08047  | 4.06  | 0.000 | 29.13332 85.27115    |
| _cons  | -4584.899   | 2664.437  | -1.72 | 0.090 | -9896.357 726.559    |

### Voorbeeld 2b: Modeluitkomsten uit R

```
Call:
lm(formula = y ~ x1 + x2, data = ols_data)
```

```
Residuals:
```

```
    Min       1Q   Median       3Q      Max
-3.9384 -1.6412 -0.4262  1.4503  5.6626
```

```
Coefficients: (1 not defined because of singularities)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.34110    0.66670   59.01  <2e-16 ***
x1           0.51066    0.03054   16.72  <2e-16 ***
x2              NA         NA      NA      NA
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.2 on 48 degrees of freedom
Multiple R-squared:  0.8535,    Adjusted R-squared:  0.8504
F-statistic: 279.6 on 1 and 48 DF,  p-value: < 2.2e-16
```



### 3. Groepsonthulling

Als voorbeeld voor het schenden van groepsonthulling bekijken we een frequentietabel over het drugsgebruik onder de jeugd in Zuiderwijk aan Zee.

Tabel 3: Inwoners van Zuiderwijk aan Zee naar leeftijd en drugsgebruik

| Drugsgebruik<br>Leeftijd | Geen | Alleen softdrugs | Harddrugs (evt.<br>ook softdrugs) | Totaal |
|--------------------------|------|------------------|-----------------------------------|--------|
| < 15 jaar                | 78   | 12               | 13                                | 103    |
| 15 – 17 jaar             | -    | 367              | 12                                | 381    |
| 18 – 21 jaar             | -    | 28               | 389                               | 417    |
| Totaal                   | 80   | 407              | 414                               | 901    |

De frequentie in de kolom 'geen' in de leeftijdscategorieën vanaf 15 jaar is onthullend, omdat hieruit volgt dat bijna iedereen in andere leeftijdsgroepen dus wél drugs gebruikt, ondanks het gebruik van onderdrukking. Als een vader van een jongen van 19 in dit dorp deze tabel onder ogen krijgt, weet hij bijna zeker dat zijn zoon drugs gebruikt. In deze tabel moet de kolom 'Geen' voor de categorie '<15 jaar' dus ook leeggemaakt worden omdat de restgroep van de andere categorieën kleiner is dan 10 respondenten.

### 4. Dominantie

Als voorbeeld voor het schenden van de dominantieregel kijken we naar de investeringen van ondernemingen naar regio en activiteit.

Tabel 4: Investerings van ondernemingen (mln euro) naar regio en activiteit.

Selectie: Alle bedrijven in Nederland uit de periode x

|              | Regio 1 | Regio 2 | Regio 3 | Totaal |
|--------------|---------|---------|---------|--------|
| Activiteit 1 | 78      | 12      | 9       | 99     |
| Activiteit 2 | 4       | 367     | 12      | 383    |
| Activiteit 3 | 10      | 28      | 389     | 427    |
| Totaal       | 92      | 407     | 410     | 909    |

In dit voorbeeld zijn onderstaande achtergrondtabellen noodzakelijk ter controle van de dominantieregel. Deze achtergrondinformatie dient aangeboden te worden in een **apart bestand** met in de naam "GEENOUTPUT".

Uit tabel 4a blijkt dat deze output voldoet aan vuistregel 1: Alle output bevat tenminste 10 eenheden (ongewogen) als grondslag voor elke cel of datapunt. Uit tabel 4c (welke volgt uit tabel 4b) blijkt dat één cel (Activiteit 3, Regio 3) niet voldoet aan de dominantieregel. Alle bedrijven die zijn gevestigd in Regio 3 investeren gezamenlijk 389 mln in activiteit 3. Het bedrijf met de meeste investering (234 mln) maakt hierbij 60% van het totaal uit. Daarmee zit het boven de 50% en domineert dit bedrijf deze cel waardoor de kans op onthulling te groot is en daarom niet in de huidige vorm mag worden vrijgegeven.



*Tabel 4a: Totaal aantal bijdragers per cel*  
*Deze tabel hoort als achtergrondinfo bij tabel 4*

|              | Regio 1 | Regio 2 | Regio 3 | Totaal |
|--------------|---------|---------|---------|--------|
| Activiteit 1 | 14      | 10      | 12      | 36     |
| Activiteit 2 | 13      | 15      | 11      | 39     |
| Activiteit 3 | 20      | 21      | 23      | 64     |
| Totaal       | 47      | 46      | 46      | 139    |

*Tabel 4b: Omvang van grootste bijdrager per cel (mln euro)*  
*Deze tabel hoort als achtergrondinfo bij tabel 4*

|              | Regio 1 | Regio 2 | Regio 3 | Totaal |
|--------------|---------|---------|---------|--------|
| Activiteit 1 | 34      | 4       | 2       |        |
| Activiteit 2 | 1       | 167     | 5       |        |
| Activiteit 3 | 2       | 10      | 234     |        |
| Totaal       |         |         |         |        |

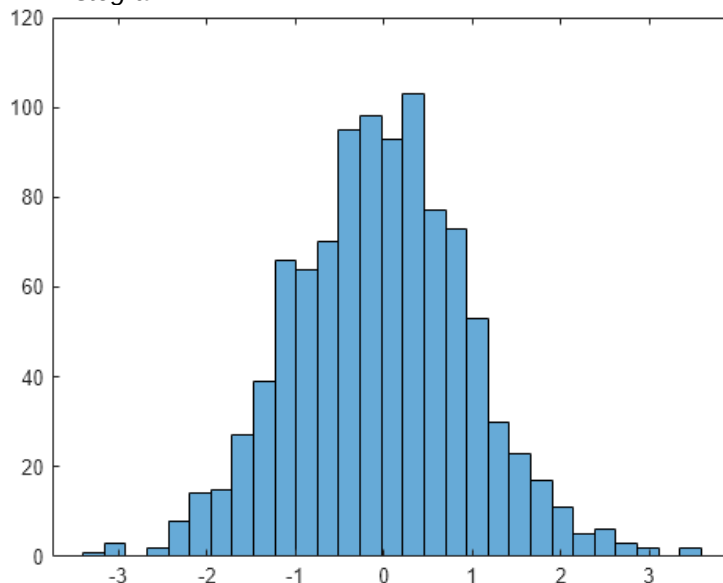
*Tabel 4c: Procentuele bijdrage van grootste bijdrager per cel*  
*Deze tabel hoort als achtergrondinfo bij tabel 4*

|              | Regio 1 | Regio 2 | Regio 3 | Totaal |
|--------------|---------|---------|---------|--------|
| Activiteit 1 | 44%     | 33%     | 22%     |        |
| Activiteit 2 | 25%     | 46%     | 42%     |        |
| Activiteit 3 | 20%     | 36%     | 60%     |        |
| Totaal       |         |         |         |        |

## 5. Aanvullend: Figuren

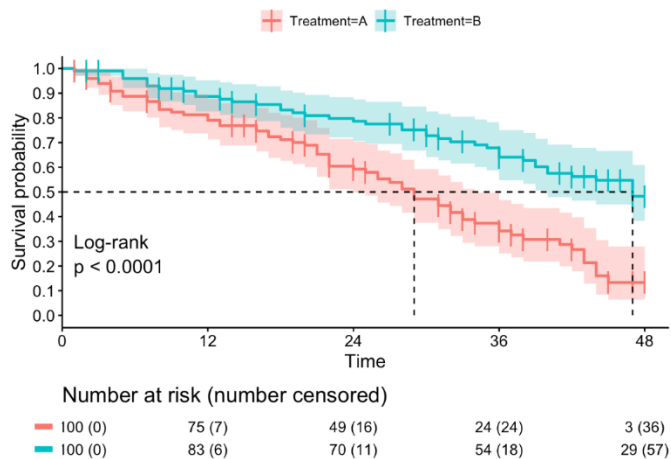
Onderzoekers kunnen veel soorten figuren aanbieden, waarbij we graag zien dat de datapunten minimaal 10 onderliggende observaties bevat. Hieronder bespreken we zowel een veelvoorkomend figuur (histogram) als een lastiger te interpreteren figuur (Kaplan Meier plot).

Figuur 1: Histogram



Bij een histogram (figuur 1) willen we van elke “bin” het bijbehorende aantal ontvangen. Bij bovenstaand figuur kan je nu aflezen dat  $N=0$  voor een aantal bins, waardoor deze niet aan de richtlijnen voldoet. Ook is de schaal klein genoeg dat je voor de andere waarden zou kunnen afmeten tot met vrij grote precisie wat de bijbehorende frequentie is, ook voor aantallen onder de 10.

Figuur 2: Kaplan Meier



Bij een Kaplan Meier plot (figuur 2) mogen wel aantallen van 0 voorkomen wanneer deze het begin van een periode representeren, omdat deze bij definitie 0 is en daardoor niet onthullende informatie bevat. Wel moet worden opgelet dat ook bij een populatie met voldoende observaties de verschillen niet alsnog leiden tot onthulling; tussenliggende aantallen lager dan 3 worden daarom hier niet toegestaan.



## Bijlage 2: Standaardformulier beschrijving output<sup>6</sup>

|                                                                                                                                                               |                                                                                                                                                                                   |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Naam <i>[naam gebruiker]</i>                                                                                                                                  | Ben Slim                                                                                                                                                                          |
| Telefoonnummer <i>[telefoonnummer waarop gebruiker gebeld kan worden als er vragen over deze output zijn]</i>                                                 | 06-12345678                                                                                                                                                                       |
| Instelling <i>[naam instelling]</i>                                                                                                                           | Universiteit van Den Haag                                                                                                                                                         |
| Datum <i>[datum waarop output is aangemaakt]</i>                                                                                                              | dd-mm-jjjj                                                                                                                                                                        |
| Project <i>[contractnummer + naam onderzoek]</i>                                                                                                              | 1234 - Doorstroom van uitkeringsgerechtigden naar de arbeidsmarkt <sup>7</sup>                                                                                                    |
| Deze output bestaat volledig uit documentatie <i>[syntax of tekstbestand]</i> Indien Ja, dan hoeft de rest van het formulier niet ingevuld te worden.         | Nee                                                                                                                                                                               |
| Deze output is een aanpassing van eerder tegengehouden output                                                                                                 | Ja, namelijk van dd-mm-jjjj / Nee                                                                                                                                                 |
| Deze output voldoet volledig aan de <a href="#">outputrichtlijnen</a>                                                                                         | Ja                                                                                                                                                                                |
| Deze output voldoet niet volledig aan de <a href="#">outputrichtlijnen</a><br>Hiervoor heb ik de volgende (uitgebreide) uitleg:                               | Nee                                                                                                                                                                               |
| Data<br><i>[gebruikte bronbestanden]</i>                                                                                                                      | BAANKENMERKEN, BIJSTANDUITKERINGTAB<br>AOTOTUITKERINGTAB, WUSUITKERING1ATAB                                                                                                       |
| Onderzoeksbestand<br><i>[beschrijving van samengesteld onderzoeksbestand]</i>                                                                                 | Het onderzoeksbestand bestaat uit alle WW-, AO- en Bijstandsgerechtigden op dd-mm-jjjj, met daaraan gekoppeld persoonsgegevens op hetzelfde peilmoment en baangegevens over jjjj. |
| Doel en doelbinding<br><i>[doel output en hoe de output past binnen de onderzoeksvraag van het project]</i>                                                   | Het doel is om in kaart te brengen hoeveel en welke uitkeringsgerechtigden binnen een jaar doorstromen naar een baan. Dit beantwoordt tevens onze onderzoeksvraag.                |
| Relatie eerder aangeleverde output<br><i>[beschrijving van de relatie met eerder aangeleverde output]</i>                                                     | In de vorige output keek ik alleen naar WW-uitkeringen. Nu dezelfde exercitie voor WW+BIJSTAND+AO                                                                                 |
| Inhoud output <i>[beschrijving van de output, komen er bijvoorbeeld tellingen, kruistabellen of analyses <sup>7</sup> in voor]</i>                            | Excelbestand met verschillende kruistabellen: - tabellen naar soort uitkering en persoonskenmerken                                                                                |
| Eventuele selecties in de output heb ik duidelijk aangegeven bij de analyses waarop deze van toepassing zijn.<br><i>[beschrijving van gemaakte selecties]</i> | Personen in GBA met een WW-, AO- of bijstandsuitkering op dd-mm-jjjj. Als er een extra selectie is gemaakt staat dit bij de betreffende tabel.                                    |

<sup>6</sup> Indien dit formulier niet goed of niet volledig is ingevuld wordt de output niet in behandeling genomen. Indien het bestand correct is ingevuld en aangeboden voor 05:00 's ochtends, zal de (reguliere) output of documentatie binnen drie werkdagen worden gecontroleerd. Output light zal dezelfde dag worden gecontroleerd.

<sup>7</sup> Indien er modeluitkomsten in de output voorkomen dan graag het onderliggende aantal observaties duidelijk vermelden.

## Bijlage 3: CBS-wet

*Hieronder de onderdelen van de CBS-wet dit hoofdzakelijk van belang zijn bij het aanbieden van output.*

### Artikel 37

1. De door de directeur-generaal in het kader van de uitoefening van de taken ter uitvoering van deze wet ontvangen gegevens worden uitsluitend gebruikt voor statistische doeleinden.
2. De in het eerste lid bedoelde gegevens worden niet verstrekt aan anderen dan degenen die belast zijn met de uitvoering van de taak van het CBS.
3. De in het eerste lid bedoelde gegevens worden slechts zodanig openbaar gemaakt dat daaraan geen herkenbare gegevens over een afzonderlijk persoon, huishouden, onderneming of instelling kunnen worden ontleend, tenzij, ingeval het gegevens met betrekking tot een onderneming of instelling betreft, er een gegronde reden is om aan te nemen dat bij de betrokken onderneming of instelling geen bedenkingen bestaan tegen de openbaarmaking.

### Artikel 41

1. In afwijking van artikel 37 kan de directeur-generaal op verzoek, ten behoeve van statistisch of wetenschappelijk onderzoek, een verzameling van gegevens met betrekking tot het gebruik waarvan passende maatregelen zijn genomen om herkenning van afzonderlijke personen, huishoudens, ondernemingen of instellingen te voorkomen, verstrekken aan een dienst, organisatie of instelling als bedoeld in het tweede lid, dan wel daartoe toegang verlenen.
2. Een verzameling van gegevens als bedoeld in het eerste lid kan worden verstrekt, dan wel daartoe kan toegang worden verleend aan:
  - a. een universiteit in de zin van de Wet op het hoger onderwijs en wetenschappelijk onderzoek;
  - b. een bij wet ingestelde organisatie of instelling voor wetenschappelijk onderzoek;
  - c. bij of krachtens de wet ingestelde planbureaus;
  - d. de communautaire en nationale instanties voor de statistiek van de lidstaten van de Europese Unie;
  - e. onderzoeksafdelingen van ministeries en andere diensten, organisaties en instellingen.

### Artikel 42

De directeur-generaal wiligt een verzoek als bedoeld in artikel 41 slechts in, indien de verzoeker naar het oordeel van de directeur-generaal voldoende maatregelen heeft getroffen om te voorkomen dat de verzameling van gegevens voor andere doeleinden dan statistisch of wetenschappelijk onderzoek wordt gebruikt.

### Artikel 42a

1. In afwijking van artikel 37 kan de directeur-generaal op verzoek, ten behoeve van wetenschappelijk onderzoek op het terrein van de volksgezondheid, gegevens verstrekken die het CBS ten behoeve van statistisch onderzoek op grond van artikel 12a van de Wet op de lijkbezorging heeft verzameld, van personen die in een wetenschappelijk onderzoek waren betrokken.
2. De directeur-generaal wiligt een verzoek als bedoeld in het eerste lid slechts in, indien de betrokkene tot een zodanige verstrekking zijn uitdrukkelijke toestemming heeft gegeven dan wel, indien de betrokkene tot een zodanige verstrekking geen uitdrukkelijke toestemming heeft kunnen geven, voor zover verzoeker voldoende aantoont dat:
  - a. het vragen van toestemming bij leven van de betrokkene in redelijkheid niet mogelijk was of kon worden vergegd;
  - b. niet gebleken is dat de betrokkene bij leven bezwaar heeft gemaakt tegen het verwerken van diens persoonsgegevens ten behoeve van wetenschappelijke onderzoek;
  - c. het onderzoek een algemeen belang dient;
  - d. het onderzoek niet zonder de desbetreffende gegevens kan worden uitgevoerd;
  - e. het onderzoek overigens voldoet aan daaraan redelijkerwijs te stellen eisen.



- 
3. De directeur-generaal kan nadere voorwaarden verbinden aan een verstrekking krachtens het eerste lid.
  4. De artikelen 41, tweede lid, en 42 zijn van overeenkomstige toepassing.
  5. Onze Minister stelt, in overeenstemming met Onze Minister van Volksgezondheid, Welzijn en Sport, gehoord het College bescherming persoonsgegevens, nadere regels met betrekking tot de in het eerste lid bedoelde bevoegdheid.