

# Kankerrisico door pesticiden decennialang ‘verkeerd’ ingeschat

Waarom de keuze voor de frequentistische statistiek  
voor beide ‘partijen’ problematisch is.

Door Dr. Marc Jacobs

MSJ Advies

## Inhoudsopgave

Samenvatting .....	4
Inleiding .....	7
De kritiek op de statistiek .....	8
Wat kunnen we hier nu uit afleiden? .....	14
Wat mag u van mij verwachten? .....	15
Kansen, verdelingen en context.....	18
De normaalverdeling .....	18
Rangorde .....	23
Het buitenbeetje .....	25
Een kans is afhankelijk van de context .....	28
Wat kunnen we tot nu toe concluderen? .....	29
Zoek de verschillen.....	31
Kansen zonder context.....	31
Verschillende groepen of niet? .....	35
Wat kunnen we hier nu uit afleiden? .....	40
Fundamenten van de frequentistische Statistiek.....	41
Grenswaardes .....	42
Hypothese(s) toetsen .....	44
One-sample t-test.....	45
Independent samples t-test.....	48
Betrouwbaarheidsinterval.....	51
Zaken die van invloed zijn op het betrouwbaarheidsinterval .....	55
Waarom is 5% de grenswaarde? .....	60
Alfa en betá .....	60
Effectgrootte .....	63
De dekkingsraad.....	70
Gluren .....	74
Wat kunnen we hier nu uit afleiden? .....	76
Eenzijdig en tweezijdig toetsen .....	78
Tweemaal eenzijdig testen (TOST) .....	82

Het probleem van meerdere testen .....	83
Wat kunnen we hier nu uit afleiden? .....	90
Glyfosaat: beschrijving en visualisatie van de data .....	91
Voorbeeld uitwerken: Knezevich and Hogan (1983).....	93
Alle studies bekijken .....	97
Wat kunnen we hieruit afleiden? .....	108
Glyfosaat: eenzijdig en tweezijdig toetsen.....	109
Het werk van Portier controleren.....	110
Wat als we nu meenemen wat we niet zagen?.....	114
Het probleem van meerdere testen .....	127
Wat kunnen we hieruit afleiden? .....	134
Glyfosaat: Dose-Response analyses .....	135
Non-lineaire dose-response analyse .....	136
Lineaire dose-response analyse .....	137
Linear Mixed Model .....	141
Generalized Linear Mixed Model .....	151
Binomiaal model .....	153
Poisson model .....	175
Negative binomial .....	182
Zero-Inflated & hurdle models.....	187
Wat kunnen we hieruit concluderen?.....	192
Glyfosaat: een Bayesiaanse analyse .....	194
Een simpele regressie als voorbeeld.....	194
De likelihood ratio .....	208
Bayesiaanse analyse van het hurdle model.....	211
Modelleren van de verandering: vóóraf vs. áchteraf.....	216
De binomiaalverdeling.....	222
Het effect van de prior .....	237
Wat kunnen we concluderen? .....	238
Conclusie en aanbevelingen .....	239
Figuren.....	243
Tabellen .....	251

## Samenvatting

In dit rapport ben ik aan de slag gegaan met de uitspraken uit de BNNVARA /ZEMBLA reportage die stelt dat de relatie tussen glyfosaat en de kans op kanker verkeerd wordt berekend. De door de meeste onderzoekers gehanteerde tweezijdige toets toetst namelijk of glyfosaat beschadigd óf beschermd. Daarmee hanteert deze toets een te hoge grenswaarde voor het bepalen van de zogenaamde statistisch significantie. Dit is een onderdeel uit de frequentistische statistiek en wordt algemeen beschouwd als een essentieel onderdeel in het vaststellen van een causaal verband. Door het vervangen van de tweezijdige toets door een eenzijdige toets zou de afkapwaarde krimpen en de statistiek mee recht doen aan de mogelijke *mode of action* van het bestrijdingsmiddel, aldus BNNVARA/ZEMBLA. Uiteindelijk zou het toepassen van de eenzijdige toets de relatie tussen glyfosaat en de kans op kanker statistisch significant maken.

In dit rapport ben ik aan de slag gegaan met de stellingen uit de BNNVARA/ZEMBLA reportage én met de bron(nen) waarop deze uitspraken zijn gebaseerd. Het ingenomen standpunt dat het vervangen van een toets meer recht doet aan de werkelijke relatie tussen glyfosaat en kanker is voor mij zo kort door de bocht dat ik eerst stil wil staan bij wat statistiek nu eigenlijk is én waar wij modelleurs het voor toepassen. Daarmee is de eerste helft van het rapport vooral een inleiding in de statistiek en een uitleg waarom het vervangen van de tweezijdige toets door een eenzijdige toets niet zo eenvoudig is. Modelleeren is geen exacte wetenschap en statistiek toepassen betekent aannames maken. Het betekent vooral dat je bewust moet zijn van de gebreken van de statistiek én de moeilijkheden rondom het rekenen met kansen. Modellen zijn voor het begin van een gesprek en niet het einde! Om dit deel minder gevoelig te maken voor de inhoud ben ik begonnen met een neutraal voorbeeld: de lengteverdeling van de Nederlandse mannen en vrouwen.

Vervolgens ben ik met de bronnen aan de slag gegaan. Het blijkt dat één studie bovenal wordt aangehaald: de [studie van Portier uit 2020](#) die een her-analyse doet van historische dierproeven voor glyfosaat.

Wat ik in feite getracht heb te doen, is het op verschillende manieren proberen te herhalen van de bevindingen van Portier door gebruik te maken van de data zoals

gerapporteerd door Portier. Dit rapport is dus bovenal een herhaling van onderzoek dat is uitgevoerd en wat wordt aangehaald als hét bewijs dat glyfosaat kankerverwekkend is.

Kort door de bocht genomen lukt het mij niet om zijn bevindingen exact te repliceren. Dat betekent dus niet dat ik geen statistisch significante verschillen kan vinden als ik dezelfde methoden gebruik als Portier. Die zie ik ook, maar wanneer ik de methodiek van Portier hanteren treden er wel problemen op die niet worden geadresseerd.

In het algemeen rapporteert Portier per geslacht per studie welke kancersoort er wel of niet is opgetreden per dosering. Dit maakt dat de studie van Portier meer dan 200 statistische toetsingen kent. Het herhaaldelijk statistisch testen van eenzelfde dataset is echter een katalysator voor het vinden van vals positieven. Dit komt door de aannames die haast inherent zijn aan de frequentistische statistiek. Wanneer ik hiervoor corrigeer verdwijnen alle statistisch significante effecten.

Wat verder mist is dat de studies verschillen in welke soorten kanker wordt gerapporteerd. Het lijkt er sterk op dat elke analyse berust op het vinden van een soort kanker in welke dosering dan ook waarnaar voor elke dosering een analyse wordt gedaan. Wanneer een kancersoort in zijn geheel uitblijft wordt dat soms wel gerapporteerd, maar men is hier niet consequent in. De analyse van Portier wordt dus bovenal gedaan op het niveau van de tumor. Niet alleen ondervinden we dan het probleem van de vals positieven, maar we werken ook met proporties die als hoger worden gerapporteerd dan ze daadwerkelijk zijn. Dit komt omdat bij het samenvoegen van de studies, wat later ook door Portier wordt gedaan, het uitblijven van kanker niet wordt meegenomen in de berekening van de kans op kanker.

In het algemeen is het uitermate lastig gebleken om een zogenaamde dose-respons analyses uit te werken. Traditionele non-lineaire analyses mislukken en met behulp van meer lineaire technieken zie ik een hoop onzekerheid. De relatie tussen glyfosaat en kanker verschilt vaak en veel. Verder is er een substantiële kans op kanker voor de nul-dosering (de controlegroep). Dit alles maakt dat het zoeken naar een model dat recht doet aan de geobserveerde data, én aan de modelassumpties, buitengewoon lastig is.

Ik heb echt moeten zoeken om die relatie te vinden. Door de bocht genomen lukt het mij niet om met behulp van de frequentistische statistiek een relatie aan te tonen tussen dosering en kanker. De bevindingen zijn vaak niet statistisch significant wanneer ik tweezijdig toets. Een uitstap naar een eenzijdige toets voegt daar weinig aan toe én maakt dat we

moeten aannemen dat het schatten van de relatie tussen dosering en kans op kanker een harde grens heeft in het schatten van de relatie. Ik voeg dan een assumptie die zich maar moeilijk laat verdedigen. Een eenzijdige toets heeft namelijk niet zo veel te maken met de richting van de relatie, maar eerder met het afkappen van de onzekerheid. In een dossier als deze, waarin de onzekerheid groot is, kan dat geen juiste methodiek zijn voor het bepalen van een relatie.

Een overstep van frequentistische statistiek naar Bayesiaanse statistiek laat zien dat modellen die dosering meenemen als verklarende variabele niet per se beter passen bij de data. Pas als we de dosering opdelen in een controlegroep en een behandelgroep lukt het mij om een relatie te tonen tussen glyphosaat en de kans op kanker. Althans, het model dat het effect van glyphosaat meeneemt wordt meer door de data ondersteund.

Uiteindelijk lijkt het erop dat alleen in de Swiss Albino ratten, op basis van één studie en dan met name bij vrouwen, de relatie tussen glyphosaat en de kans op kanker duidelijk is: een toename van 8% vanuit het model. De toename op kanker, vanuit de data, bedraagt dan 4%. Beide getallen kennen een aanzienlijke onzekerheidsmarge. Daarmee kunnen ze niet doorslaggevend zijn voor het gehele dossier.

We kunnen denk ik met dit rapport concluderen dat onderzoek naar glyphosaat beter moet en beter kan, maar daarvoor moet de data ook op het niveau van het dier worden gemeten waarbij ook wordt gekeken naar de factor tijd. Dat ontbreekt nu. Verder hebben we het hier over dierproeven en niet over menselijke studies.

Deze bevindingen zijn een stuk milder dan de uitspraken vanuit de BNNVARA/ZEMBLA reportage waarin zo sterk werd opgeroepen om de tweezijdige toets te vervangen door de eenzijdige toets. Het uitblijven van die vervanging zou wijzen op het moedwillig negeren van bewijs dat glyphosaat kankerverwekkend is. Nu ben ik geen toxicoloog en heb dus voornamelijk gekeken naar de statistiek achter de studie van Portier die zo vaak werd aangehaald in deze reportage en vervolgartikelen. Over de biologische *mode of action* laat ik mij in dit rapport niet uit. Mij ging het er boven alles om, om de relatie tussen data, statistiek en gemaakte uitspraken beter te duiden. Dit rapport is dus maar één enkel onderdeel in een groter dossier.

## Inleiding

In dit rapport ga ik aan de slag met het BNNVARA rapport van 16 september 2024, getiteld “Kankerrisico door pesticiden decennialang ‘verkeerd’ ingeschat”<sup>1</sup>. In dit artikel, en de bijbehorende Zembla reportage, worden een aantal uitspraken gedaan die betrekking hebben op de statistiek zoals deze wordt toegepast in onderzoek naar pesticiden. Specifiek gaat het over het onterecht toepassen van zogenaamde tweezijdige toetsen. In het artikel wordt gesproken van een ‘verkeerde rekenmethode die het risico op kanker kan versluieren’:

*Alle in Nederland toegelaten bestrijdingsmiddelen zouden opnieuw moeten worden getoetst om te kijken of ze kankerverwekkend zijn. Dat zegt hoogleraar Milieubiologie Geert de Snoo (Universiteit Leiden), ook oud-bestuurder bij de Nederlandse bestrijdingsmiddelenautoriteit Ctgb, tegen Zembla. Zijn oproep wordt ondersteund door andere toonaangevende deskundigen. Volgens De Snoo wordt er momenteel bij de toelating “een verkeerde rekenmethode gebruikt, die het risico op kanker kan “versluieren”.*

In deze inleiding zal ik meerdere uitspraken de revue laten passeren en deze presenteren als quotes. Deze quotes geven de aanleiding voor dit rapport, omdat er onterecht wordt gesuggereerd dat een ‘onjuiste’ methode kan worden vervangen door een ‘juiste’ methode. Ook wordt onterecht gesuggereerd dat die ‘juiste’ methode de enige ‘objectieve’ methode is.

Die objectieve rekenmethode bestaat niet. Rekenen in het algemeen, en zeker statistiek in het bijzonder, kan niet zonder aannames en het is belangrijk om de gemaakte aannames transparant en zorgvuldig te communiceren. De stelling die door Prof. Dr. Geert de Snoo en anderen wordt gemaakt, kan dus niet zomaar gemaakt worden. Sterker nog: een gedegen motivatie is vereist, maar die motivatie wordt niet gegeven. Daarmee wordt, wederom, de suggestie gemaakt dat de oplossing makkelijk is en daarmee voor de hand ligt. In dit rapport zal ik door middel van voorbeelden aantonen dat het toepassen van de ‘juiste’ statistiek dezelfde kritiek met zich mee brengt als het toepassen van de ‘onjuiste’ statistiek. Het uiteindelijk doel is om helder te maken wat de invloed van aannames is op welke statistiek dan ook.

---

<sup>1</sup> <https://www.bnnvara.nl/zembla/artikelen/kankerrisico-door-pesticiden-decennialang-verkeerd-ingeschat>

## De kritiek op de statistiek

Voordat een bestrijdingsmiddel in de landbouw mag worden gebruikt is een pesticide-fabrikant verplicht om met proefdieronderzoek aan te tonen dat de stof niet kankerverwekkend is. De industrie levert de gegevens van het onderzoek aan, en de toelatingsautoriteiten - in Nederland is dat het Ctgb - beoordelen de resultaten. Om te onderzoeken of een bestrijdingsmiddel kankerverwekkend is, worden proefdieren (vaak ratten of muizen) aan het middel blootgesteld. Na een groot aantal maanden wordt de proef gestopt en wordt gekeken of de blootgestelde dieren vaker kanker hebben gekregen dan de niet-blootgestelde dieren (de zogeheten controlegroep).

In het programma wordt het statistisch onderdeel van deze procedure als volgt beschreven:

*“Om toeval uit te sluiten (proefdieren kunnen ook ziek worden door andere oorzaken), worden de resultaten statistisch geanalyseerd. Daarbij kan gekozen worden voor de zogeheten eenzijdige of tweezijdige test. Bij eenzijdig testen wordt alleen gekeken naar een toename aan ziekte bij de proefdieren. Bij tweezijdig testen wordt zowel een toename als een afname onderzocht. Tweezijdig testen wordt ook wel bij medicijnonderzoek gebruikt, om zo de voor- en nadelen van een medicijn te wegen. Maar als je alleen wilt weten of iets kanker kan veroorzaken, is eenzijdig testen de enige juiste keuze, legt hoogleraar De Snoo uit. Volgens De Snoo “slaat het nergens op” om pesticiden tweezijdig te testen: “We gaan er niet vanuit dat de bestrijdingsmiddelen die we buiten gebruiken een genezende werking hebben”. Het gebruik van de eenzijdige test is volgens de hoogleraar veel preciezer. En het gebruik van de verkeerde test kan de indruk wekken dat er niks aan de hand is, terwijl het landbouwgif in kwestie wel degelijk kankerverwekkend is. Met de tweezijdige test zie je het kankerverwekkende effect niet, zegt De Snoo, terwijl dat juist is waar de test voor is bedoeld. De tweezijdige test is gunstig voor de pesticide-industrie, omdat minder snel aan het licht zal komen dat een stof kankerverwekkend is. Zo kan een gevvaarlijk bestrijdingsmiddel toch worden toegelaten.*

Vervolgens vraagt men zich in de rapportage hardop af waarom de tweezijdige test toch gebruikt wordt als deze onjuist blijkt te zijn? Volgens de Snoo komt dit omdat de tweezijdige

test in lijn is met de studieprotocollen van de beschikbare studies. Studieprotocollen worden opgesteld voor aanvang van een dierproef en beschrijven hoe deze zal worden uitgevoerd. De uitspraak die dan volgt vormt de kern van de kritiek die Geert de Snoo wil aanvoeren om al het onderzoek opnieuw te analyseren:

*“Als tweezijdig testen is voorgeschreven, deugen de protocollen niet”, zegt Geert de Snoo. “We zijn er niet in geïnteresseerd of glyfosaat ook als geneesmiddel op de akker kan worden gespoten”, zegt De Snoo, en dus moet je eenzijdig toetsen.*

De fout is volgens Geert de Snoo dus zo ingrijpend dat ook opname in het studieprotocol niet voldoende is. Wederom wordt niet heel duidelijk waarom dit fout is. De stelling dat glyfosaat geen positieve bijdrage kan leveren mag dan wel relevant zijn vanuit een chemisch / biologisch standpunt, maar er zijn ook statistische consequenties aan verbonden. Daar wordt helaas niet op ingegaan. Uit alles blijkt dat het toepassen van de juiste statistiek helder moet maken wat eigenlijk al helder moet zijn: glyfosaat is kankerverwekkend.

Vervolgens komt er een andere expert aan het woord: de Duitse toxicoloog Peter Clauzing. Deze expert in de toxicologie nam, in opdracht van milieuorganisatie HEAL<sup>2</sup>, het toelatingsrapport voor glyfosaat onder de loep. Hij ontdekte dat bij vier proefdierstudies statistisch significant verbanden met kanker werden gevonden wanneer de eenzijdige test wordt gebruikt. Wordt echter de tweezijdig test gebruikt, zoals de autoriteiten hebben gedaan, dan verdwijnt dit effect<sup>3</sup>. Dit schiet in het verkeerde keelgat bij de Brusselse advocaat Antoine Bailleux die namens milieuorganisaties de Europese Commissie voor de rechter om de toelating van glyfosaat aan te vechten. Ook hij wijst op de tekortkomingen in de rekenmethode:

*Het wordt nog schimmiger. Het Ctgb stelt dat er is gekozen voor tweezijdig testen, omdat dit in lijn zou zijn met de studieprotocollen. Maar na herhaaldelijk doorvragen van Zembla blijkt dat de tweezijdig test helemaal niet staat vermeld in de studieprotocollen: “In de studieprotocollen staat niet explicet dat er sprake is van tweezijdig testen”, erkent het Ctgb*

---

<sup>2</sup> <https://www.env-health.org/>

<sup>3</sup> <https://www.env-health.org/wp-content/uploads/2022/06/HEAL-How-the-EU-risks-greenlighting-a-pesticide-linked-to-cancer-2022.pdf>

*uiteindelijk. Op dit punt is het rapport dus niet correct. De onderbouwing van het Ctgb voor de tweezijdige test is volgens Bailleux “een serieuze vergissing, of een poging om te misleiden”. Bailleux stelt dat er “geen flard” wetenschappelijk bewijs is dat glyfosaat zou kunnen beschermen tegen kanker. De keuze voor tweezijdig testen is volgens hem daarom ook onbegrijpelijk. Ctgb: tweezijdig testen is ‘standaard’*

Dit moet dus anders en wat volgt is een alternatief voorstel van Geert de Snoo en collega’s:

*Ons onderzoek roept de vraag op hoe het dan zit met al die andere bestrijdingsmiddelen waaraan we worden blootgesteld. Het probleem blijkt veel groter. Want een woordvoerder van het Ctgb schrijft in een e-mail als wij doorvragen over de testmethode bij glyfosaat: “Er is een goede wetenschappelijke uitleg voor de vraag naar eenzijdig of tweezijdig toetsen. Er wordt immers conform Europese afspraken altijd standaard tweezijdig getoetst [...]”. Hoogleraar De Snoo noemt dit “slechte statistiek” en “slechte wetenschap” en het maakt hem “ongerust”. De Snoo roept op om alle toegelaten pesticiden opnieuw tegen het licht te houden. Zijn oproep wordt gesteund door andere toonaangevende deskundigen. Emeritus-hoogleraar Toxicologie Martin van den Berg (Universiteit Utrecht) deelt de zorgen van De Snoo. Volgens hem werkt de tweezijdige test “in het voordeel van de pesticide-industrie” omdat bewijs voor kanker wordt “verdund”.*

Geert de Snoo staat trouwens niet alleen in zijn oproep om alle pesticiden opnieuw te toetsen. Los van de Duitse toxicoloog Peter Clausing, advocaat Antoine Bailleux en Martin van den Berg, zouden ook Hoogleraar Risicobeoordeling Ad Ragas (Radboud Universiteit) hoogleraar Ecotoxicologie Martina Vijver (Universiteit Leiden) graag een herbeoordeling door middel van een “onafhankelijk panel” zien:

*Om zo afstand te creëren tot “de politiek en de economie”, want volgens Ad Ragas is de gang van zaken rondom glyfosaat en het tweezijdig testen “schimmig”.*

De uitspraken die gedaan worden in deze rapportage winden er geen doekjes om: de industrie bagatelliseert willens en wetens de kankerverwekkend aart van pesticiden zoals glyfosaat. Maar de industrie is hier volgens BNNVARA/Zembla niet alleen in. Ook het Ctgb

doet volgens hen een spreekwoordelijk oogje dicht. Maar in plaats van letterlijk weg te kijken is men van mening dat het Ctgb negatieve gevolgen wegredeneert door gebruik te maken van een standaard protocol van het Ctgb<sup>4</sup>:

*Deskundigen stellen tegenover Zembla dat het Ctgb bij deze beoordeling ‘vooringenomen’ te werk is gegaan, en dat aanwijzingen dat glyfosaat kanker kan veroorzaken ‘systematisch zijn weggedeneerd’. In dit artikel leggen we stap voor stap uit hoe dat zit.*

Dit zijn ernstige beschuldigingen. Om hun punt duidelijk te maken worden uitknipsels getoond die laten zien hoe de Europese burger juridisch en uiteindelijk ook door middel van wetenschap beschermd wordt tegen kankerverwekkende stoffen:

*Om te bepalen of een stof ‘verondersteld’ kankerverwekkend is voor mensen, worden dierproeven gebruikt. In de Europese verordening staat dat als er ‘voldoende bewijs’ voor kanker bij proefdieren is, ervan uit wordt gegaan dat de stof ook voor mensen kankerverwekkend is. Dat leidt dan in principe tot een verbod.*

De definities van ‘voldoende bewijs’ komen van het Internationaal Agentschap voor Kankeronderzoek (IARC) dat zelf in 2015 onderzoek heeft gedaan naar glyfosaat<sup>5</sup>:

*Het IARC-panel had toen inzage in twee proefdierstudies met muizen. Deze studies bevatten volgens IARC ‘voldoende bewijs’ voor kanker. In andere woorden: het onderzoek van IARC toont aan dat er genoeg bewijs is voor een Europees glyfosaatverbod.*

Uit dat onderzoek kan en mag volgens Ad Ragas dus maar één conclusie volgen en dat is een algeheel verbod:

*Het gaat volgens hoogleraar risicobeoordeling Ad Ragas (Radboud Universiteit) om statistische significante uitkomsten, die met name bij de proefdieren die hogere doseringen*

---

<sup>4</sup> <https://www.bnnvara.nl/zembla/artikelen/hoe-het-ctgb-aanwijzingen-voor-kanker-door-glyfosaat-wegredeneert>

<sup>5</sup> <https://www.iarc.who.int/wp-content/uploads/2018/07/MonographVolume112-1.pdf>

*kregen voorkomen. ‘En volgens de regels moet dat worden gezien als veroorzaakt door glyfosaat’.*

Maar het Ctgb oordeelde anders<sup>6</sup>:

*Maar het Ctgb heeft geoordeeld dat de vijf muizenstudies waarin een verband met kanker is vastgesteld ‘onvoldoende’ bewijs opleveren voor de classificatie van glyfosaat. Volgens het Ctgb zit er ‘geen bewijs’ voor kanker in deze proefdierstudies. Hoe kan dat?*

In die standaard zit volgens BNNVARA/Zembla het systematische probleem van wegredeneren:

*Het gaat volgens hoogleraar risicobeoordeling Ad Ragas (Radboud Universiteit) om statistische significante uitkomsten, die met name bij de proefdieren die hogere doseringen kregen voorkomen. ‘En volgens de regels moet dat worden gezien als veroorzaakt door glyfosaat’. Maar het Ctgb heeft geoordeeld dat de vijf muizenstudies waarin een verband met kanker is vastgesteld ‘onvoldoende’ bewijs opleveren voor de classificatie van glyfosaat. Volgens het Ctgb zit er ‘geen bewijs’ voor kanker in deze proefdierstudies. Hoe kan dat?*

Wat volgt is een opsomming van kritiek die ik niet allemaal kan bespreken in dit rapport, gewoonweg omdat de kennis mij hier ontbreekt. Zo wordt er gesproken over ‘limietdosering’, ‘spontaniteit van tumoren’, een ‘virus onder de proefdieren’ en ‘verkeerde controlegroepen’. Ik laat mij in dit rapport niet uit over de correctheid van deze kritiek, omdat ik geen expert ben op het gebied van de toxicologie. Waar ik wel kennis van heb is statistiek en modellering en op dat vlak is er heel veel kritiek. Daar zal ik mij dus voornamelijk op richten. We gaan verder:

*Bij de muizenstudies worden grofweg twee statistische methoden gebruikt: de ‘trend test’ en de ‘pairwise comparison’. Volgens de OECD-handleidingen die van toepassing zijn kunnen beide methoden worden gebruikt om uit te sluiten ‘dat toeval de oorzaak van de uitkomst’ is.*

*In andere woorden: met beide methoden kan met zekerheid worden vastgesteld dat de*

---

<sup>6</sup> [https://food.ec.europa.eu/document/download/ba487ec2-3e60-4db2-8bc8-e095cf6e792e\\_en?filename=pesticides\\_aas\\_agg\\_report\\_202106.pdf](https://food.ec.europa.eu/document/download/ba487ec2-3e60-4db2-8bc8-e095cf6e792e_en?filename=pesticides_aas_agg_report_202106.pdf)

*tumoren bij de proefdieren door de blootstelling (glyfosaat in dit geval) komen. De rapporterende lidstaten leggen de lat echter hoger. Wanneer niet beide testen een zogeheten ‘statistisch significant’ resultaat opleveren, ziet het Ctgb dit als reden om de waargenomen tumoren minder zwaar mee te wegen.*

Dit is niet dezelfde kritiek als de eerder genoemde kritiek over het een-of tweezijdig toetsen. Hier gaat het er namelijk om of het zien van een verschil in een bepaalde richting al voldoende aanleiding kan geven om te spreken van een oorzaak-gevolg relatie. Ook wordt er een behoorlijk stevige uitspraak gemaakt over het ‘uitsluiten van toeval’. Over het gebruik van het woord ‘toeval’ kom ik later nog uitgebreid op terug. Uiteindelijk gaat ook dit rapport van BNNVARA/Zembla over tweezijdig toetsen:

*Bij de statistische analyse van de uitkomsten, moet nog een andere keuze worden gemaakt. De resultaten kunnen ‘enkelzijdig’ of ‘tweezijdig’ worden getoetst. Een enkelzijdige toets kijkt alleen naar een toename van tumoren, terwijl de tweezijdige toets zowel naar toename als afname van kanker kijkt. ‘Met gebruik van de tweezijdige toets zwak je de statistische significantie af’, legt toxicoloog Peter Clusing uit. ‘Je hebt hierdoor twee keer zoveel tumoren nodig om statistische significantie te bereiken.’*

Verder worden de OECD richtlijnen aangehaald die stelt dat een eenzijdige toets meer gepast zou zijn:

*Dat beaamt ook een woordvoerder van de OECD desgevraagd: ‘In relatie tot tumorincidentie in een toxicologisch onderzoek, is het logisch om de eenzijdige toets te gebruiken.’ ... ‘Toch gebruikt het Ctgb de tweezijdige toets. Het argument dat hiervoor wordt opgevoerd is dat dit zou staan in de ‘protocollen’ van de muizenstudies. Studieprotocollen worden opgesteld voor aanvang van een proef, en beschrijven hoe een test zal worden uitgevoerd en bevatten details over hoe de uitkomsten ervan zullen worden geanalyseerd.’*

Maar dan zijn we weer terug bij de originele kritiek die nu ook door toxicoloog Paul Scheepers wordt benadrukt:

*Paul Scheepers noemt het echter 'onzinnig' dat door het Ctgb tweezijdig is getoetst. 'Omdat je bij dit soort studies juist een toename verwacht. Daar wil je met je onderzoek naar kijken.'*

Het Ctgb lijkt ook niet helder te kunnen maken waarom er nu wel voor de tweezijdige toets is gekozen en niet voor de eenzijdige toets:

*De woordvoerder stelt nu dat de tweezijdige toets zou zijn gekozen omdat dit in lijn is met 'Europese afspraken'. Ondanks herhaalde vragen van Zembla, kan de woordvoerder niet vertellen welke Europese afspraken dit zijn.*

Emeritus-hoogleraar toxicologie Martin van den Berg (Universiteit Utrecht) en hoogleraar Ad Ragas vinden er uiteindelijk geen doekjes om. Volgens hen kunnen we gewoonweg het volgende concluderen:

*'Ze zijn vooral bezig met het wegschaffen van bewijs', concludeert emeritus-hoogleraar toxicologie Martin van den Berg (Universiteit Utrecht) die op verzoek van Zembla ook heeft meegelezen. 'Daar kun je niet omheen als je het dossier bekijkt. Ze zoeken naar argumenten om het niet te hoeven classificeren als een kankerverwekkende stof.' ... 'Ik zie dat ze wetenschappelijke aanwijzingen dat glyfosaat mogelijk kankerverwekkend is, systematisch wegredeneren', concludeert hoogleraar Ad Ragas wanneer hij het beoordelingsrapport overziet. 'En in ieder geval worden de protocollen, die moeten garanderen dat je één voor één netjes alle stappen zet, geschonden.'*

## **Wat kunnen we hier nu uit afleiden?**

Wat ik u tot nu toe heb voorgesloten is een aaneenschakeling van citaten en stukken uit de twee BNNVARA/Zembla rapportages. Mijn reden voor het aanreiken van deze ietwat droge stof is dat ik met de lezer wil delen op welke uitspraken ik mijn eigen rapport zal baseren. Er zijn namelijk een flink aantal zaken genoemd die van belang zijn om te onderzoeken, maar ik kan mij niet op alles richten.

De bulk van het commentaar gaat echter over de statistiek en daar kan en wil ik mij wel op richten. Volgens de wetenschappers die aan bod zijn gekomen gaat het hier niet om

schoonheidfoutjes, of andere aannames, maar is er sprake van ‘wegmoffelen’, ‘systematisch wegredeneren’ en ‘schimmige praktijken’. Volgens deze heren is de wetenschap geschonden. Hun oproep om de data opnieuw te analyseren met de ‘juiste techniek’ is daarmee een logisch gevolg.

Maar wat is de juiste techniek dan? Volgens de onderzoekers zelf is het antwoord kraakhelder: eenzijdig testen. Toch denk ik dat over deze vervanging te simpel wordt gedacht. Om tweezijdig testen (of toetsen zoals het ook wel wordt genoemd) te vervangen door eenzijdig testen zul je namelijk een aantal aannames moeten maken. En die aannames reiken verder dan de stelling dat “we niet zijn geïnteresseerd of glyfosaat ook als geneesmiddel op de akker kan worden gespoten”. Geen van beide rapportages rept over die (statistische) aannames. De stelling dat het glashelder wordt dat glyfosaat kankerverwekkend is door het vervangen van tweezijdig door eenzijdig testen kent helemaal geen statistische basis. De statistiek zoals die hedendaags wordt beschreven kent sowieso geen sterke basis: deze staat vol van aannames.

Deze toch wel gedurfde uitspraak vergt van mij de nodige uitleg en dat komt niet omdat de statistiek zelf zo lastig is. Het heeft er juist alles mee te maken dat statistisch testen<sup>7</sup> geen glasharde wetenschap is. Statistisch testen bestaat namelijk grotendeels uit keuzes maken. Toch is veel onderzoekers geleerd dat statistisch toetsen harde wetenschap is en dit is, zo zal ik betogen, een van de redenen waarom er zo makkelijk gesproken wordt over het wisselen tussen één- of tweezijdig testen. In de BNNVARA/Zembla rapportage lijkt het haast alsof de keuze voor een toets natuurlijk valt, maar dat is natuurlijk niet zo (ook al is veel onderzoekers dit wel geleerd)<sup>8</sup>. Van mij mag u verwachten dat ik precies zal gaan uitleggen waarom dat is en waarom de keuze voor deze vorm van statistiek sowieso problematisch is. Over statistiek wordt in het algemeen te makkelijk gedacht en dat gevoel brengt soms meer schade toe dan het bespreken van enige twijfel in dossiers.

## **Wat mag u van mij verwachten?**

Statistiek is geen eenvoudig vak, maar dat is niet direct het gevolg van de onderliggende materie. Wat de statistiek zo lastig maakt zijn de aannames die er aan ten grondslag liggen.

---

<sup>7</sup> Ik gebruik testen en toetsen afwisselend.

<sup>8</sup> In de statistiek zijn veel beslisbomen te vinden die onderzoekers krijgen aangereikt om de keuze voor een soort statistiek te rechtvaardigen. Ironisch genoeg is die beslisboom gebaseerd op de aard van de data en het wel of niet voldoen aan bepaalde aannames. Hier ontbreekt het steeds weer aan.

We hoeven een uitleg over statistisch testen dus ook niet moeilijker te maken dan het is, maar het is wel zaak om het duidelijk uit te leggen en te laten zien wat de gevolgen zijn als aannames veranderen. Een zachte introductie kan hier wonderen verrichten en dat is precies wat ik hier zal nastreven. Toch kan ik niet geheel voorkomen dat de materie op een bepaald moment wat complex wordt. Vandaar dat dit document in stappen is opgedeeld en daarmee bestaat uit twee grote delen. Deel 1 gaat over statistiek, kansen en aannames. Deel 2 gaat over de glyfosaat data zoals gerapporteerd door Portier.

Zo wil ik eerst laten zien hoe wij wetenschappers aan de slag gaan met data nadat wij deze hebben ontvangen. Dat het van belang is om deze data te beschrijven en te visualiseren. Dat het goed is om te weten hoe de data is opgedeeld en hoe de samenstelling is. Dit helpt ons om te bepalen welke waarden we vaak zien en welke waarden meer zeldzaam zijn. Ik zal dit alles doen op basis van een voorbeeld waar geen waardeoordeel aan verbonden is: de lengteverdeling van Nederlandse mannen en vrouwen.

Vervolgens zal ik laten zien wat een kansverdeling is en waarom deze belangrijk zijn. Dan zal ik tonen hoe deze kansverdelingen worden gebruikt om cijfers te labelen zodat wetenschappers een antwoord kunnen presenteren op de vraag of een bevinding ‘normaal’ is of toch ‘anders’. Dan volgt het onderwerp van ‘statistische significantie’ en hoe deze term wordt verward met toeval en causaliteit. We zagen het al in de Zembla rapportage:

*Bij de muizenstudies worden grofweg twee statistische methoden gebruikt: de ‘trend test’ en de ‘pairwise comparison’. Volgens de OECD-handleidingen die van toepassing zijn kunnen beide methoden worden gebruikt om uit te sluiten ‘dat toeval de oorzaak van de uitkomst’ is.*

Het feit dat statistische significantie zo belangrijk is geworden in onze hedendaagse wetenschap maakt dat we ons **moeten** afvragen wat deze term nu eigenlijk betekent. Wat betekent het als iets significant is? Berekent significantie iets anders in de statistische wereld dan in de ‘echte’ wereld. Zijn er kansen verbonden aan deze significantie of gaat het hier om een dichotome beslissing? Wat is de voorspellende waarde van een statistisch significant effect? Berekent dit automatisch dat er dan ook sprake is van een causaal verband? En hoe zeker mag je zijn dat een volgend onderzoek niet een tegenovergesteld verband laat zien?

In Deel 2 ga ik aan de slag met de glyfosaat data en zal ik bovenal proberen te repliceren wat Portier gedaan heeft. Daarnaast zal ik de data analyseren op de manier zoals ik het zelf zou doen als ik de data voor het eerst zou zien. Om transparant te zijn in mijn methodiek heb ik de gebruikte data en codes op een [GitHub pagina](#) staan.

Voor wie geen behoefte heeft aan Deel 1 kan direct door naar Deel 2. De rest laat ik aan u, de lezer, over.

## Kansen, verdelingen en context

Zoals ik al beschreef zal ik beginnen met een voorbeeld waar geen waardeoordeel aan verbonden is. Niet alleen leest dit luchtiger, maar het zal ook echt helpen om de materie beter te begrijpen. Wij mensen zijn namelijk niet vrij van oordelen en de kennis die wij tot ons nemen is afhankelijk van onze overtuigingen en dus ook van de eerdere kennis die wij tot ons hebben genomen. Daarom zal ik hier nog niet schrijven over de glyfosaat data waar zoveel kritiek op is, maar over de lengte van mensen. Hoewel geen enkel onderwerp 100% vrij van context is, ben ik mij er zeker van dat rondom de lengte van mensen een vele kleinere negatieve connotatie hangt dan aan glyfosaat. Belangrijker nog is dat het onderwerp lengte u als lezer alles zal bieden om kennis te maken met de fundamentele aannames van het statistisch toetsen.

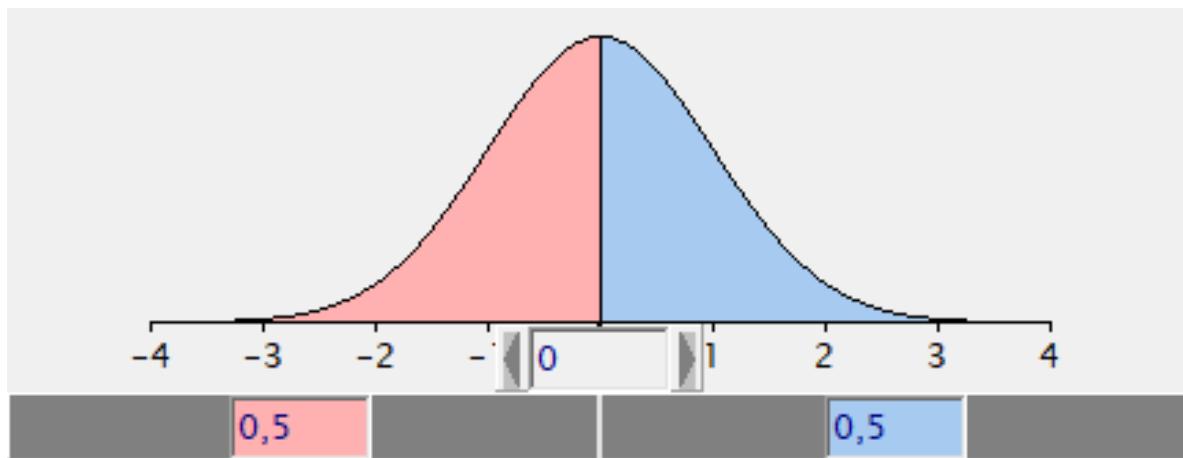
## De normaalverdeling

Laten we dus beginnen met data over lengte. Het vinden van gegevens over de lengte van Nederlandse mannen en vrouwen is trouwens nog niet zo makkelijk, want ik krijg vooral gemiddelde waarden. Dat is niet voldoende om een zogenaamde normaalverdeling (**Figuur 1**) te beschrijven: deze bestaat uit een gemiddelde en een spreidingsmaat (vaak de standaarddeviatie). De normaalverdeling is een kansverdeling die beschrijft hoe vaak een bepaalde waarde voorkomt. Elke waarde krijgt dus een frequentie en die frequentie wordt geschaald van 0 tot 1. Daarmee kunnen we een figuur maken zoals **Figuur 1** wat een gemiddelde van 0 heeft en een standaard deviatie van 1.

Hoewel er veel meer kansverdelingen zijn dan de normaalverdeling wordt deze verdeling het meest gebruikt<sup>9</sup>. Veelvuldig gebruik maakt het niet de beste verdeling die er is, maar voor ons voorbeeld is het toereikend en om een normaalverdeling te maken moet ik dus twee parameters weten te vullen: het gemiddelde en de standaarddeviatie.

---

<sup>9</sup> <https://nl.wikipedia.org/wiki/Kansverdeling>



**Figuur 1.** Een theoretische normaalverdeling met gemiddelde 0 en standaard deviatie 1. Het kenmerk van de normaalverdeling is zijn 'bel'-vorm.

Uiteindelijk lukt het mij om, via een website van het Radboud UMC over groeistoornissen<sup>10</sup>, het volgende te lezen:

*Nederland heeft de langste bevolking ter wereld. Mannen zijn gemiddeld 184 cm lang, vrouwen 170,6 cm. De langste 2,5 procent van de bevolking is langer dan 198 cm (mannen) of 184 cm (vrouwen).*

Dit zijn alleen gegevens over het gemiddelde. Informatie over de standaard deviatie volgt later:

*Rondom de gemiddelde lengte bestaat een spreiding, zowel naar beneden als naar boven. Deze drukken we uit in een gestandaardiseerde maat voor spreiding in de bevolking, de zogenaamde Standaard Deviatie (SD). 95 procent van de bevolking bevindt zich met zijn lengte tussen de lijnen - 2 SD en + 2 SD. Voor mannen is dit tussen de 170 cm en 198 cm. En voor vrouwen tussen de 158 cm en 184 cm.*

Op basis van deze gegevens kunnen we een normaalverdeling simuleren. Dit is eigenlijk niet zoals het normaliter gaat. Vaak start je met een dataset van lengtes en visualiseer je de verdeling. Dan kijk je welke theoretische verdeling<sup>11</sup> deze gegevens het best beschrijft. Die

<sup>10</sup> <https://www.radboudumc.nl/patientenzorg/aandoeningen/groeistoornis>

<sup>11</sup> <https://nl.wikipedia.org/wiki/Kansverdeling>

gegevens hebben we nu niet, maar alleen de parameters. Voor nu is dat prima, want lengte laat zich prima tonen met een normaalverdeling zoals ik zal laten zien. Laten we de gegevens uit de tekst in **Tabel 1** zetten:

Geslacht	Gemiddelde lengte (cm)	2x Standaard deviatie lengte (cm)
Mannen	184	170 - 198
Vrouwen	170,6	158 - 184

**Tabel 1.** De gegevens zoals verkregen uit de berichtgeving van het Radboud UMC, maar dan in tabelvorm verwerkt.

Vervolgens kunnen we wat gaan rekenen. Als ik namelijk weet dat 2x de standaard deviatie een afstand heeft van 14 cm links ( $184 - 170$ ) en 14 cm rechts ( $198 - 184$ ) dan lijkt het er sterk op dat de verdeling waarschijnlijk normaal verdeeld is. Dit is omdat een normaal verdeling een gemiddelde heeft én een standaard deviatie<sup>12</sup>. Met die aanname weet ik vervolgens ook:

1. dat 1x de standaard deviatie bij mannen 7 cm is,
2. dat 1x de standaard deviatie bij de vrouwen links 6.3 cm ( $170.6 - 158 / 2$ ) is en rechts 6.7 cm ( $184 - 170 / 2$ ). Deze verdeling lijkt dus minder op een normaalverdeling maar zou waarschijnlijk voldoende kunnen worden samengevat door een normaalverdeling.

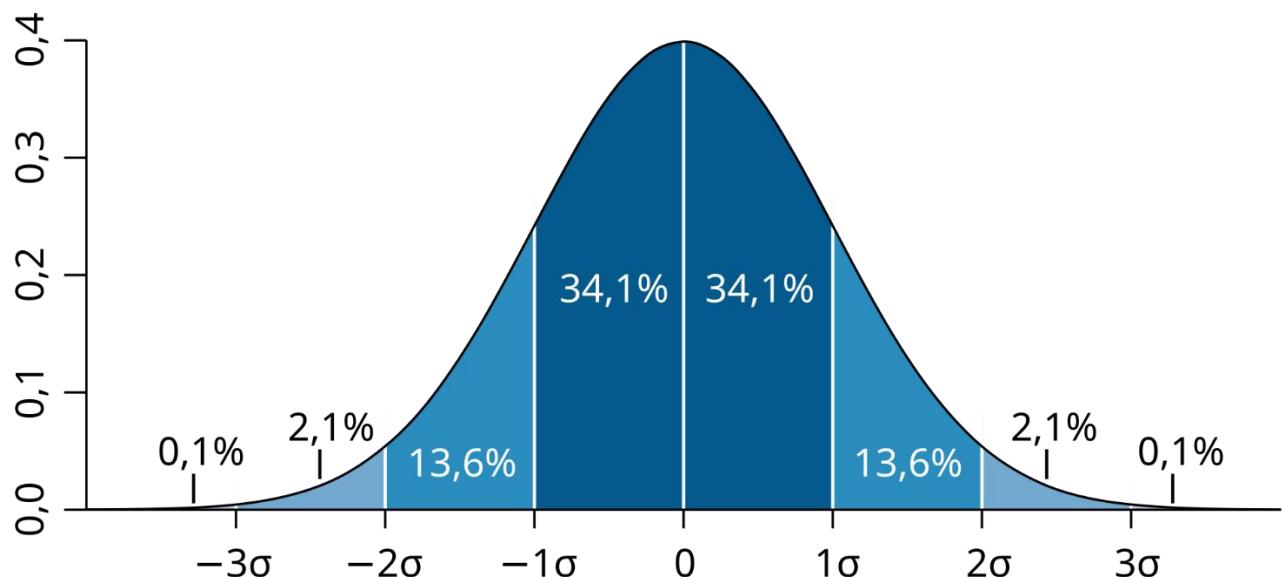
Laten we de proef op de som nemen door de gegevens in een theoretische simulatie te gieten: ik gebruik een theoretische normaalverdeling om een reeks aan waarden te creëren. Als de normaalverdeling passend is, zal ik op basis van het gemiddelde en de standaard deviatie ook een normaalverdeling terugkrijgen. Hoewel ik nu letterlijk terug krijg wat ik erin stop<sup>13</sup>, kan ik wel zien of ik de informatie uit **Tabel 1** kan beschrijven met een normaalverdeling. Als mij dat lukt, dan is mijn vertrouwen groter dat de normaalverdeling een passende verdeling is.

---

<sup>12</sup> Sommige andere verdelingen (er zijn er een hoop!) gebruiken deze parameters ook, maar zijn vaak niet even groot aan weerszijden van het gemiddelde. Omdat de spreiding een even getal is, lijkt het er sterk op dat dit wel het geval is hier.

<sup>13</sup> Ik gebruik een normaalverdeling om een normaalverdeling te maken.

Ik begin met het maken van een figuur<sup>14</sup> voor mannen waarbij het gemiddelde 184 cm is en de standaard deviatie 7 cm. We kunnen testen of deze verdeling passend is door op zoek te gaan naar de waarde die we vinden bij 2x de standaard deviatie. De theoretische verdeling van de normaalverdeling is zichtbaar in **Figuur 2**. Wat ook zichtbaar is, is hoeveel % van de waarden zich bevinden op één, twee én drie standaard deviaties afstand van het gemiddelde. Twee standaard deviaties van het gemiddelde behelst ongeveer 95.4% van de gegevens.



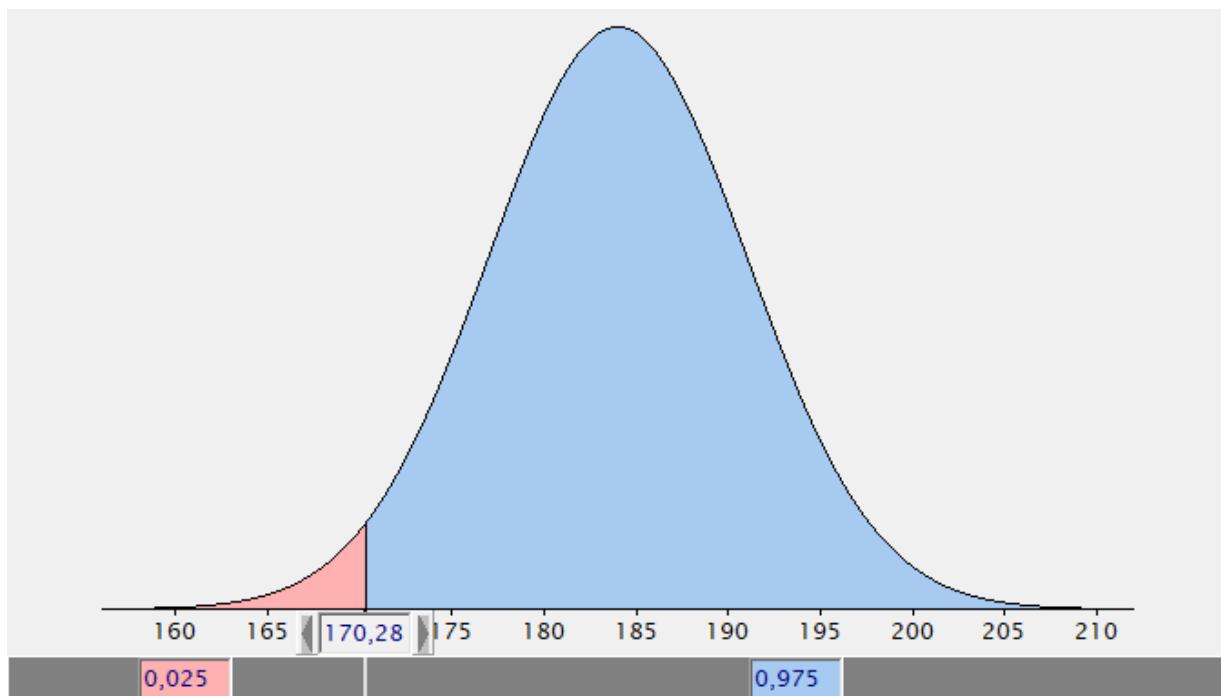
**Figuur 2.** De theoretische verdeling van massa in een normaalverdeling. Ongeveer 68% van de gegevens vallen binnen één standaard deviatie van het gemiddelde. Bij twee standaarddeviaties is dat ongeveer 95%.

Om te zien of de normaalverdeling passend is, moet ik dus op zoek naar de waarde die past bij (ongeveer) 2.5% van de verdeling en 97.5% van de verdeling. **Figuur 3** laat inderdaad zien dat 170 cm de waarde is die past bij twee keer de standaarddeviatie in de verdeling bij mannen. **Figuur 4** toont dat ook voor vrouwen de normaalverdeling passend lijkt te zijn: bij twee keer de standaarddeviatie vinden we de waarde 158 cm<sup>15</sup>. We hebben hiermee dus kunnen laten zien dat, als we de normaalverdeling gebruiken, we de door het Radboud UMC

<sup>14</sup> Hiervoor gebruik ik de PyQRS tool (<https://pyqrs.eu/sk/>; <https://pyqrs.gitlab.io/PyQRS/>)

<sup>15</sup> Dit gebeurt als we kiezen voor een standaard deviatie van 6.3.

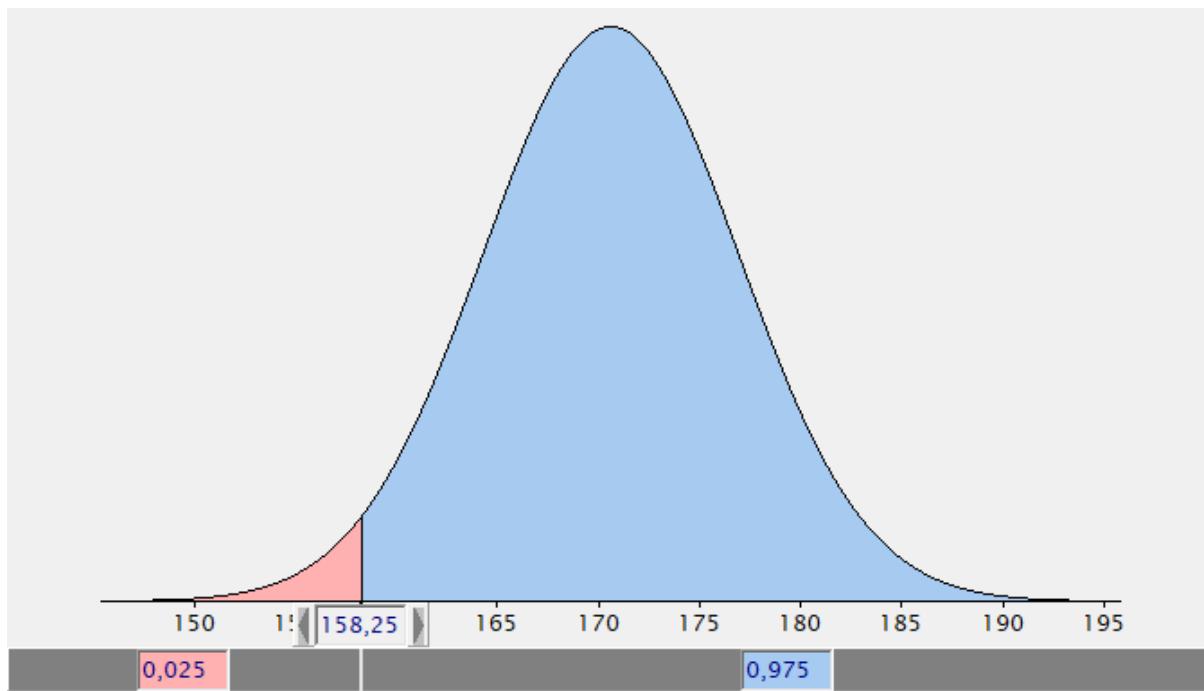
gepresenteerde waarden voor twee keer de standaarddeviatie terugkrijgen. De normaalverdeling lijkt dus passend.



**Figuur 3.** De normaalverdeling met een gemiddelde van 184 cm en een standaard deviatie van 7 cm. Dit is waarschijnlijk hoe de verdeling van lengte bij de Nederlandse mannen verdeeld is.

Wat kunnen we nou met deze informatie? Nou, eigenlijk best veel, want het lijkt erop dat we een wiskundige kansverdeling hebben gevonden die de lengte van de Nederlandse vrouwen en die van de mannen adequaat beschrijft. Op basis van die gegevens kunnen achterhalen hoe vaak een bepaalde waarde voorkomt. We krijgen als het ware een rangorde van gegevens<sup>16</sup>.

<sup>16</sup> Daarmee wordt dan weer vaak getracht om te bepalen of er waarden zijn die we kunnen bestempelen als 'normaal' of 'bijzonder'. Let wel! Dit zijn subjectieve labels aan cijfers en dus wat 'normaal' of 'bijzonder' is, is voor iedereen anders.



**Figuur 4.** De normaalverdeling met een gemiddelde van 170 cm en een standaard deviatie van 6,3 cm. Dit is waarschijnlijk hoe de verdeling van lengte bij de Nederlandse vrouwen verdeeld is.

## Rangorde

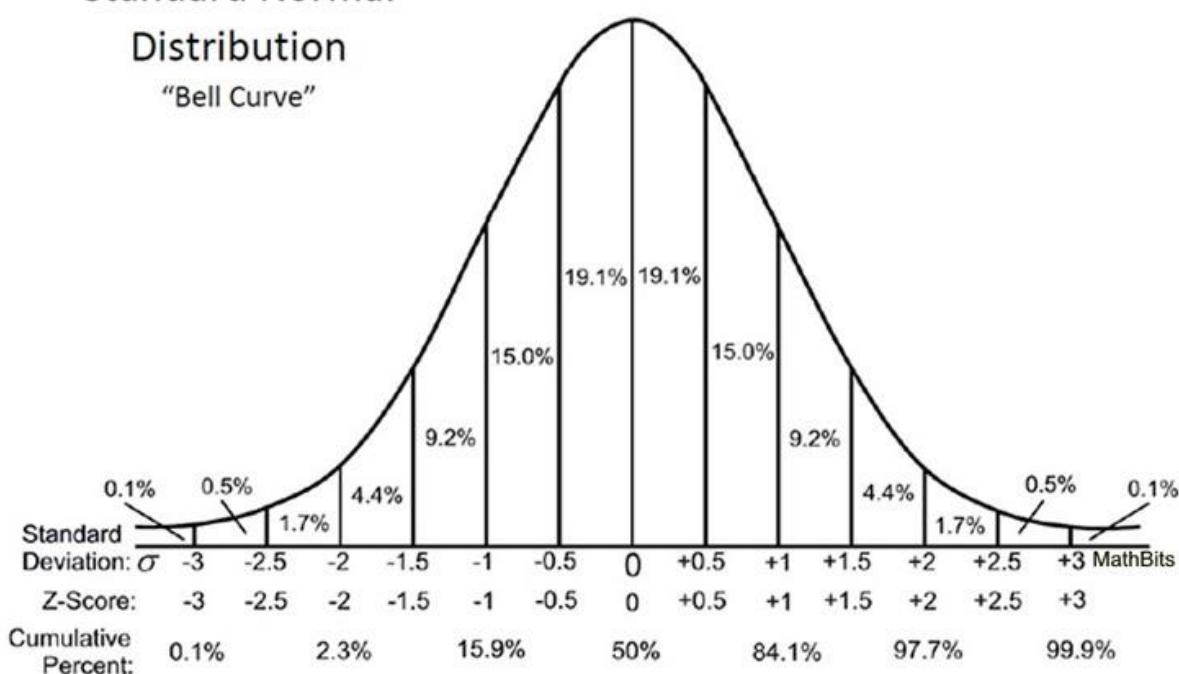
We hebben van zowel de Nederlandse mannen als de Nederlandse vrouwen een kansverdeling. We kunnen nu een aantal dingen gaan doen, waaronder bepalen welke rang een bepaalde lengte heeft. De rang van een waarde bepaalt hoeveel procent van een verdeling onder die waarde zit. Zie ook **Figuur 5** die, in het Engels, laat zien wat de verdeling van massa is per halve standaarddeviatie. Onderaan zie je de cumulatieve verdeling.

Om het concept van rang wat meer duiding te geven kunnen we het volgende doen: mijn lengte (186 cm) nemen. Door mijn lengte te plaatsen in de kansverdeling van de lengte van de Nederlandse man laat **Figuur 6** zien wat het percentage mannen is dat kleiner is dan 186 cm. Dat is 61% van de mannen in Nederland. Daarmee weten we ook het percentage wat groter is: 39%. Mijn lengte heeft dus als rang het 61ste percentiel. We kunnen deze waarde ook anders tonen, namelijk door een cumulatieve verdeling op te zetten. Een cumulatieve verdeling laat beter zien wat de rangorde van waarden is in een kansverdeling en wordt getoond in **Figuur 7**.

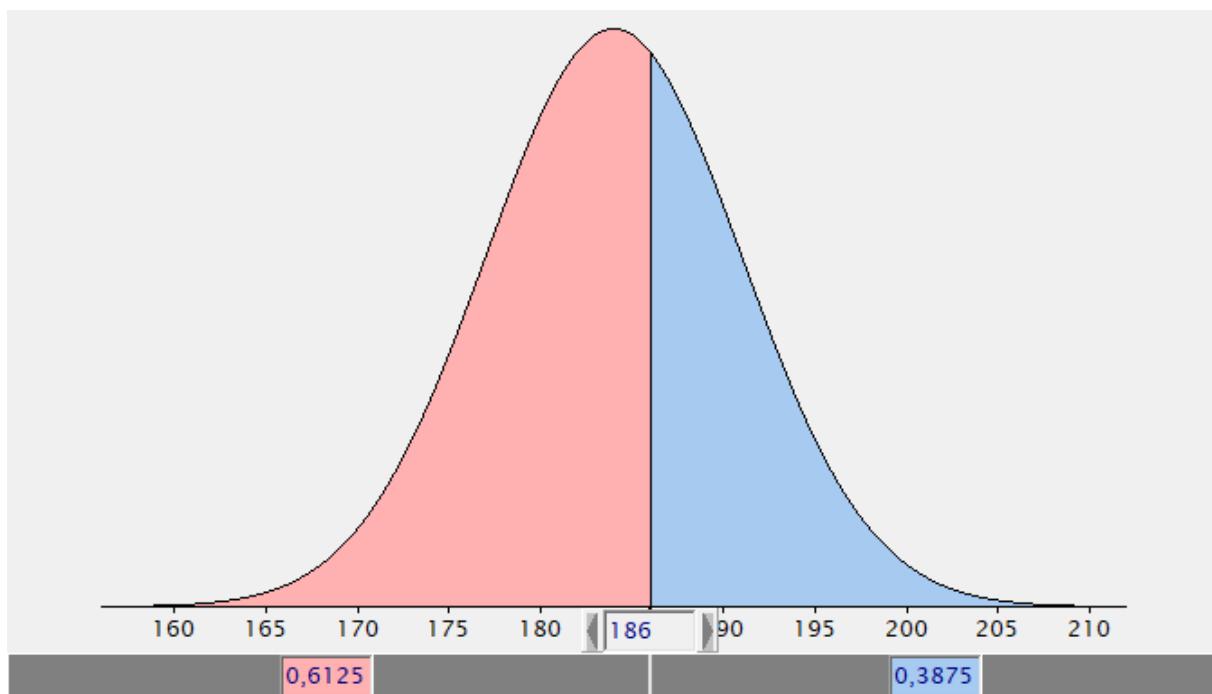
## Standard Normal

### Distribution

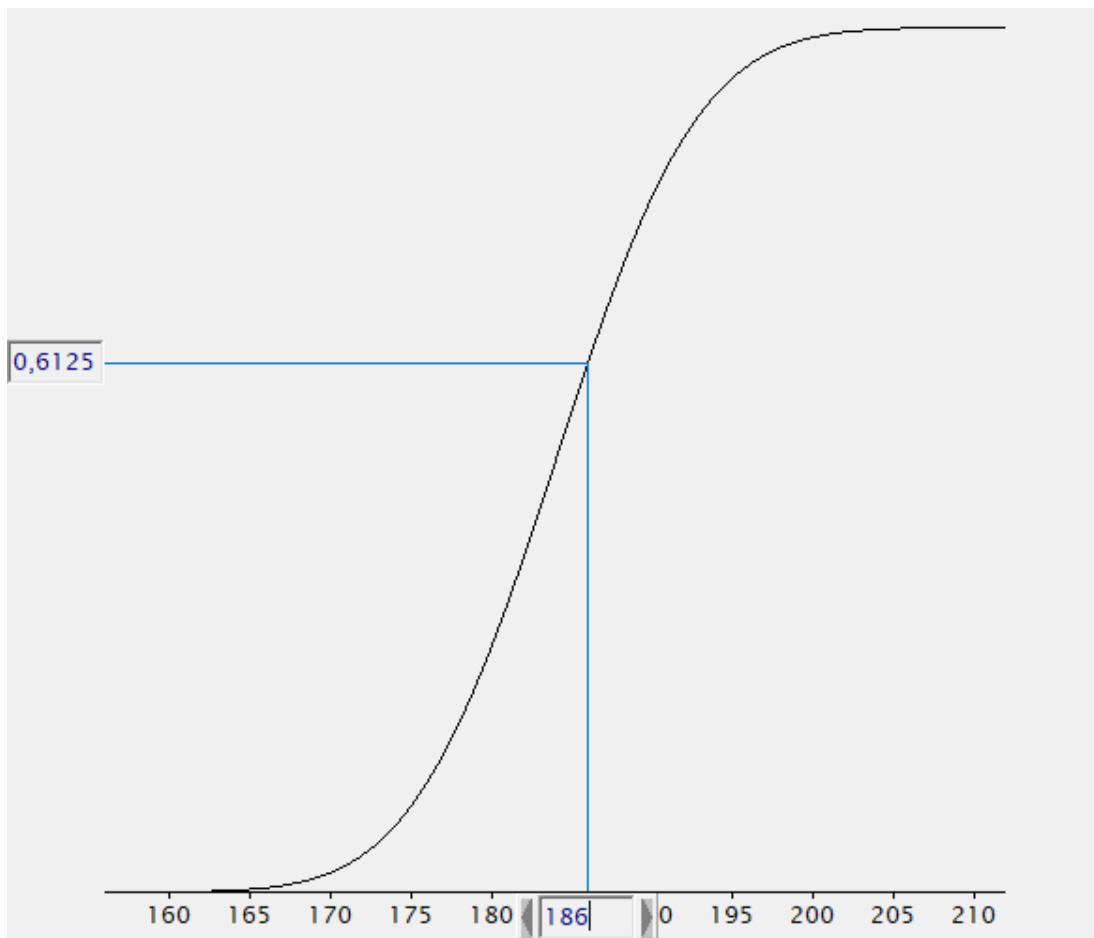
“Bell Curve”



**Figuur 5.** Theoretische kansverdeling van de normaalverdeling met een gemiddelde 0 en standaard deviatie 1. Onderaan staan de cumulatieve percentages.



**Figuur 6.** De plek waar een lengte van 186 cm past in een verdeling van lengtes van Nederlandse mannen.



**Figuur 7.** De cumulatieve kansverdeling bij een gemiddelde van 184 cm met 7 cm standaarddeviatie.

De lengte 186 cm beslaat het 61<sup>ste</sup> percentiel.

## Het buitenbeetje

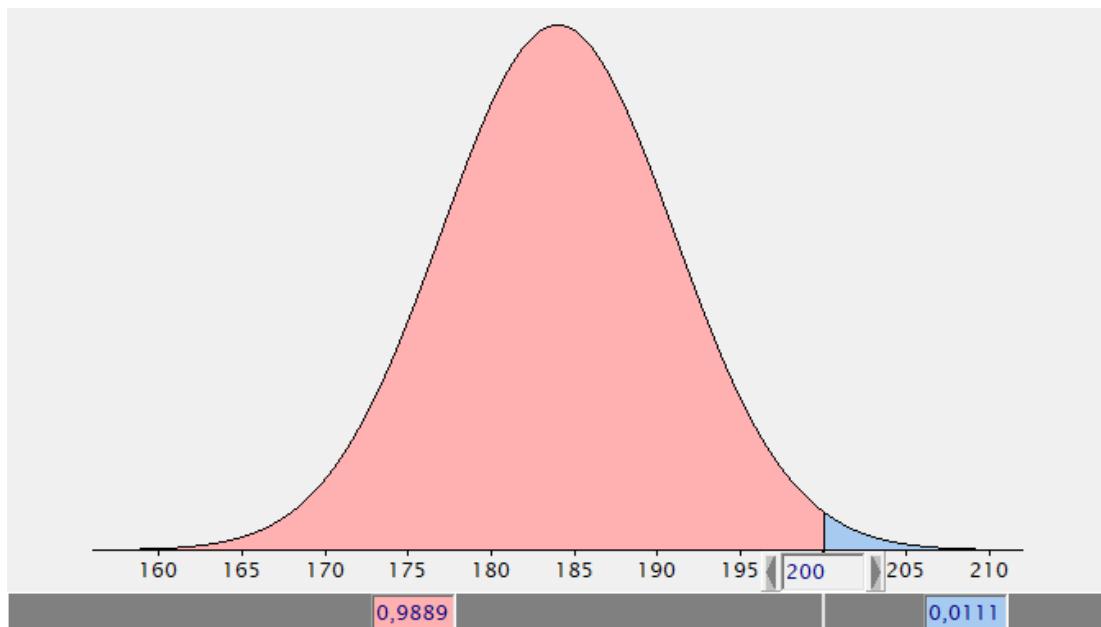
We kunnen dus zeggen dat mijn eigen lengte heel mooi past in de verdeling. De lengte zit gevoelsmatig dicht bij het gemiddelde (2 cm verschil), maar is in rangorde toch 10% verwijderd van dat gemiddelde<sup>17</sup>. Dat komt omdat de grootste massa van onze gegevens zich bevindt rondom het gemiddelde: dat is de aard van de normaalverdeling. We kunnen dus niet zomaar zeggen dat een bepaalde waarde ‘normaal’ is of ‘gek’ en het is belangrijk om de verdeling van waarden te nemen zoals deze zijn. Met 186 cm val ik in het 61<sup>ste</sup> percentiel bij een normaalverdeling van 184 cm met standaard deviatie 7 cm. Dat is wat we nu weten.

Zodra we een kansverdeling hebben kunnen we gaan ‘spelen’ met cijfers. Want hoe zit het met iemand die 200 cm is? Of 150 cm? Waar komen deze lengtes uit in de

<sup>17</sup> In een perfecte normaalverdeling beslaat het gemiddelde het 50<sup>ste</sup> percentiel. Omdat we nu werken met een theoretische normaalverdeling is dit ook zo. Wanneer we metingen doen op basis van observaties is dit zeker niet altijd het geval.

kansverdeling van Nederlandse mannen? Laten we de proef op de som nemen door beide getallen in deze verdeling van lengtes in te voeren en te bepalen wat het percentage mannen is dat kleiner is dan deze lengte.

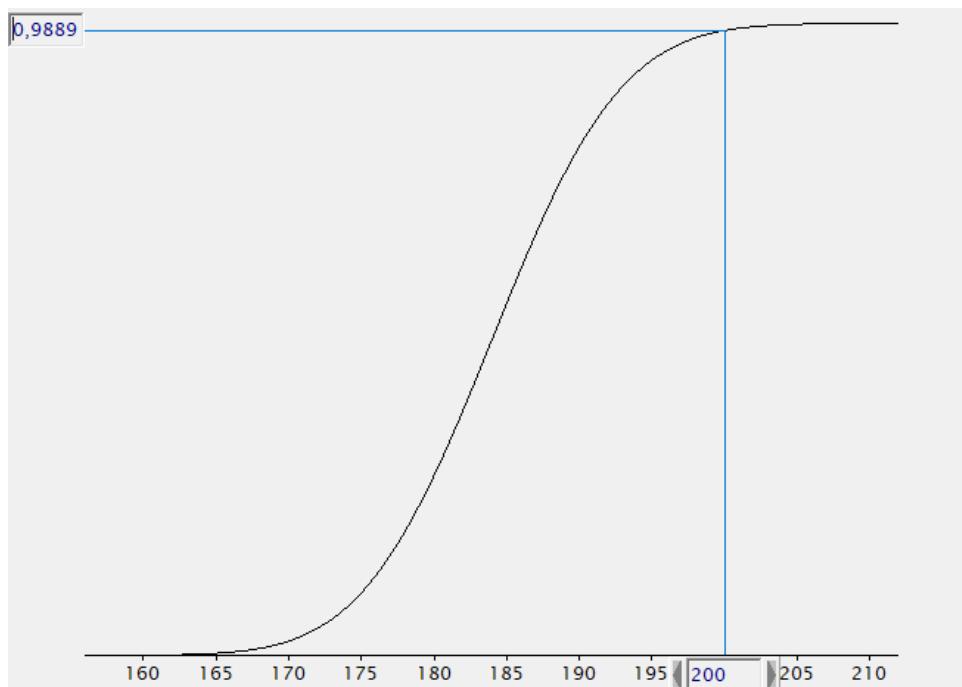
Eerst laat ik het resultaat voor 200 cm zien (**Figuur 8** en **Figuur 9**). Wat direct opvalt is dat 200 cm behoorlijk aan de rechterkant van de verdeling zit. Ben je 200 cm dan ben je langer dan 98,9 % van de Nederlandse bevolking. Alleen 1,1% is groter dan deze 200 cm.



**Figuur 8.** De plek waar een lengte van 200 cm past in een normaalverdeling van lengtes van Nederlandse mannen.

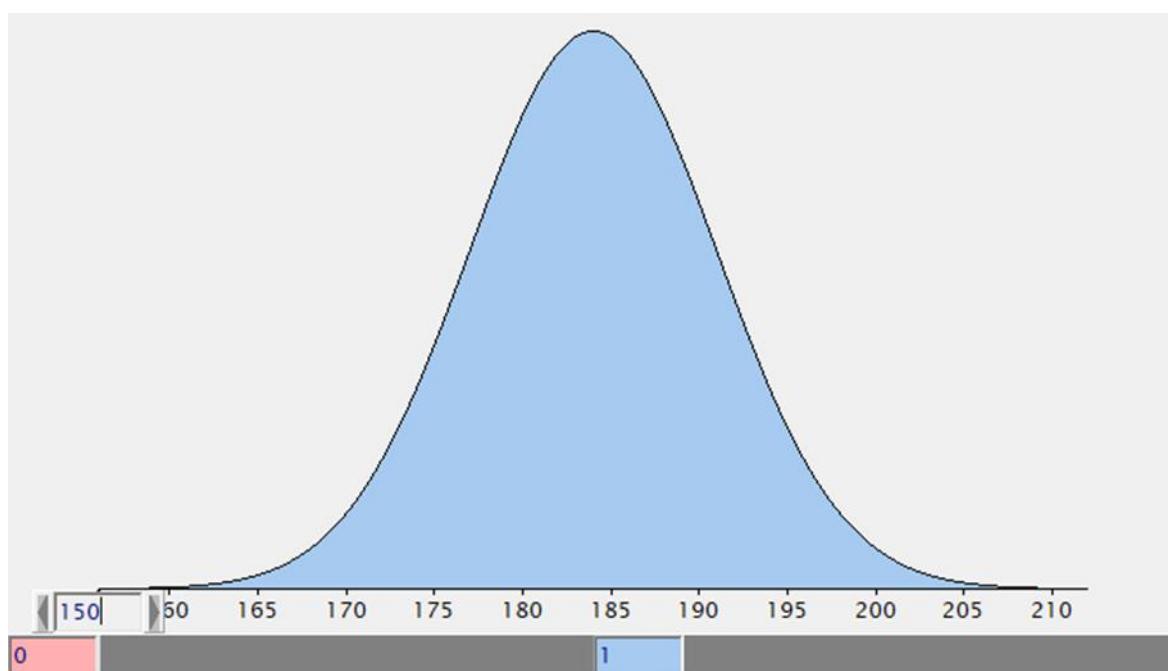
Hoe zit het met de 150 cm (**Figuur 10**)? Nou, dat ziet er heel anders uit. Uit **Figuur 9** bleek al dat een lengte van 150 cm niet zichtbaar is in de verdeling van Nederlandse mannen. Daarmee lijkt het alsof 150 cm bijna niet voorkomt. Toch weten we dat er wel degelijk mannen zijn die kleiner zijn dan 150 cm, bijvoorbeeld door een groeistoornis<sup>18</sup>. Maar dit zien we niet in de kansverdeling.

<sup>18</sup> <https://www.bvkm.nl/thema/groeistoornis/>



**Figuur 9.** De plek waar een lengte van 200 cm past in een normaalverdeling van lengtes van Nederlandse mannen. Deze keer afgebeeld in een cumulatieve verdeling.

Dat is een hele interessante bevinding, want dat een waarde een extreem lage frequentie kent hoeft nog niet te betekenen dat het niet voorkomt. De kans dat je iemand aantreft die 150 cm groot is, is alleen erg klein: volgens de verdeling zelfs tegen de 0 aan (maar we weten dus dat dit niet waar is). Dit gegeven alleen al maakt dat een kansverdeling niet heilig is.

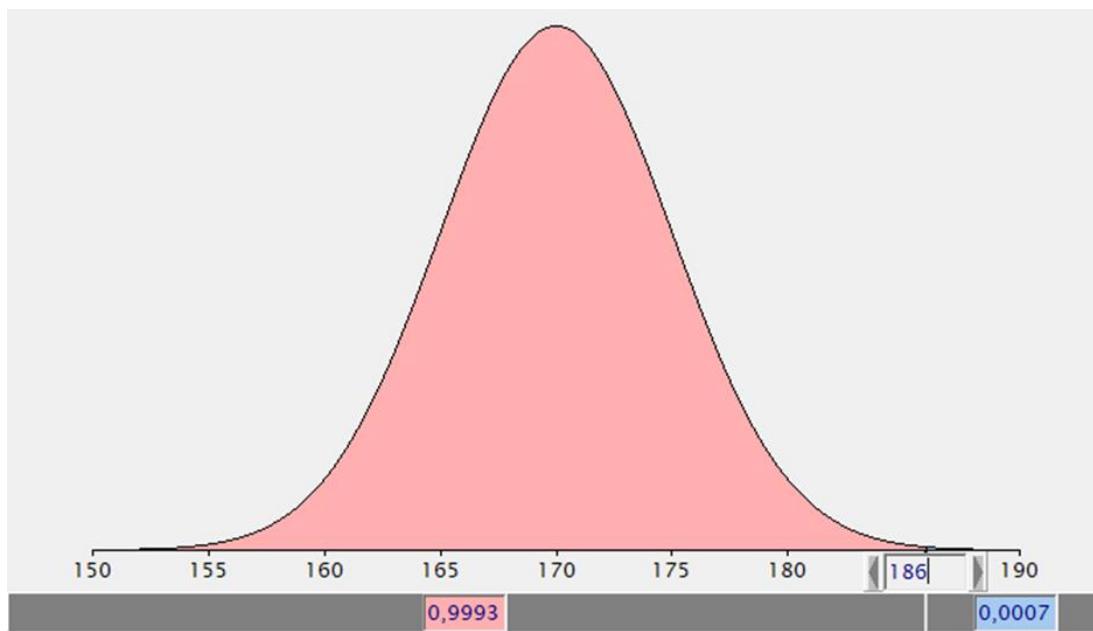


**Figuur 10.** De plek waar een lengte van 150 cm past in een normaalverdeling van lengtes van Nederlandse mannen.

## Een kans is afhankelijk van de context

Als we eenmaal een rangorde hebben dan kunnen we gaan rekenen met kansen, maar dat rekenen dient waakzaam te gebeuren: wij mensen zijn van nature niet zo goed in kansrekening. Natuurlijk zijn de nodige kansen al de revue gepasseerd. Zo hebben we uit de normaalverdelingen bepaald welk percentiel een bepaalde waarde beslaat én hoeveel gegevens boven of onder een bepaalde waarde vallen. Deze rangorde is ook een kans want het geeft de plaatsing van een waarde aan én die plaatsing is weer gebaseerd op de verdeling van kansen.

Stel nou dat we weten dat een groep mannen een gemiddelde lengte heeft van 170 cm met een standaard deviatie van 5 cm. En stel nou dat we mijn eigen lengte (186 cm) hierin zouden opnemen (**Figuur 11**). Dan zou je zien dat mijn lengte niet vaak voorkomt. Sterker nog, 99,9% van de mannen is kleiner dan ik ben. Dat is een andere bevinding dan wat we zagen in **Figuur 6**: hier was maar 61% van de mannen kleiner dan 186 cm in een verdeling van 184 (7) cm<sup>19</sup>.



**Figuur 11.** De plek waar 186 cm past in een normaalverdeling van lengte van Nederlandse mannen met een gemiddelde van 170 cm en een standaard deviatie van 5 cm.

<sup>19</sup> De notatie 184 (7) cm betekent een gemiddelde van 184 cm met een standaard deviatie van 7 cm.

Dus daar waar 186 cm eerder geen wenkbrauwen deed fronsen (**Figuur 6**) zouden we nu ons best moeten doen om iemand te vinden die 186 cm is (**Figuur 11**). De rangorde van 186 cm is daarmee compleet verschoven hoewel het getal zelf niet is veranderd. Aan de waarde (186 cm) is niets veranderd, maar de ‘waarde’ van het getal (uitgedrukt in kansen) is wel veranderd. Je hebt nou een eenmaal een verdeling (of rang) nodig om aan een waarde een kans toe te kennen. Maar nogmaals, mensen van 186 cm zullen er echt wel zijn. Een verdeling is dus geen weerspiegeling van de werkelijkheid: het is een beschrijving van verkregen gegevens in de vorm van een rangorde. Het is maar de vraag of we ook echt alle gegevens hebben.

## Wat kunnen we tot nu toe concluderen?

Met glyfosaat heeft bovenstaande nog weinig te maken, maar wat hopelijk wel helder is geworden is hoe handig kansverdelingen zijn. Ik heb getracht te laten zien dat wanneer informatie binnenkomt het handiger rekenen is wanneer er een kansverdeling overheen kan worden gelegd. In het geval van de lengte van Nederlandse mannen en vrouwen hebben we de ruwe data nooit gezien, maar konden we op basis van de beschrijvingen wel twee normaalverdelingen maken. Deze verdelingen stellen ons vervolgens in staat om uit te rekenen hoe vaak een bepaalde lengte voorkomt. Met die kennis kun je weer verder rekenen.

Toch is dit voor de gemiddelde wetenschapper niet voldoende. Als we weten dat ongeveer 61% van de mensen kleiner is dan 186 cm, en 39% is groter, dan is dat voor velen gewoonweg informatie. Maar het zegt niks over de waarde van een getal. Het zegt niet of 186 cm ‘normaal’ is of ‘vreemd’. Het enige wat we kunnen zeggen is dat 186 cm het 61<sup>ste</sup> percentiel is in een verdeling met gemiddeld 184 cm en standaarddeviatie 7 cm. Daarmee valt 186 cm binnen één standaarddeviatie van het gemiddelde.

Maar mensen willen graag meer en wat men graag wil weten is of een waarde toebehoort aan een bepaalde verdeling. Dit wordt ook wel ‘statistisch significant’ genoemd. Het bepalen van die significantie, zo zal ik u later laten zien, is een essentieel onderdeel in de wetenschap. Het predicaat ‘statistisch significant’ is helaas synoniem geworden voor ‘echt’ of ‘causaal’. Een waarde die statistisch significant wordt namelijk als zo afwijkend gezien dat deze we aan een andere groep moet toebehoren dan verwacht. Verwachtingen spelen

sowieso een grote rol in statistiekland, maar dit wordt vaak vergeten. Zo blijft onderbelicht in de reportage van BNNVARA/Zembla dat juist de verwachting dat glyfosaat schadelijk is maakt dat de onderzoekers zich sterk inzetten voor eenzijdig testen. Met het gebruik van die toets is het vinden van een statistisch significant verschil makkelijker te duiden en zodra iets statistisch significant is dan moet het wel ‘waar’ zijn.

In het volgende hoofdstuk wil ik dieper ingaan op het duiden van verschillen (tussen waarden en / of tussen groepen) en welke rol verschillen spelen in de frequentistische statistiek.

## Zoek de verschillen

In het vorige hoofdstuk heb we gezien wat de waarde is van een frequentieverdeling. Ook hebben we gezien hoe, op basis van diezelfde verdeling, geprobeerd wordt om waarde toe te kennen aan getallen. Een getal wat netjes in de verdeling valt zal weinig wenkbrauwen doen fronsen, maar een getal wat haast niet voorkomt in diezelfde verdeling doet dat wel.

De vraag die dan uiteindelijk vaak wordt gesteld is: “wat betekent dit getal?”

We zullen zien dat in de frequentistische statistiek de waarde van een getal vaak afhankelijk is van de afstand tot een ander getal. We zagen dit al in de vorige voorbeelden waarbij we de afstand tot het gemiddelde nemen, of de rangorde in een verdeling. Nu gaan we kijken naar verschillen tussen groepen. En ondanks dat we al twee groepen hebben, mannen en vrouwen, wordt de stof eenvoudiger als we data eerst samen te voegen. Laten we daarom een nieuw voorbeeld nemen op basis van de voor ons bekende data.

## Kansen zonder context

Stel nou dat iemand achter een groot zwart doek staat en op een blaadje schrijft: ‘200 cm’. Dit is de lengte van de persoon waarvan we niet weten of het een man of een vrouw is<sup>20</sup>. Wat is dan de kans dat diegene überhaupt uit Nederland komt op basis van de lengtegegevens alleen?<sup>21</sup> Zouden we er überhaupt een kans aan durven te verbinden? We hebben namelijk gezien dat bij de vrouwen deze lengte niet in de kansverdeling voorkomt (**Figuur 11**)<sup>22</sup>. Bij de mannen was dit wel het geval, maar de kans was niet heel groot (**Figuur 8**). Is het dan zo dat de persoon achter het zwarte doek wel een Nederlandse man moet zijn? Of durven we zelfs dat niet aan en weten we niet eens of wel kunnen spreken over iemand afkomstig uit de Nederlandse verdeling?

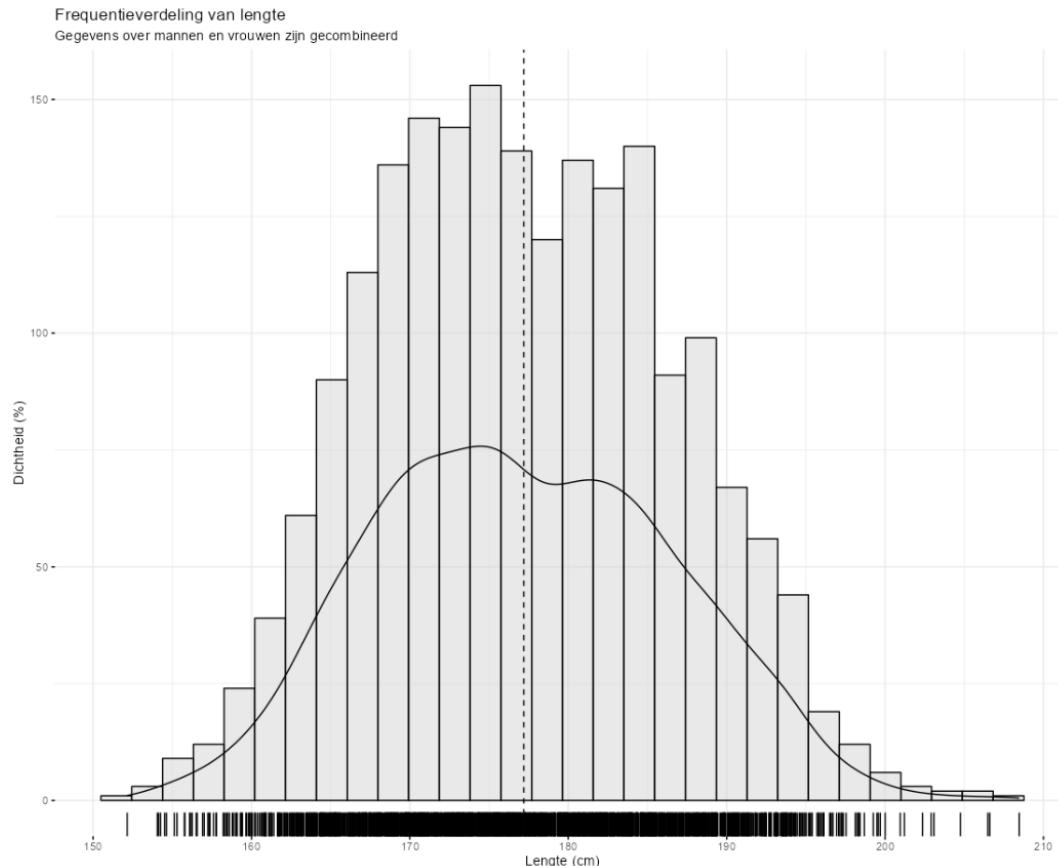
---

<sup>20</sup> Het enige wat we weten (of mogen aannemen) is dat diegene de waarheid spreekt en echt 200 cm is.

<sup>21</sup> De variabelen die lengte bepalen zijn niet allemaal bekend, maar genetica speelt absoluut een rol, net als geslacht. De gegevens over de Nederlandse bevolking en haar gemiddelde lengte per geslacht is dus, voor een deel, genetisch bepaald. Als ik hier spreek over land van herkomst dan spreek ik over genetische afkomst, niet de ware nationaliteit van die persoon die natuurlijk nog steeds Nederlands kan zijn ook al ligt de genetische oorsprong ergens anders.

<sup>22</sup> Wat dus betekent dat de kans heel klein is op basis van de verdeling. Het betekent niet dat er geen Nederlandse vrouwen zijn die zo lang zijn.

Stel nou dat we geen weet hebben van de verdeling van mannen en vrouwen. En dat de enige informatie die we hebben een reeks getallen is die zich grafisch laat tonen in **Figuur 12**<sup>23</sup>.



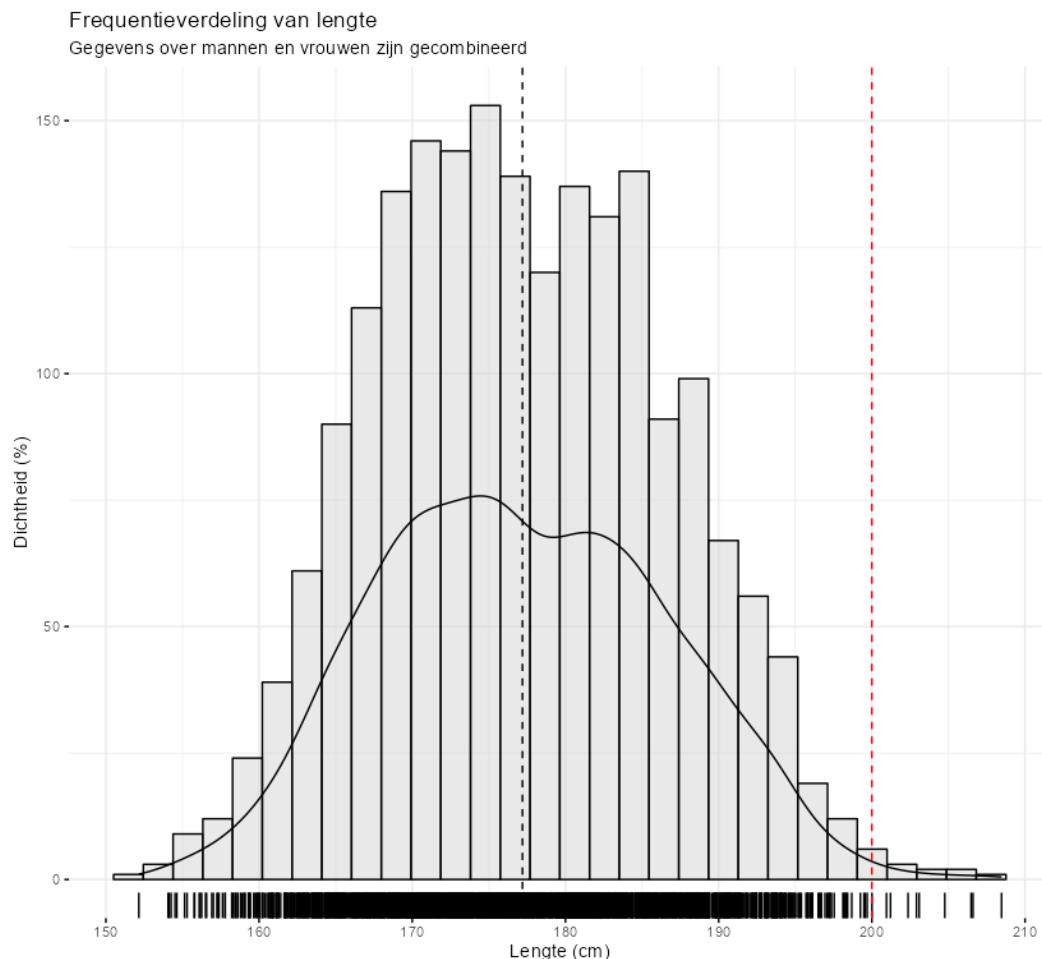
**Figuur 12.** De verdeling van lengte waarbij de gegevens van mannen en vrouwen zijn gecombineerd. De gegevens zijn gesimuleerd.

Wat direct opvalt is dat deze verdeling niet meer de mooie klokvorm laat zien die we voorheen hebben gezien. Uiteraard komt dit omdat we twee verdelingen met elkaar gecombineerd hebben<sup>24</sup>, maar het komt regelmatig voor dat dit de vorm van de data is waarmee je begint. Als we namelijk niks weten van de oorsprong van de data, deze verdeling maken, en kijken hoe vaak 200 cm voorkomt dan zien we in **Figuur 13**Error! Reference

<sup>23</sup> De oplettende lezer zal hier kunnen betogen dat we nu omgekeerd werken: we maken namelijk een verdeling op basis van waarden die we kennen dus we werken met voorkennis. We weten namelijk dat deze verdeling bestaat uit twee verdelingen. Toch komt het vaak voor dat een data-analist gegevens ziet die lijkt op de verdeling uit **Figuur 12**. Dit noemen we ook wel een multimodale verdeling: een verdeling met meerdere bulten. Het is dan zaak om te bepalen of deze verdeling bruikbaar is of niet.

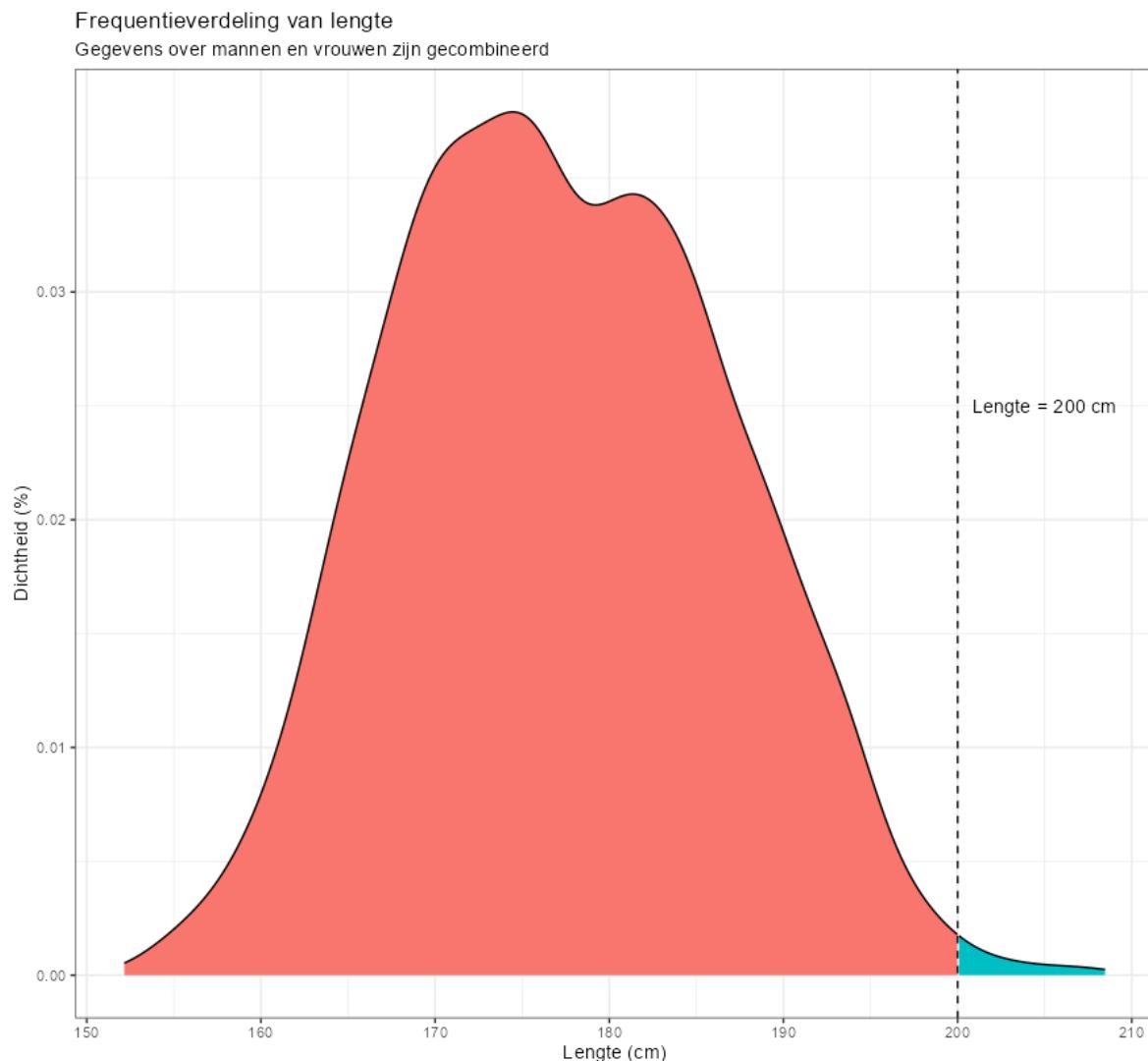
<sup>24</sup> Dit komt omdat we nu met gesimuleerde data werken en niet met een theoretische vorm. Ook zien we nu twee ‘bulten’ wat vaak een indicatie is dat we te maken hebben met een combinatie aan verdelingen. Dit alles laten we even terzijde.

source not found. dat dit nog steeds een lengte is die weinig voorkomt. Uitgerekend blijkt dat ongeveer 99.5% van de data minder is dan 200 cm. Daarmee beslaat 200 cm in dit voorbeeld het 99<sup>ste</sup> percentiel.



**Figuur 13.** De verdeling van lengte waarbij de gegevens van mannen en vrouwen zijn gecombineerd. De gegevens zijn gesimuleerd. De rode stippellijn geeft aan waar de 200 cm valt in die verdeling.

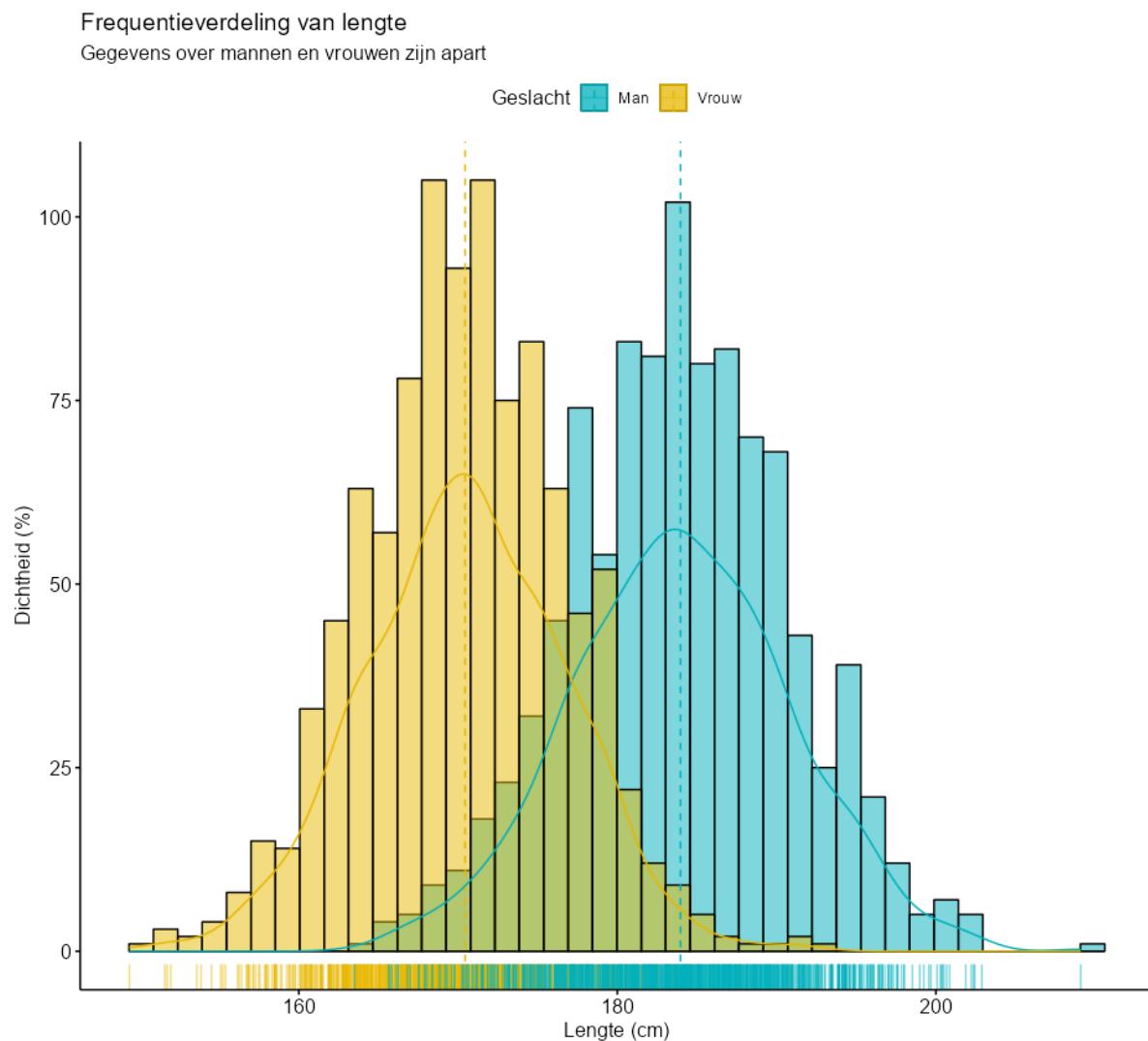
Een meer eenvoudige manier om dit te laten zien is zichtbaar in **Figuur 14**. Hier zien we dat bij de verdeling van de data er maar een heel klein deel is dat groter is dan 200 cm. Betekent dit nu dat we kunnen zeggen dat de persoon achter het doek wel of niet een Nederlandse man is? Wie bepaalt of 0.5% te klein is om een uitspraak te doen? Wie zich nu afvraagt of we ons die vraag wel moeten stellen doet een juiste observatie, maar helaas is het in de frequentistische statistiek zo dat we wel degelijk met afkapwaardes (moeten) werken.



**Figuur 14.** De verdeling van lengte waarbij de gegevens van mannen en vrouwen zijn gecombineerd. De gegevens zijn gesimuleerd. Het turquoise gedeelte laat zien wat het vlak is wat 200cm of meer is.

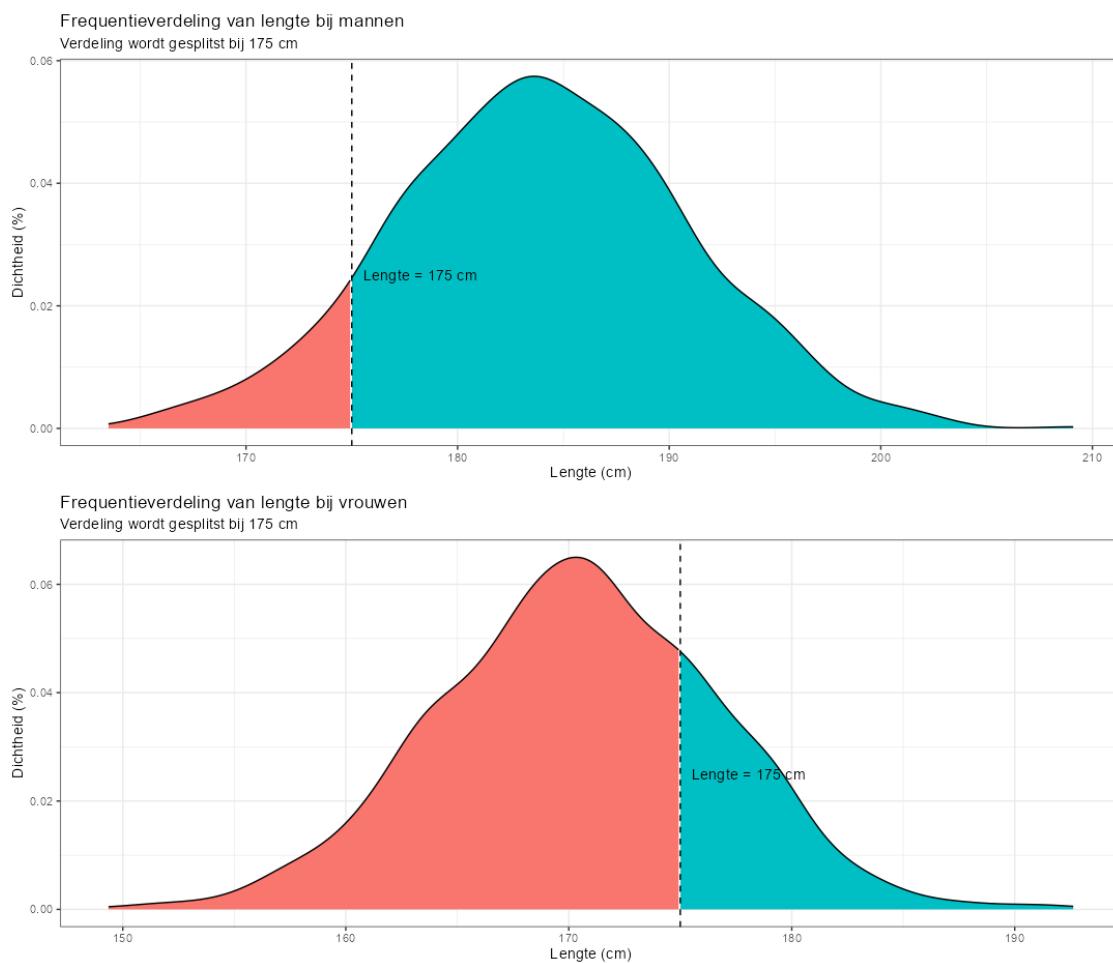
## Verschillende groepen of niet?

Laten we gemakshalve aannemen dat 0.5% voldoende is om te zeggen de kans aannemelijk is dat we hier met een Nederlander te maken hebben. Er is voor nu geen basis om te bepalen bij welk percentage we dat niet kunnen. Kunnen we dan ook iets zeggen of we te maken hebben met een vrouw of met een man? Om deze uitspraak te doen moeten we data splitsen naar hun oorspronkelijke verdeling. Het wordt dus tijd om de groepen visueel te splitsen. Het resultaat zien we in **Figuur 15** waarbij in het blauw de verdeling van lengtes voor de mannen zichtbaar is en in het geel die van de vrouwen. We zien dus nu voor het eerste beide verdelingen in dezelfde figuur.



**Figuur 15.** De verdeling van Nederlandse mannen en vrouwen in eenzelfde grafiek. Zichtbaar zijn de verschillende pieken en de overlap. Beiden zijn van betekenis.

Wat direct opvalt is dat beide groepen een piek hebben die wat uit elkaar staan. Dit is in de lijn der verwachting omdat het gemiddelde verschil 13,4 cm bedraagt<sup>25</sup>. Maar wat ook opvalt is de grote mate van overlap. Het visualiseren van deze data geeft dus heel wat te denken. Zo lijkt het alsof de groepen verschillend zijn, en dat is misschien ook wel het geval als we iemand van 200 cm meten. Maar wat als we nou iemand blind zouden meten en die persoon zou 175 cm groot zijn? Zouden we dan zomaar iemand aan de eerste of de tweede verdeling kunnen toewijzen (en daarmee stellen dat deze persoon wel een man of vrouw moet zijn?) Laten we de proef op de som nemen door voor beide verdelingen te bepalen wat de frequentie van lengtes is die kleiner is dan 175 cm. Als we dit visualiseren krijgen we **Figuur 16**.

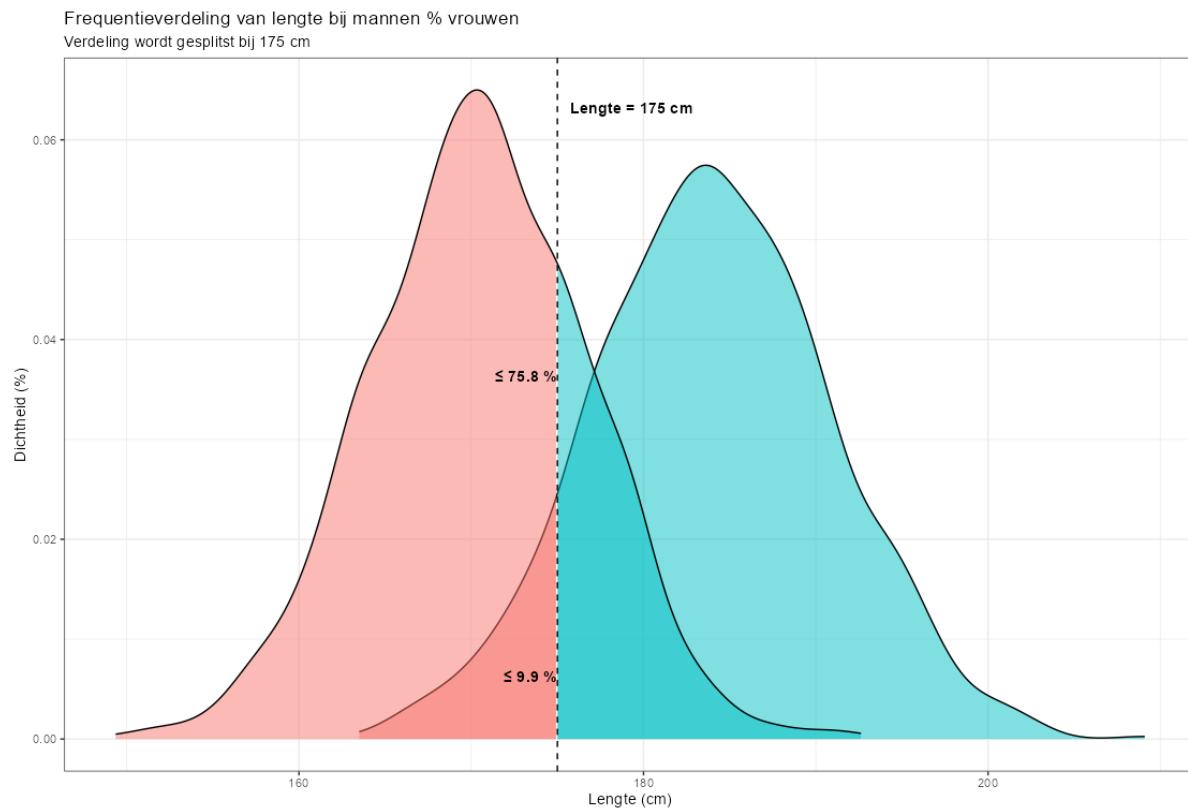


**Figuur 16.** De plek van 175 cm in de verdeling van Nederlandse mannen en van vrouwen.

<sup>25</sup> We zullen later zien dat de mate van spreiding bepaalt of een verschil tussen groepen van betekenis of niet. Hoewel een gemiddeld verschil van 13,4 cm daadwerkelijk een verschil is van 13,4 cm ziet dat verschil er anders uit als de spreiding 5 cm, 15 cm of 25 cm is. De reden dat mensen graag met de normaalverdeling werken is omdat het gemiddelde en de spreiding los van elkaar worden geschat: dat maakt het een aantrekkelijke verdeling om mee te werken.

De bovenste grafiek laat de zien wat het percentage is wat kleiner is dan 175 cm bij mannen en de onderste wat het percentage is bij vrouwen. Als we gaan rekenen dan ontdekken we dat ongeveer 10% van de mannen kleiner is dan 175 cm en bij vrouwen is dat ongeveer 76%. Omgekeerd betekent dit dat 90% van de mannen en 24% van de vrouwen groter is dan 175 cm. De lengte 175 cm beslaat bij mannen dus het 10<sup>de</sup> percentiel en bij vrouwen het 76<sup>ste</sup> percentiel.

Kunnen we dan met droge ogen beweren dat de kans groter is dat iemand die we meten op 175 cm een grotere kans heeft om een man te zijn dan om een vrouw te zijn? Om een beter beeld te vormen kunnen we ook beide verdelingen combineren in dezelfde grafiek en dan laten zien waar 175 cm past bij beide verdelingen. Dit zien we in **Figuur 17** waarin ik de twee verdelingen over elkaar heen leg, een splitsing maak bij 175 cm, en de vlakken inkleur die groter (of gelijk) of kleiner (of gelijk) zijn dan die waarde. De percentages heb ik er ook bijgezet.



**Figuur 17.** De lengte van 175 in de verdeling van mannen en die van vrouwen. Het rode gedeelte is het percentage mannen of vrouwen die kleiner of gelijk zijn aan 175 cm. Het turquoise vlak is het percentage dat groter of gelijk is. Op basis van deze massa kun je niet bepalen of iemand een man of vrouw is. Je kunt hoogstens het percentage uitrekenen waarin we iemand met deze lengte zien gegeven de verdeling van lengtes.

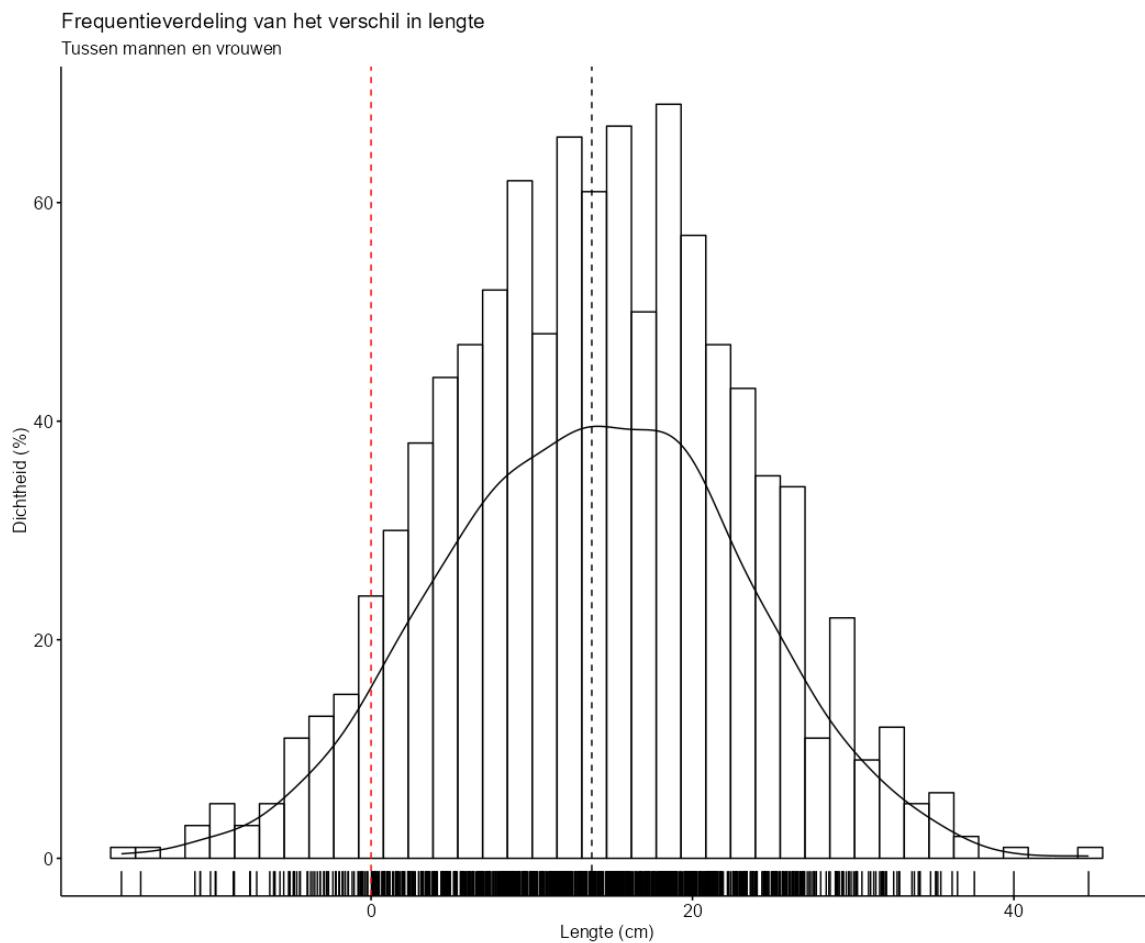
Wat wederom opvalt is dat 175 cm in beide verdelingen voorkomt, maar een stuk minder bij de ene verdeling dan de andere. De vraag is nu of dit voldoende informatie biedt om te bepalen dat de kans groter is dat de gemeten persoon een vrouw is en geen man. We zouden dit in theorie kunnen voorspellen. Maar dit is niet het belangrijkste wat we kunnen doen. Het belangrijkste is bepalen of de groepen bij elkaar horen. We kunnen en mogen onszelf dus de vraag stellen: zijn mannen en vrouwen, gemiddeld genomen, even lang?

Deze vraag lijkt wellicht wat absurd, want wie naar **Figuur 15** kijkt kan duidelijk zien dat de gemiddelde lengte bij mannen 13,4 cm groter is dan de gemiddelde lengte bij vrouwen. Om dit getal wat meer duiding te geven is het wellicht heilzaam om een verdeling te maken van de verschillen tussen de groepen. Dit doen we simpel door de waarden van elkaar af te trekken. Wat overblijft wordt zichtbaar in **Figuur 18** waarin we gemakshalve niet alleen het gemiddelde verschil laten zien (zwarte stippellijn), maar ook de plek van ‘geen verschil’ tonen (rode stippellijn). Wat direct opvalt is dat de verdeling van verschillen ook een normaalverdeling heeft<sup>26</sup>. Het gemiddelde van die verdeling is uiteraard 14, maar de 0 (wat neerkomt op geen verschil) is ook een mogelijkheid in die verdeling.

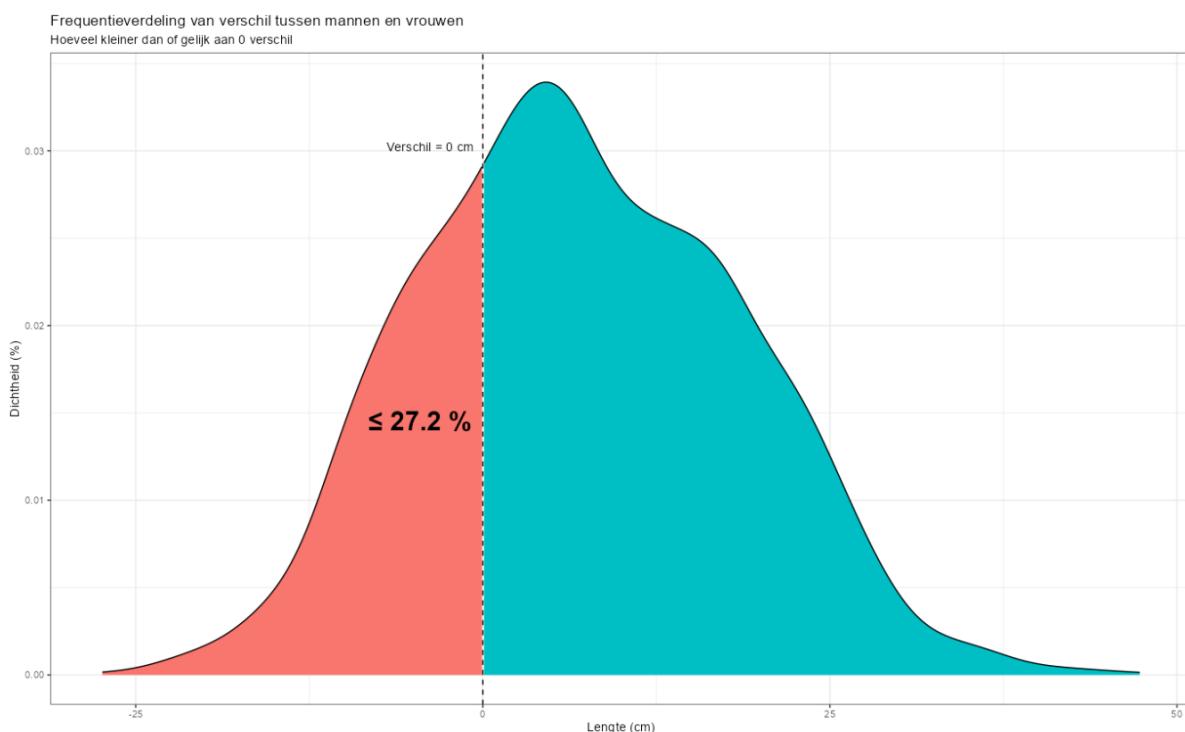
Om dit beter te tonen is het netjes om nog een laatste grafiek te maken, waarbij ik het percentage uitreken én toon wat kleiner of gelijk is aan nul. Dat wordt dan **Figuur 19**. Wat **Figuur 19** laat zien is dat het verschil tussen de groepen kleiner of gelijk kan zijn aan 0. Ook toont de grafiek dat 27.2% van de getallen kleiner is dan 0. Daarmee is 72.8% groter dan 0. Het getal, wat zegt dat er geen verschil is, is dus wel degelijk een mogelijkheid. Toch is dit niet voldoende om te bepalen dat de groepen niet verschillend zijn.

---

<sup>26</sup> Het mag eigenlijk niet verbazen dat deze verdeling ook de normaalverdeling volgt: het verschil tussen tweezelfde verdelingen behoudt vaak de originele verdeling.



**Figuur 18.** De verdeling van verschillen tussen de groep Nederlandse mannen en de groep Nederlandse vrouwen.



**Figuur 19.** De verdeling van verschillen. Ongeveer 27% van de waarden zijn negatief.

## Wat kunnen we hier nu uit afleiden?

In dit hoofdstuk hebben we het nog steeds niet over glyfosaat gehad, maar zijn we gebleven bij de lengtes van mannen en vrouwen. Ook heb ik nog niet gesproken over statistische significantie en eenzijdig of tweezijdig toetsen. Wat ik wel heb gedaan is (wederom) laten zien dat een verdeling een handig instrument kan zijn om te tonen welke waarden we mogen verwachten. Wanneer we meerdere delingen hebben kunnen we het verschil tussen die delingen opmerken en kijken of een bepaalde waarde voorkomt en hoe vaak. Tot nu toe hebben we ons onthouden van waardeoordeelen en conclusies. We zijn strikt gebleven bij observaties en het berekenen van percentages zelfs als we kijken naar de verschillen tussen groepen. Het ontbrak ons tot nu toe aan een kwalitatieve component.

Toch gaat de statistiek hier wel over. De vraag of vrouwen en mannen verschillende groepen zijn op basis van lengte is een vraag die past in de frequentistische statistiek. Het is een vraag die eigenlijk niet zonder waardeoordeel te plaatsen is. Wie namelijk wil weten of groepen verschillend zijn en daarbij statistiek moet toepassen zal ook een manier moeten vinden om te bepalen wanneer een verschil groot genoeg is om als verschil te worden bestempeld. Daarvoor zijn grenswaarden nodig. Die grenswaarden hebben we tot nu toe nog niet genoemd en dat komt omdat ze de basis vormen van de frequentistische statistiek. Na twee inleidende hoofdstukken zullen we ons nu dus keren tot de fundamenteiten van deze statistiek. Met alle mitsen en maren die daarbij horen.

## Fundamenten van de frequentistische Statistiek

---

*It should always be appreciated that a statistical analysis has its limitations. Statistical analysis cannot rescue poor data resulting from a flawed design or a poorly conducted study. Good experimental design, again the ‘strategy’ (paragraph 266), is the critical role of statistical analysis in a study. An appropriate data analysis will follow directly from a correct experimental design (including the selection of statistical methods to be applied) and implementation<sup>27</sup>.*

---

We zijn nu aangekomen bij het belangrijkste hoofdstuk van dit rapport. Om hier te komen moesten we samen de vorige twee hoofdstukken doorlopen, want wie iets wil met frequentistische statistiek ontkomt niet aan frequentieverdelingen én aan verschillen.

In dit hoofdstuk wil ik u, de lezer, eerst uitleggen wat de frequentistische statistiek nu precies is. Om dit te doen zal ik gebruik maken van de al getoonde voorbeelden, waarbij ik langzaamaan nieuwe terminologie introduceer om de terminologie uit de BNNVARA/Zembla reportage te duiden. Zo kom ik te spreken over ‘nulhypothese’, ‘alternatieve hypothese’, ‘betrouwbaarheidsinterval’, ‘vals positief’, ‘alfa waarde’ en ‘kracht’. Hoewel dit hoofdstuk het meest technisch zal zijn, zal ik geen formules gebruiken.

Wat ik vooral wil doen is visualiseren hoe de frequentistische statistiek tracht om waarde te koppelen aan getallen. Dit gebeurt door middel van grenswaarden en het is deze grenswaarde die een verdeling van kansen opdeelt in ‘echt’ of ‘niet echt’. Het woord ‘echt’ moeten we met een korrel zout nemen, maar strikt genomen gaat het erom dat we een verschil zien dat groter is dan een bepaalde statistische grenswaarde. Als dat gebeurt dan spreken we in de frequentistische statistiek van een verschil dat we niet kunnen ‘toekennen aan toeval’<sup>28</sup>. Het verschil moet daarom wel iets betekenen en dus wel echt zijn.

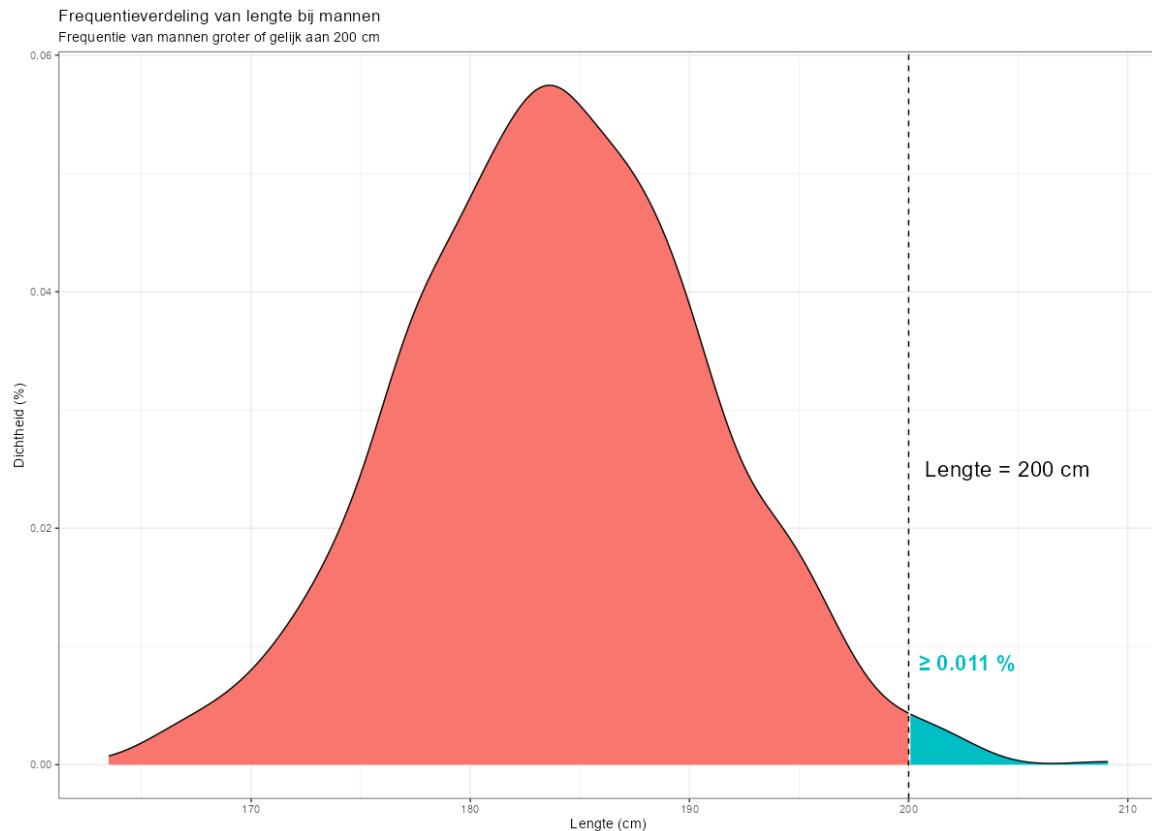
Tot nu toe hebben we ons voornamelijk bezig gehouden met het uitrekenen van kansen. Zodat als we een man vinden die 200 cm is we ons de vraag kunnen stellen: ‘wat is de kans dat we zo iemand tegenkomen in de verdeling van Nederlandse mannen?’. We zien

---

<sup>27</sup> [https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453\\_9789264221475-en](https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453_9789264221475-en).

<sup>28</sup> We zagen het gebruik van dit woord al een aantal keren terugkomen in de BNNVARA/Zembla rapportage.

het resultaat in **Figuur 20** wat ons toont dat 0.011 % van de mannen groter of gelijk is aan 200 cm. Dus, vinden we iemand van 200 cm, dan is dat aardig zeldzaam en de vraag die we ons nu kunnen stellen is: “is deze persoon wel een Nederlandse man?”<sup>29</sup>. Daarvoor hebben we een grenswaarde nodig.



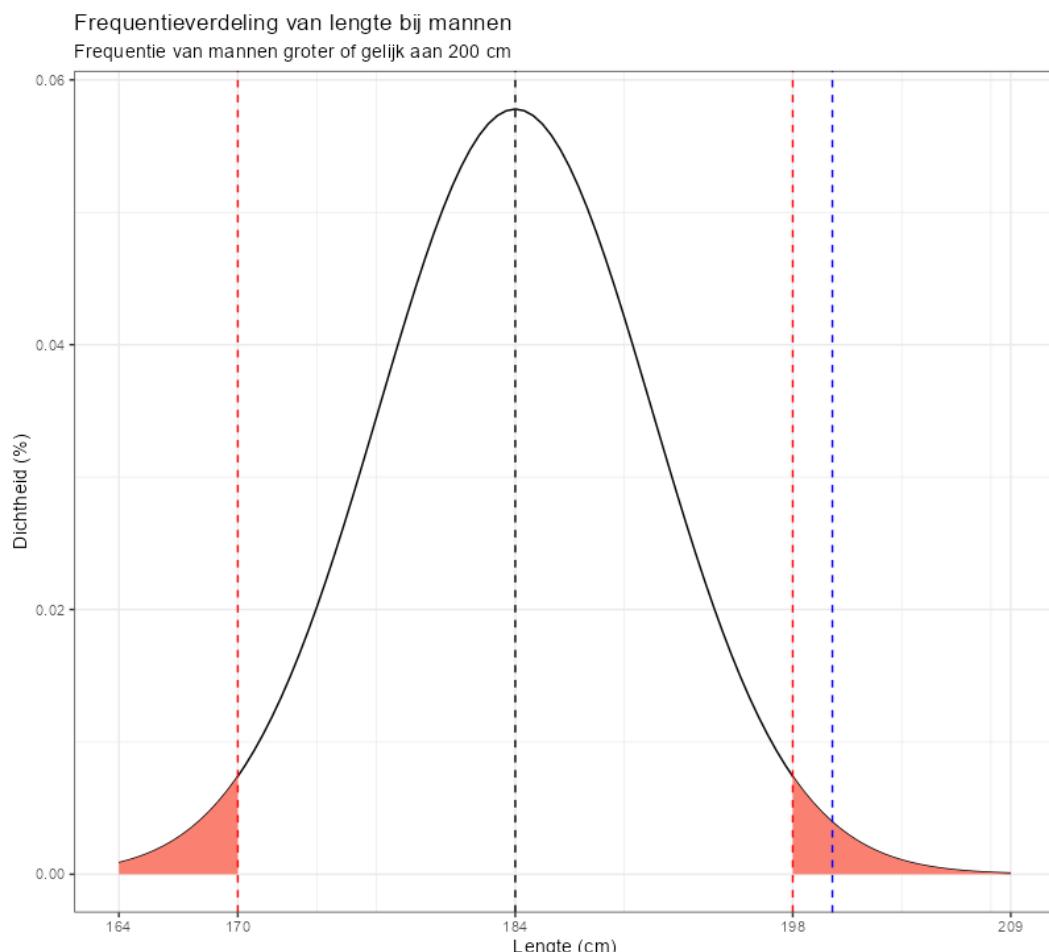
**Figuur 20.** De lengte 200 cm in de frequentieverdeling van Nederlandse mannen. De waarde 200 cm komt minder dan 0.011% voor.

## Grenswaarden

De grenswaarde is dus een essentieel onderdeel van de frequentistische statistiek en bepaalt in feite of een gevonden waarde te vreemd is om te negeren. We zagen al in **Figuur 20** dat 200 cm het 99% percentiel beslaat, maar dat zegt niet of een waarde ‘significant’ is. Het opstellen van eens grenswaarde maakt dit wel mogelijk en is daarmee een hele belangrijke kwestie in de frequentistische statistiek. Het verandert één enkele waarde in een binaire classificatie.

<sup>29</sup> Deze vraag is dus ook een verschil. Als wij op zoek zijn naar de betekenis van 200 cm in de bekende verdeling van Nederlandse mannen dan trachten wij te ontdekken of het verschil tussen 200 cm en de andere metingen zo groot is dat deze wel afkomstig moet zijn uit een andere verdeling. We vragen ons dus af of we hier wel te maken hebben met een Nederlandse man.

Om redenen die niet heel duidelijk zijn, is in de frequentistische statistiek bepaalt dat 5% een geaccepteerde grenswaarde is. Een getal is dus ‘statistisch significant’ als het percentiel buiten de 5% valt. Deze 5% is alleen van toepassing als we naar één zijde van de verdeling kijken. Wanneer we naar beide zijdes kijken dan is de grenswaarde 2.5%. **Figuur 21** laat zien hoe dit er grafisch uit ziet.



**Figuur 21.** De 2,5% grenswaarden links en rechts van het gemiddelde afgevinkt met een rode stippellijn. De zwarte stippellijn is het gemiddelde en de blauwe stippellijn laat zien waar de 200 cm valt.

We zien namelijk twee rode vlakken en elk rood vlak beslaat 2,5% van de massa van de verdeling. Bij de verdeling van Nederlandse mannen zit de grens op 170 cm en 198 cm<sup>30</sup>. We zeggen dan dat als een waarde buiten een van die grenzen valt, deze waarde statistisch significant is. Dat klinkt wellicht wat gek, want we kunnen duidelijk zien dat dat ook die

<sup>30</sup> We zagen in **Tabel 1** al dat dit de waardes zijn die horen bij 2x de standaard deviatie. In een perfecte normaalverdeling valt 95% van de massa binnen 2 standaard deviaties van het gemiddelde. Dit is waarom we deze waarden nu weer terug zien.

waarden mogelijk zijn. Alleen is hun frequentie zo klein dat we bepalen dat deze waarden vreemd zijn. Nogmaals: dit is geen exacte wetenschap, maar een afspraak die ooit is gemaakt.

## Hypothese(s) toetsen

Om te toetsen of een waarde groot genoeg is om interessant te zijn hebben we niet alleen grenswaardes nodig, maar ook hypotheses. In de frequentistische statistiek is er altijd sprake van twee hypotheses: de nulhypothese én de alternatieve hypothese. De nulhypothese is de hypothese waarmee je begint en die je probeert te falsificeren op basis van bewijs. Daarmee valt hypothese toetsen zoals we deze kennen in de frequentistische statistiek veelal samen met het falsificatieprincipe van Karl Popper<sup>31</sup>. Deze beroemde wetenschapsfilosoof zei dat het beter is om een bepaalde hypothese eenmalig te falsificeren dan er meermaals bevestiging voor te zoeken. Wie wil stellen dat alle zwanen wit zijn kan beter op zoek naar gaan een zwarte zwaan dan ontelbaar witte zwanen verzamelen. In de kansrekening is het echter zelden ‘alles of niets’, waardoor het proberen te ‘falsificeren’ van de nulhypothese ook problematisch is. Daarom hebben we dus grenswaardes geïntroduceerd, maar die kennen weer hun eigen problemen zoals we straks zullen zien.

De nulhypothese is een stelling van een bepaalde aart waarin heel vaak wordt gesteld dat een theoretisch gemiddelde gelijk is aan bepaalde waarde. Zo zou de nulhypothese hier kunnen zijn dat het gemiddelde uit de populatie mannen 184 cm is. Maar we mogen ook stellen dat het populatiegemiddelde 200 cm is. Of we stellen dat het verschil tussen mannen en vrouwen 0 cm is. Daarmee zou de nulhypothese zijn dat de lengteverdeling tussen mannen en vrouwen nagenoeg gelijk is. Uiteindelijk gaat het erom dat er een bepaalde stelling wordt gedaan (de nulhypothese) en dat er vervolgens gegevens worden verzameld. Dan wordt gekeken of de nulhypothese overeind blijft staan op basis van die gegevens en op basis van de gekozen grenswaardes.

---

<sup>31</sup> [https://en.wikipedia.org/wiki/Karl\\_Popper](https://en.wikipedia.org/wiki/Karl_Popper).

## One-sample t-test

In de wereld van de frequentistische statistiek zijn er dus testen ontwikkeld die wetenschappers kunnen inzetten om te bepalen of een verschil groot genoeg is of niet<sup>32</sup>. Afhankelijk van het soort vraag valt de keuze voor een specifiek soort test<sup>33</sup>. Willen we uitrekenen wat de kans is dat de gemiddelde lengte van Nederlandse mannen 184 cm als wij iemand van 200 cm observeren dan kunnen we gebruik maken van de zogenaamde *one-sample t-test*. Deze test kijkt of een steekproef past bij de populatie van waaruit die theoretisch afkomstig zou moeten zijn. Met andere woorden: met deze test kunnen we analyseren of het gemiddelde van een steekproef significant verschilt van een bepaalde waarde. Met nog andere woorden: met deze test kunnen we uitrekenen wat de kans is dat we 200 cm zien als het gemiddelde echt 184 cm is. Het gaat dus om het verschil tussen twee waarden! We kunnen deze test uitvoeren met de gegevens zichtbaar in **Tabel 2**<sup>34</sup>:

Gemiddelde	Standaard deviatie	Aantal in populatie	Nulhypothese
184 cm	7 cm	1000	200 cm

**Tabel 2.** De vier waardes die ik invoer om te bepalen of de nulhypothese (200 cm) past bij een observatie van 184 (7) cm.

Alvorens we dieper duiken in het resultaat is verstandig om in stappen uit te leggen wat ik nu precies gedaan heb:

1. Ik heb aan het rekenprogramma een rij getallen ( $N=1000$ ) aangeboden met een gemiddelde van 184 en een standaard deviatie van 7.
2. Ik heb de nulhypothese meegegeven dat de lengteverdeling van mannen een gemiddelde heeft van 200 cm. Nu weet ik dat het gemiddelde 184 cm, maar ik vraag het programma om net te doen alsof dit 200 cm<sup>35</sup>. Door dit zo te stellen vraag ik wat

---

<sup>32</sup> Wat dan helaas weer wordt verward met de uitspraak dat een verschil echt is of niet. Eigenlijk gaat het erom of men het geobserveerde verschil wel of niet mag verwachten op basis van de gegevens van de verdeling, maar gek genoeg wordt er in de frequentistische statistiek maar heel weinig met verwachtingen gedaan.

<sup>33</sup> Soms zijn meerdere testen toepasbaar. De keuze voor een test hangt af van de assumpties van de test en de data.

<sup>34</sup> De oplettende lezer zal al hebben ontdekt dat de manier van toetsen vaak omgekeerd gaat: we hebben een theoretisch gemiddelde van 184 cm en doen vervolgens observaties. Als we dan ontdekken dat onze observaties een gemiddelde hebben van 200 cm met een standaard deviatie van 7 cm, dan kunnen we ons afvragen of die 184 cm wel een accurate beschrijving is van het gemiddelde. Andere testen kijken verder en doen uitspraken over het verschil tussen groepen: de observaties met een gemiddelde van 200 cm behoren tot een andere groep dan de groep met een gemiddelde van 184 cm.

<sup>35</sup> De computer zal ons in deze nooit tegenspreken.

de kans is dat we een gemiddelde van 184 cm zien als het echte gemiddelde 200 cm is.

3. Ik heb de alternatieve hypothese aangeboden dat het gemiddelde van de populatie niet 200 cm is. Ik geef niet aan of deze groter of kleiner moet zijn (daar komen we nog op).
4. Ik heb de kans bepaalt dat mijn bevinding ook vals positief kan zijn. Die kans heb ik op 5% gezet. Dit heet ook wel de alfa waarde. Omdat ik geen richting wil bepalen splits ik de alfa waarde: 2.5% links en 2.5% rechts (zie **Figuur 21**). Dat zijn mijn grenswaarden.
5. Ik kies ervoor om te werken met een T-verdeling<sup>36</sup> in plaats van een normaalverdeling. De T-verdeling is iets strenger, maar lijkt met grote N heel erg op de normaalverdeling.

Zoals we zullen zien is dit alles vooral een theoretische exercitie. Wat ik namelijk geobserveerd heb is een verdeling van lengtes met een gemiddelde van 184 cm en een standaard deviatie van 7 cm. Ik vraag nu om te bepalen wat de kans is dat de ‘echte’ verdeling een gemiddelde van 200 cm heeft. Dit vraag ik omdat ik daadwerkelijk iemand heb gevonden die 200 cm is. De uitkomst van mijn exercitie we in **Tabel 3**.

T-waarde	Vrijheidsgraden	p-waarde
-73.473	999 (1000 - 1)	< 0.001

**Tabel 3.** De uitkomst van de one-sample t-toets.

Wat staat hier nu precies? Laten we elke cel nagaan:

1. Ik toets wat de kans is dat 200 cm het ‘echte’ gemiddelde is en niet 184 cm zoals geobserveerd. De waarde die we vinden op de T-verdeling is -73.473.
2. De kritieke T-waarden zijn -1.96 links en 1.96 rechts bij 999 vrijheidsgraden. Het aantal vrijheidsgraden wordt gezet op de grootte van de steekproef (1000) minus 1. We vinden een T-waarde van -73. Deze waarde is 37x groter dan de kritieke grenswaarde.

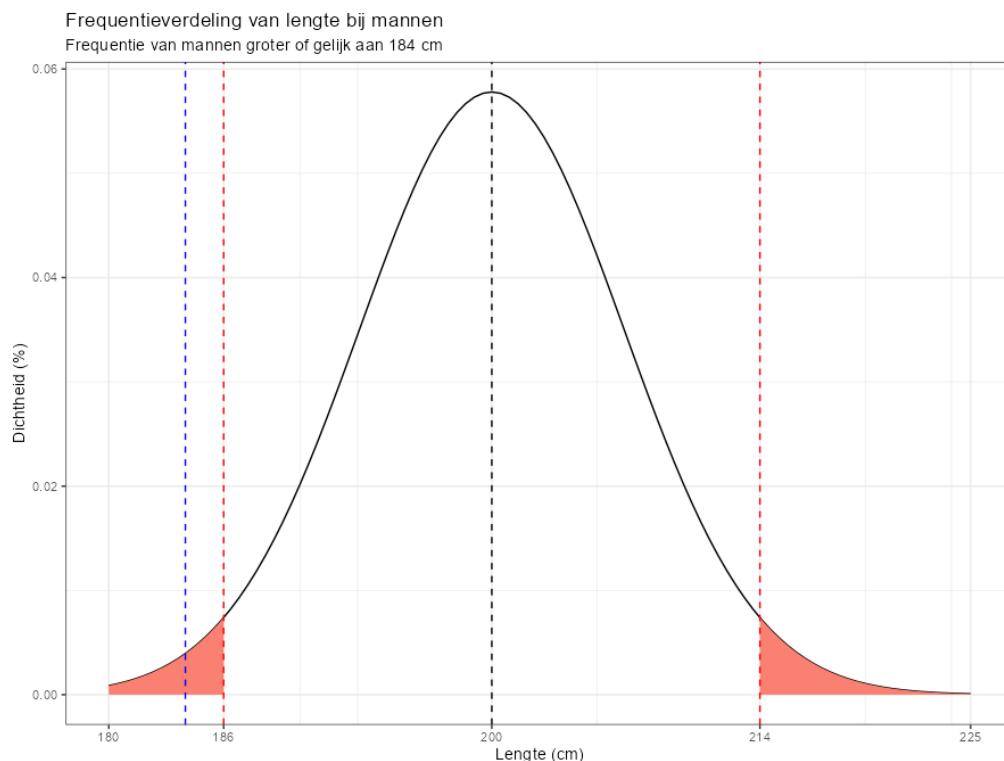
---

<sup>36</sup> [https://en.wikipedia.org/wiki/Student%27s\\_t-distribution](https://en.wikipedia.org/wiki/Student%27s_t-distribution)

3. De p-waarde die uit dit alles volgt is <0.001. Dit is de kans dat de bevinding door toeval komt en niet echt is. We hebben de totale grenswaarde op 0.05 wat neerkomt op 0.025 link en 0.025 rechts. We hebben nu dus een statistisch significante bevinding wat zoiets betekent als dat het geobserveerde gemiddelde van 184 cm zo verschillend is van 200 cm dat het gemiddelde niet 200 cm kan zijn.

De oplettende lezer zal nu misschien de wenkbrauwen fronsen want hebben we wel de juiste toets toegepast? In **Figuur 21** lieten we zien dat een geobserveerde waarde van 200 cm niet vaak voorkomt in een verdeling met een gemiddelde van 184 cm met standaard deviatie 7 cm. Maar wat we nu hebben berekend is eigenlijk het omgekeerde, namelijk of de geobsedeerde waarden wel afkomstig kunnen zijn als het gemiddelde van alle Nederlandse mannen 200 cm is. Is dat niet iets heel anders?

Laten we het voorbeeld omdraaien. Stel, we meten 1000 mannen en we vinden dat deze mannen gemiddeld 200 cm lang zijn met een standaard deviatie van 7 cm. Wat is dan de kans dat het echte gemiddelde 184 cm is? We zien dit grafisch in **Figuur 22**.



**Figuur 22.** De plek van 184 cm in een frequentieverdeling van 200 cm bij een standaarddeviatie van 7 cm.

Wanneer we dezelfde toets toepassen vinden we dezelfde t-waarde, maar nu met een minus teken. Ook vinden we dezelfde p-waarde<sup>37</sup>. In het licht van de frequentistische statistiek stellen we dus eigenlijk dezelfde vraag, maar dan omgekeerd. Toch is dat niet helemaal zo, want waar we in de eerste vraag zeiden dat de nulhypothese een gemiddelde van 200 cm zeggen we nu dat de nulhypothese een gemiddelde van 184 cm is. Wat blijft is de alternatieve hypothese die steeds hetzelfde is: het getal is niet gelijk aan het getal van de nulhypothese.

### Independent samples t-test

Voordat we dieper ingaan op de p-waarde, wil ik nog één voorbeeld nemen. En dat is het voorbeeld wat ik liet zien in **Figuur 15**: zijn mannen en vrouwen significant van elkaar verschillend. Ik heb het via **Figuur 18** en **Figuur 19** kort aangestipt, maar ik wil de frequentistische statistiek nu tastbaar maken door middel van hypothese toetsen. We doen het volgende:

1. Ik stel een onderzoeksvraag op: zijn mannen en vrouwen significant van elkaar verschillend in lengte?
2. Ik stel de nulhypothese op: mannen en vrouwen zijn even lang.
3. Ik stel de alternatieve hypothese op: mannen en vrouwen zijn niet even lang.
4. Ik zet de grenswaarde vast op alfa 5%.

Met de waarden die we hebben verzameld kunnen we nu een *independent samples t-test* doen. Als we deze toets toepassen, krijgen we het volgende resultaat zoals getoond in Tabel 4:

T-waarde	Vrijheidsgraden	p-waarde
45.716	1985.4	< 0.001

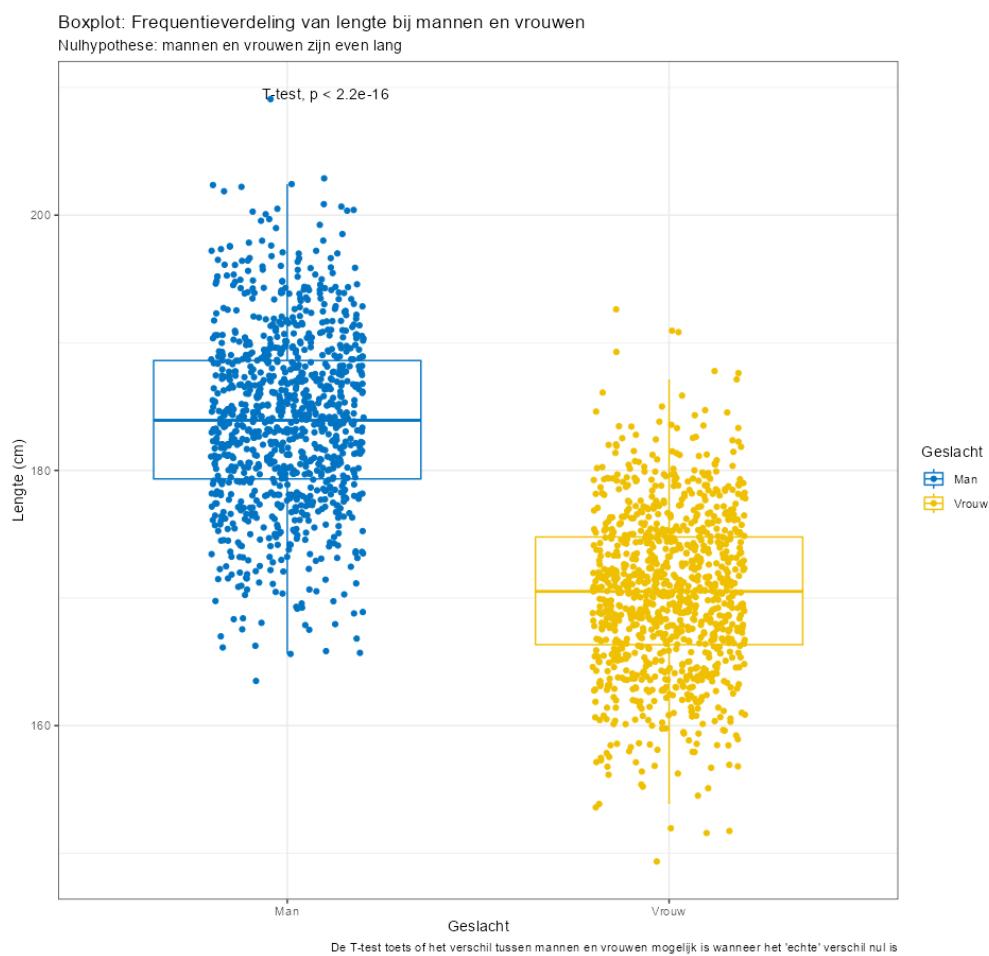
**Tabel 4.** Resultaten van een independent samples t-test.

---

<sup>37</sup> De p-waarde is bij tweezijdig toetsen afhankelijk van de grootte en niet de richting van de t-waarde.

Wat heel duidelijk blijkt (volgens deze toets) is dat de nulhypothese ('het echte verschil tussen mannen en vrouwen is gelijk aan 0') verworpen kan worden. De T-waarde, die het verschil weerspiegelt tussen beide groepen, is zo groot dat het volgens de frequentistische statistiek haast geen toeval kan zijn om dit verschil te vinden mochten beide groepen toch echt hetzelfde zijn. Met andere woorden: als het echt zo is dat het verschil tussen mannen en vrouwen nul is, dan is de kans dat we een verschil als deze zien, kleiner is dan 0.001%. Deze kans is kleiner dan de grenswaarde van 0.05% en dus mag geconstateerd worden dat de groepen wel verschillend moeten zijn<sup>38</sup>.

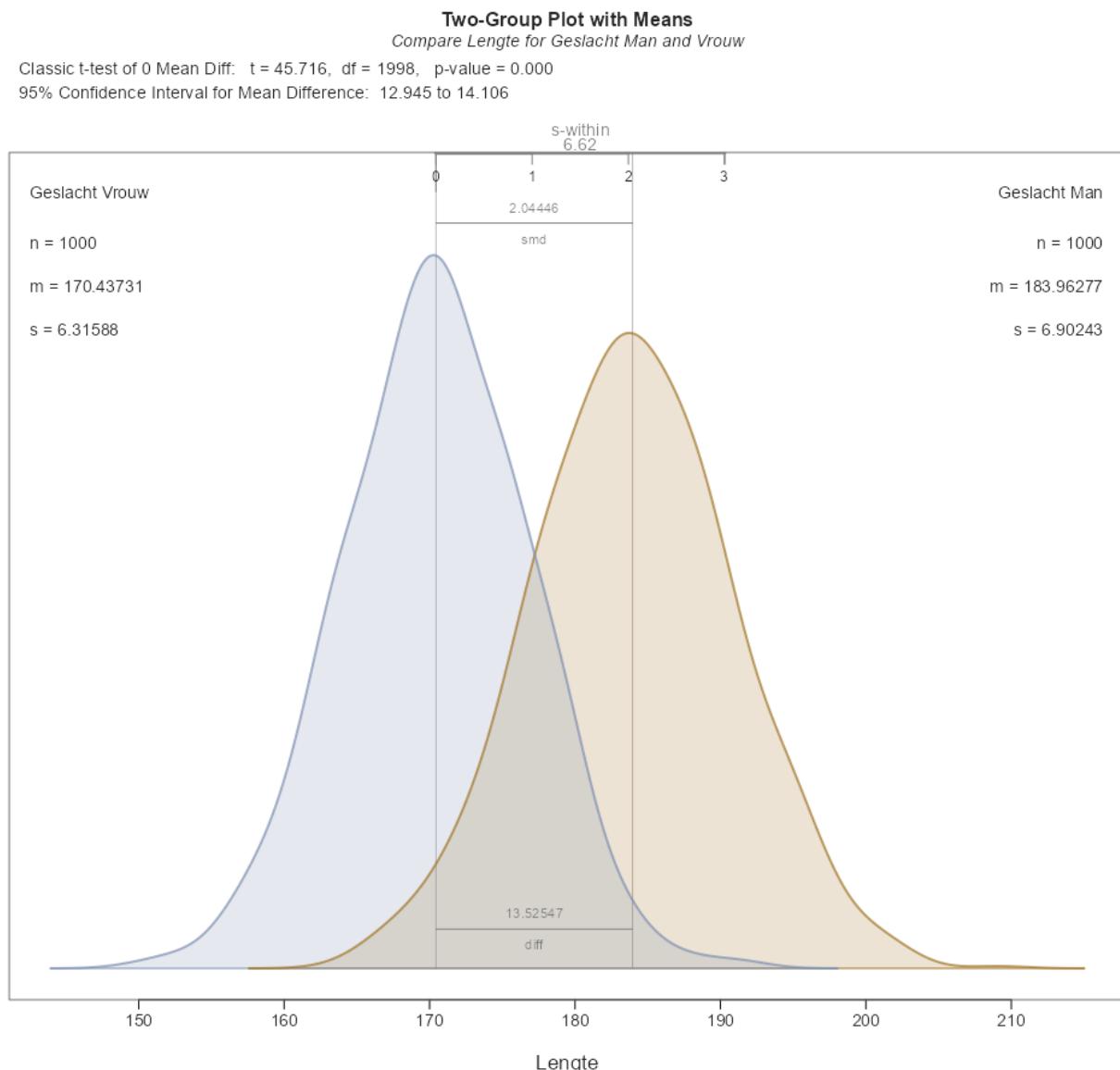
We kunnen dit resultaat grafisch tonen. Ten eerste kunnen we laten zien hoe de verdelingen er tussen mannen en vrouwen uitzien met daarboven de p-waarde (**Figuur 23**).



**Figuur 23.** De verdeling van lengtes bij mannen en vrouwen afgebeeld in een boxplot. Linksboven zien we de p-waarde die hoort de bij independent samples t-test.

<sup>38</sup> Voor wie nu de vraag stelt dat het gaat om één enkele observatie van groepen stelt een juiste vraag die we adresseren in het onderdeel rondom betrouwbaarheidsintervallen.

Hoewel enigszins informatief laat deze figuur niet goed zien waarom de nulhypothese uiteindelijk toch verworpen wordt. Daarom is het misschien beter om een ander figuur te tonen wat lijkt op **Figuur 15****Figuur 19**, maar een flinke statistische verdiepingsslag maakt (**Figuur 24**).



**Figuur 24.** De verdeling van lengtes voor mannen en vrouwen, met daarbij informatie over het gemiddelde (m), de standaard deviatie (s), het verschil (diff) en het gestandaardiseerde verschil (smd).

Om de gegevens uit deze figuur enigszins te duiden is het belangrijk om de focus te houden op het verschil in gemiddelde. Dat verschil is ongeveer 13.5 cm<sup>39</sup> en dat is groter dan de 0 uit

---

<sup>39</sup> Theoretisch zou het verschil 13,4 cm moeten zijn, maar omdat we de data gesimuleerd hebben met elk 1000 personen wijkt dit wat af.

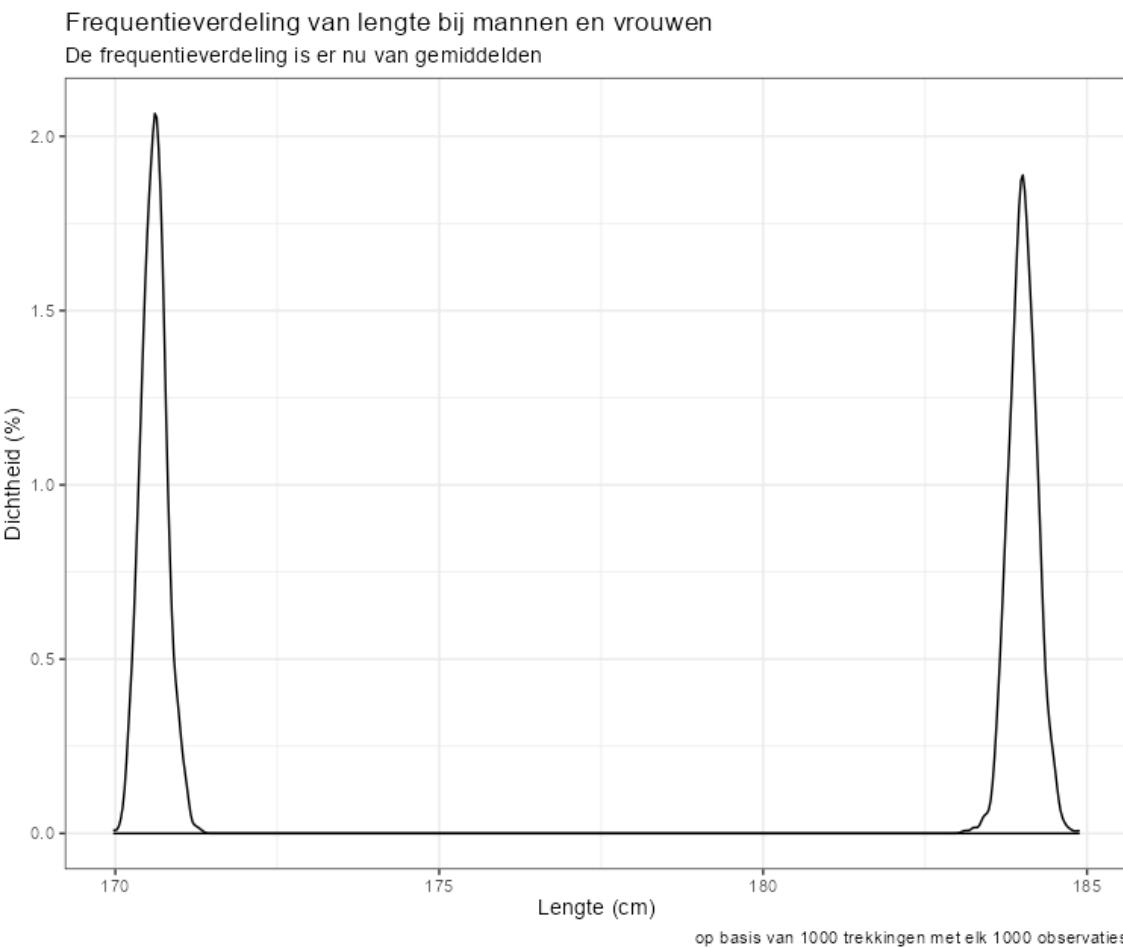
de nulhypothese. Het is nu zaak om te zien of een gemiddeld verschil van 13,5 past in de nulhypothese dat de groepen gemiddeld gelijk zijn aan elkaar. Alvorens dit te doen (het antwoord staat al in de grafiek) is het verstandig om eerst het concept ‘betrouwbaarheidsinterval’ te introduceren.

## Betrouwbaarheidsinterval

Wat we in de statistiek heel graag willen weten is hoe betrouwbaar de observaties zijn die we hebben gemaakt. Als we een dataset krijgen en daaruit opmaken dat twee groepen, indien met elkaar vergeleken, een verschil van 13,5 cm laten zien dan willen we graag weten of dit ver afwijkt van het ‘ware’ getal. Let wel: het ‘ware’ getal valt nooit te observeren en in de frequentistische statistiek heerst de gedachte dat hoe meer data er verzameld is hoe meer bewijs er is voor een bepaald gemiddelde. Dit is ook waarom het frequentistische statistiek heet: de observaties staan voorop.

We hebben ons tot nu toe beperkt dat een dataset afkomstig van een enkel moment. Op basis van die dataset kun je wel een uitspraak doen, maar we willen graag weten hoe betrouwbaar het verschil van 13,5 cm is. Misschien dat bij een volgende steekproef het verschil wel -13,5 cm is. Dit zou betekenen dat het gemiddelde van beide verschillen 0 is wat de nulhypothese van geen verschil ondersteund. Het is daarom te voorbarig om te stellen dat het gevonden verschil van 13,5 cm voldoende bewijs is om te stellen dat de nulhypothese van ‘geen verschil’ verworpen kan worden.

Wat we nu kunnen doen is een simulatie uitvoeren. Stel dat we 10,000x een steekproef doen van 1,000 mannen en 1,000 vrouwen uit de bekende verdeling uit **Tabel 1**. Welke gemiddelde waarden per groep zouden we dan vinden? En hoe zou het verschil tussen die groepen er dan uitzien? Laten we deze exercitie uitvoeren waarvan we de resultaten zowel in de vorm van een tabel als ook een grafiek zullen tonen. Om de resultaten gemakkelijker te verteren is het beter om eerst te grafieken te tonen. **Figuur 25** laat de verdeling van gemiddeldes zien. Links zie je de verdeling van 1000 gemiddeldes voor vrouwen en rechts voor de mannen. In tabelvorm hebben we de volgende gegevens (**Tabel 5**).



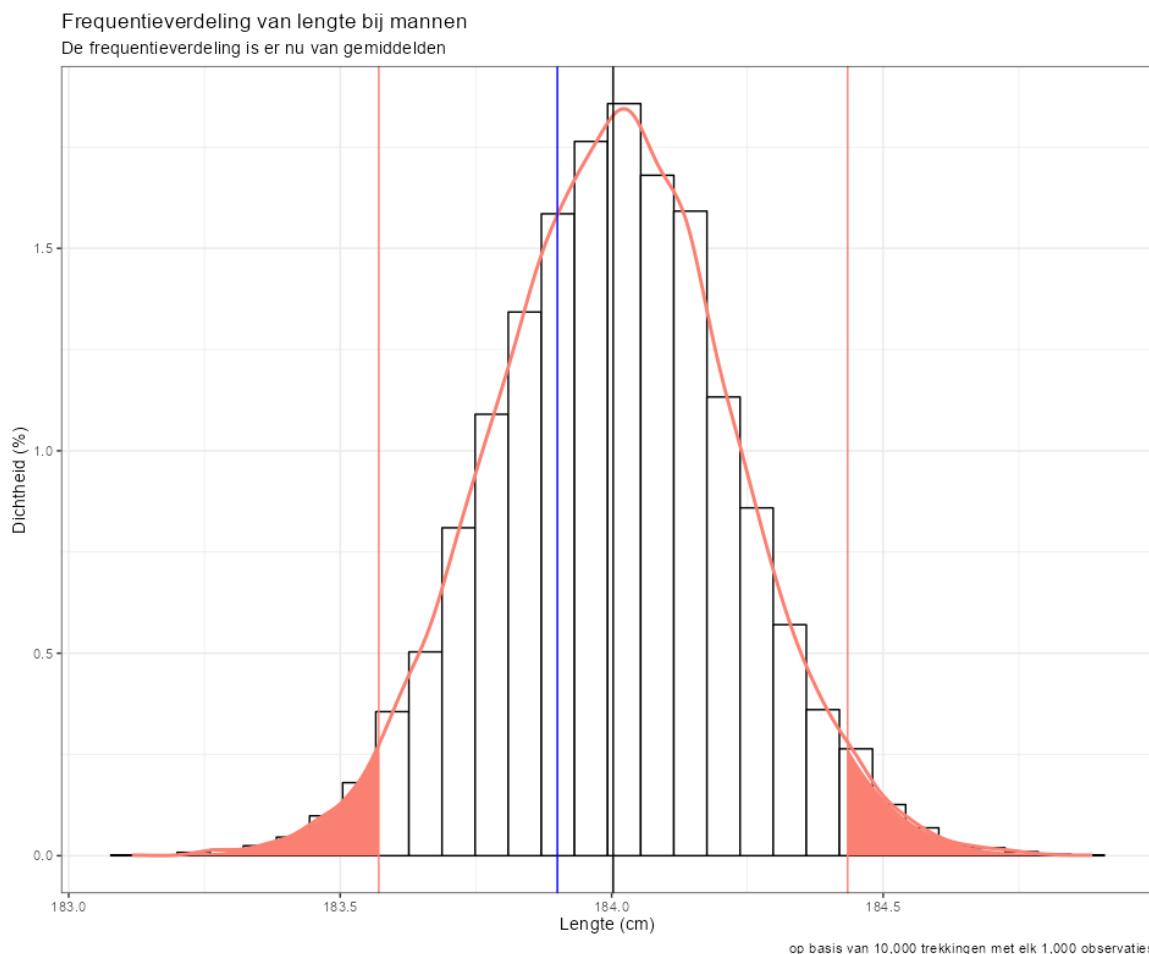
**Figuur 25.** De verdeling van gemiddeldes voor mannen en vrouwen. Het valt duidelijk te zien dat er tussen die gemiddeldes geen observaties zijn. Er is daadwerkelijk een kloof tussen twee torens.

<b>Geslacht</b>	<b>Gemiddelde lengte (cm)</b>		
	<i>populatie</i>	<i>steekproef 1</i>	<i>10,000 steekproeven</i>
Mannen	184	183.9	184
Vrouwen	170.6	170.4	170.6
Verschil	13.4	13.5	13.5

**Tabel 5.** De waarden zoals verkregen uit het Radbouw UMC afgezet tegen de waarden verkregen uit simulaties. Één enkele steekproef bevat 2000 simulaties: 1000 gesimuleerde observaties voor mannen en 1000 gesimuleerde observaties voor vrouwen.

Duidelijk is dat dat het gemiddelde van één enkele steekproef kan afwijken van het gemiddelde, maar dat als we 10,000 gemiddeldes berekenen op 1,000 trekkingen elk we terugkomen bij de originele populatie. Dat is niet gek, maar misschien goed om toch nog een keer te laten zien in een figuur. Het volstaat om het voor alleen de mannen te doen, want voor de vrouwen gaat hetzelfde principe op. Het resultaat is **Figuur 26** en wat dit figuur laat zien is de verdeling van 10,000 gemiddeldes. Die verdeling heeft een gemiddelde van 184.

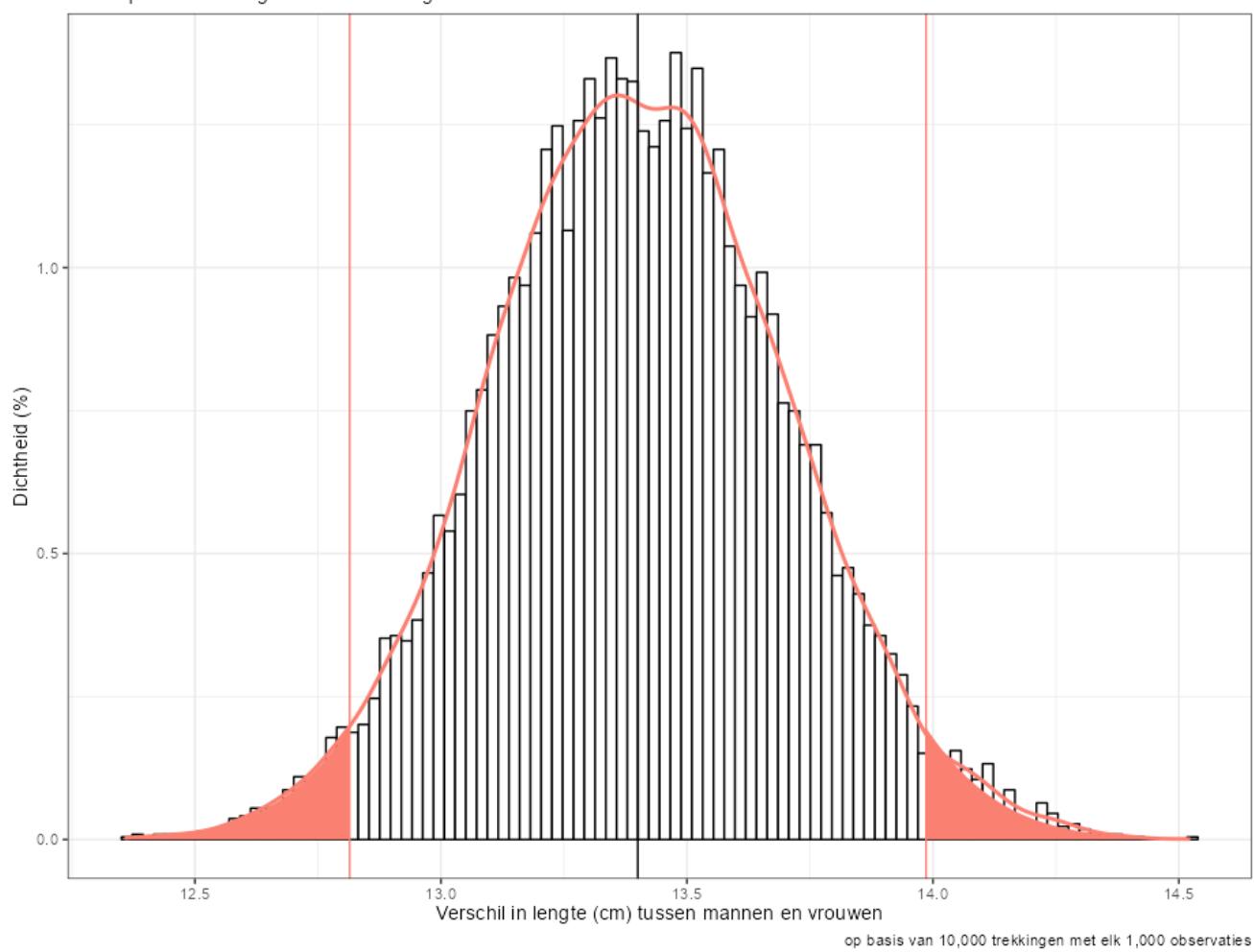
De blauwe lijn is het gemiddelde uit een enkele steekproef: 183.9 en die valt mooi in die verdeling. Wat ook belangrijk is zijn de grenzen waarbinnen 95% van de observaties vallen: 183.5 en 184.4. Dat betekent dat als we 10,000 keer een steekproef nemen en die gemiddeldes bij elkaar zetten 95% van de observaties tussen de 183.5 en 184.4 valt. Deze spreiding noemen we ook wel het 95% betrouwbaarheidsinterval.



**Figuur 26.** Gesimuleerde verdeling van gemiddeldes. De blauwe lijn is één willekeurig gekozen gemiddelde. De zwarte lijn laat het gemiddelde van alles gemiddeldes zien (184 cm) en de rode lijnen tonen het 95% betrouwbaarheidsinterval.

Doen we hetzelfde voor de vrouwen dan komen we uit op een 95% spreiding van: 170.2 en 170.9. **Figuur 25** laat dit eigenlijk al zien. Wat ons dan nog rest is een grafiek te maken van de spreiding van de verschillen (**Figuur 27**). Wat we dan zien is dat het gemiddelde verschil 13.4 is met een spreiding tussen de 12.8 en de 13.9. Een gemiddeld verschil van 0 vinden we geen enkele keer.

Frequentieverdeling van verschil in lengte tussen mannen en vrouwen  
De frequentieverdeling is er nu een van gemiddelen



**Figuur 27.** De verdeling van gesimuleerde verschillen met in het midden de zwarte lijn die het gemiddelde verschil toont (13.4 cm). De rode lijnen zijn de 95% betrouwbaarheidsintervallen. We zien geen enkele keer de waarde 0.

Wat betekent dit nu allemaal? Ten eerste dat wanneer we twee groepen met elkaar vergelijken het wel degelijk zo kan zijn dat er overlap is tussen die groepen (**Figuur 15**). Dit wil niet zeggen dat de nulhypothese dat er geen verschil is blijft behouden. Om dit te onderzoeken is het belangrijk om op zoek te gaan naar het 95% betrouwbaarheidsinterval van het verschil. Volgens de frequentistische statistiek wordt het vervolgens evident (**Figuur 24**) dat de verschillen tussen mannen en vrouwen, zoals waargenomen in de steekproef, geen toeval kan zijn. Want, zo stelt men, de kans dat een verschil van 13.5 gevonden wordt terwijl het echte verschil 0 is, is kleiner dan 0.001%. Onder de nulhypothese van 'geen verschil' komt een bevinding als die van ons zo weinig voor dat het haast niet kan kloppen dat de nulhypothese van 'geen verschil' klopt. De alternatieve hypothese, die van 'er is een verschil', moet dus wel correct zijn. De observaties laten dit namelijk zien.

Deze manier van redeneren vormt het hart van de frequentistische statistiek waarbij er altijd wordt gewerkt met een nulhypothese en een alternatieve hypothese. Wanneer het gaat om het aantonen van verschillen tussen groepen is de nulhypothese haast altijd dat er ‘geen verschil is’. Mochten we een verschil ontdekken in de geobserveerde waarden dan volgt de berekening of dat verschil mogelijk is onder de nulhypothese. Dit is de p-waarde die we al een aantal keren hebben benoemd. Als die p-waarde kleiner is dan de vooraf gestelde grens van wat een acceptabele kans is, dan vervalt in de frequentistische statistiek de aannemelijkheid van de nulhypothese.

### Zaken die van invloed zijn op het betrouwbaarheidsinterval

Het lijkt nu wellicht dat een groot verschil tussen groepen voldoende is, maar dat is niet zo. De spreiding is ook belangrijk, net als de grote van de groep en natuurlijk de vooraf bepaalde grenswaarde.

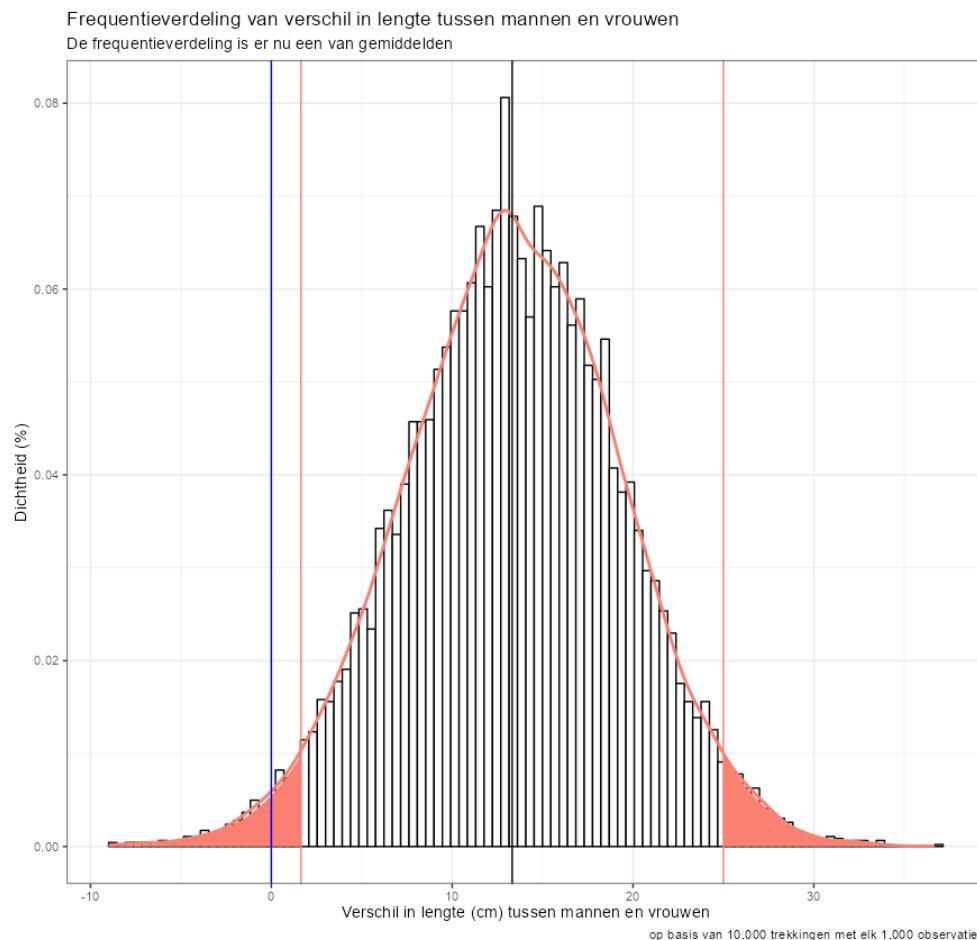
In **Tabel 6** maak ik zichtbaar wat het 95% betrouwbaarheidsinterval is als functie van de spreiding (standaard deviatie). Duidelijk te zien is dat het betrouwbaarheidsinterval groter wordt naarmate de spreiding groter wordt.

<b>Factor</b>	<b>Spreiding (cm)</b>		<b>95% Betrouwbaarheidsinterval verschil</b>	
	<i>Man</i>	<i>Vrouw</i>	<i>Links</i>	<i>Rechts</i>
1	7	6.3	12.9	14
2	14	12.6	12.5	14.5
5	35	31.5	10.5	16.3
10	70	63	7.5	19.4
20	140	126	1.5	25.3

**Tabel 6.** Het 95% betrouwbaarheidsinterval bij een gelijk gemiddelde, maar een steeds groter wordende spreiding. Naarmate de spreiding van één of beide groepen groter wordt zal ook het betrouwbaarheidsinterval toenemen. Hoe groter de spreiding rondom een gemiddelde hoe minder betrouwbaar is dat gemiddelde.

Dit is niet gek, want hoe meer onzeker we zijn over het gemiddelde, hoe minder zeker we kunnen we zijn over het betrouwbaarheidsinterval van het verschil tussen de groepen. Grafisch ziet het laatste voorbeeld (factor 20) er uit zoals in **Figuur 28**. We zien nu voor het eerst wel de blauwe lijn van ‘geen verschil’. Het verschil is toch significant zoals je kunt zien, want de blauwe lijn valt in het rode gebied: de vooraf bepaalde grenswaarde. Wanneer we

willen aantonen dat er echt een verschil is tussen groepen wil je de lijn van ‘geen verschil’ buiten je kritieke gebieden hebben. Dit is dus omgekeerd zoals bijvoorbeeld in **Figuur 22**.



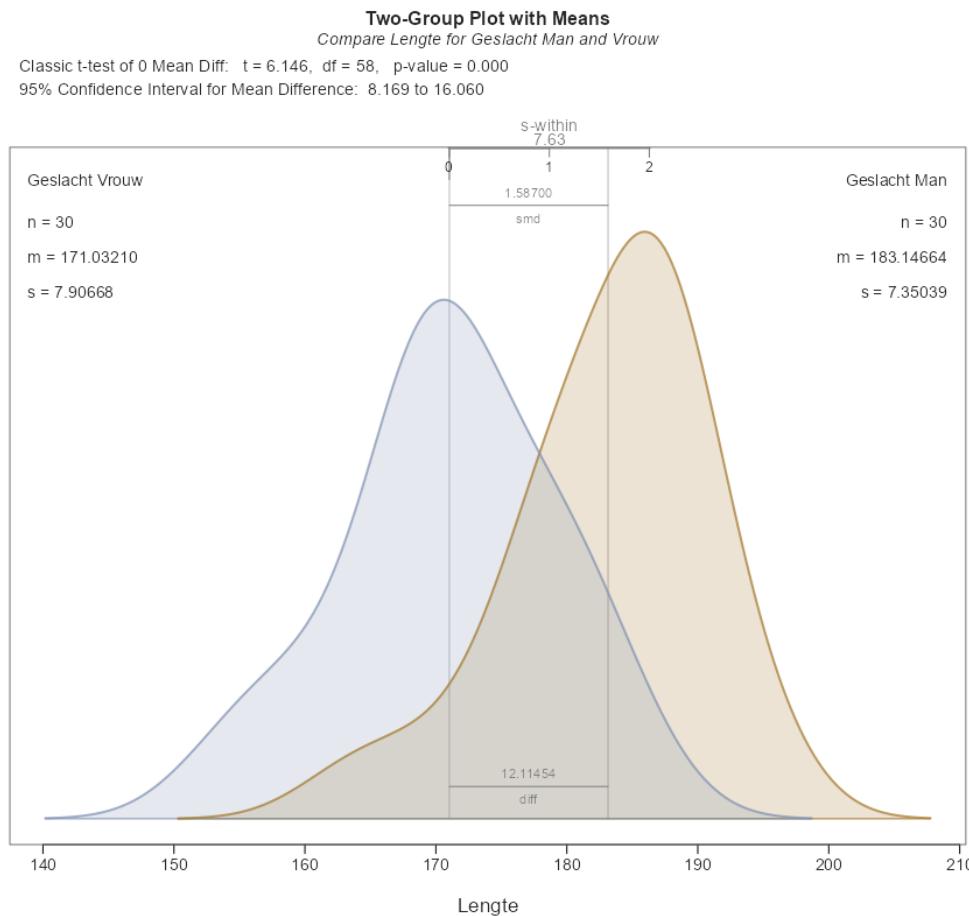
**Figuur 28.** De verdeling van verschillen wanneer de spreiding van beide groepen een factor 20 is van de originele spreiding. We zien nu voor het eerste de blauwe lijn van ‘geen verschil’ maar deze valt buiten de grenswaarde zodat we alsnog de nulhypothese van ‘geen verschil’ verwerpen.

Nu is niet alleen de spreiding van belang, maar ook de groepsgrootte: het aantal observaties dat we maken. De grootte van de steekproef is misschien wel de grootste invloed die er is. Niet alleen kan de grootte het gemiddelde beïnvloeden, maar het beïnvloedt absoluut ook de spreiding rondom een gemiddelde.

Stel dat we geen 1000 observaties hadden om een gemiddelde en een spreiding te bepalen, maar alleen 30 observaties. Als we 30 observaties simuleren uit een bekend gemiddelde en standaard deviatie dan zou het goed kunnen zijn dat we een figuur zien zoals

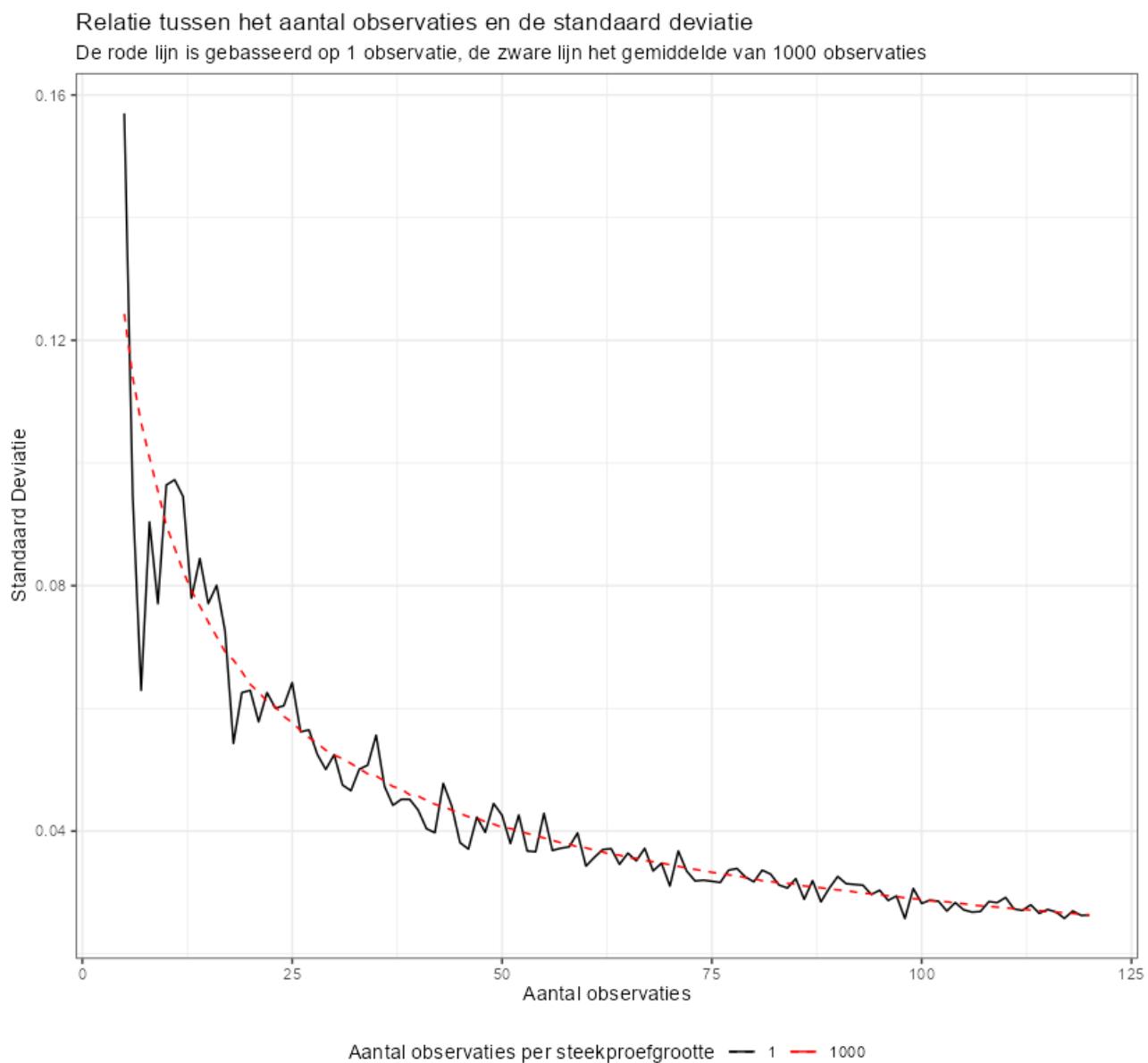
**Figuur 29.** Wat direct opvalt is dat de vorm niet meer mooi klokvormig is. Ook zien we

gemiddeldes die afwijken van het theoretisch gemiddelde en ook de spreiding wijkt af. Toch is ook in dit voorbeeld het verschil groot genoeg om in de frequentistische statistiek te zeggen dat de nulhypothese van 'geen verschil' niet kan worden behouden.



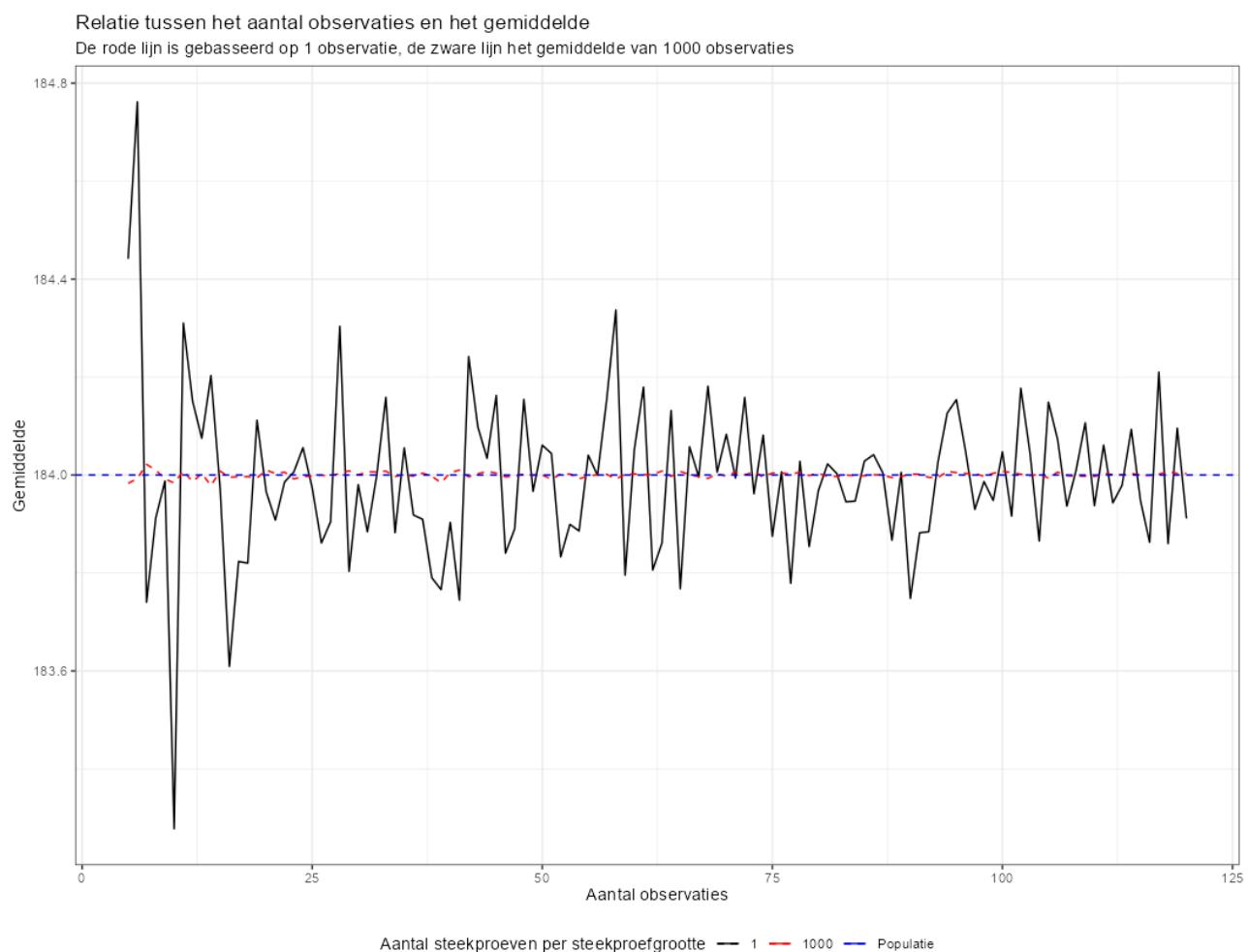
**Figuur 29.** De verdeling van lengte bij mannen en vrouwen zoals bezien wanneer we 30 trekkingen doen uit een voor ons bekende gemiddelde én spreiding. Het geobserveerde gemiddelde en spreiding wijken af van de bekende waarden.

Een simulatie die laat wat de relatie is tussen steekproefgrootte en standaard deviatie zien we in **Figuur 30**. Wat we duidelijk zien is dat de standaard deviatie globaal afneemt naarmate de steekproefgrootte toeneemt. Ook zijn er schommelingen wanneer we het houden bij één enkele observatie. Dit alles toont hoe breekbaar deze puntschattingen zijn.



**Figuur 30.** Relatie tussen standaard deviatie en aantal observaties per steekproefgrootte.

Hetzelfde kunnen we trouwens snel doen voor het gemiddelde wat er dan vervolgens zo uit ziet (**Figuur 31**). Dit alles mag eigenlijk niet verbazen: hoe meer observaties hoe betrouwbaarder de schatting<sup>40</sup>.



**Figuur 31.** Relatie tussen de schatting van het gemiddelde en de steekproefgrootte. De blauwe lijn is het theoretisch gemiddelde (184 cm). De zwarte en rode lijn zijn schattingen gebaseerd op 1 steekproef per steekproefgrootte en het gemiddelde van 1000 steekproeven, respectievelijk.

Nu we dit alles weten rest is het tijd om het belangrijkste onderdeel van de frequentistische statistiek te bespreken: de keuze voor 5% als de grenswaarde.

<sup>40</sup> Daargelaten dat er geen andere contextuele informatie nodig is die ontbreekt.

## Waarom is 5% de grenswaarde?

Ik zal maar direct met de deur in huis vallen: hier heb ik geen passend antwoord op. Ergens is er ooit besloten dat 5% de grenswaarde is die gehanteerd kan worden om te bepalen dat een gevonden waarde wel of niet bij de nulhypothese past. Dit alleen is een onwenselijke situatie omdat deze 5% bepaalt of een waarde wel of niet het predicaat statistisch significant krijgt. Zo belangrijk is die 5% dat het overschrijden ervan vaak hét verschil maakt tussen wel of niet kunnen publiceren. De kritiek van BNNVARA/Zembla op de manier van toetsen is daarmee een begrijpelijke klacht: mocht het toepassen van een andere, verdedigbare, toets wel leiden tot een statistisch significant verschil dan opent dit direct een aantal deuren in de wetenschappelijke wereld. Een geobserveerde waarde wordt vanaf dat moment als een ‘echt’ effect gezien. Laten we daarom eens kijken wat die 5% nou precies betekent.

### Alfa en betá

Ik schreef al dat niemand evident kan aantonen waarom 5% de gebruikte grens is en niet 4% of 10%. Of waarom we überhaupt een grens moeten hebben. Toch is 5% de grens en het lijkt daarmee meer op een gentlemen’s agreement dan op een wetenschappelijke bepaling.

Maar wat betekent die grens nou precies? In feite is het volgende waar: als we een 5% grens hanteren waarbinnen we een waarde zo bijzonder vinden dat deze niet van de nulhypothese afkomstig kan zijn (en deze daarmee dus moet worden vervangen), dan accepteren we 5% ook als de kans op een vals-positieve waarde. Dat betekent dat als we 20x een toets doen, en 20x een waarde (bijvoorbeeld een verschil) als statistisch significant bestempelen, ongeveer één van die verschillen geen echt verschil is<sup>41</sup>. Dit komt omdat we de grens op 5% zetten en ergens hebben we bepaald dat dit de grens moet zijn. Die grens noemen we ook wel de alfa ( $\alpha$ ) waarde.

Maar die 5% staat niet alleen. In de frequentistische statistiek heb je ook nog de betá ( $\beta$ ) waarde die synoniem staat voor de vals-negatieve waarde die we acceptabel vinden. Deze  $\beta$  waarde wordt ook vaak wat ad-hoc gekozen. Wat dan nog rest is de  $1 - \beta$  waarde en dit is de kracht van een studie wat neerkomt op de frequentie succes. Met andere woorden: als ik 100x een experiment doe, en mijn kracht is 80%, dan verwacht ik dat als er echt een

---

<sup>41</sup> Dit is géén hard gegeven: we spreken hier over kansen.

verschil is ik dit in 80% van de gevallen zal gaan zien. Deze drie waarden: de alfa ( $\alpha$ ), betá ( $\beta$ ) en kracht ( $1-\beta$ ) worden ook wel gebruikt om vooraf te bepalen hoe groot de steekproefgrootte moet zijn om de nulhypothese te verwerpen. Dus hoewel de  $\alpha$  een belangrijk kenmerk is, staat deze niet alleen in het toetsen van een hypothese: de  $\beta$  is ook belangrijk. We kunnen dit gegeven laten zien in **Tabel 7**.

		De nulhypothese is	
		Correct	Niet correct
De nulhypothese werd	Verworpen ( $P < \alpha$ )	Correct negatief $\alpha$	Vals positief $\beta$
	Niet verworpen ( $P \geq \alpha$ )	Vals negatief $1 - \alpha$	Correct positief $1 - \beta$

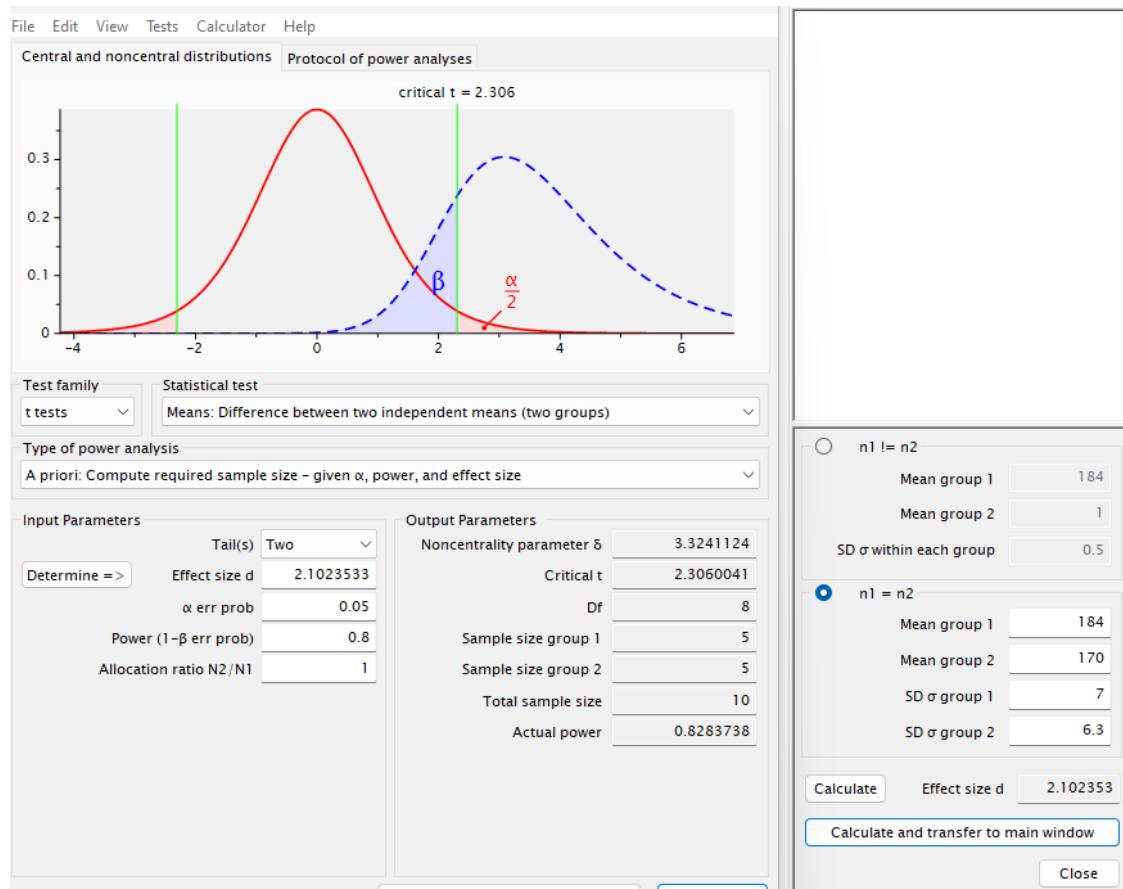
**Tabel 7.** Relatie tussen het verwerpen of behouden van de nulhypothese in relatie tot  $\alpha$  en  $\beta$ .

Hoewel de tabel geen antwoord geeft op de vraag waarom bijna elke studie kiest voor een  $\alpha$  van 5% én een  $\beta$  van 20% (en dus voor een kracht van 80%), kunnen we wel doorlopen (en visualiseren) wat deze keuzes betekenen. Om dit te doen zal ik eerst laten zien hoe deze waarden met elkaar in verhouding staan door gebruik te maken van een klein stukje software<sup>42</sup>. Dit doe ik door data van de lengte van Nederlandse mannen en vrouwen te simuleren.

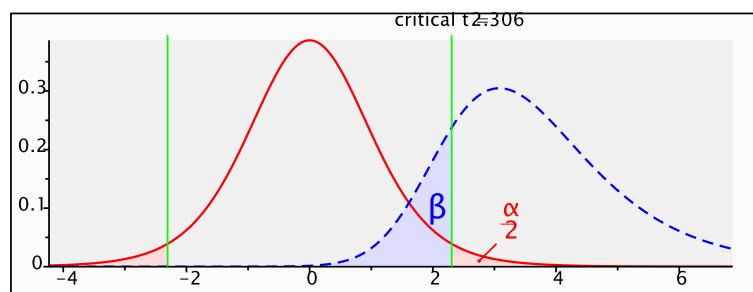
We hebben in **Figuur 27** gezien hoe sterk het verschil was tussen de mannen en vrouwen. Dat verschil was zo sterk dat als we een studie zouden ontwerpen met een  $\alpha$  van 5% én een  $\beta$  van 20% (en dus voor een kracht van 80%), we ongeveer 5 mensen per groep nodig zouden hebben om het verschil op te merken. Dit zie je in **Figuur 32**. Wat het programma doet is uitrekenen welke t-waarde groter is dan de grenswaarde gegeven het gemiddelde van beide groepen, hun spreiding en de gekozen  $\alpha$  en  $\beta$  waarden. Omdat de relatie tussen deze waarden in een sluitende formule te vangen is, is het eenvoudig om ergens te starten en vervolgens de rest in te vullen. Laten we gemakshalve de grafiek uit **Figuur 32** groter weergeven en omzetten tot **Figuur 33**. De groene lijnen zijn de waarden die toebehoren aan een  $\alpha/2$ . Het gedeelte wat blauw gearceerd is, staat links van één van deze

<sup>42</sup> G\*Power 3 <https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower>

lijnen en toont het gebied waarbij we kunnen spreken van een vals negatieve ( $\beta$ ): het onterecht behouden van de nulhypothese.



**Figuur 32.** Een weergave van het G\*Power programma waarmee we kunnen laten zien wat de kracht van een studie is om een effect te vinden, gegeven het gemiddelde van twee groepen, hun standaard deviatie, de groeps grootte en de gekozen  $\alpha$  en  $\beta$  waarden.



**Figuur 33.** De theoretische verdeling van twee groepen onder een vooraf bepaald gemiddelde, spreiding en groeps grootte.

## Effectgrootte

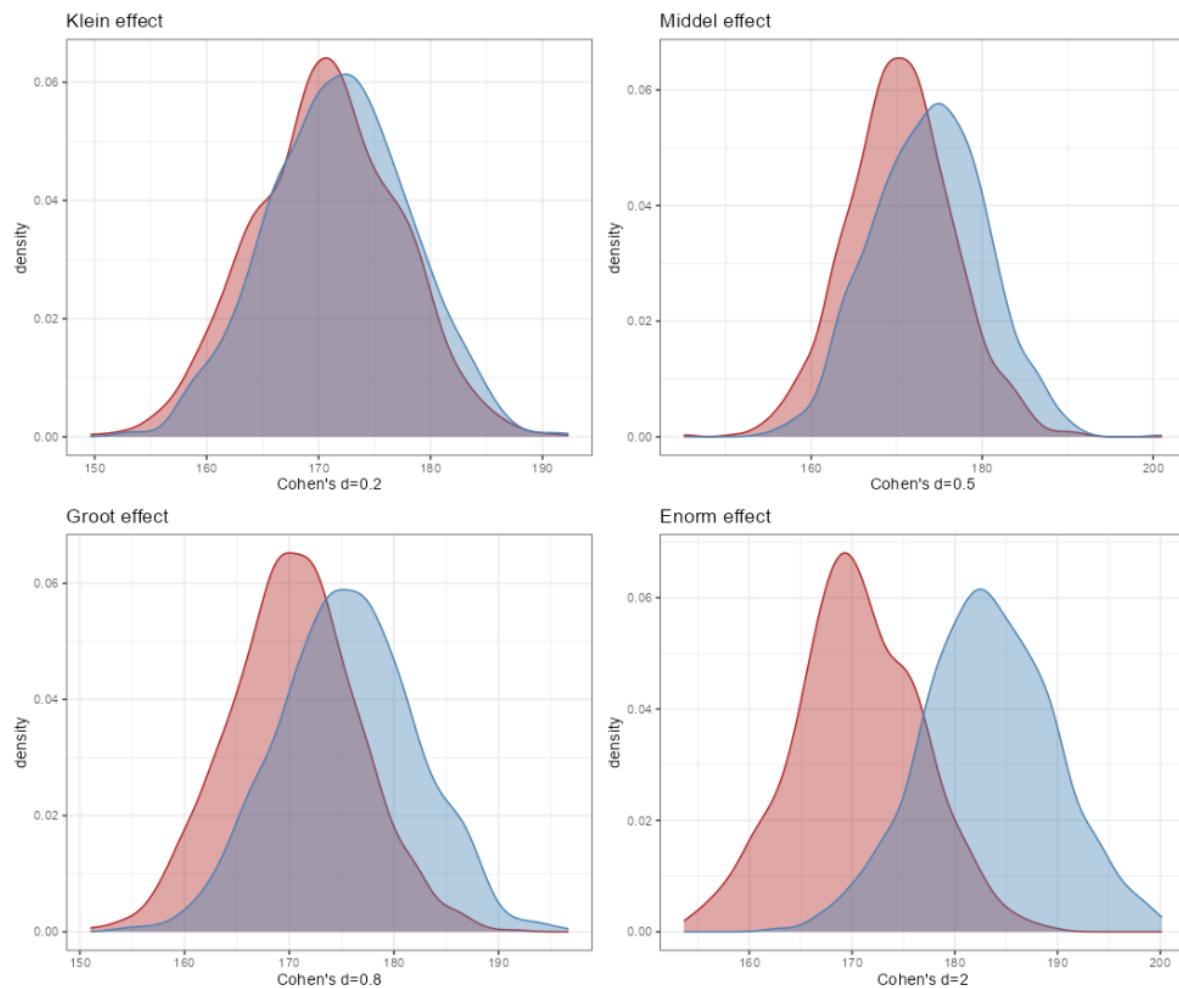
Naarmate de spreiding van de groepen kleiner wordt (vaak door toename van de groepsgrootte) en / of het verschil tussen de groepen groter wordt, zal de  $\beta$  waarde ook afnemen. Het verschil tussen groepen wordt ook wel eens de effectgrootte genoemd. In **Figuur 32** zie je dat deze waarde 2.1 is.

De effectgrootte is handig om te weten, of naartoe te werken, omdat het hier gaat om een gestandaardiseerd verschil. Het is dus communiceerbaar over verschillende onderzoeken heen en niet afhankelijk van onderzoeksfield. Een gewenste effectgrootte van 2 is overal hetzelfde. Wel kan een effectgrootte op verschillende manieren berekend worden. Zo hebben we onder andere Cohen’s  $d$  en Hedges’  $g$ <sup>43</sup> maar er zijn er nog veel meer. In het kort komt het erop neer dat de waarden gestandaardiseerde verschillen meten die ook weer gerangschikt worden op grootte. Een  $d$  waarde van 2 wordt gezien als een enorm verschil. In ons geval zitten we daar zelfs boven.

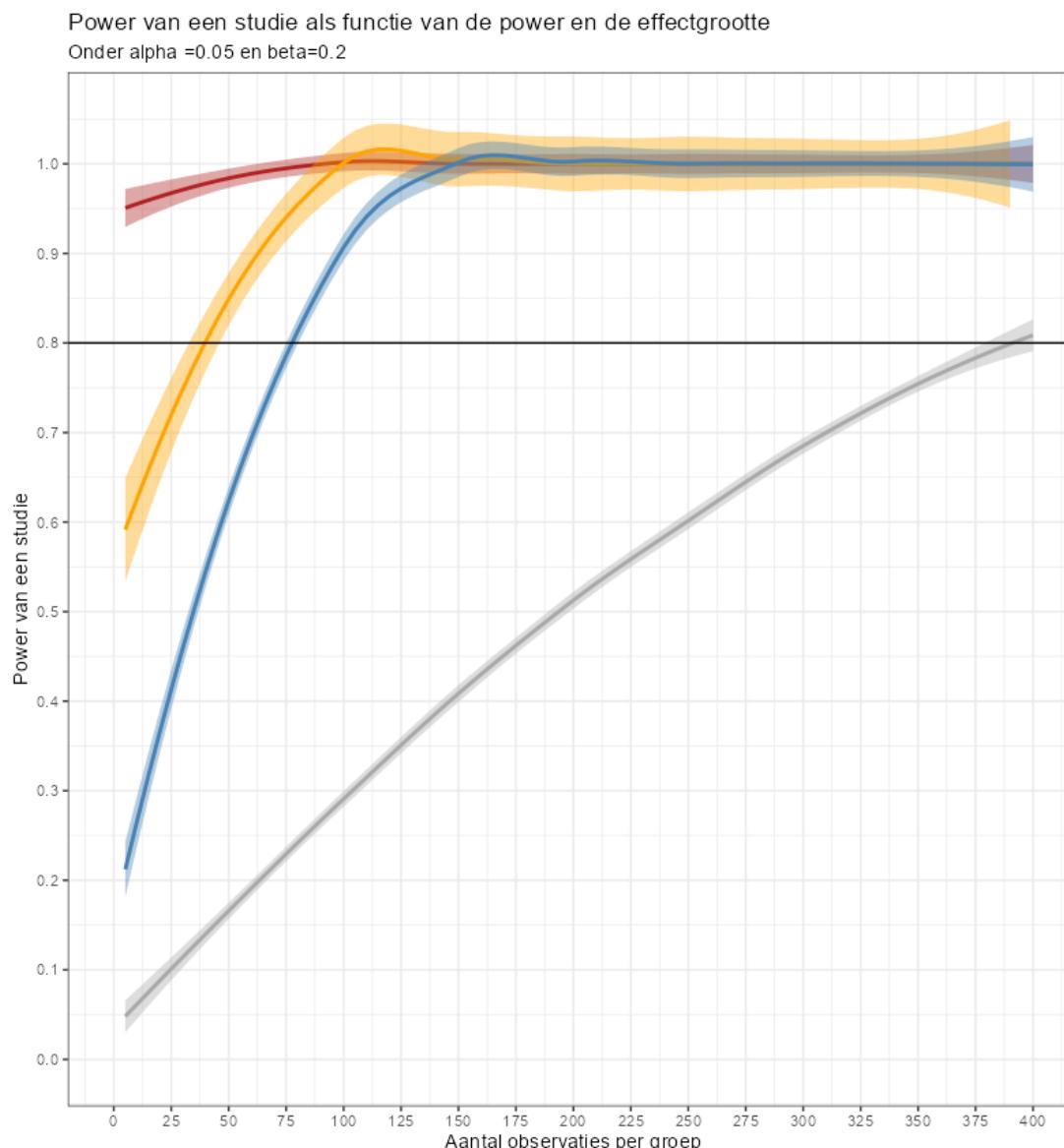
Wellicht is het goed om te tonen hoe nu de verschillende onderdelen, zoals  $\alpha$ ,  $\beta$ , effectgrootte en steekproefgrootte met elkaar samenhangt. Het resultaat zien we in **Figuur 34**. We kunnen nu de relatie tussen effectgrootte en het aantal benodigde observaties ook tonen. In **Figuur 35** zien we dat het benodigd aantal observaties afneemt naarmate de effectgrootte toeneemt. We hebben hier gemakshalve de  $\alpha$  en  $\beta$  waarde gelijk gehouden. De effectgrootte tussen mannen en vrouwen is zo groot dat het haast lijkt alsof één enkele observatie per groep genoeg is. We kunnen dit ook laten zien door middel van **Figuur 36**. Eigenlijk laat deze figuur hetzelfde zien, namelijk dat hoe groter de effectgrootte is, hoe minder observaties er nodig zijn om de nulhypothese te vervangen.

---

<sup>43</sup> [https://en.wikipedia.org/wiki/Effect\\_size](https://en.wikipedia.org/wiki/Effect_size)

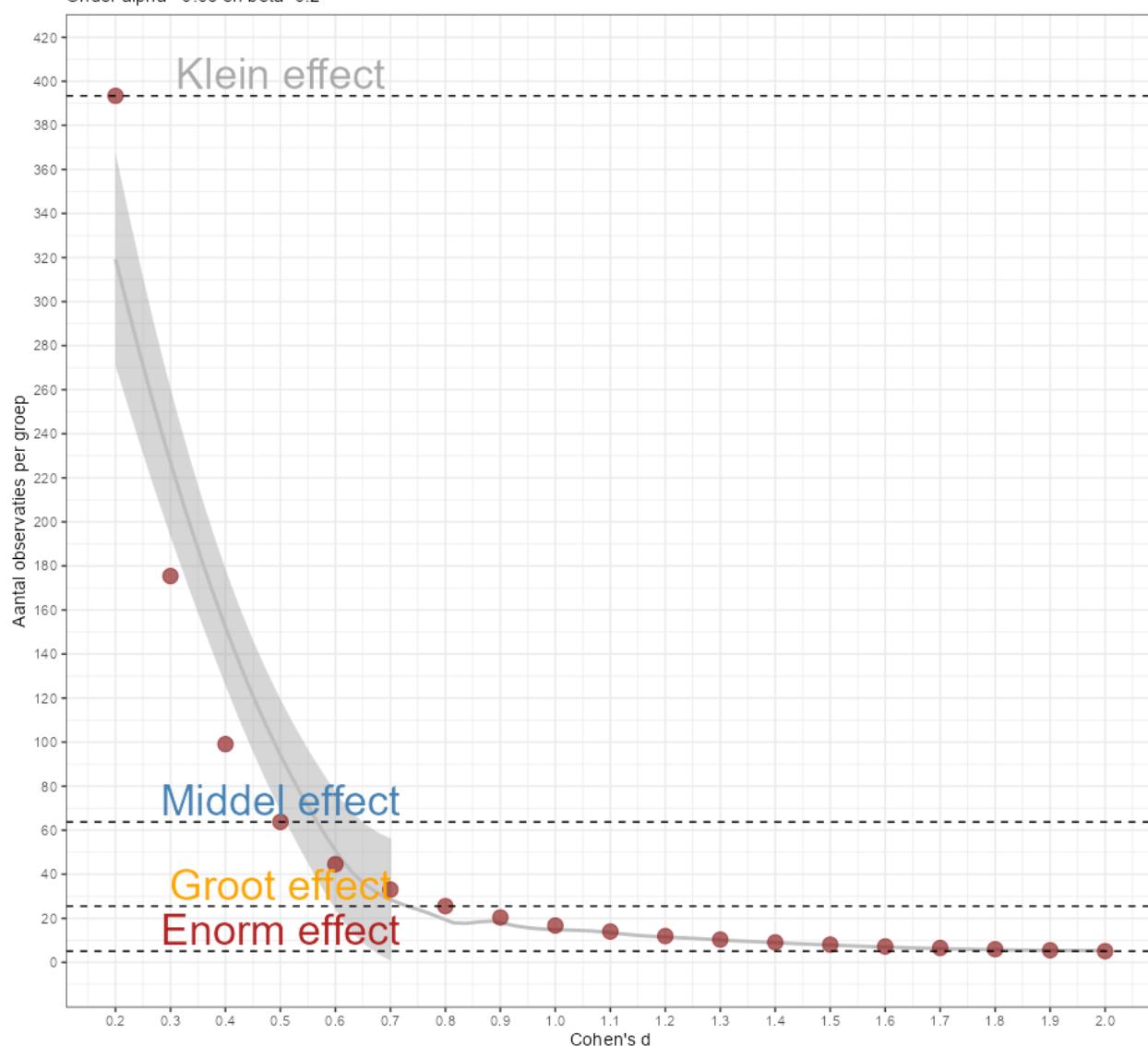


**Figuur 34.** Het verschil tussen twee groepen op basis van de effectgrootte. De plot rechtsonder lijkt het meest op de data zoals daadwerkelijk verzameld. Dit komt omdat ik ben begonnen met de verdeling links (rood) en daar een effectgrootte van twee heb opgeteld. Dit is ook de effectgrootte die we zelf hebben berekend.



**Figuur 35.** Relatie tussen aantal observaties per groep, de effectgrootte tussen groepen en de kracht van een studie bij een  $\alpha$  waarde van 0.05 en een  $\beta$  waarde van 0.2.

Benodigde steekproefgrootte als functie van Cohens'd  
Onder alpha =0.05 en beta=0.2

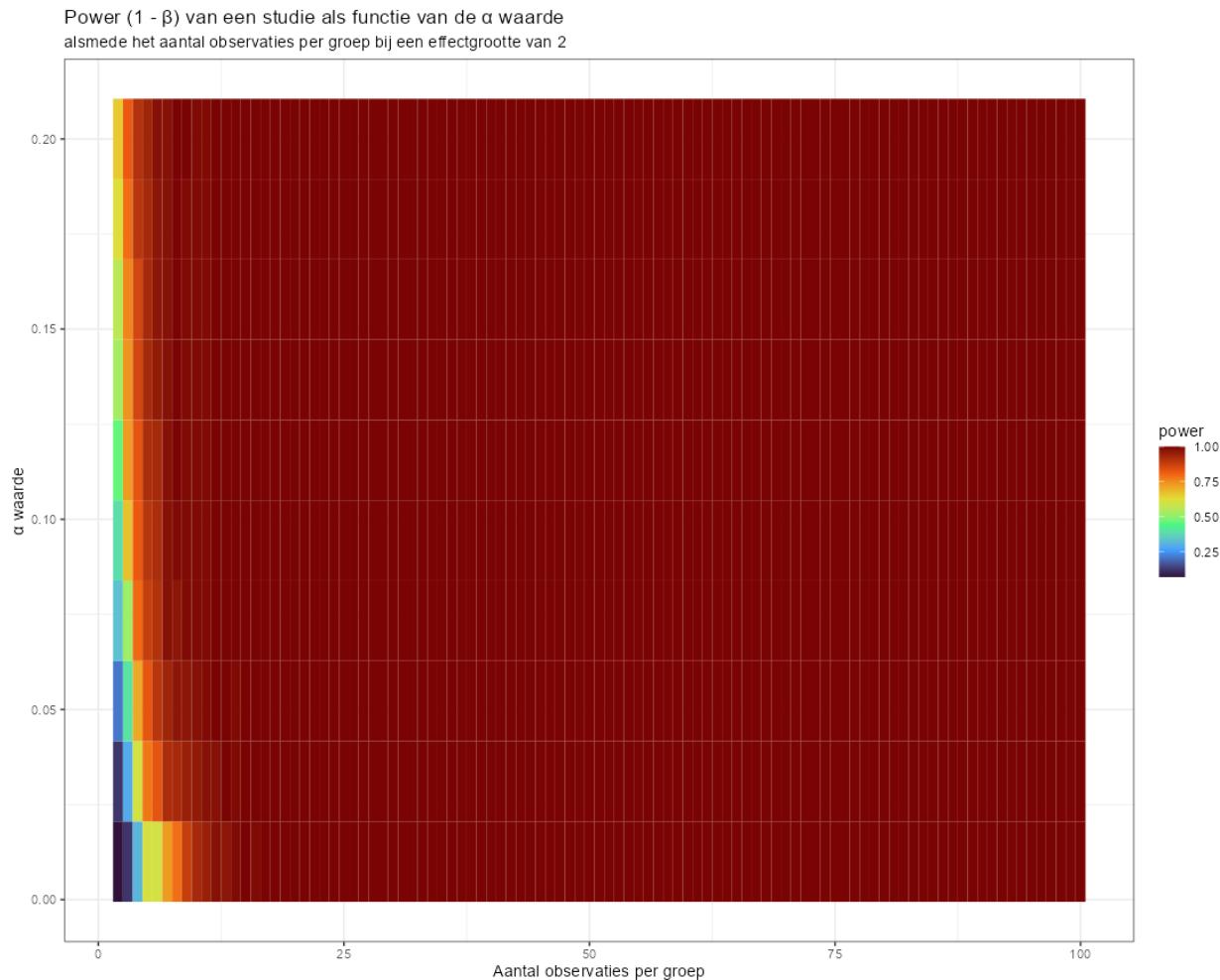


**Figuur 36.** Relatie tussen effectgrootte (Cohen's d) en het aantal observaties per groep. De stippeellijnen laten zien wat de minimale grootte per groep moet zijn onder een  $\alpha$  van 0.05 en een  $\beta$  van 0.2.

Het wordt nu tijd om de invloed van  $\alpha$  en  $\beta$  te laten zien. We beginnen met **Figuur 37** waarin we duidelijk kunnen zien dat de  $\alpha$  waarde een rol speelt. Bij een extreem kleine  $\alpha$  waarde vindt zelfs bij een effectgrootte van 2 een verschuiving van de power plaats. Echter gaat het hier om een  $\alpha$  waarde kleiner dan 0.0001 wat betekent dat we een hele klein kans op vals positieven accepteren. Het verschil is gewoonweg te groot. Wellicht dat een beter voorbeeld zou zijn waarbij we de effectgrootte kleiner maken, bijvoorbeeld 0.38<sup>44</sup>. Het resultaat zien

<sup>44</sup> Hierbij heeft groep 1 een gemiddelde van 170.6 met een standaard deviatie van 6.3. Groep 2 heeft een gemiddelde van 173 met een standaard deviatie van 7.

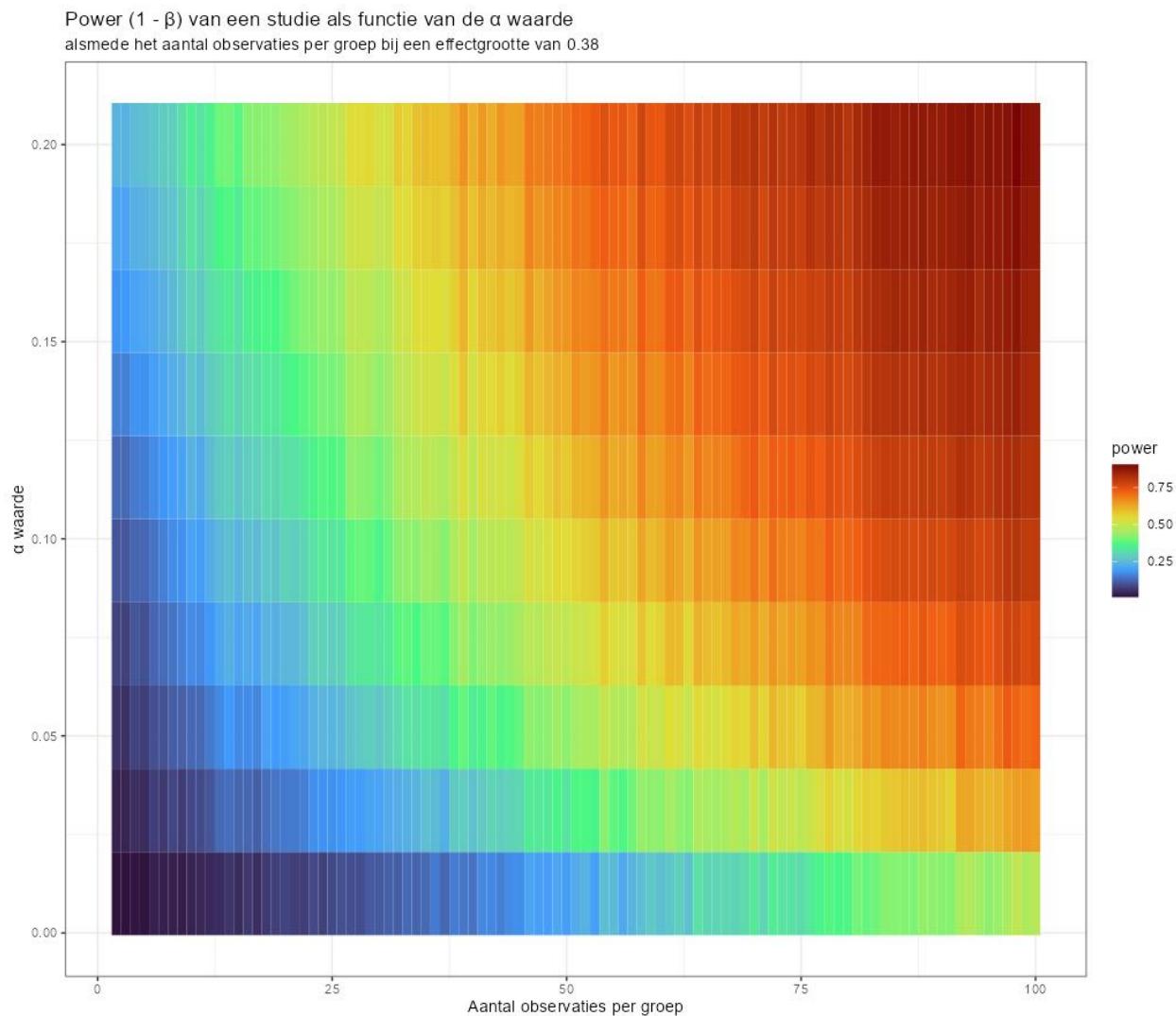
we in **Figuur 38**. Nu is duidelijk te zien hoe de  $\alpha$  waarde en de groeps grootte de power van de studie bepaalt. De keuze voor de  $\alpha$  waarde maakt dus wel degelijk uit.



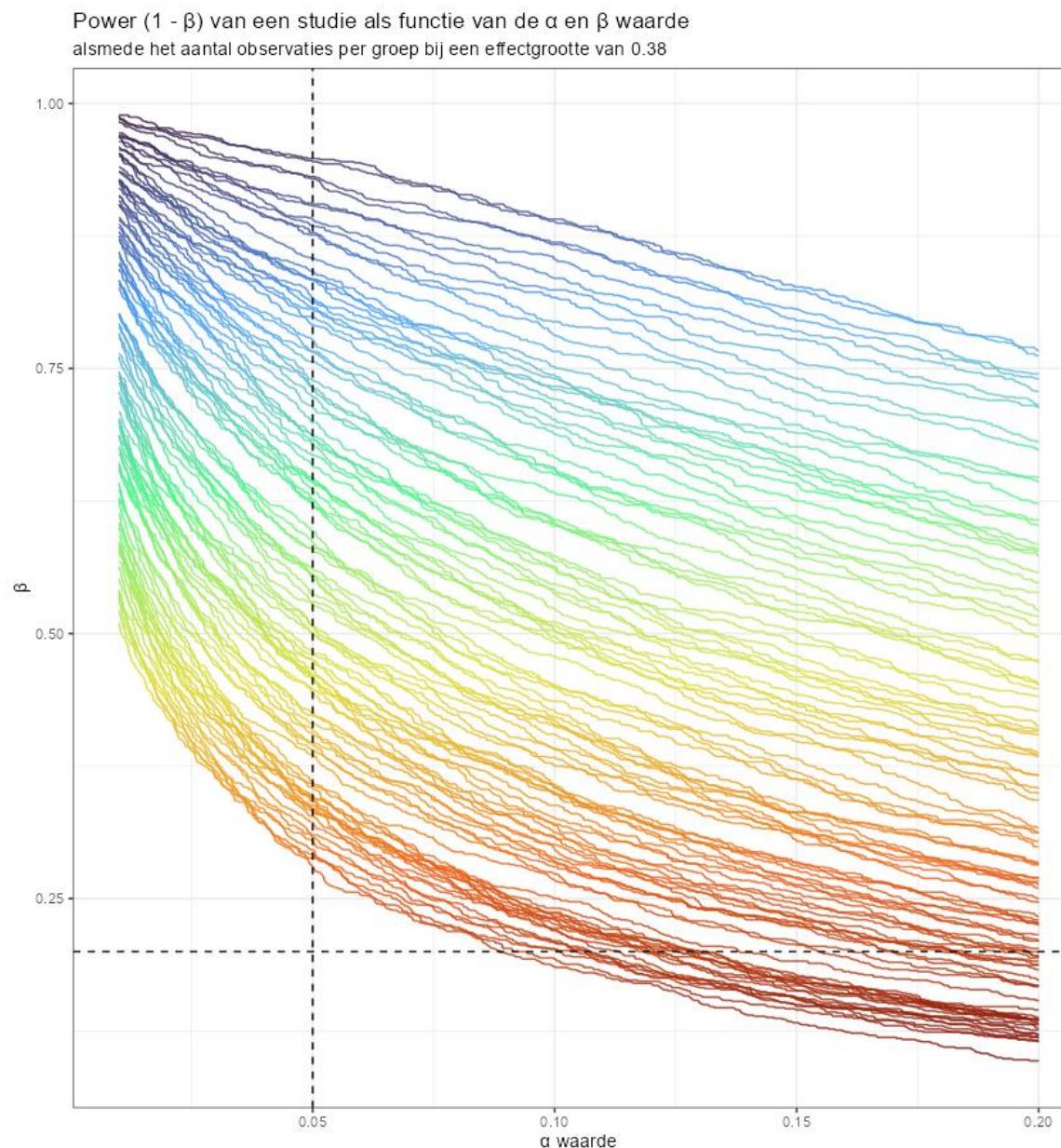
**Figuur 37.** De relatie tussen de  $\alpha$  waarde en de power ( $1 - \beta$ ) van een studie bij verschillende aantal observaties per groep. Omdat de effectgrootte zo groot is (2) maakt de keuze voor de  $\alpha$  waarde eigenlijk niet meer uit.

Wat we nog niet gezien hebben is het effect van de  $\beta$  waarde. Die tonen we nu in **Figuur 39**. In principe toont deze figuur hetzelfde als **Figuur 38**, maar dan met andere assen. In beide figuren valt te zien dat een steekproefgrootte van 200 personen niet genoeg is om een power van 80 % te krijgen bij een  $\alpha$  waarde van 0.05. De zwarte stippellijn in **Figuur 39** laat duidelijk zien dat geen lijn onder de kruising van de stippellijnen ligt. Wil men dus toch gaan rekenen met deze waarden, dan is het hopen op een toevalstreffer. Omdat we met een  $\alpha$  waarde werken valt nooit uit te sluiten dat dit niet kan. Wat we ook kunnen doen is de  $\alpha$

waarde veranderen. Laten we eens kijken wat er met ons betrouwbaarheidsinterval gebeurt als we de  $\alpha$  waarde aanpassen.



**Figuur 38.** De relatie tussen de  $\alpha$  waarde, het aantal observaties per groep en de power van de studie.



**Figuur 39.** De relatie tussen de  $\alpha$  en  $\beta$  waarde van een studie en de kracht om een effect te vinden.

Ter herinnering: het betrouwbaarheidsinterval bestaat uit twee getallen waartussen het zogenaamde ‘echte’ getal zich zogenaamd moet bevinden. In **Tabel 6** zagen we al hoe het betrouwbaarheidsinterval een functie is van de spreiding van de verdelingen. Nu wordt het tijd om te bezien wat er met dit interval gebeurt als we de  $\alpha$  waarden gaan aanpassen. In **Tabel 8** zien we twee betrouwbaarheidsintervallen: voor 95% ( $\alpha = 0.05$ ) en 99% ( $\alpha = 0.01$ ). We zien duidelijk hoe het kader rondom het ‘echte’ verschil steeds groter wordt. Schijnbaar is het zo dat wanneer onze observaties ver uit elkaar liggen (veel spreiding) én meer

zekerheid eisen het moeilijker wordt om definitief vast te stellen of de nulhypothese van ‘geen verschil’ wel behouden kan worden. We willen graag meer bewijs.

Spreiding (cm)		95% Betrouwbaarheidsinterval (van het verschil)		99% Betrouwbaarheidsinterval (van het verschil)	
Man	Vrouw	Links	Rechts	Links	Rechts
7	6.3	12.9	14	12.7	14.29
14	12.6	12.5	14.5	12.1	15.2
35	31.5	10.5	16.3	10.2	17.8
70	63	7.5	19.4	7	22.3
140	126	1.5	25.3	0.7	31.2

Tabel 8. Betrouwbaarheidsinterval als functie van de spreiding van twee groepen en de  $\alpha$  waarden.

Deze drang naar meer bewijs is goed en noodzakelijk. In de frequentistische statistiek leeft namelijk een groot probleem: de dekkingsgraad.

## De dekkingsraad

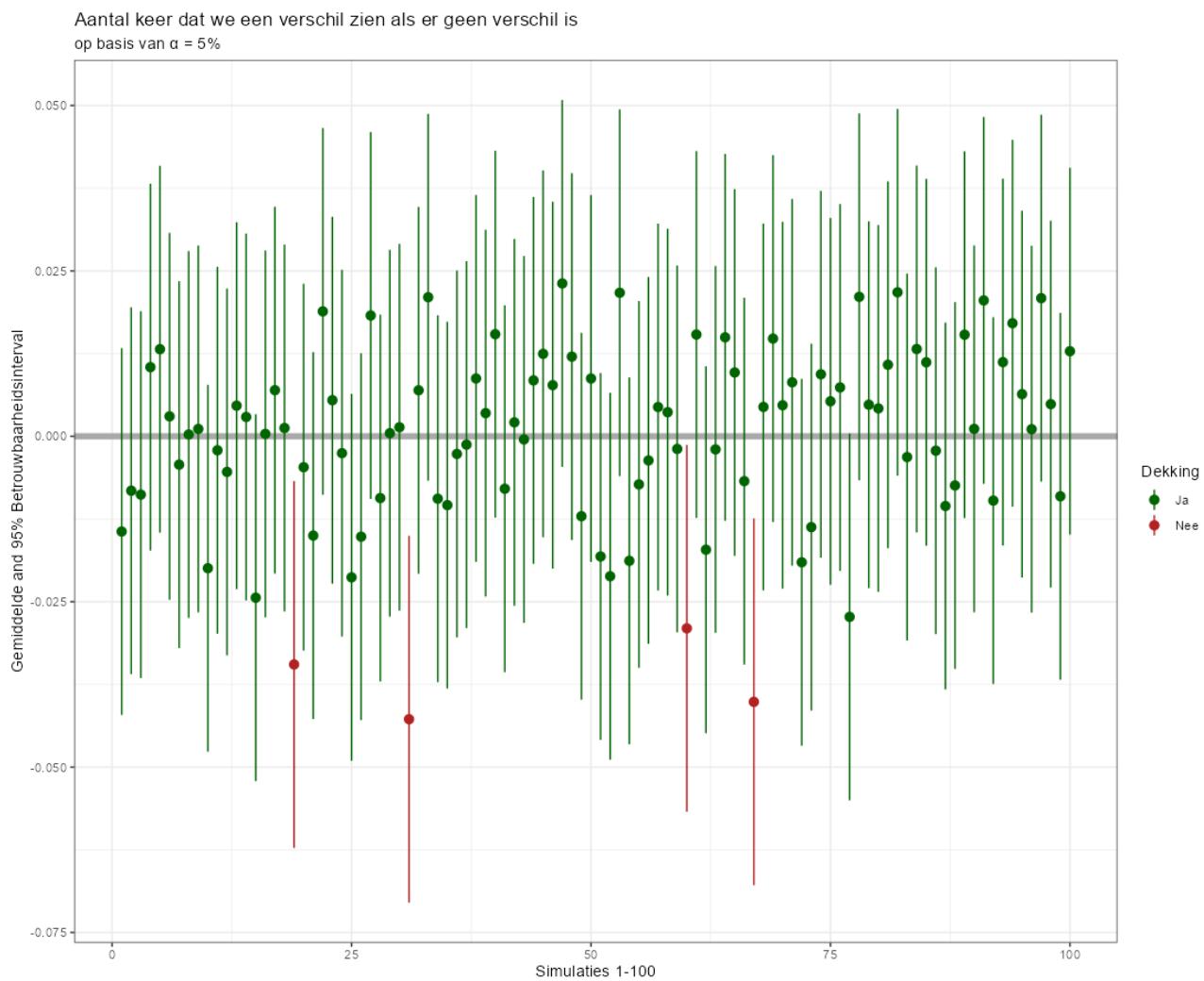
We hebben al genoemd dat de keuze voor een grenswaarde vaak makkelijk wordt genomen, maar niet zonder consequenties is. Kiezen we voor een  $\alpha$  waarde van 5% dan accepteren we dat we observaties als statistisch significant beoordelen ook al zijn ze dat niet (een vals positieve). Hoe dat er precies uitziet laat **Figuur 40** zien. Wat we hier zien is een simulatie van 100 steekproeven met waarden die komen uit een normaalverdeling met gemiddelde 0 en standaard deviatie 1<sup>45</sup>. Vervolgens ben ik gaan simuleren onder de  $\alpha$  van 5%. Wat ik zie is stiekem ook wat ik verwacht: in vier simulaties vind ik een statistisch significant verschil wordt gevonden<sup>46</sup>. Er vindt dan geen ‘dekking’ plaats. Zouden we maar één steekproef doen dan lopen we dus kans dat we een bevinding doen die niet ‘echt’ is. De dekkingsgraad is hier dus geen 100%, maar 96%<sup>47</sup>.

<sup>45</sup> Een gesimuleerd verschil van 0. Als ik dan toch een verschil zie dan kan dat eigenlijk niet kloppen.

<sup>46</sup> Dat het niet exact vijf is heeft te maken met de aard van de simulatie. Had ik 10,000 simulaties gemaakt dan was de kans groter geweest dat er 500 niet gedekt zouden worden. Met deze kleine aantallen kan het heel goed zijn dat een volgende reeks van 100 simulaties zes afwijkingen laat zien. In de echte wereld doen we vaak maar één enkele studie: dat zet tot denken.

<sup>47</sup> 96x geen verschil gevonden wanneer er ook geen verschil is. Wat dus afwijkt van de theoretische 95%.

Om wellicht wat meer gevoel te krijgen bij de dekkingsgraad kunnen we een voorbeeld met een munt laten zien. Nou is het zo dat eenieder die een munt ziet het gevoel heeft dat die munt zuiver is en daarmee 50% kans toekent aan 'kop' en 50% kans toekent aan 'munt'. Die 50/50 verdeling van 'kop' en 'munt' is het gevolg van de geometrie van de munt zelf. Daarom spreken we ook wel van een geometrische kans.

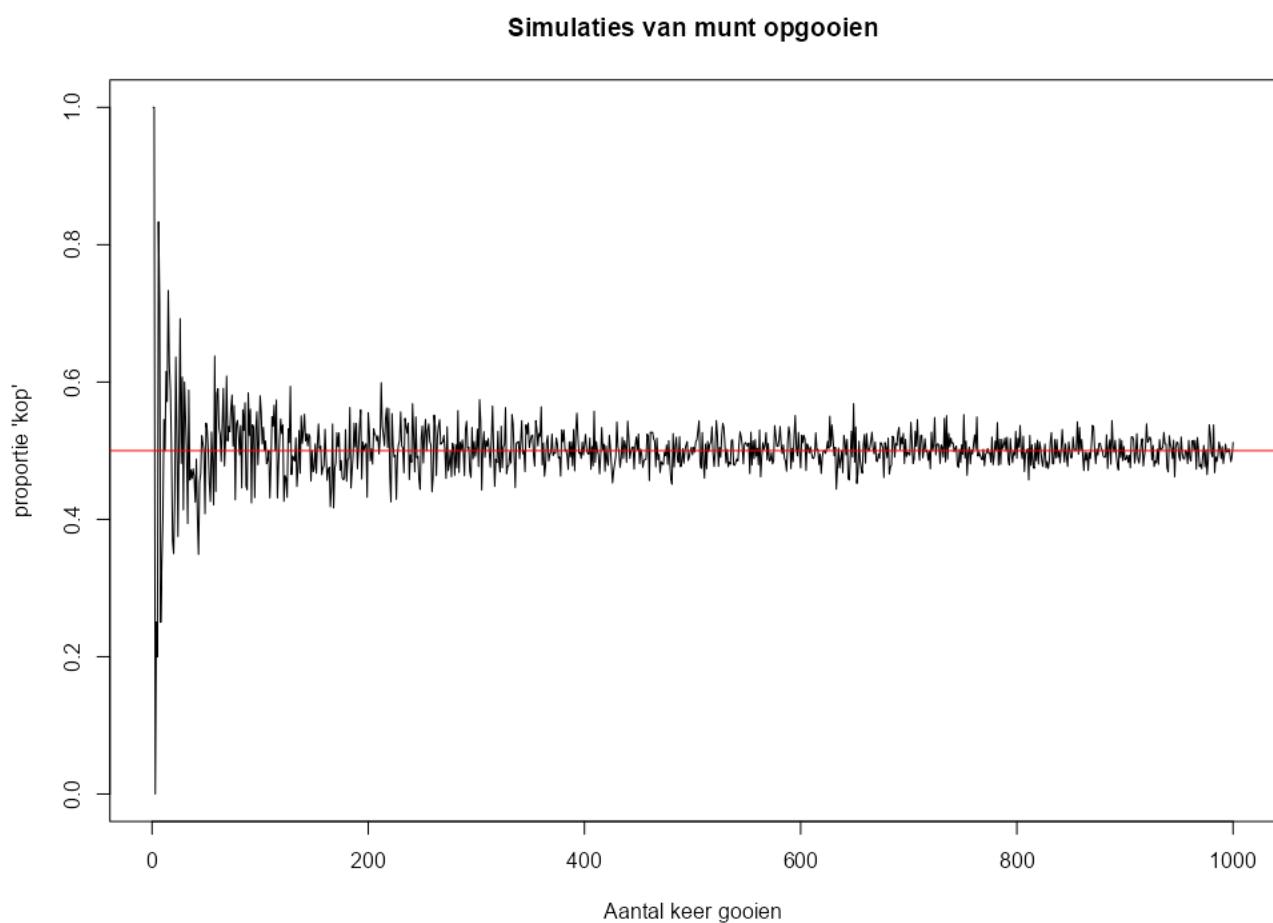


**Figuur 40.** Dekkingsgraad bij een  $\alpha$  van 5%. De simulatie laat zien dat we een dekkingsgraad van 96% hebben (4/100 simulaties is niet gedeekt).

Toch moeten we ook hier spreken van een model. Wie de tijd neemt om een aantal keer een munt op te gooien zal ontdekken dat deze 50/50 verdeling niet altijd opgaat. Soms is de ratio kop iets meer of iets minder dan 50% en juist in het begin wil dit nog wel eens extreem

schommelen<sup>48</sup>. We moeten dus een heel aantal keer een munt opgooien om daadwerkelijk de theoretische verdeling van ‘kop’ en ‘munt’ te vinden. Dit is ook een dekkingsgraad, want de dekkingsgraad is hier het aantal keren munt opgooien dat we moeten voltooien om ook daadwerkelijk de beoogde theorie van 50/50 te ‘zien’. **Figuur 41** laat dit duidelijk zien met de rode horizontale streep die synoniem staat voor een gelijke kans.

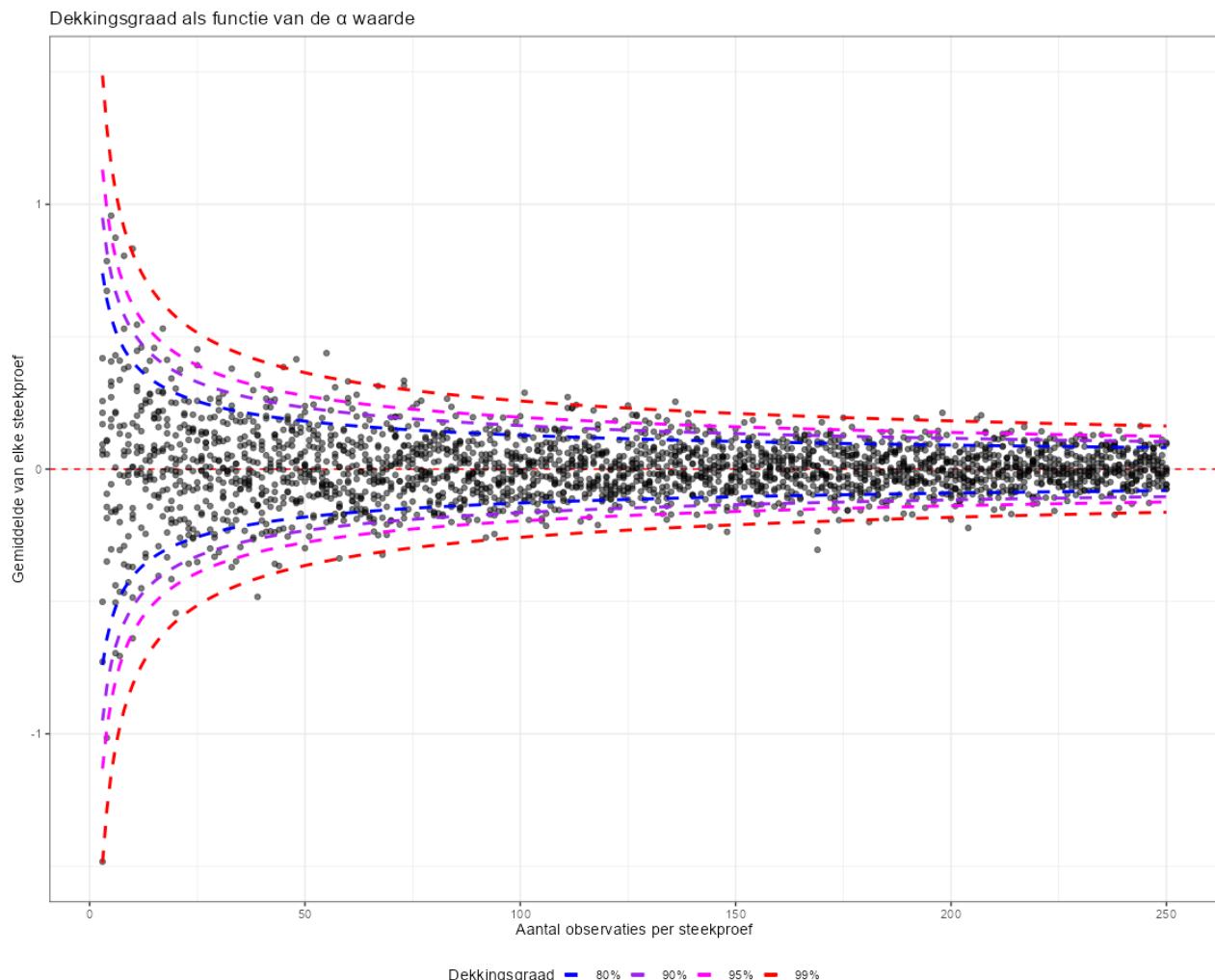
Wat belangrijk is om te onthouden is dat het 95% betrouwbaarheidsinterval OOK een model is. Dat is wat de dekkingsgraad laat zien en dit speelt voornamelijk een rol wanneer we meerdere studies doen. Maar ook als we maar één enkele studie doen speelt het een rol, want veel van de berekeningen uit de frequentistische statistiek is een combinatie van observaties met theoretische aannames. Die aannames kunnen we niet zomaar negeren.



**Figuur 41.** De dekkingsgraad van een simulatie van munt opgooien. Hoewel de theoretische verdeling 50/50 is duurt het een tijd voordat de simulatie ook in de buurt komt. Dit laat zien dat 50/50 een model is en niet per se de werkelijkheid representeren. Uiteindelijk, zo laat de theorie ook zien, zal een munt bij oneindig veel keer munt opgooien de beoogde 50/50 zien. Tenzij de munt niet zuiver is.

<sup>48</sup> Een kans uitrekenen op een kleine deler maakt dat elke keer munt opgooien een groter percentage beslaat.

Een andere manier om de invloed van de dekkingsgraad te laten zien<sup>49</sup> is door verschillende dekkingsgraden over elkaar heen te leggen in een simulatie waarin ik weet dat de ‘echte’ verdeling een gemiddelde van 0 heeft met een standaard deviatie van 1. **Figuur 42** laat zien dat de dekkingsgraad een functie is van het aantal observaties per steekproef en de beoogde graad zelf.



**Figuur 42.** De dekkingsgraad van het betrouwbaarheidsinterval bij 80, 90, 95 en 99%. Wat opvalt is dat deze dekkingsgraad ook afhankelijk is van de steekproefgrootte. Voor elke steekproefgrootte hebben we 10 herhaalde steekproeven genomen.

Wat we nog niet direct zien is dat de dekkingsgraad ook een functie is van het aantal herhaalde steekproeven. Dit zagen we deels al in **Figuur 41** omdat we hier het percentage kop/munt cumulatief verwerkt hebben. Dat gebeurt in **Figuur 42** niet dus daarom laat ik het zien in **Tabel 9**. Wat ik hier heb gedaan is het simuleren van een reeks observaties met

<sup>49</sup> Die dus een functie is van het aantal herhaalde onderzoeken én de  $\alpha$  waarde

gemiddelde 0 en standaard deviatie 1. Vervolgens ben ik het aantal observaties in één enkele steekproef én het aantal steekproeven gaan vergroten. Wat duidelijk te zien is, is dat de gesimuleerde dekkingsgraad de theoretische dekkingsgraad benaderd als we een groot aantal steekproeven doen. Helaas hebben we maar zelden zoveel observaties en dus komen veel van deze oefeningen voort uit simulaties en niet uit observaties uit de ‘echte’ wereld.

Samenvatten is het dus van belang om waakzaam te zijn in het beoordelen van de nauwkeurigheid van het 95% betrouwbaarheidsinterval. Deze geeft past echt een dekking van 95% bij een groot aantal herhalingen<sup>50</sup>. Vaak hebben we die herhalingen niet tot onze beschikking en wat rest is het ‘vertrouwen’ op de theoretische dekkingsgraad. Deze theoretische dekkingsgraad is echter een model op zichzelf.

<b><math>\alpha</math></b>	<b>Aantal observaties in één steekproef</b>	<b>Aantal herhaalde steekproeven</b>	<b>Theoretische dekkingsraad</b>	<b>Gesimuleerde dekkingsgraad</b>
0.05	1	1	0.95	Niet mogelijk
0.05	10	1	0.95	1
0.05	10	10	0.95	0.9
0.05	100	10	0.95	1
0.05	10	100	0.95	0.97
0.05	100	100	0.95	0.93
0.05	1,000	10	0.95	1
0.05	10	1,000	0.95	0.946
0.05	1,000	1,000	0.95	0.942
0.05	10,000	10	0.95	1
0.05	10	10,000	0.95	0.9505
0.05	10,000	10,000	0.95	0.9478
0.05	10	100,000	0.95	0.9503
0.05	10	1,000,000	0.95	0.9497

**Tabel 9.** Gesimuleerde dekkingsgraad als functie van het aantal observaties in één steekproef en aantal herhaalde steekproeven.

## Gluren

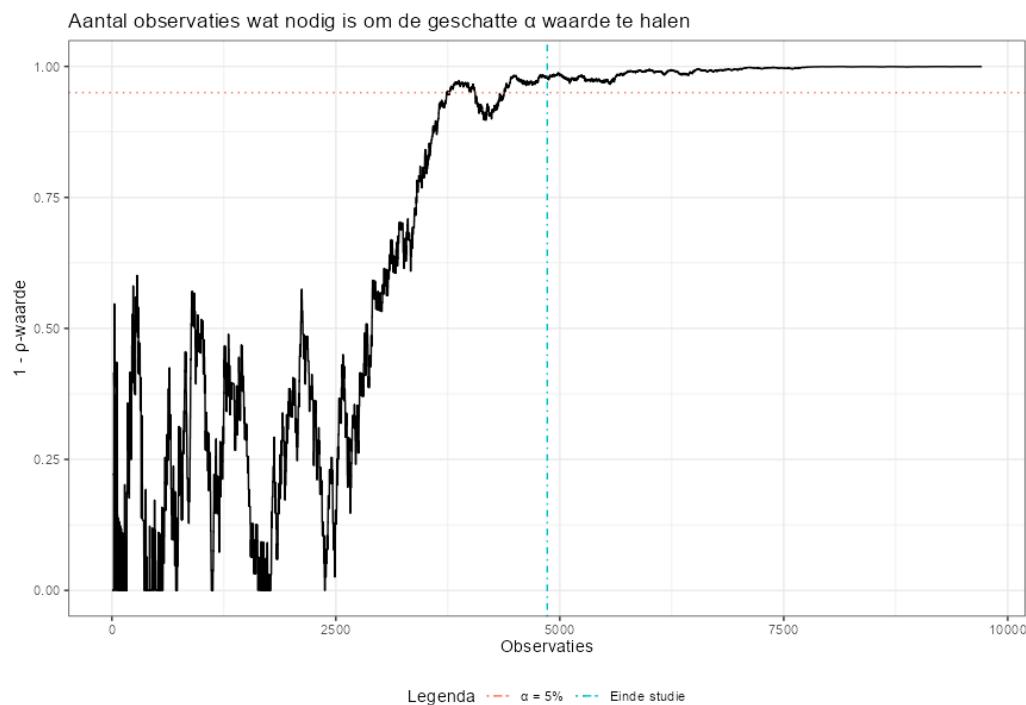
Veel van de hierboven genoemde voorbeelden gaat uit van theoretische kennis en niet geobserveerde ‘werkelijkheid’. Vaak weten we dus helemaal niet wat de ‘echte’ waarde is<sup>51</sup>.

<sup>50</sup> Een oneindig aantal herhalingen. Dit is namelijk de theorie achter de frequentistische statistiek.

<sup>51</sup> In veel gevallen simuleren we met theoretische verschillen en waardes: we doen dat zodat we kunnen zien of de simulatie toont wat we verachten te zien op basis van de assumpties van de frequentistische statistiek en dus ook van de

We observeren juist zaken om een schatting te kunnen maken van dat 'echte' getal zou kunnen spreken, als we al kunnen spreken van een 'echt' getal. Verder wordt in veel gevallen de data lukraak verzameld, zonder dat er een goede berekening vooraf is gedaan om een inschatting te krijgen van de benodigde steekproefgrootte. Dit zijn allemaal zaken die een rol spelen in het per ongeluk toe kennen van een statistisch significant effect dat er wellicht helemaal niet is. Als we namelijk de data analyseren wanneer deze binnenkomt, en dat wellicht meerdere keren achter elkaar doen, dan kan het heel goed zijn dat we een vals positieve toe kennen. Dit wordt ook wel 'gluren' genoemd. Laten we dit 'gluren' eens uitwerken met een voorbeeld.

Laten we beginnen met een voorbeeld waarin we uitrekenen hoeveel observaties we nodig hebben om een verschil van 2,5% te vinden in proporties. Bij een  $\alpha$  van 5% en een  $\beta$  van 20% is dat 4860 observaties per groep. Vervolgens simuleren we het steeds groter worden van de groepsgrootte om te zien of we inderdaad het beoogde verschil zien. Het resultaat van deze simulatie zien we in **Figuur 43**.



**Figuur 43.** Voorbeeld van gluren. We zien hier het aantal observaties dat nodig is om de geschatte waarde  $\alpha$  te halen.

---

statistische testen die we gebruiken. In het echt hebben we vaak geen flauw benul wat de 'echte' waarde is en of we überhaupt kunnen spreken van een 'echte' waarde: daarover verschillen de stromingen van de statistiek ook nog fundamenteel. Dat is nu niet het thema: belangrijker is het om te laten zien wat er nodig is om de theorie vanuit de statistiek ook daadwerkelijk zichtbaar te maken.

Maar dit is niets anders dan het testen van één enkele studie waarbij we de steekproefgrootte steeds groter maken. Het is een bevestiging van de eerdere berekening hoeveel observaties we nodig hebben om te zien wat we denken te gaan zien. We zullen nu dit proces moeten gaan herhalen om te zien wat we krijgen wat we willen krijgen. Als we dit 1000x doen dan zien we inderdaad de  $\alpha$  en  $\beta$  waarde terug zoals beoogt. Dit is eigenlijk helemaal niet zo gek, want we hebben van tevoren bepaald hoeveel observaties we nodig hebben om een beoogd verschil te zien. We hebben toen dat verschil gesimuleerd en zien terug wat we ingevoerd hebben. Eigenlijk hebben we dus niets anders gedaan dan gekeken of de statistische formules wel doen wat ze beogen te doen.

Maar als we nou niet wachten tot het einde van de proef, of niet eens een einde hebben omdat we nooit de berekening hebben gedaan op basis van de geschatte steekproefgrootte. Als we nu continue gaan kijken wanneer we een significante  $p$  waarde krijgen dan zien we dat we uiteindelijk een  $\alpha$  waarde van rond de 30% krijgen! Dit terwijl de originele condities waarmee de data zijn verzameld hetzelfde is gebleven. Eigenlijk zou die 30% niet mogen verbazen: in **Figuur 43** valt duidelijk te zien dat de  $\alpha$  waarde heel lang rond dit getal blijft zweven. Het is belangrijk om tot je te nemen dat grenswaarden niet vast staan: ze zijn afhankelijk van context.

## Wat kunnen we hier nu uit afleiden?

We hebben in dit hoofdstuk een heleboel zaken de revue laten passeren. Zo zijn we begonnen met het benoemen van de grenswaarden die bepalen wanneer een bevinding als ‘statistisch significant’ wordt bestempeld of niet. We hebben laten zien dat deze grenswaarde vaak gebruikt wordt in relatie tot hypothese toetsen: wanneer een bevinding de grenswaarde overschrijdt wordt de nulhypothese verworpen ten gunste van de alternatieve hypothese. Hiervoor zijn verschillende testen in het leven geroepen zoals de one-sample t-test (het toetsen of een steekproef afwijkend is van het theoretisch gemiddelde) of een independent samples t-test (zijn groepen daadwerkelijk verschillend?). Uiteindelijk hebben we gezien dat het betrouwbaarheidsinterval meer informatie biedt dan alleen een uitdrukking van de kans dat een bevinding op basis ‘van kans’ is geobserveerd.

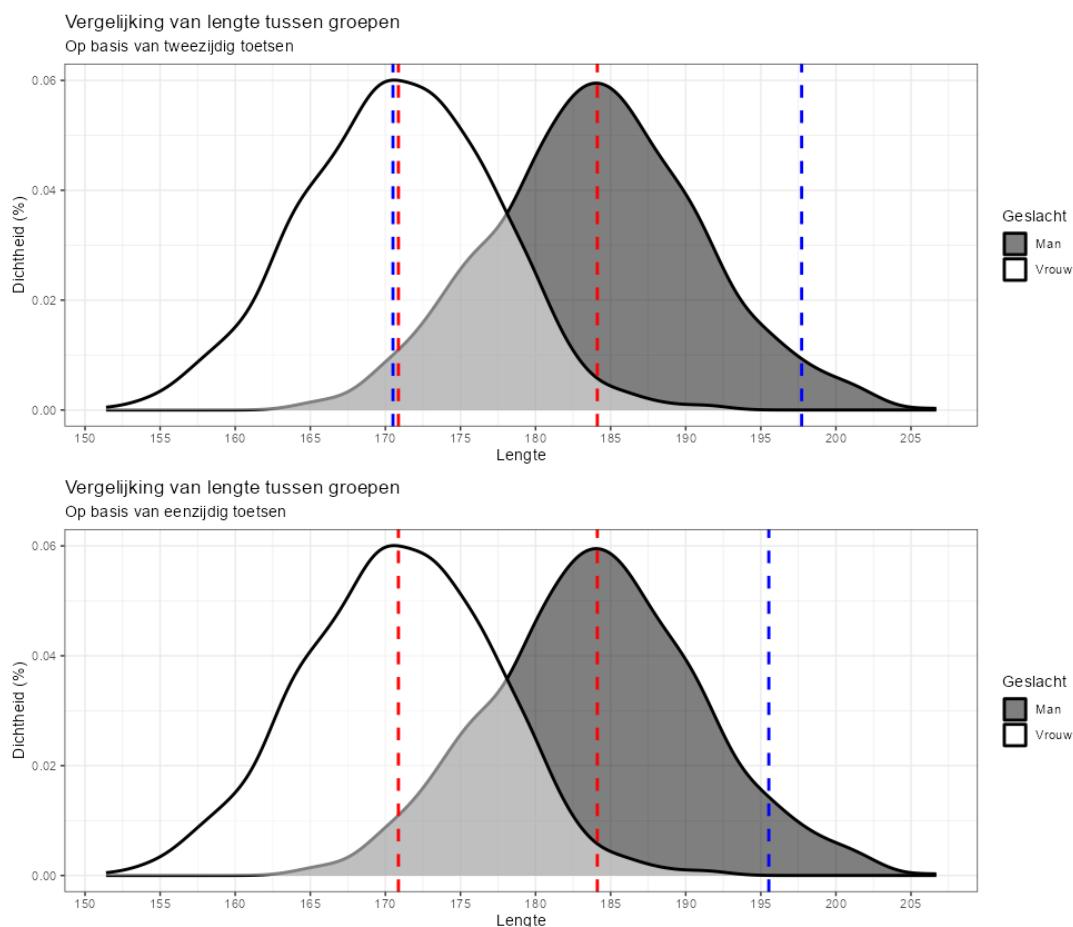
Al deze exercices zijn echter geen exacte wetenschap en elke stap in het proces van toetsen of de nulhypothese nog gangbaar is berust op aannames. Zo hebben we de  $\alpha$  en  $\beta$

waarde. Verder van invloed is de steekproefgrootte, de spreiding rondom het gemiddelde en het aantal herhalingen van een steekproef. De effectgrootte tussen twee groepen is essentieel, maar de test of de nulhypothese kan worden vervangen door een alternatieve hypothese berust zelf op een model. De dekkingsgraad van deze modellen is erg slecht in kleine groepen en vooral bij weinig herhalingen.

Ondanks deze kritiek, en deze kritiek is al jaren bekend, wordt de frequentistische statistiek nog steeds veelvuldig toegepast. Dat kan en mag, mits die beperkingen worden begrepen en er rekening mee wordt gehouden. Het gemak waarmee in de BNNVARA/Zembla rapportage wordt gerept van het vervangen van een tweezijdige toets door een eenzijdige toets lijkt niet aan te tonen dat er een volledig begrip is van die begrepen. Om dit toch enigszins aantoonbaar te maken volgt nu het laatste hoofdstuk zonder glyfosaat gegevens: de uitleg waarom de keuze voor tweezijdig en eenzijdig toetsen helemaal niet zo makkelijk inwisselbaar is.

## Eenzijdig en tweezijdig toetsen

Om de voorbeelden hier wat meer tot de verbeelding te laten spreken kunnen we gebruik maken van de vorige voorbeelden, maar wel met een kleine aanpassing. We hebben namelijk gezien dat de verschillen tussen mannen en vrouwen zo groot zijn dat dit in de meeste omstandigheden tot ‘statistisch significant’ wordt gerekend<sup>52</sup>. We kunnen dit ook grafisch tonen (**Figuur 44**). Wat deze figuur laat zien zijn de voor ons bekende verdelingen van mannen en vrouwen. De rode stippellijnen vormen het gemiddelde voor elke groep. De blauwe lijnen zijn de grenswaarden bij een  $\alpha$  van 5%. Wat er nu gebeurd is dat wanneer wij van eenzijdig naar tweezijdig toetsen gaan de rechter blauwe lijn verschuift. De grenswaarde wordt minder streng: van 1.96 naar 1.64<sup>53</sup>. Dit komt omdat de we de  $\alpha$  niet meer door twee delen.



**Figuur 44.** Grenswaarde (blauwe stippellijn) als functie van tweezijdig en eenzijdig toetsen.

<sup>52</sup> Wat dan weer vaak synoniem staat voor een ‘echt’ verschil.

<sup>53</sup> 1.96 en 1.64 zijn de grenswaarden voor de t-verdeling bij 999 vrijheidsgraden.

De vraag die we ons nu direct moeten stellen is of dit wel gerechtvaardigd is. Door het aanpassen van de toets is de kans op een vals positieve ook direct gestegen: we hebben namelijk het gebied waarbuiten we iets statistisch significant vinden vergroot. We kunnen dit laten zien door een tabel te maken waaruit blijkt dat de grenswaarde naar binnen is geschoven (**Tabel 10**):

Nulhypothese	Alternatieve hypothese	$\alpha$ -waarde	95% betrouwbaarheidsinterval
Geen verschil	Wel een verschil	0.05	3.13 - 26.17
Geen verschil	Mannen < vrouwen	0.05	$\infty$ - 24.32
Geen verschil	Mannen > vrouwen	0.05	4.99 - $\infty$

**Tabel 10.** De uitkomsten met dezelfde data maar met drie verschillende alternatieve hypotheses: wel een verschil, mannen kleiner dan vrouwen en mannen groter dan vrouwen. Het betrouwbaarheidsinterval is altijd oneindig ( $\infty$ ) aan de andere kant waarvoor getoetst wordt.

Wat zou er nu gebeuren als we eenzijdig toetsen, maar de  $\alpha$  waarde bij het eenzijdig toetsen toch door twee delen? Zoals we in **Tabel 11** kunnen zien verandert dan het betrouwbaarheidsinterval. Deze staan nu gelijk aan het tweezijdig toetsen bij  $\alpha / 2$ :

Nulhypothese	Alternatieve hypothese	$\alpha$ -waarde	betrouwbaarheidsinterval
Geen verschil	Wel een verschil	0.05	3.32 - 26.17
Geen verschil	Mannen < vrouwen	0.025	$\infty$ - 26.17
Geen verschil	Mannen > vrouwen	0.025	3.13 - $\infty$

**Tabel 11.** Betrouwbaarheidsinterval bij eenzijdig toetsen wanneer de  $\alpha$  waarde handmatig wordt aangepast naar  $0.05 / 2$ . De p-waarde blijft hetzelfde, maar het betrouwbaarheidsinterval schikt zich naar de tweezijdige toets.

De keuze om een of tweezijdig te toetsen is dus niet zonder gevolgen, maar nergens vinden we een rationale om te kiezen voor een bepaalde grenswaarde. De keuze om een tweezijdige toets om te zetten in een éénzijdige toets gebeurt zelden en als het al gebeurt dan is dat alleen wanneer de onderzoeker op voorhand weet dat het verschil tussen twee groepen maar één kant op kan gaan. Daarmee zet een onderzoeker zich ook vast: als het verschil een niet verwachte kant op gaat dan is de p-waarde per definitie groter dan 0.05. Dit is een van de redenen waarom veel onderzoekers kiezen voor de tweezijdige toets.

Laten we eens zien wat er gebeurt als we de  $\alpha$  waarde aanpassen. We zien het resultaat in **Tabel 12** wat laat zien dat ook bij een verkleining van de  $\alpha$  (van 5% naar 1%) een

tweezijdige toets niet significant is, maar een éénzijdige toets wel. Dit alles benadrukt de kritiek van BNNVARA/Zembla: het statistisch significant maken van observaties is soms maar verwijderd van één enkele keuze.

Nulhypothese	Alternatieve hypothese	$\alpha$ -waarde	betrouwbaarheidsinterval
Geen verschil	Wel een verschil	0.05	3.13 - 26.17
Geen verschil	Mannen < vrouwen	0.05	$\infty$ - 24.32
Geen verschil	Mannen > vrouwen	0.05	4.99 - $\infty$
Geen verschil	Wel een verschil	0.01	-0.49 – 29.79
Geen verschil	Mannen < vrouwen	0.01	$\infty$ - 28.32
Geen verschil	Mannen > vrouwen	0.01	0.97 - $\infty$

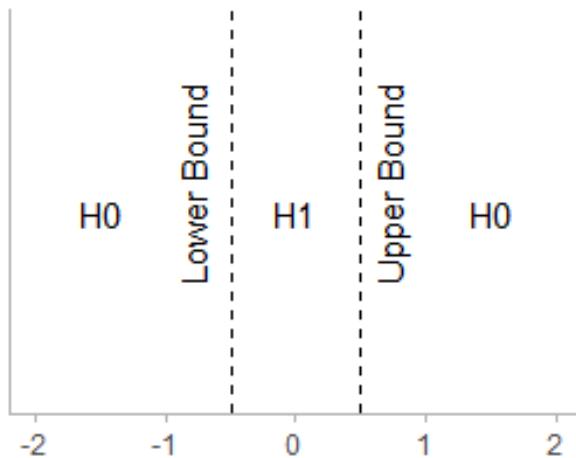
Tabel 12. Betrouwbaarheidsinterval als een functie van de  $\alpha$  waarde en de richting van de toets.

Eigenlijk kan ik niet veel meer toevoegen aan bovenstaande. We hebben gezien dat de dekkingsgraad niet optimaal is bij kleine steekproeven en met weinig herhalingen, dus dat zal hier niet anders zijn. Want we wellicht wel kunnen doen is anders kijken naar de manier van eenzijdig of tweezijdig toetsen en dat heeft dan weer alles te maken met onze nulhypothese. We gaan er namelijk nu steeds vanuit dat de nulhypothese de stelling draagt dat er ‘geen verschil’ is. Maar deze stelling maakt direct dat de alternatieve hypothese wel een verschil moet zijn, waardoor de gehele toetsing een binaire natuur krijgt: niet óf wel.

Deze zwart-wit blik op verschillen is vaak niet geheel in lijn met wat we in de natuur vinden waarbij het heel goed kan zijn dat er wel verschillen zijn tussen groepen, maar we deze verschillen niet interessant genoeg vinden. Zo zal een gemiddeld verschil van 1 cm tussen mannen en vrouwen wel als verschil bestempeld worden, maar nauwelijks zichtbaar zijn. We missen dus de klinische relevantie van het verschil. Wie dit wil introduceren kan gaan werken met zogenaamde equivalentie testen<sup>54</sup>: de nulhypothese zegt nu dat het verschil groter is dan een vooraf bepaalde klinische grens (**Figuur 45**). De alternatieve hypothese stelt daarmee dat dit verschil niet zo is, en de manier waarop er nu getoetst wordt is door twee eenzijdige toetsen te combineren. Het grootste verschil is trouwens niet die combinatie van toetsen, maar het omdraaien van de nulhypothese: niet langer is de alternatieve hypothese dat er wél een verschil is. Nee, de alternatieve hypothese is nu dat dat het verschil kleiner is dan de klinische grens en grafisch ziet dit er zo uit (**Figuur 46**).

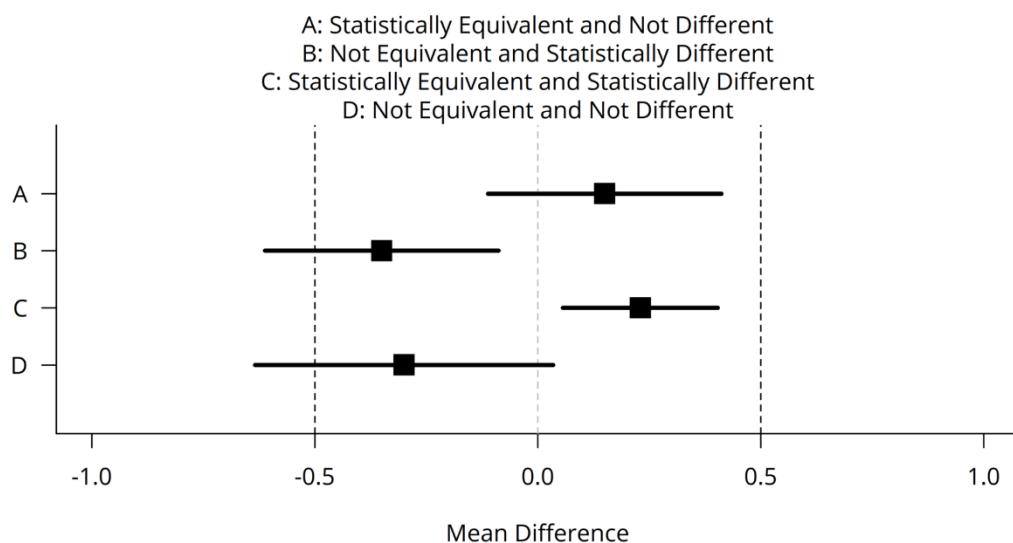
<sup>54</sup> [https://en.wikipedia.org/wiki/Equivalence\\_test](https://en.wikipedia.org/wiki/Equivalence_test)

## Equivalence Test



H1 = Alternative Hypothesis  
H0 = Null Hypothesis

**Figuur 45.** De alternatieve hypothese is bij tweemaal eenzijdig toetsen er een van 'geen verschil' waarbij 'geen verschil' betekent dat er een verschil is wat binnen de klinische grenzen valt.



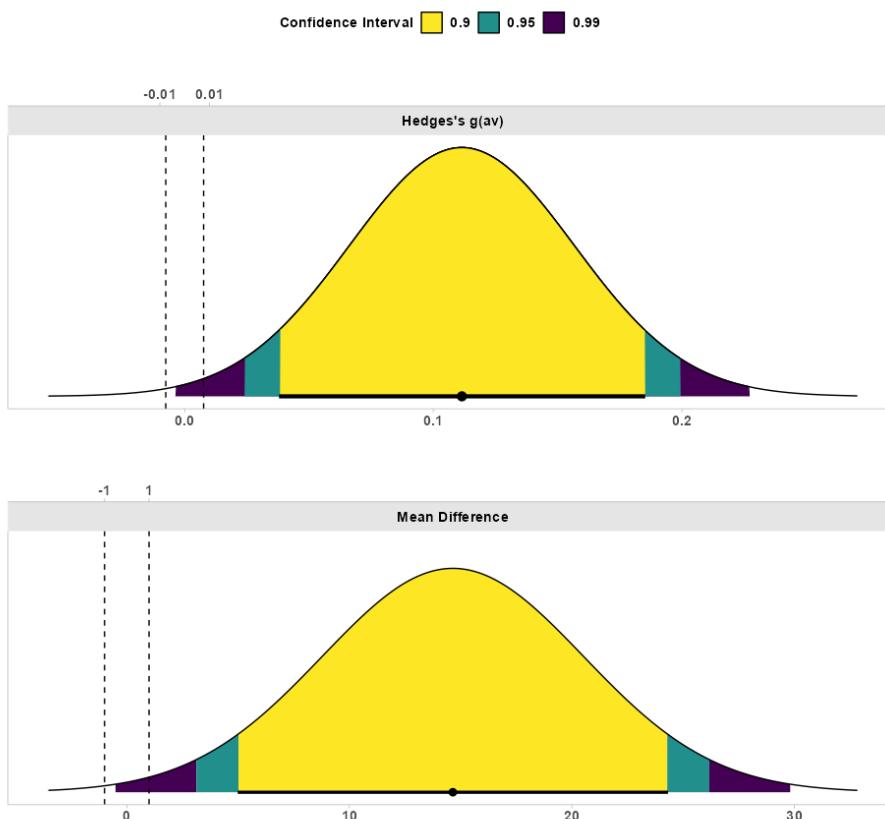
**Figuur 46.** Verschillen vormen van testen op basis van het betrouwbaarheidsinterval van een verschil. De grenswaarde is nu niet alleen de grens van 'geen verschil', maar ook de grens die gezet wordt op -0.5 en 0.5. Het wordt er niet makkelijker op zo. ([https://en.wikipedia.org/wiki/Equivalence\\_test#/media/File:Equivalence\\_Test.png](https://en.wikipedia.org/wiki/Equivalence_test#/media/File:Equivalence_Test.png))

Laten we hier dieper op ingaan door wederom gebruik te maken van de lengtedata zoals we deze kennen.

## Tweemaal eenzijdig testen (TOST)

Het toetsen op equivalentie is één manier om te tonen zien dat het niet kunnen vervangen van de nulhypothese nog niet hoeft te betekent dat er geen noemenswaardig verschil is. Aan de andere kant wordt er nu ook de andere kant op getest waardoor het bewijs moet komen dat de stelling dat er één verschil is kan worden verworpen (**Figuur 45**).

Het beste wat we nu kunnen doen is een voorbeeld nemen. Nu hebben we al veelvuldig gezien dat het verschil tussen mannen en vrouwen gemiddeld 13,4 cm is. Maar wat als we die kennis niet hebben en zeggen dat we de klinische grens op 1 cm leggen. Dan krijgen we als resultaat **Figuur 47** en wat deze figuur laat zien is het gemiddeld verschil ('mean difference') tussen de groepen. Ook zien we de betrouwbaarheidsintervallen. Alleen het 99<sup>ste</sup> betrouwbaarheidsinterval gaat over de grens van 0 (zagen we al in **Tabel 12**) én de grens van 1. Dat bekent dat voor het 99<sup>ste</sup> percentiel het verschil niet significant is én niet equivalent. De linker staart bevindt zich namelijk tussen de grens van -1 en 1.



**Figuur 47.** Verschil tussen mannen en vrouwen in lengte waarbij de klinische grens op -1 én 1 is gezet. Het resultaat in de onderste grafiek laat zien dat het 99% betrouwbaarheidsinterval zowel de 1 als de 0 overschrijdt. Daarmee is het verschil tussen mannen en vrouwen niet statistisch significant én niet statistisch equivalent. We kunnen dus niet zomaar de nulhypothese van 'geen verschil' aannemen ook al is het verschil statistisch significant.

## Het probleem van meerdere testen

We hebben tot nu benoemd welke problemen we kunnen ervaren met het zoeken naar statistisch significantie. We hebben vooral gezien dat er ook een heleboel aannames worden gemaakt die we niet goed kunnen staven. Ook hebben we gezien dat één enkele toets vaak geen goede dekkingsgraad heeft: het is de herhaling die belangrijk is.

Het laatste probleem wat ik hier wil introduceren lijkt wat paradoxaal gegeven mijn eerdere opmerking dat herhalingen belangrijk zijn (**De dekkingsraad**). In het deel over de dekkingsgraad ging het erom of de  $\alpha$  waarde die we toepassen in een statistische test ook wel zichtbaar is: het is de proportie gevallen waarin de test een vals positieve laat zien. Wat we zagen was dat de  $\alpha$  waarde een theoretische waarde is die alleen behaald wordt wanneer we heel veel herhalingen doen van exact hetzelfde. De  $\alpha$  waarde is daarmee vooral een theoretische waarde.

Ik wil het in deel hebben over iets anders, namelijk het probleem van meervoudig testen (of toetsen). Vaak is het zo dat in een studie meerdere testen worden gedaan om te bezien of een waarde ‘statistisch significant’ is. Zo kan het goed zijn dat er in een studie niet één test maar 20 testen worden gedaan. Zo kan het zijn dat wanneer we groepen met elkaar willen vergelijken we niet alleen de groepen vergelijken, maar ook de groepen op verschillende momenten in de tijd. Of onder andere omstandigheden. Of misschien willen we groepen wel vergelijken op basis van verschillende uitkomsten. Zo kan één test als snel oplopen tot 20 verschillende testen waarbij exact dezelfde data wordt gebruikt. Dit is anders dan in het voorbeeld van de dekkingsgraad waarin we simulaties doen onder dezelfde theoretische omstandigheden (maar wel met andere data), en zien dat die omstandigheden ook pas echt behaald worden bij grote herhalingen: de aannames van de frequentistische statistiek vallen pas bij veel herhalingen op hun plek. In dit geval is het doen van een groot aantal herhalingen problematisch. Zo zou je het mogen vergelijken met het kopen van niet één lot, maar 100 loten in de loterij: de kans om iets te winnen neemt toe.

We hebben al eerder benoemd dat wij bij een  $\alpha$  waarde van 0.05 verwachten dat 1 van 20 testen statistisch significant is terwijl deze eigenlijk niet statistisch significant is. Alleen, dit principe is niet van toepassing op **meerdere** testen<sup>55</sup>. Wat blijkt namelijk: de  $\alpha$

---

<sup>55</sup> [https://en.wikipedia.org/wiki/Multiple\\_comparisons\\_problem](https://en.wikipedia.org/wiki/Multiple_comparisons_problem)

waarde van 5% is niet meer 5% per test als we meerdere testen doen. Net als de kans om te winnen bij één lot niet gelijk is aan de kans bij het kopen van meerdere loten<sup>56</sup>

Wellicht dat de lezer het nu even niet meer snapt. Laten we daarom visualiseren wat er gebeurt met de grenswaarde van 5% als we meerdere testen doen (op dezelfde data). Om dat enigszins behapbaar te doen gaan we met drie gesimuleerde groepen werken: drie groepen maakt dat we zes verschillende vergelijken kunnen maken ( $3!$  oftewel  $3 \times 2 \times 1$ ). Omdat we niet een derde geslacht kunnen toevoegen, is het wellicht handig om nu met leeftijden de werken over tijd. Op de website van het CBS zien we de gemiddelde lengte van mannen en vrouwen over tijd<sup>57</sup>. Getabuleerd ziet dit er als volgt uit (Tabel 13):

<b>Geslacht</b>	<b>Jaar</b>			
	<b>1930</b>	<b>1960</b>	<b>1980</b>	<b>2001</b>
<i>Man</i>	175.6 (6.67)	181.7 (6.90)	183.9 (6.99)	182.9 (6.95)
<i>Vrouw</i>	165.4 (6.12)	168.5 (6.23)	170.7 (6.32)	169.3 (6.26)

Tabel 13. De gemiddelde (en de standaard deviatie) lengte van mannen en vrouwen voor 1930, 1960, 1980 en 2001.

Om de groepen met elkaar te vergelijken hebben we ook een standaard deviatie nodig. Die wordt niet gegeven, maar kunnen we wellicht afleiden uit de data zoals eerder gebruikt waarbij we de ratio van de standaard deviatie ten op zichtte van het gemiddelde toepassen. Bij mannen is dit  $7/184 = 0.038$  en bij vrouwen is dat  $6.3 / 170.6 = 0.037$ . Ongeveer hetzelfde. Deze fracties gebruiken we om de 8 groepen te maken (Tabel 13) én te visualiseren. We kunnen dus nu een heleboel groepen met elkaar vergelijken.

Bij zoveel vergelijken kunnen we al direct uitrekenen dat de  $\alpha$  waarde niet meer 5% kan zijn. Dit is namelijk de  $\alpha$  waarde bij één enkele vergelijking (en we hebben gezien dat de steekproefgrootte genoeg moet zijn om die waarde te halen). Maar als we dus uit diezelfde steekproef meerdere testen doen, dan is de  $\alpha$  waarde niet meer 5%. Deze is nu afhankelijk van het aantal testen  $\tau$ <sup>58</sup>. Bij acht vergelijkingen zou de  $\alpha$  niet meer 5% zijn, maar 33.6% wat betekent dat de theoretische kans op een vals-positieve gestegen is van 5% naar 33%. In Tabel 14 kunnen we laten zien hoe snel de kans op een vals positieve toeneemt als het aantal testen (op dezelfde data!) toeneemt. We zien ook deze toename groter is bij een

<sup>56</sup> Hoewel bij meerdere loten kopen de kans nog steeds extreem klein blijft.

<sup>57</sup> <https://longreads.cbs.nl/nederland-in-cijfers-2022/hoe-lang-zijn-nederlanders/>

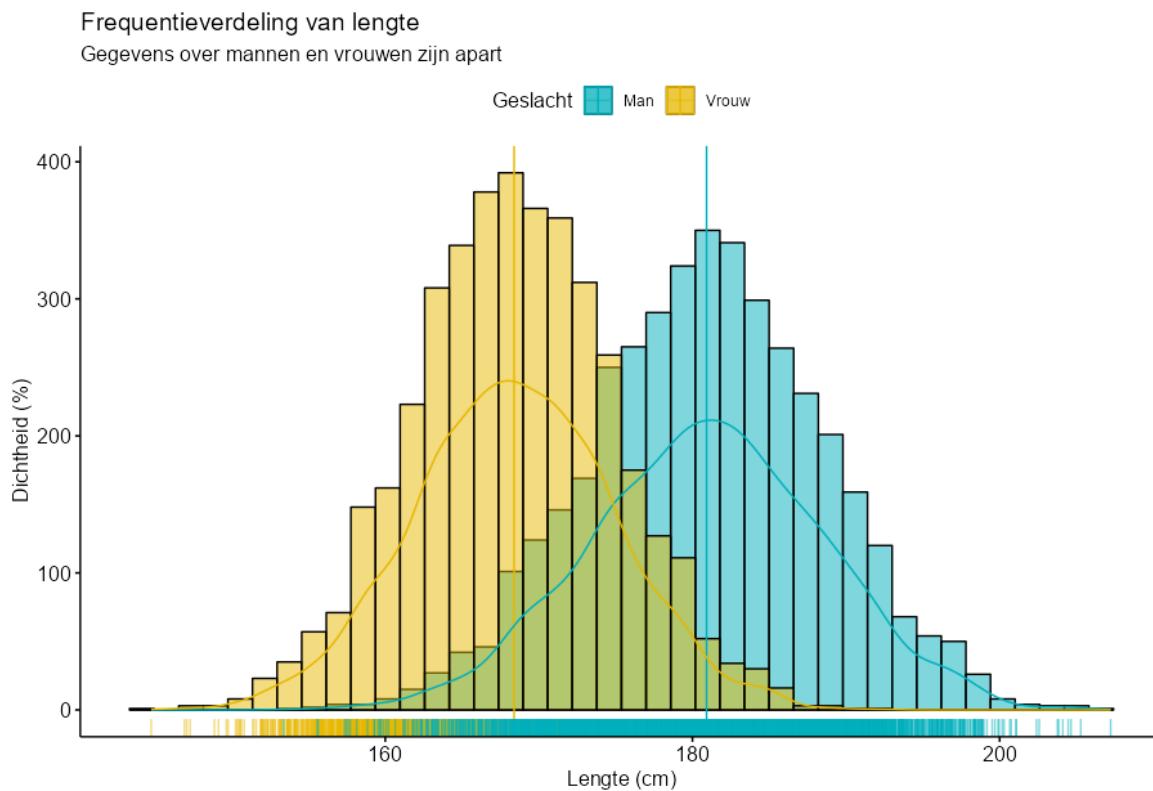
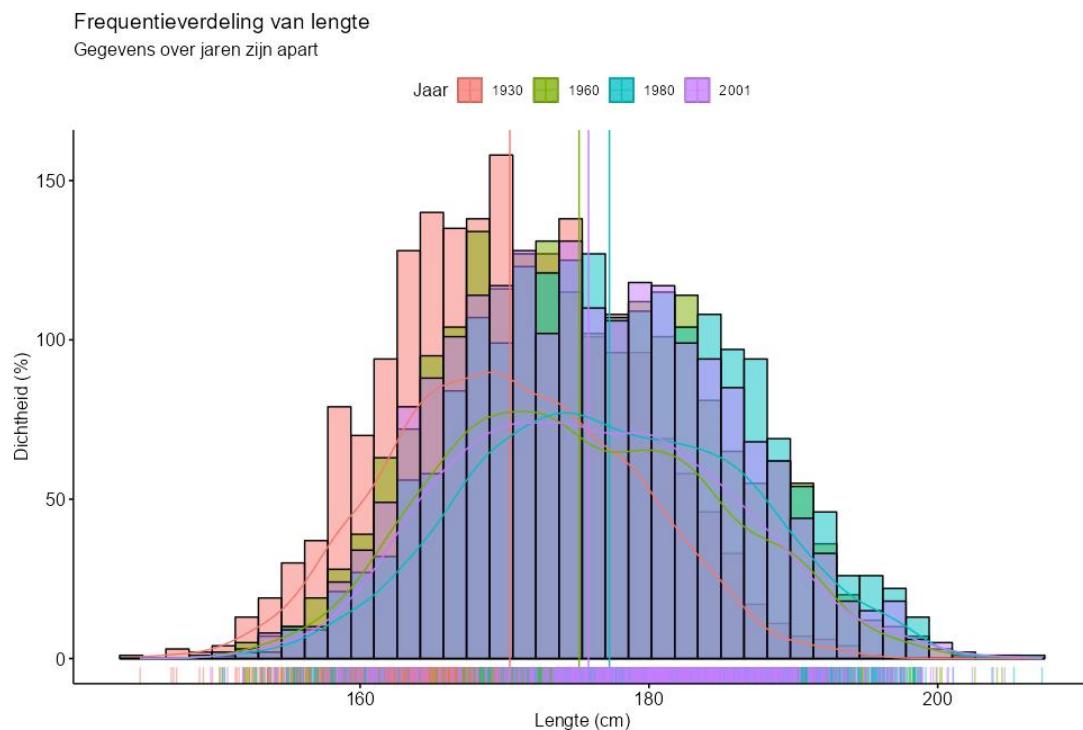
<sup>58</sup>  $1 - ((1 - \alpha)^{\tau})$ ;  $1 - ((1 - 0.05)^8)$

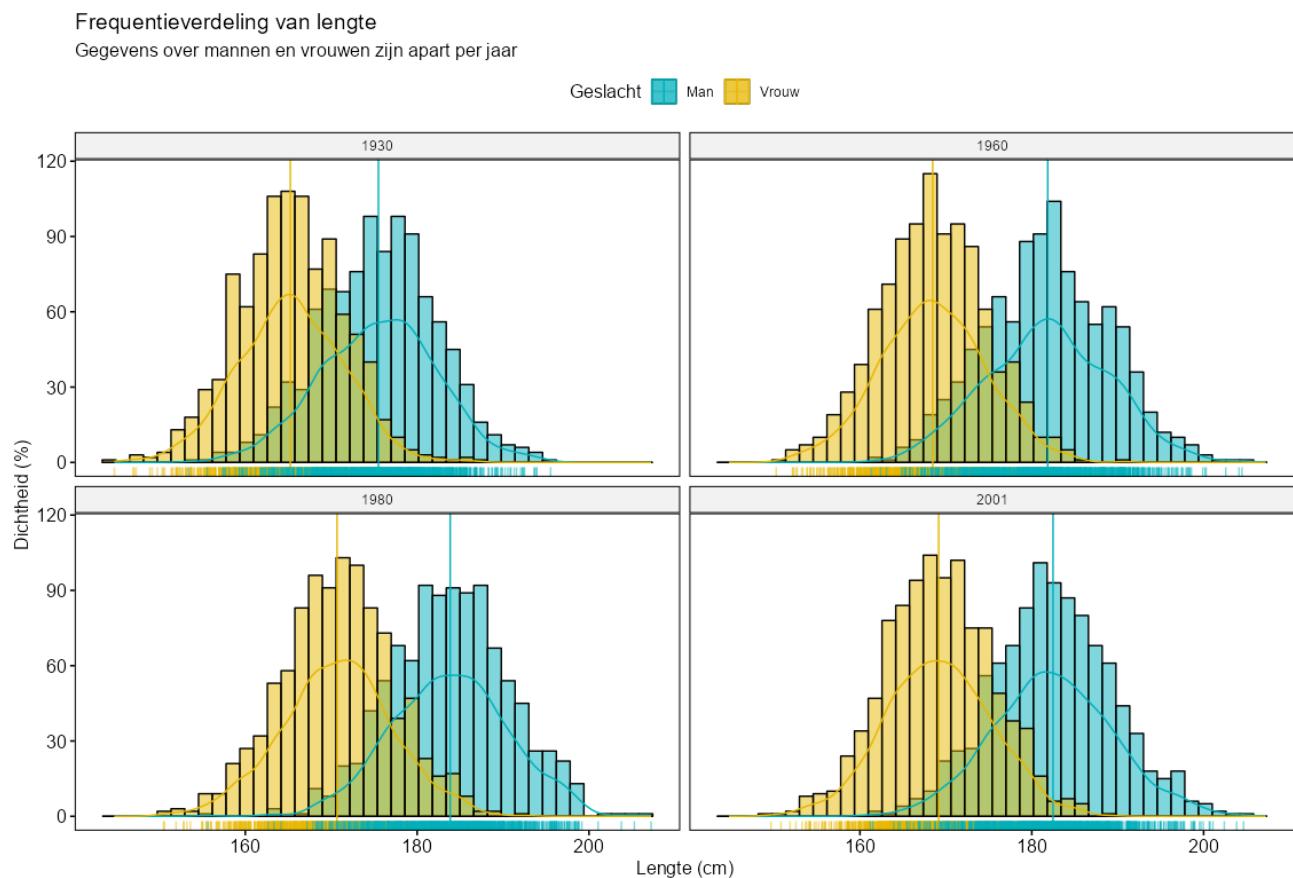
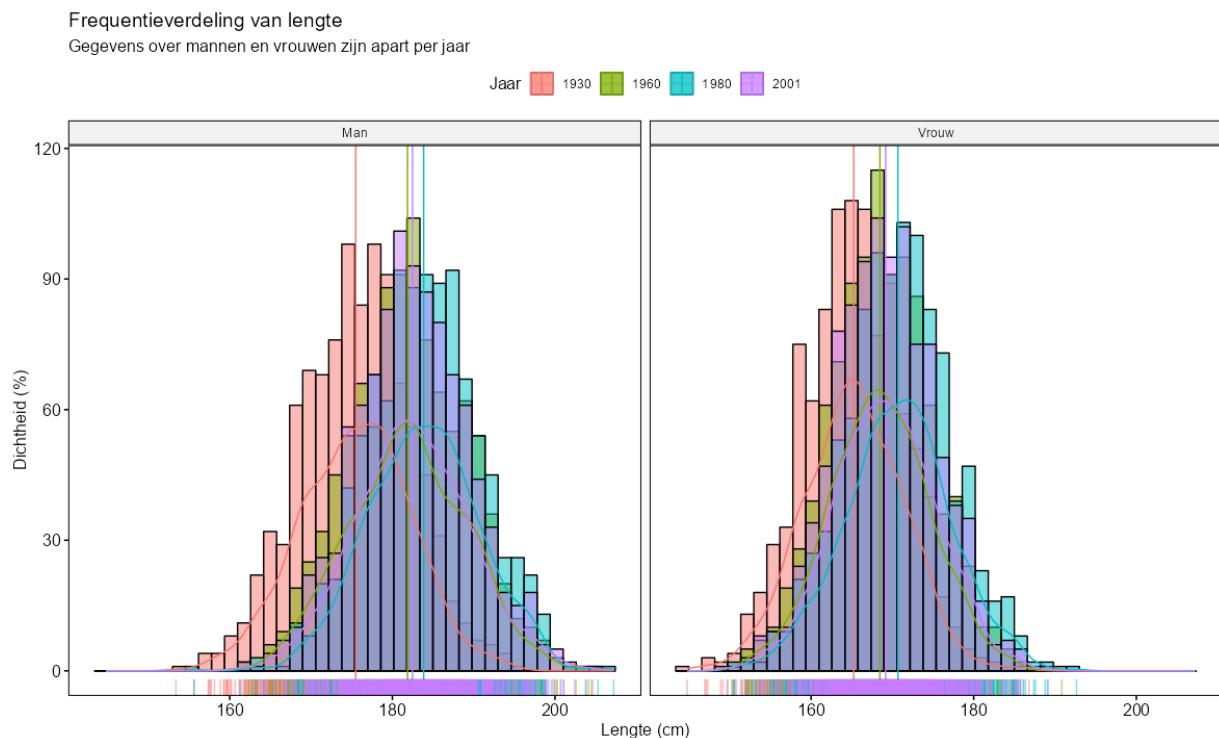
grottere  $\alpha$ : een grenswaarde van 5% stapt nou eenmaal sneller dan een grenswaarde van 1%.

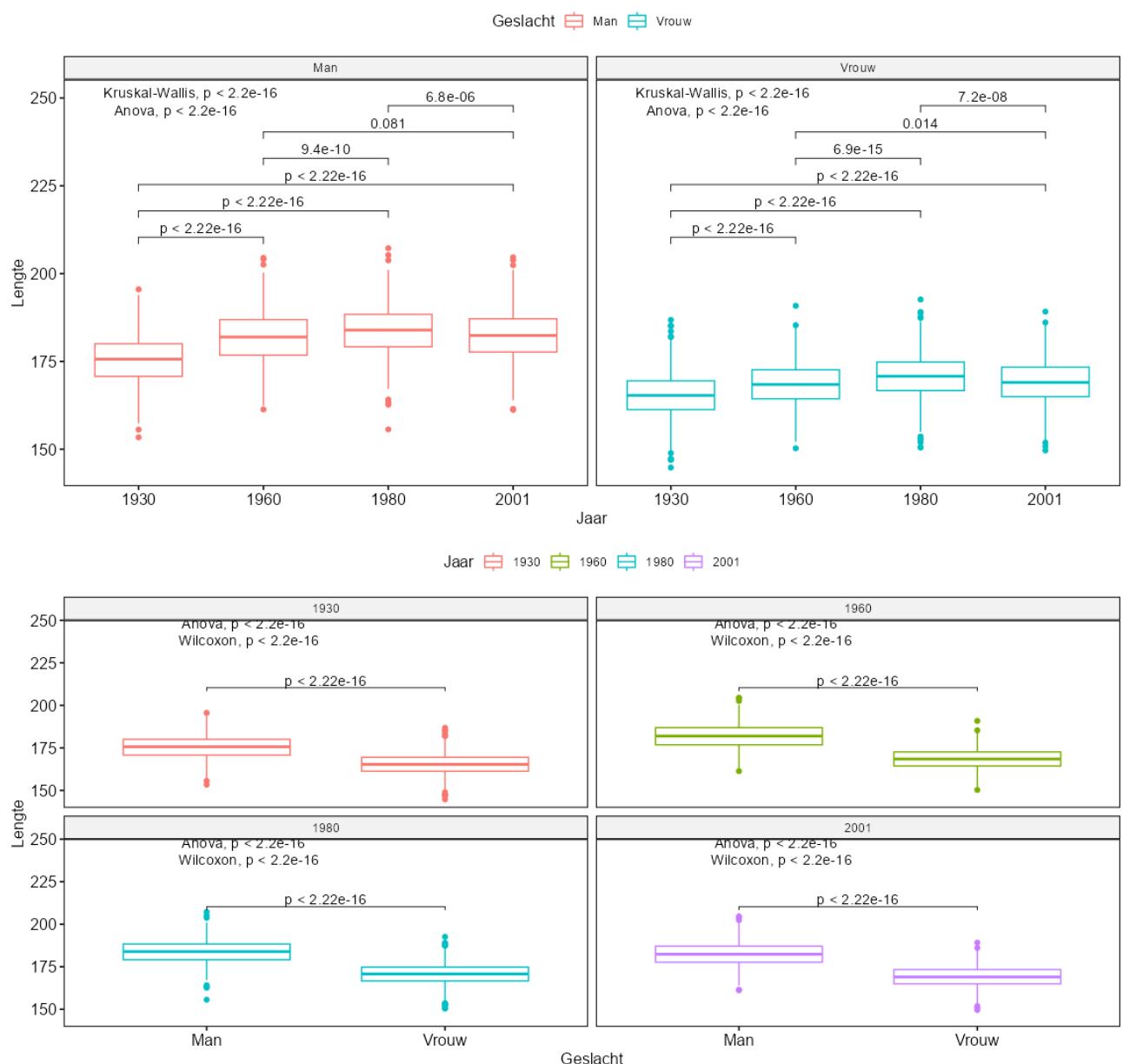
<b><math>\alpha</math> - waarde</b>	<b><math>\tau</math> - waarde</b>	<b>Aantal vals positieven</b>
0.05	1	0.05
0.05	5	0.23
0.05	10	0.40
0.05	20	0.64
0.01	1	0.01
0.01	5	0.049
0.01	10	0.09
0.01	20	0.18

**Tabel 14.** Kans op een vals-positieve als functie van de  $\alpha$  waarde en het aantal testen  $\tau$  op dezelfde data.

Laten we nu gemakshalve de lengtegegevens over de jaren heen voor elk geslacht visualiseren en dan met elkaar vergelijken. We kunnen verschillende figuren maken op basis van het vergelijk wat we willen: Geslacht (**Figuur 48**), Jaar (**Figuur 49**), Geslacht \* Jaar (**Figuur 50**) of Jaar \* Geslacht (**Figuur 51**). Dat zijn een hoop mogelijke vergelijkingen. Laten we deze grafieken nu eens optuigen met het aantal mogelijke vergelijkingen die we kunnen maken (**Figuur 52**). Dat zijn er een heleboel. Laten we nu deze data gebruiken om te zien hoeveel verschillende  $p$ -waarden we kunnen krijgen (**Tabel 15** en **Figuur 53**). Elke  $p$  is afkomstig van een t-test zoals we die al eerder hebben gezien.

**Figuur 48.** Verdeling man en vrouwen over alle jaren heen.**Figuur 49.** Verdeling per jaar over de geslachten heen.

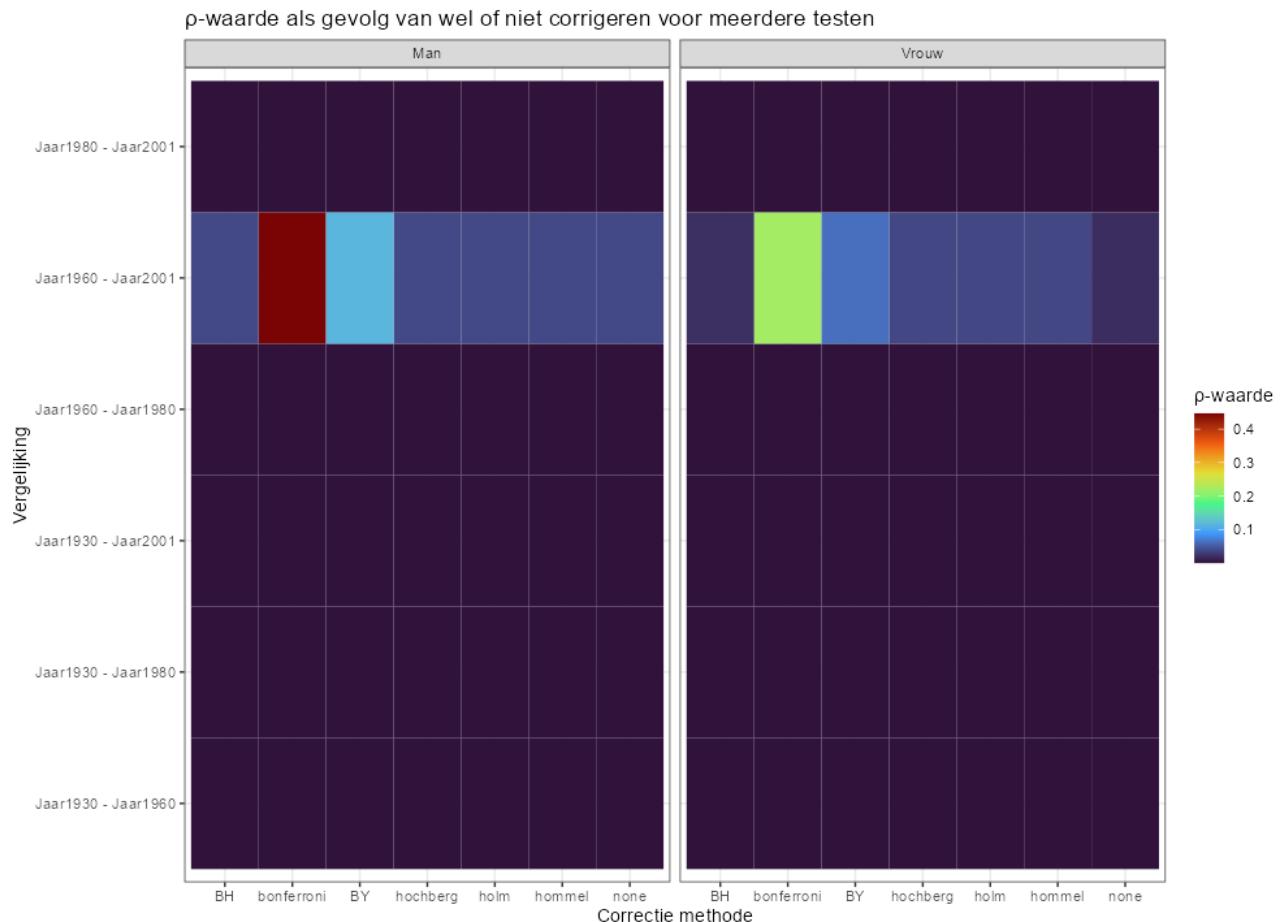
**Figuur 50.** Verdeling man en vrouw per jaar.**Figuur 51.** Verdeling jaren per geslacht.



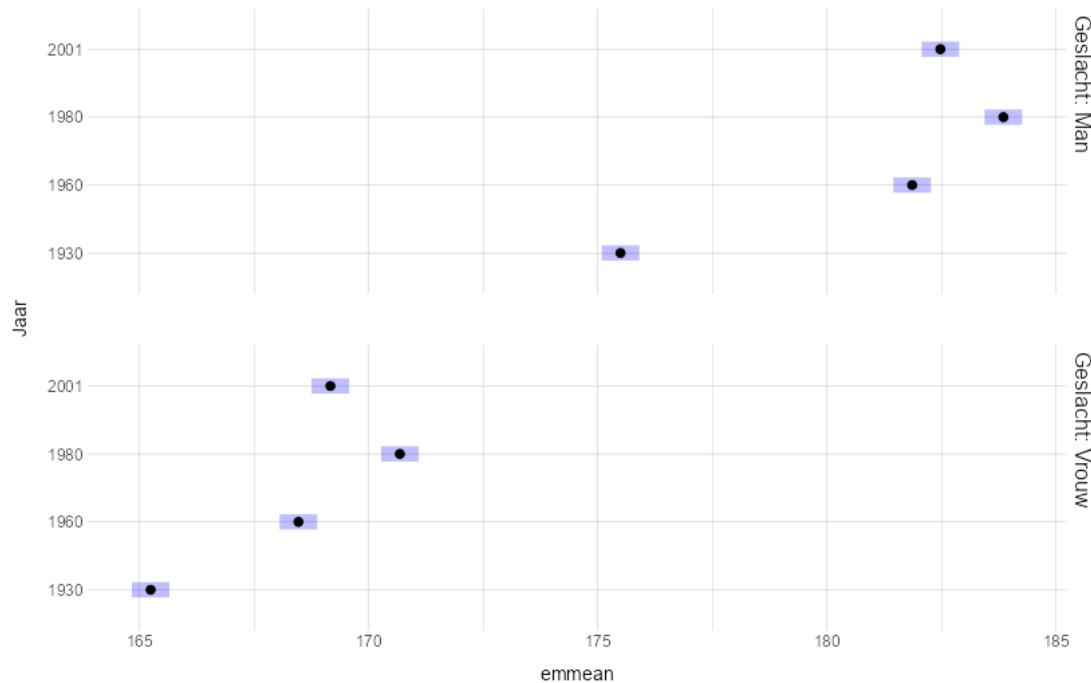
**Figuur 52.** De verdeling van lengte (als boxplot) per jaar per geslacht op twee verschillende manieren met de mogelijke testen ertussen. Het zijn een hoop vergelijkingen die we kunnen maken.

Geslacht	Vergelijking	t-test	Tukey	Bonferroni	Benjamin-Hochberg	FDR
Man	1930 vs 1960	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Man	1930 vs 1980	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Man	1930 vs 2001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Man	1960 vs 1980	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Man	1960 vs 2001	0.0372	0.1585	0.2233	0.0370	0.0370
Man	1980 vs 2001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Vrouw	1930 vs 1960	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Vrouw	1930 vs 1980	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Vrouw	1930 vs 2001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Vrouw	1960 vs 1980	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Vrouw	1960 vs 2001	0.0184	0.0853	0.1101	0.020	0.020
Vrouw	1980 vs 2001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

**Tabel 15.** p-waarden als functie van de vergelijking en de test.**Figuur 53.** p-waarden als functie van de vergelijking en test.

Wat opvalt uit **Tabel 15** en **Figuur 53** is dat de verschillen tussen groepen zo groot is dat de p-waarde bijna overal overeind blijft. We kunnen dit verschil visualiseren in **Figuur 54**. Met zulke verschillen maken p-waarden eigenlijk niets meer uit. Toch is het zeker niet nutteloos om correcties te doen. Zo zien we in **Figuur 53** dat we wel degelijk één p-waarde over de grens van 5% hebben getrokken. Niet vaak zullen we verschillen zien die zo groot zijn als dit.



**Figuur 54.** Gemiddelde waarde en 95% betrouwbaarheidsinterval.

## Wat kunnen we hier nu uit afleiden?

Bovenstaande heeft hopelijk laten zien hoe eenzijdig en tweezijdig toetsen vooral te maken heeft met het verlengen óf inkorten van de afkapwaarden. In de wereld van de frequentistische statistiek zijn de onderdelen zoals de  $\alpha$ -waarde,  $\beta$ -waarde,  $p$ -waarde en het betrouwbaarheidsinterval allemaal met elkaar versmolten. Maar onderliggend zijn het allemaal middelen, gereedschap, om te bepalen of een bepaalde waarde een significante betekenis heeft of niet. Zo maakt het eenzijdig toetsen het de onderzoeker een stuk makkelijker om een statistisch significant verschil te vinden puur en alleen omdat de afkapwaarde verkleind wordt. De assumptie is dat er maar naar één kant van de verdeling hoeft te worden gekeken. Al die keuzes, en meer, hebben invloed op een vaak binaire conclusie: er is wél of géén effect. Terwijl de wereld een heel stuk complexer is.

Het wordt daarom nu tijd om Deel 2 van dit rapport te betreden en eindelijk met de glyfosaat data aan de slag te gaan. Zodat de lezer hopelijk zelf kan zien hoe complex het analyseren van data is, en het interpreteren van die analyse.

## Glyfosaat: beschrijving en visualisatie van de data

---

*Visual representation of data is also an important aspect of the analysis, relying on inspection of the data for outliers, trends, goodness of fit and checks of assumptions. Care should, therefore, be taken in carrying out statistical analyses using flowcharts. The usual considerations for the interpretation of statistical analyses should always be kept in mind.<sup>59</sup>*

---

In de BNNVARA/ZEMBLA rapportage<sup>60</sup> waarin wordt gesproken over het verkeerd toetsen wordt veelvuldig geleund op het rapport van HEAL<sup>61</sup> wat weer leunt op het werk van Dr. Portier<sup>62</sup>. Ondanks dat dit niet het enige werk is wat heeft gekeken naar het effect van glyfosaat bij dieren én mensen, zal ik mij beperkingen tot de gegevens die voor mij voor handen zijn. Op die manier krijgt dit rapport een directe connectie met eerder werk én de rapportage van BNNVARA/Zembla. De focus ligt trouwens niet exclusief op het eenzijdig en tweezijdig toetsen. Ook andere onderdelen die problematisch zijn, zal ik adresseren.

Het artikel is openlijk beschikbaar en in de supplementen vinden we in tabelvorm de data zoals gebruikt door Portier. Er wordt gekeken naar veel verschillende vormen van kanker in relatie tot verschillende doseringen. De manier waarop de data zijn verwerkt staat toe dat ik grotendeels kan repliceren wat in het artikel van Portier staat. Daarom heb ik hem via mail verzocht om de codes te sturen.

Daarbovenop kan ik laten zien hoe de statistische significantie veranderd zodra ik aannames verander. We hebben reeds gezien dat in de frequentistische statistiek sowieso al veel aannames worden gedaan.

In de studie van Portier zijn 13 studies meegenomen. Ook zijn er redenen opgegeven om een aantal studies niet mee te nemen, maar daar kan ik niks over zeggen, want ik ben geen toxicoloog. Ik ga er dus gemakshalve vanuit dat deze exclusie op goede gronden genomen is en zal rekenen met het werk waarmee Portier ook rekent. We zouden dus met hetzelfde materiaal moeten werken.

---

<sup>59</sup> [https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453\\_9789264221475-en](https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453_9789264221475-en)

<sup>60</sup> <https://www.bnvara.nl/zembla/artikelen/kankerrisico-door-pesticiden-decennialang-verkeerd-ingeschat>

<sup>61</sup> <https://www.env-health.org/wp-content/uploads/2022/06/HEAL-How-the-EU-risks-greenlighting-a-pesticide-linked-to-cancer-2022.pdf>

<sup>62</sup> <https://doi.org/10.1186/s12940-020-00574-1>

Allereerst heb ik de gegevens, zoals opgenomen uit *Supplementary Material 2* van de studie, verwerkt in een spreadsheet zodat ik deze daarna statistisch kan analyseren. De gegevens staan zijn per studie ingedeeld en worden in tabelvorm gepresenteerd (**Figuur 55**). In elke tabel staat de: (1) gebruikte dosering, (2) het aantal muizen of ratten dat is opgenomen, (3) het aantal voorvallen van een bepaalde kankersoort bij een bepaalde dosering, (4) de lengte van de studie en (5) de resultaten van de trend testen zoals gebruikt door Portier. We tellen in **Figuur 55** in totaal 18 p waarden: 9 in de mannen groep én 9 in de vrouwen groep. Dit is dus voor één studie.

Wanneer we deze gegevens uit **Figuur 55** in onze spreadsheet hebben kunnen we er direct mee gaan rekenen. Het is trouwens altijd zinvol om klein te beginnen en dan uit te breiden. Daarom zal ik starten met de gegevens in **Figuur 55** om te zien of ik kan nabootsen wat hier staat.

**Table S1.** Tumors of interest in male and female CD-1 mice from the 24-month feeding study of Knezevich and Hogan (1983) [11] – Study A

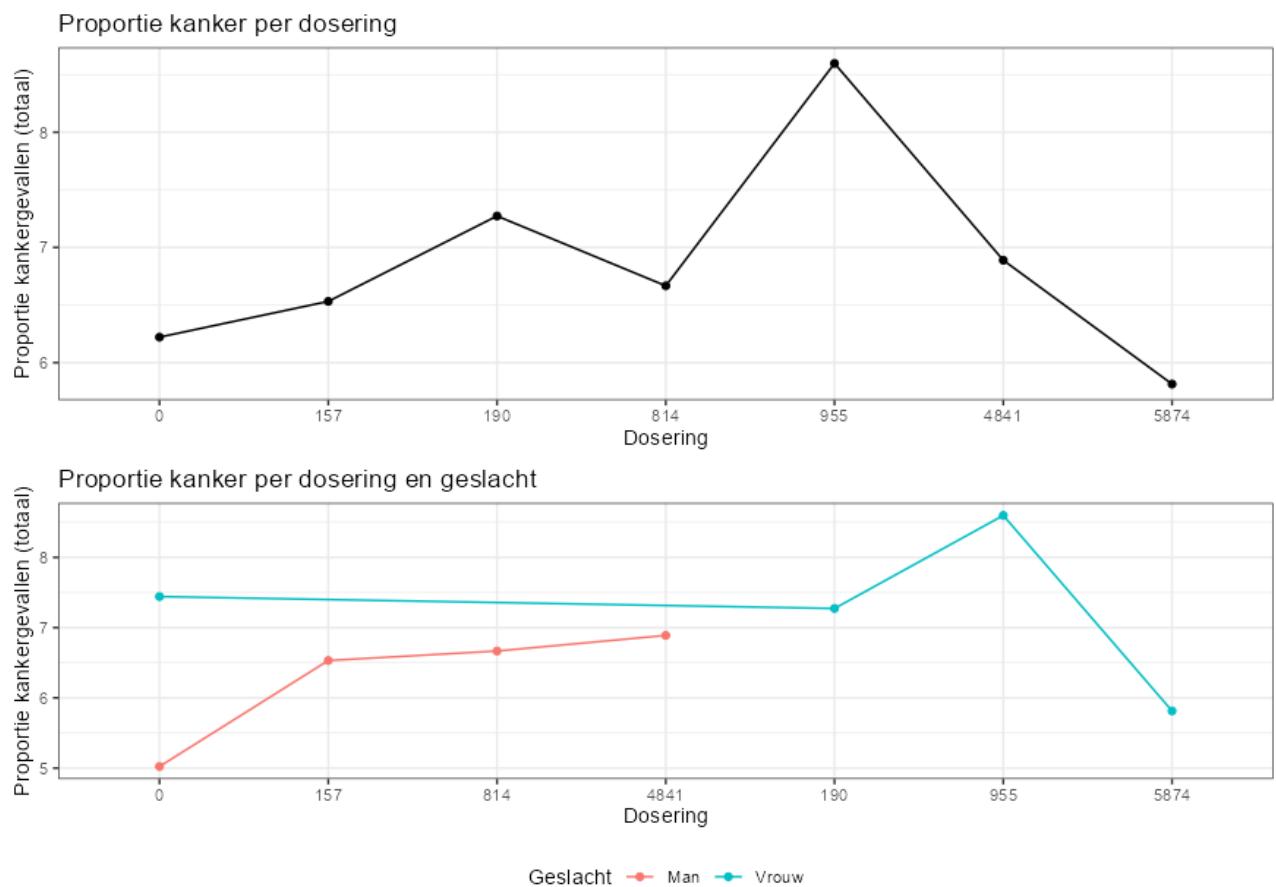
Tumor	Doses (mg/kg/day) or Tumor Incidence <sup>1</sup>				Trend Test p-value
Males	0	157	814	4841	
Kidney Adenomas (original pathology)	0/49	0/49	1/50	3/50	0.019
Kidney Adenomas <sup>2</sup>	1/49	0/49	0/50	1/50	0.442
Kidney Carcinomas <sup>2</sup>	0/49	0/49	1/50	2/50	0.063
Kidney Adenomas and Carcinomas <sup>2</sup>	1/49	0/49	1/50	3/50	0.065
Malignant Lymphomas	2/49	5/49	4/50	2/50	0.754
Hemangiosarcomas <sup>3</sup>	0/49	0/49	1/50	0/50	0.505
Alveolar-Bronchiolar Adenomas	5/48	9/50	9/50	9/50	0.294
Alveolar-Bronchiolar Carcinomas	4/48	3/50	2/50	1/50	0.918
Alveolar-Bronchiolar Adenomas and Carcinomas	9/48	12/50	11/50	10/50	0.576
Females	0	190	955	5874	
Hemangiomas	0/49	1/49	1/50	0/50	0.631
Harderian Gland Adenomas	2/45	0/48	1/49	0/44	0.877
Harderian Gland Carcinomas	0/45	0/48	0/49	0/44	---
Harderian Gland Adenomas and Carcinomas	2/45	0/48	1/49	0/44	0.877
Alveolar-Bronchiolar Adenomas	10/49	9/50	10/49	1/50	0.999
Alveolar-Bronchiolar Carcinomas	1/49	3/50	4/49	4/50	0.183
Alveolar-Bronchiolar Adenomas and Carcinomas	11/49	12/50	14/49	5/50	0.985
Spleen Composite Lymphosarcoma	1/50	1/48	1/49	5/49	0.016
Malignant Lymphomas	5/49	6/49	6/49	10/49	0.070

1 – Doses are given in the rows marked "Males" and "Females", tumor counts appear on the rows with the individual tumors; 2 – tumor counts obtained from EPA]; \* 0.01<p≤0.05 for Fisher's Exact Test; \*\* p≤0.01 for Fisher's Exact Test

**Figuur 55.** Voorbeeld van de data van één enkele studie zoals gerapporteerd door Portier. OP basis van deze tabel, en de andere 12 tabellen, zal ik de analyses trachten na te bootsen.

## Voorbeeld uitwerken: Knezevich and Hogan (1983)

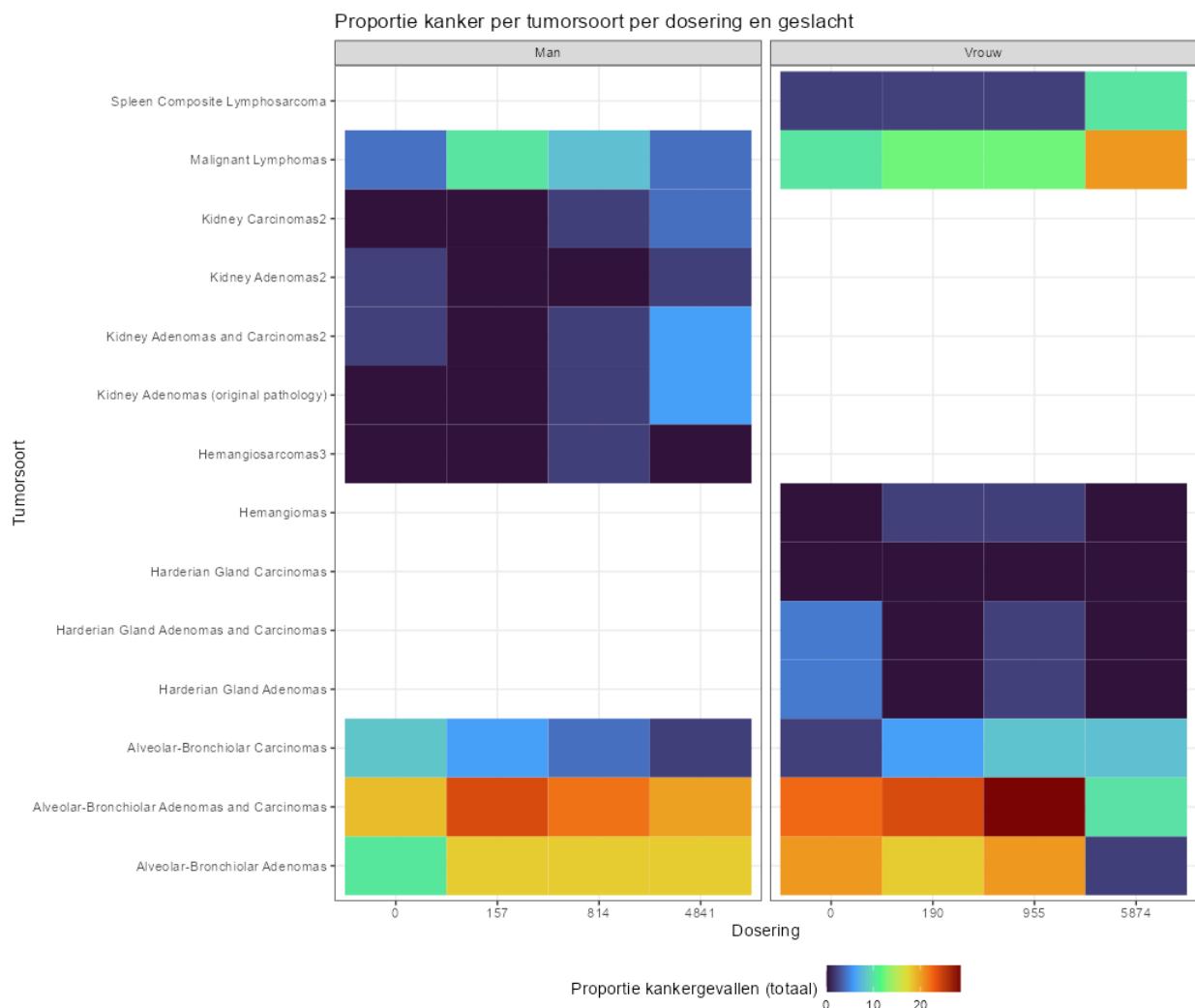
Het voorbeeld van Knezevich and Hogan (1983) kunnen we op verschillende dimensies bekijken<sup>63</sup>. Wellicht zou het goed zijn om de data eerst samen te voegen op verschillende niveaus, zoals dosering (of dosering én geslacht) en dan de proportie kanker te tonen. We zien de resultaten terug in **Figuur 56** waarin we direct beide grafieken laten zien. Het samenvoegen van data is trouwens een precaire zaak<sup>64</sup>, dus het is goed om terug te keren naar de originele opdeling per kancersoort **Figuur 57**.



**Figuur 56.** Proportie kanker per dosering, en per dosering én geslacht.

<sup>63</sup> Voor wie nu had verwacht dat ik gewoonweg de gegevens in een test zou gooien gaat voorbij aan de essentie van data-analyse en dat is dat je eerst moet zien wat er is geobserveerd.

<sup>64</sup> On andere vanwege [Simpson's paradox](#).



**Figuur 57.** Proportie kanker zoals verkregen vanuit **Figuur 55**.

Het volgende wat we kunnen doen is één enkele rij nemen voor één geslacht. Laten we kijken naar de mannen uit de Knezevich and Hogan studie én dan de rij nemen waarin staat *Kidney Adenomas (original pathology)*. Uit de grafiek lezen we dat de *Trend Test p-value* 0.019 is. Dat is lager dan de normale grens van 0.05 en daarmee dus ‘statistisch significant’. De exacte test die gebruikt is, valt niet af te leiden uit de grafiek zelf, maar in het artikel lezen we dit:

*Individual tumor counts for the individual studies are reanalyzed using the exact form of the Cochran-Armitage (C-A) linear trend test in proportions [37]. Reanalyses are conducted on all primary tumors where there are at least 3 tumors in all of the animals in a sex/species/strain combination (regardless of dosing). In addition, any tumor where a positive finding ( $p \leq 0.05$ ,*

*one-sided C-A trend test) is seen in at least one study is also evaluated, regardless of number of animals with the tumor, in all studies of the same sex/species/strain. When adenomas and carcinomas are seen in the same tissue, a combined analysis of adenomas and carcinomas is also conducted. The minimum of three tumors is used since the exact version of the C-A test cannot detect tumors in studies of this size with less than at least 3 tumors. Additional file 2: Tables S1–S13 provide the tumor count data for all tumors with a significant trend test ( $p \leq 0.05$ ) in at least one study of the same sex/species/strain along with the doses used (mg/kg/day) and the number of animals examined microscopically in each group. Pairwise comparisons between individual exposed groups and control are conducted using Fisher's exact test [37] and are provided for comparison with other reviews.*

Het is dus niet geheel duidelijke welke test er uiteindelijk gerapporteerd is. Het artikel spreekt over de Cochrane-Armitage (CA) test<sup>65</sup>, maar de tabel per studie rept vooral over de Fisher Exact Test<sup>66</sup> (**Figuur 55**). Om ervoor te zorgen dat ik hier nu niet de verkeerde statistische test neem om de gepresenteerde p-waarden na te bootsen, is het wellicht waakzaam om beide testen te nemen. Zowel eenzijdig als tweezijdig, wanneer dat behulpzaam is. De resultaten van die exercitie zien we in **Tabel 16**.

Statistische test	R procedure <sup>67</sup>	Alternatieve hypothese	$\alpha$	p - waarde
Gerapporteerde CA test	N/A	Eenzijdig	0.05	0.019
Test voor gelijke proporties	prop.test	Tweezijdig	0.05	0.111
Chi-kwadraat test	prop.trend.test	Tweezijdig	0.05	0.0250
Fisher Exact Test	fisher.test	Tweezijdig	0.05	0.196
CA test	CochranArmitageTest	Tweezijdig	0.05	0.0250
CA test	CochranArmitageTest	Eenzijdig	0.05	0.0125

**Tabel 16.** De p-waarde zoals gerapporteerd in de Portier studie voor Knezevich & Hogan, mannelijke muizen én Kidney Adenomas (original pathology) staat bovenaan. Ook rapporteren we de p-waarde van vijf verschillende testen zoals uitgevoerd in het statistiek programma R. Geen van de uitkomsten is gelijk aan de gerapporteerde p -waarde uit het artikel.

Zichtbaar wordt dat we niet de p-waarde vinden zoals gerapporteerd. Van de vijf testen die ik heb uitgevoerd komt alleen de eenzijdige CA test in de buurt. Drie van de vijf testen zijn

<sup>65</sup> [https://en.wikipedia.org/wiki/Cochran%20Armitage\\_test\\_for\\_trend](https://en.wikipedia.org/wiki/Cochran%20Armitage_test_for_trend)

<sup>66</sup> [https://en.wikipedia.org/wiki/Fisher%27s\\_exact\\_test](https://en.wikipedia.org/wiki/Fisher%27s_exact_test)

<sup>67</sup> R is het statistiek programma wat gebruikt wordt om de figuren en voorbeelden te maken zoals we die hier zien in het document. R kent een aantal basis modules, maar ook aanvullende bibliotheken. Wat ik steeds zal benoemen is exact welke procedure ik gevuld heb zodat het navolgbaar is voor de lezer hoe elke procedure in elkaar steekt.

significant, twee andere testen niet. De CA-test zoals door mij berekent komt in de buurt van de gerapporteerde CA-test. Toch zitten nu met het probleem dat we niet kunnen evenaren (althans voor dit voorbeeld) wat er exact gedaan is. Wat we wel kunnen zien in de CA test is dat de keuze om eenzijdig of tweezijdig te toetsen bij een  $\alpha$  waarde de p-waarde halveert. We zagen dit al eerder (**Tabel 10, Tabel 11, Tabel 12**) en deze resultaten zijn uiteraard geheel in lijn met de theorie van de frequentistische statistiek.

We kunnen een p-waarde zien die we niet kunnen repliceren. Laten we de proef op de som nemen en hetzelfde doen voor de vrouwen in hetzelfde artikel met dezelfde tests. En laten we dan gemakshalve ook de eerste kancersoort nemen waarvoor de bevinding statistisch significant is in het artikel: *Spleen Composite Lymphosarcoma*. De resultaten die we dan krijgen staan in **Tabel 17**. Ook hier zien we dat de gerapporteerde p-waarde niet wordt gevonden. Wel vinden we een p-waarde voor de eenzijdige CA test die dicht in de buurt komt van de gerapporteerde p-waarde. Ook zien we hier het principe zoals al eerder aangesneden door Gerard De Snoo: wanneer we van tweezijdig toetsen naar eenzijdig toetsen gaan wordt een niet statistisch significant effect wel statistisch significant.

Statistische test	R procedure	Alternatieve hypothese	$\alpha$	p - waarde
Gerapporteerde CA - test	N/A	Eenzijdig	0.05	0.016
Test voor gelijke proporties	prop.test	Tweezijdig	0.05	0.099
Chikwadraat test	prop.trend.test	Tweezijdig	0.05	0.052
Fisher Exact Test	fisher.test	Tweezijdig	0.05	0.197
CA test	CochranArmitageTest	Tweezijdig	0.05	0.052
CA test	CochranArmitageTest	Eenzijdig	0.05	0.026

**Tabel 17.** De p-waarde zoals gerapporteerd in de Portier studie voor Knezevich & Hogan, vrouwelijke muizen én Spleen Composite Lymphosarcoma. Ook rapporteren we de p-waarde van vijf verschillende testen zoals uitgevoerd in het statistiek programma R. Geen van de uitkomsten evenaart de gerapporteerde p -waarde uit het artikel.

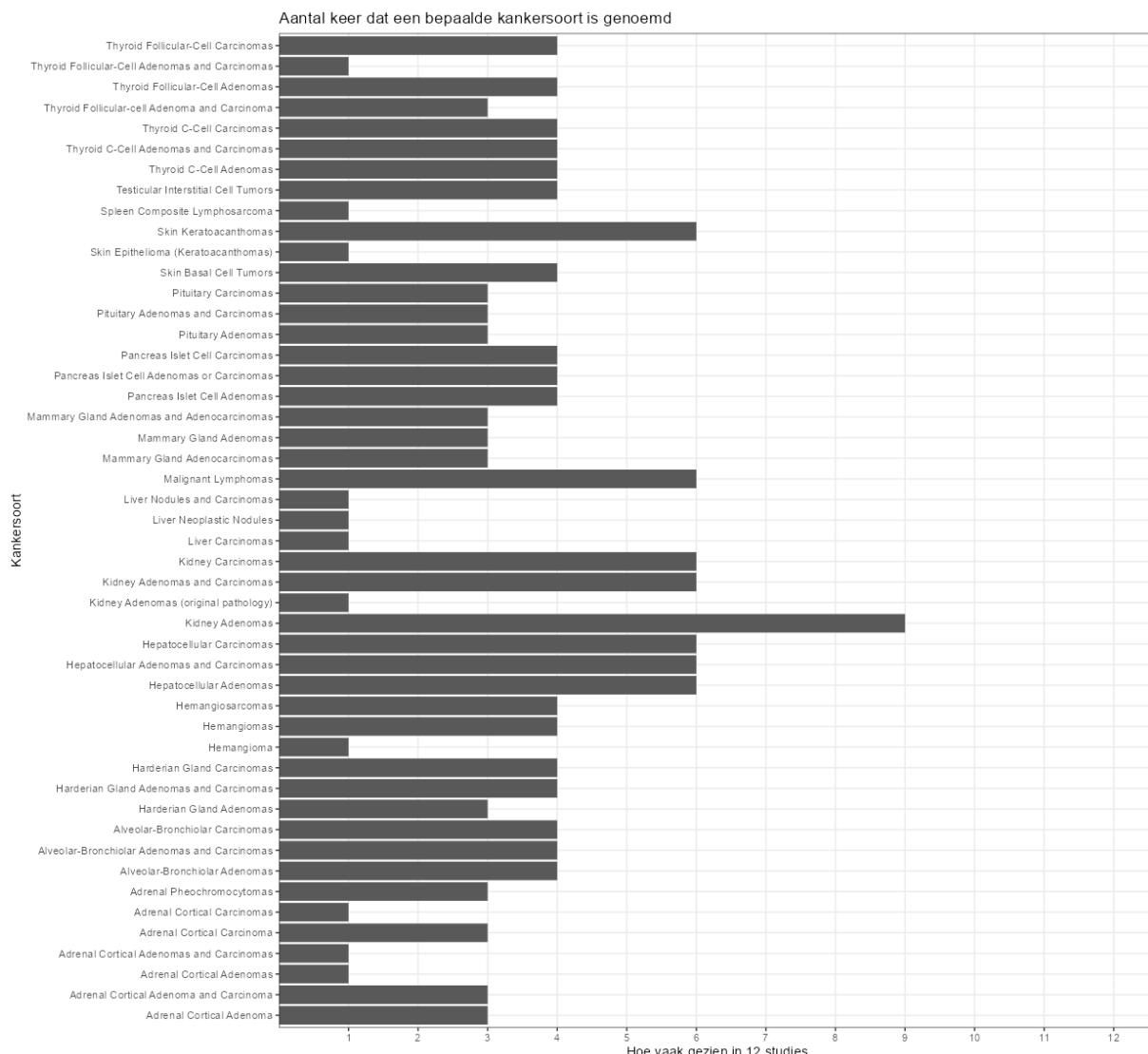
Het is nu tijd om naar alle studies te gaan kijken: apart én samen, want dat is wat ze in de studie van Portier ook hebben gedaan.

## Alle studies bekijken

Laten we naar alle studies gaan kijken. Voordat ik de nodige grafieken zal maken én berekeningen zal uitvoeren, is het essentieel dat de lezer begrijpt dat één enkele studie nooit doorslaggevend kan zijn. Nu beschrijft Portier meerdere studies, maar de studie van Portier is zelf eigenlijk ook maar één enkele studie: er is één iemand die op één enkele manier kijkt naar de data. De gemaakte keuzes die dan plaatsvinden hebben invloed op de resultaten. Hetzelfde mag ook gezegd worden van wat ik nu doe en zal laten zien: ook dit is het werk van één iemand.

Keuzes maken is een belangrijk onderdeel in het doen van onderzoek en met de data van Portier zijn er veel keuzes te maken. Zo zien we dat de studies dat de studies verschillen in dosering, het gebruik van het soort proefdieren én de kancersoorten die werden gedetecteerd. Ik kan de data dus op verschillende manieren tot mij nemen.

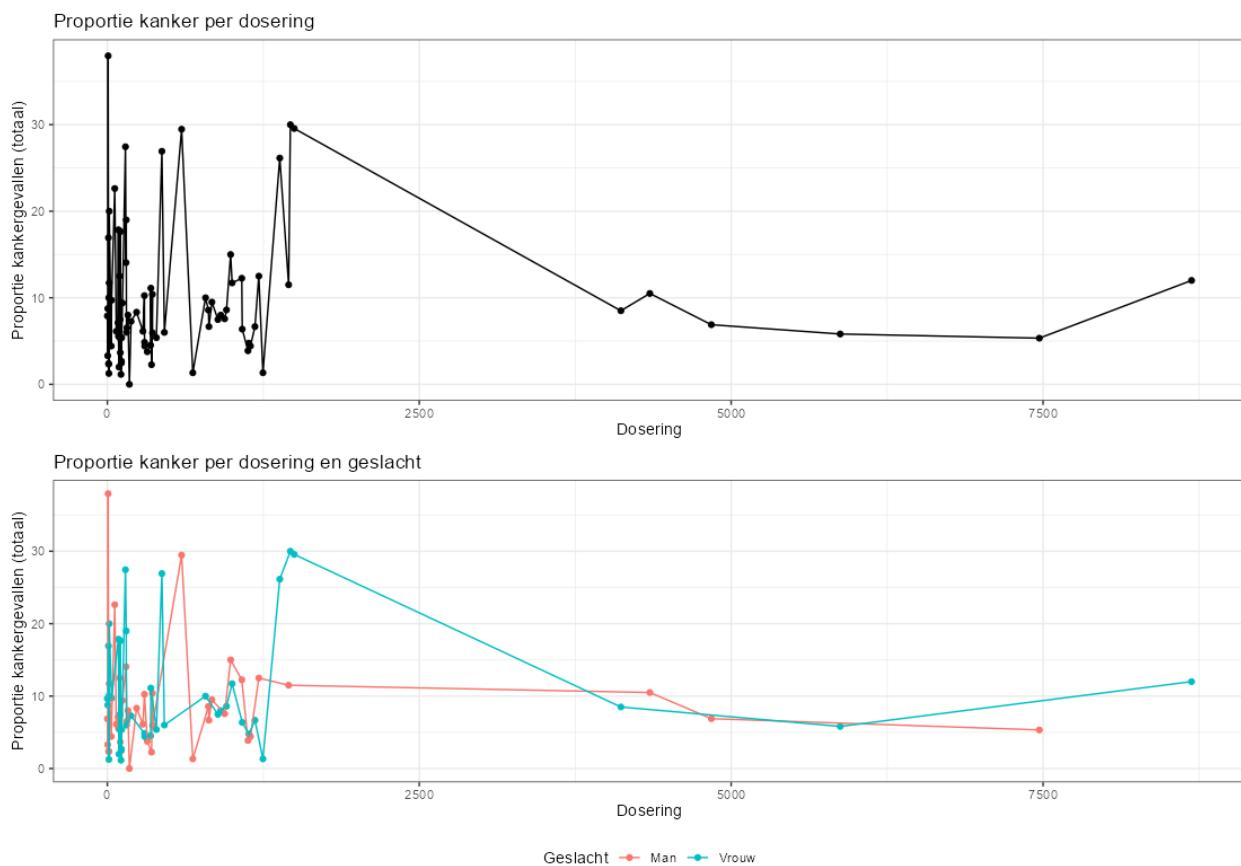
Verder beschrijft de studie van Portier duidelijk per studie welke kancersoorten er zijn gevonden, maar het verschilt wel per studie. Als in één enkele studie een bepaalde kancersoort wordt gevonden bij een bepaalde dosis, dan wordt bijgehouden of dat eerder of later ook gebeurt (**Figuur 55**). Kancersoorten die niet worden gevonden worden niet genoemd. Dat betekent dus dat we per studie alleen de lijst krijgen van kancersoorten (of tumorsoorten) die voorkomen in één enkele studie, maar niet over alle studies heen. Laten we daar eens mee beginnen om te zien hoe vaak een bepaalde kancersoort voorkomt. We kijken dus niet naar de hoeveelheid in een studie, maar over alle studies heen. Als we deze groeperen, dan krijgen we de volgende grafiek (**Figuur 58**). Wat direct opvalt is dat niet elke kancersoort even vaak wordt gezien. Wat ook opvalt is dat, als we kijken naar de berekeningen van Portier, er geen rekening is gehouden met studies waarin geen kanker is gezien. Als bijvoorbeeld *Mammary Gland Adenomas* in drie studies wordt gezien betekent dit dat het in 9 andere studies niet wordt gezien. Die data worden niet meegenomen in een eventuele dose-response analyse of analyse waarbij proporties worden vergeleken. Zouden we over alle studies heen gaan kijken dan zouden we op zijn minst rekening moeten houden dat een bepaalde soort kanker niet overal zichtbaar is. We komen hier op terug in de paragraaf over de Bayesiaanse statistiek, maar voordat we deze overwegingen mee gaan nemen in verdere analyses is het zinvol om eerst een aantal grafieken te maken. We willen namelijk zeker weten dat we de data goed begrijpen.



**Figuur 58.** Het aantal keer dat één kankersoort voorkomt in 12 studies. Bijvoorbeeld: Mammary Gland Adenomas is in 3 studies gezien. Dit kan verschillen van de 3 studies waarin Mammary Gland Adenocarcinomas is gezien.

In **Figuur 59** zien we de proportie kanker per dosering én per dosering en geslacht. Deze keer heb ik alle aantallen meegenomen zoals gerapporteerd en afgezet tegen het aantal dieren. Het is dus letterlijk de proportie per dosering zoals genoemd. Ik ga er gemakshalve vanuit dat elke dosering andere dieren heeft<sup>68</sup>. Wat we dan kunnen doen is proberen om een zogenaamde dose-response curve te maken: dit is een wiskundige formule om te bepalen of het aantal kankergevallen verschuift over tijd.

<sup>68</sup> Eigenlijk hoef ik hier niet eens vanuit te gaan, want dit is exact hoe dierproefstudies die kijken naar een zogenaamde 'dose-response' functie eruit moet zien. Alleen, ik kan het niet exact lezen in de studie van Portier.

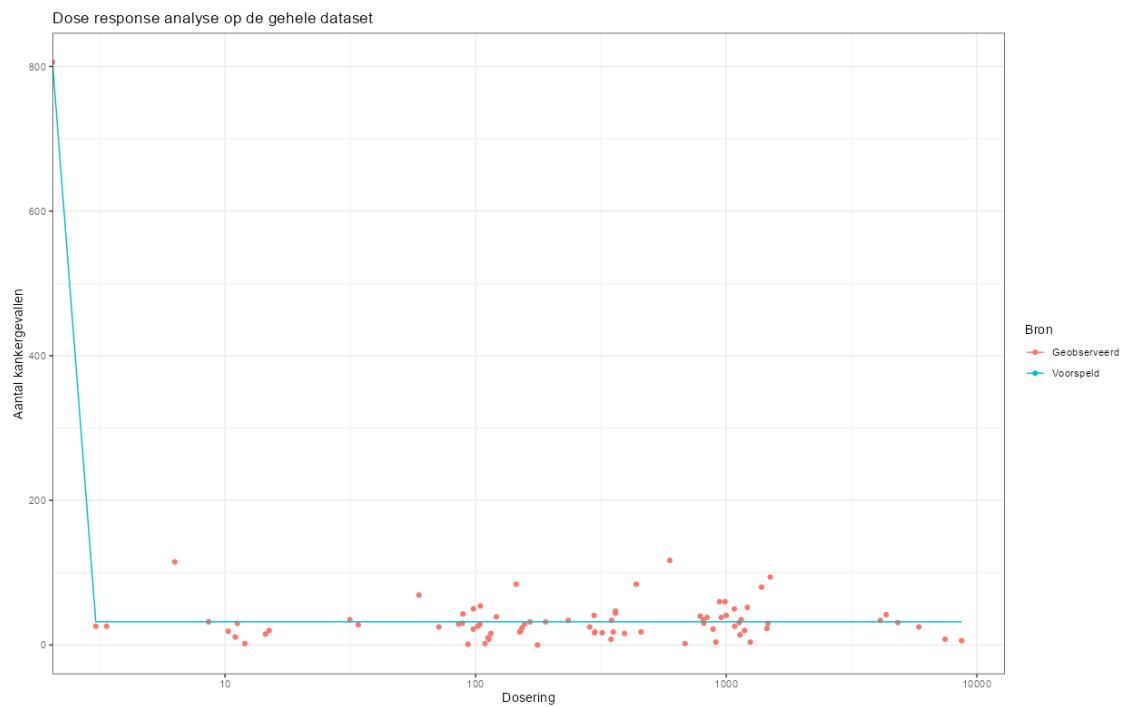


**Figuur 59.** Proportie kanker per dosering én proportie kanker per dosering en geslacht. Duidelijk te zien dat het gros van de studies data heeft verzameld tot ongeveer 2000 mg/kg/dag. De lineaire schaal is waarschijnlijk niet de beste schaal om mee te meten.

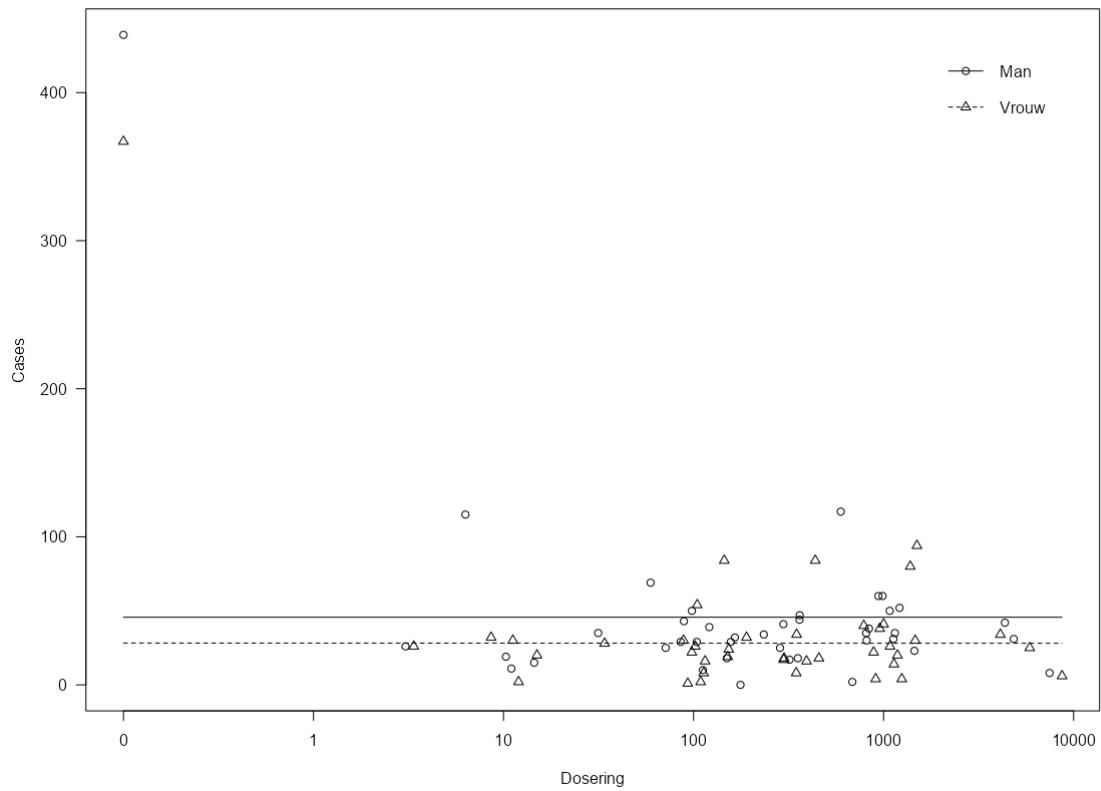
Omdat elke studie ook het aantal kankergevallen toont bij een dosering van 0 mg/kg/dag is het essentieel dat dit ons startpunt wordt (**Figuur 60**). Verder zijn de groepen per geslacht en per dosering ongeveer even groot zodat er geen zware vertekening zal zijn als we met proporties rekenen.

Wat uit de resultaten direct opvalt is dat van een dose-response geen sprake lijkt te zijn: de meeste kankergevallen treden op bij een dosering van 0. Wanneer we de grafiek opsplitsen per geslacht zien we hetzelfde (**Figuur 61**)<sup>69</sup>.

<sup>69</sup> Een eerste kanttekening die we kunnen maken is dat we niet uit de data kunnen vernemen wanneer een dier overleden is. Het enige wat we weten is hoeveel dieren er nog leven aan het einde van hun respectievelijke groep waarbij verwacht mag worden dat dieren, zoals ratten en muizen, in twee jaar tijd sterven van ouderdom.

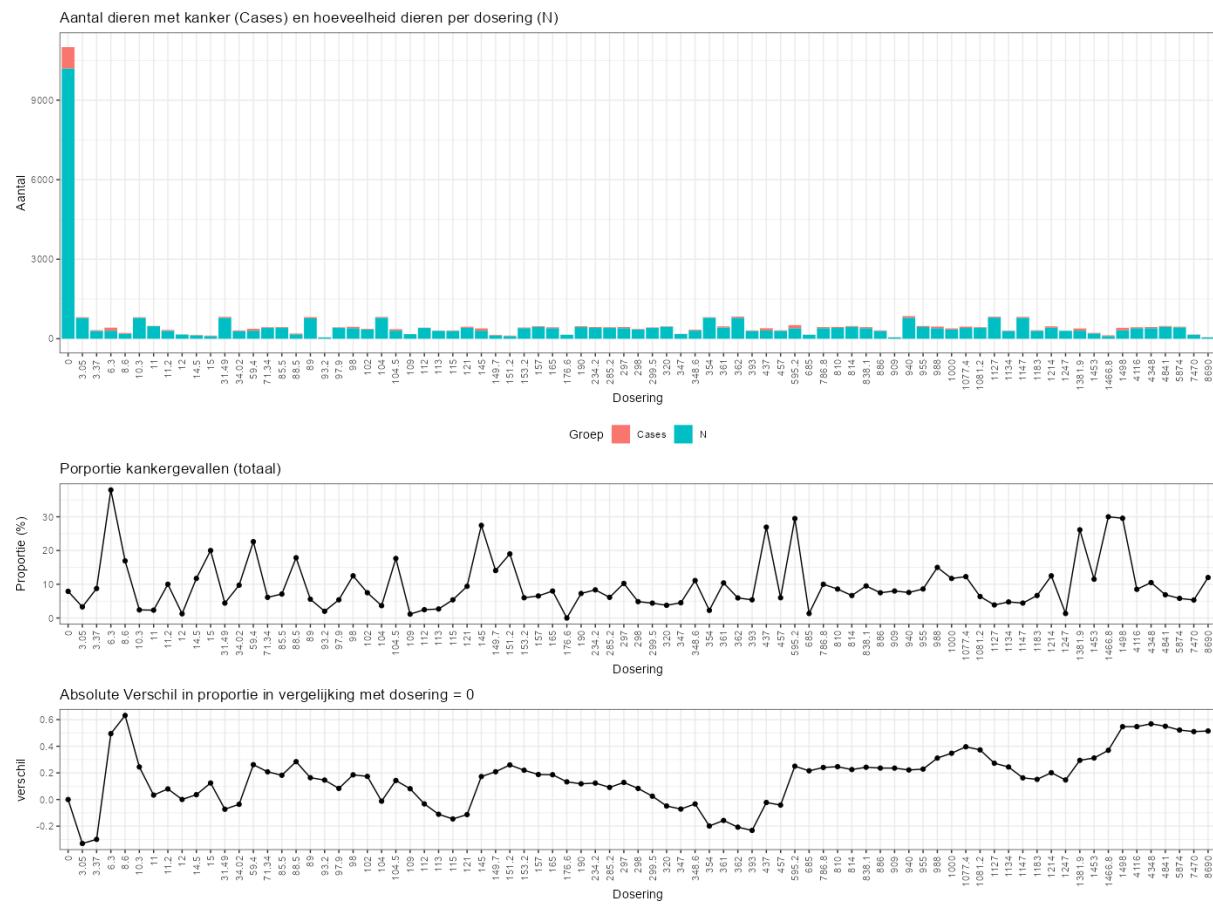


**Figuur 60.** Dose-response curve waarbij de dosering wordt afgezet tegenover het aantal kankergevallen. De blauwe lijn is de lijn afkomstig van een wiskundig model met als data de rode bolletjes: dit zijn alle observaties.



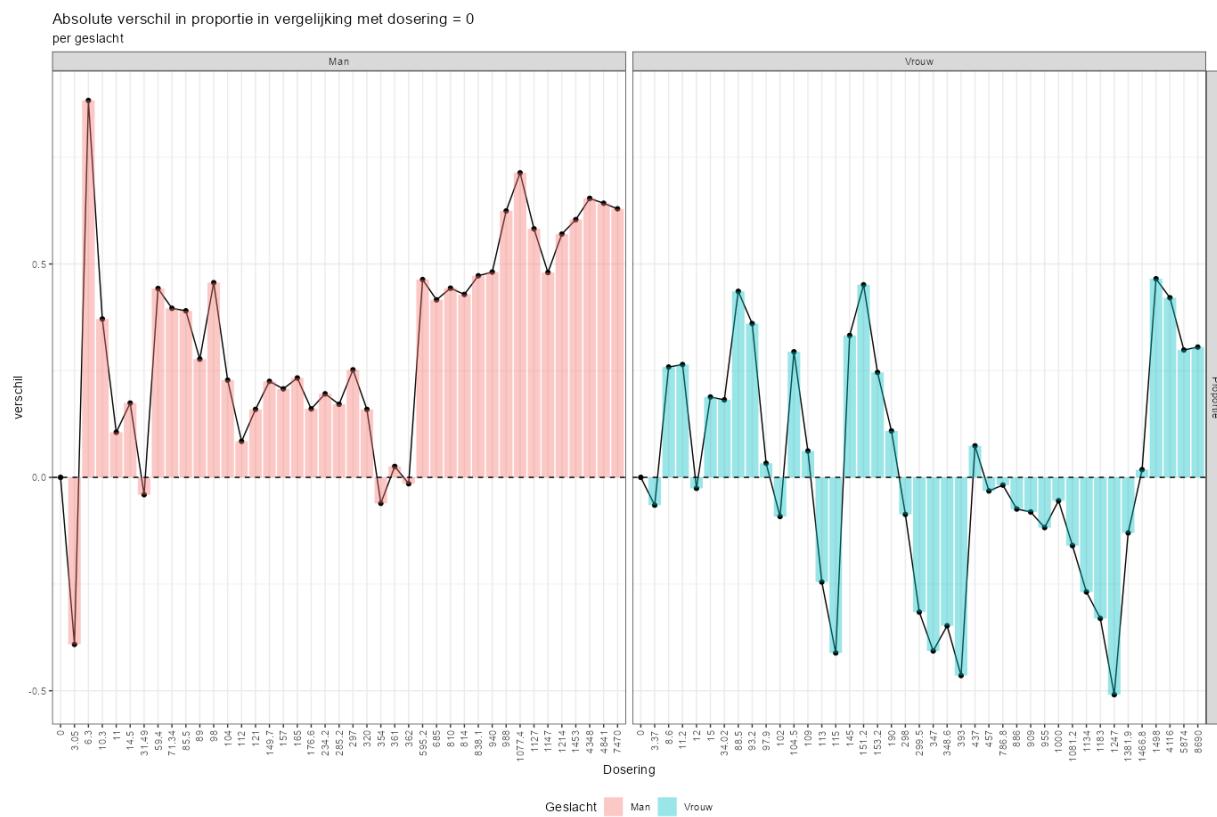
**Figuur 61.** Dose response per geslacht – exercitie is gelijk aan **Figuur 60**.

Nu het mij niet lijkt te lukken om een klassieke dose-response curve te maken, is het wellicht beter als mij richt op de ontwikkeling van het aantal kankergevallen op basis van het verschil ten opzichte van de nul-dosering. We zouden er zelfs voor kunnen kiezen om de nul-dosering niet eens mee te nemen en dan te bezien of het aantal kankergevallen inderdaad stijgt naarmate de dosis groter wordt. Maar als we **Figuur 62** tot ons nemen zien we dat dit weinig goeds doet: we krijgen keer op keer geen klassieke dose-response curve te zien. Althans, niet zoals we deze verwachten. Want hoewel de proportie ten opzichte van de nul-dosering lijkt te stijgen bij 306 mg/kg/dag zakt deze daarna weer in om uiteindelijk weer te stijgen. Als deze grafieken dus iets laten zien dan is het wel dat over alle studies heen kijken door ze simpelweg te groepen geen goede manier is om een beeld te krijgen van een mogelijke dose-response curve.



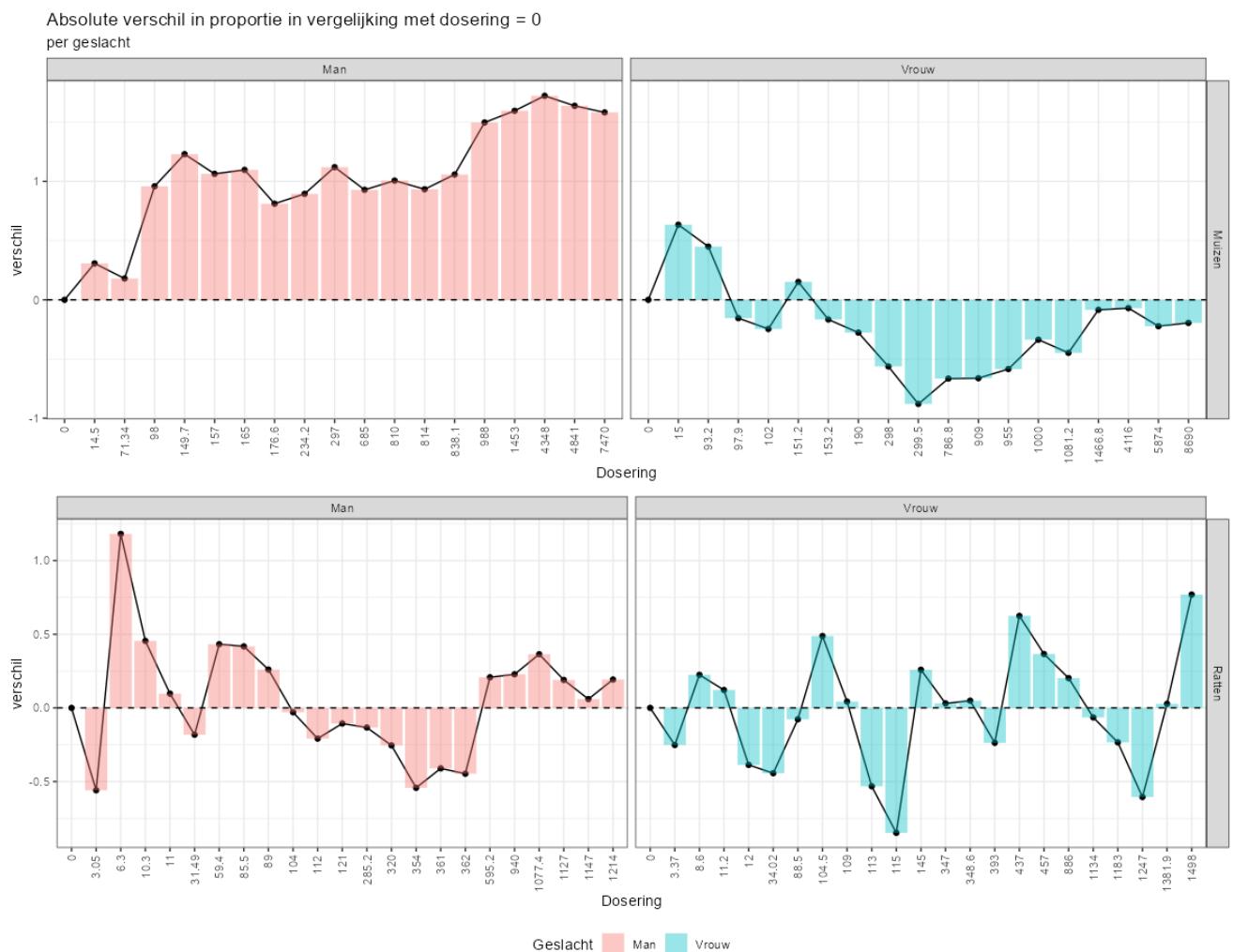
**Figuur 62.** Aantal dieren met kanker en hoeveelheid dieren per dosis (bovenste figuur);proportie kankergevallen per dosis (middelste figuur); én absolute verschil in proporties tussen de nuldosering en de andere doseringen (onderste figuur).

Een eerste volgende stap is om de data dan maar op te splitsen op basis van geslacht en dat is exact wat je kunt zien in **Figuur 63**. Vooral bij de vrouwen zien we gekke ontwikkelingen.



**Figuur 63.** Absolute verschil in proporties tussen de nuldosering en de andere doseringen per geslacht.

We kunnen nog een splitsing maken, namelijk door ook te kijken per diersoort (ratten of muizen). Als we dit doen zien we direct een interactie tussen geslacht en diersoort (**Figuur 64**).



**Figuur 64.** Absolute verschil in proporties tussen de nuldosering en de andere doseringen per geslacht én diersoort.

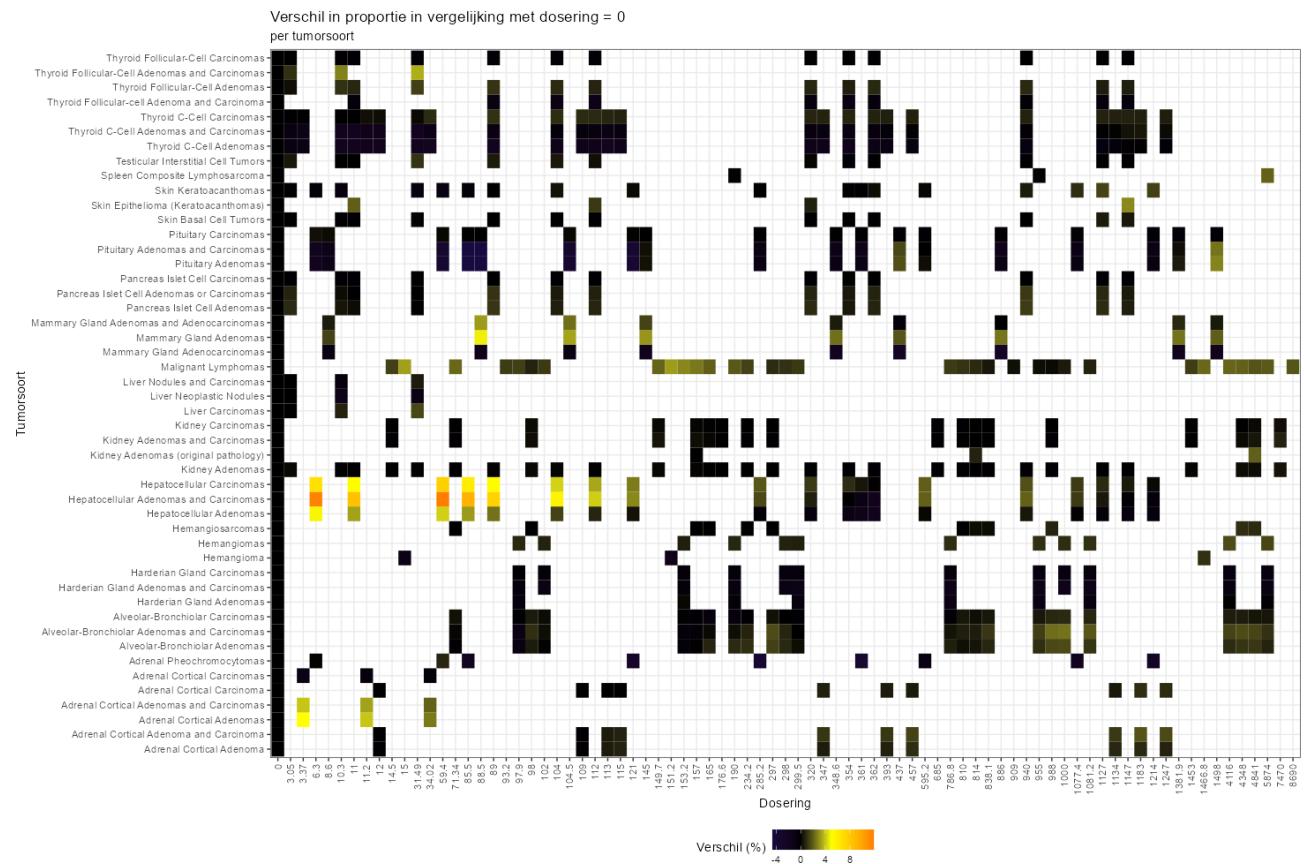
We hebben in **Figuur 58** al gezien hoe belangrijk de tumorsoort (of kancersoort<sup>70</sup>) is en het zou goed zijn om daar ook wat beter naar te gaan kijken, bijvoorbeeld door een splitsing te maken per dosering (**Figuur 65**). Ook hier zien dat het verschil in proportie alleen bij enkele tumorsoorten extreem is in vergelijk met de nul-dosering.

Deze grafieken tonen daarmee nog een andere mooie bevinding en dat is dat we niet weten van elk dier wanneer deze nou exact kanker heeft gekregen. Hadden we dit wel geweten dan hadden we zogenaamde survival analyses kunnen uitvoeren<sup>71</sup>. Nu weten we alleen per studie en per dosering hoeveel dieren er aan het einde wel of geen kanker

<sup>70</sup> Ik gebruik de termen wat afwisselend

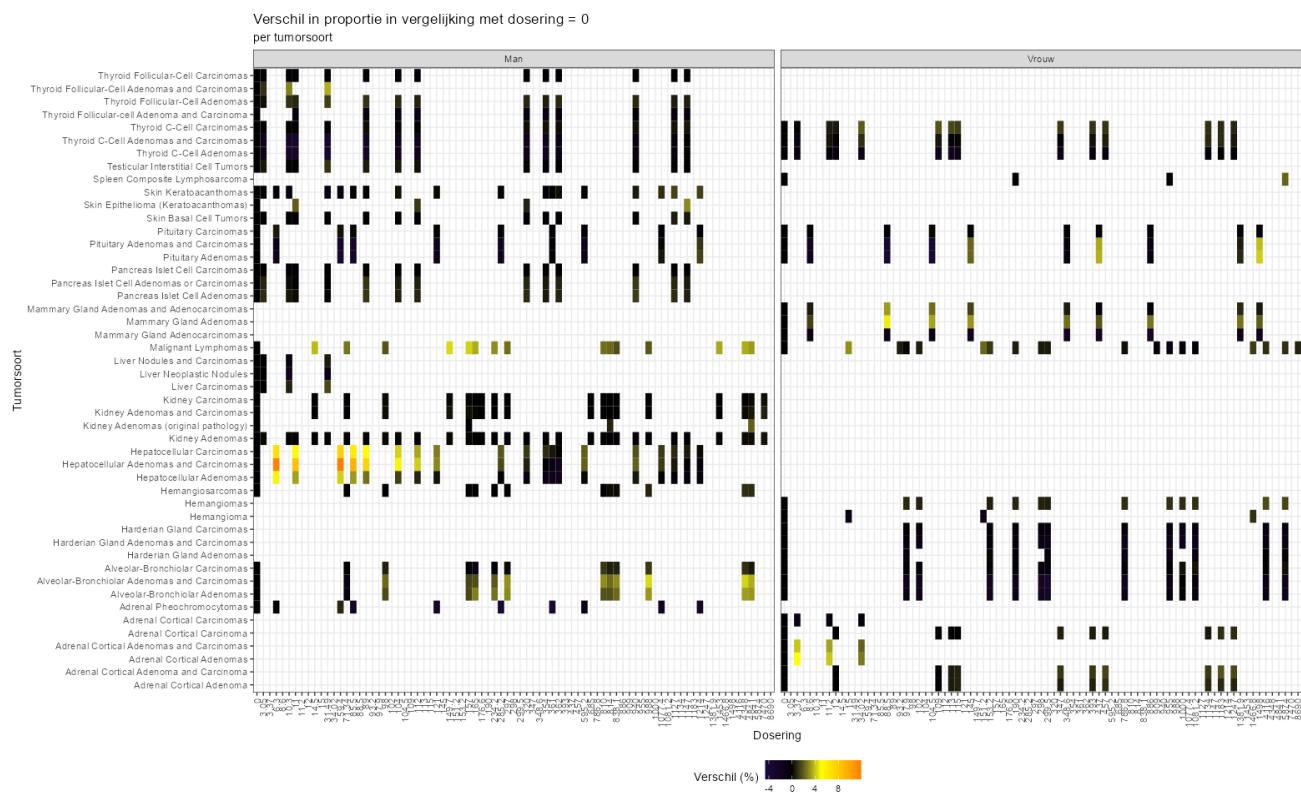
<sup>71</sup> [https://en.wikipedia.org/wiki/Survival\\_analysis](https://en.wikipedia.org/wiki/Survival_analysis)

kregen. Deze beperkingen speelt vooral wanneer we proberen om over studies heen te kijken en is minder relevant wanneer we de vergelijking maken in één enkele studie.



**Figuur 65.** Absolute verschil in proporties tussen de nuldosering en de andere doseringen per tumorsoort.

Kijken we per geslacht dan zien we hetzelfde (**Figuur 66**), namelijk dat het gros van het verschil met de proportie op de nul-dosering helemaal niet zoveel afwijkt. We zouden nu eenzelfde grafiek kunnen maken per geslacht én diersoort, maar dat zou de grafiek nodeeloos complex maken: het is nu al moeilijk om zaken visueel van elkaar te onderscheiden.



**Figuur 66.** Absolute verschil in proporties tussen de nul-dosering en de andere doseringen per tumornoort en per geslacht.

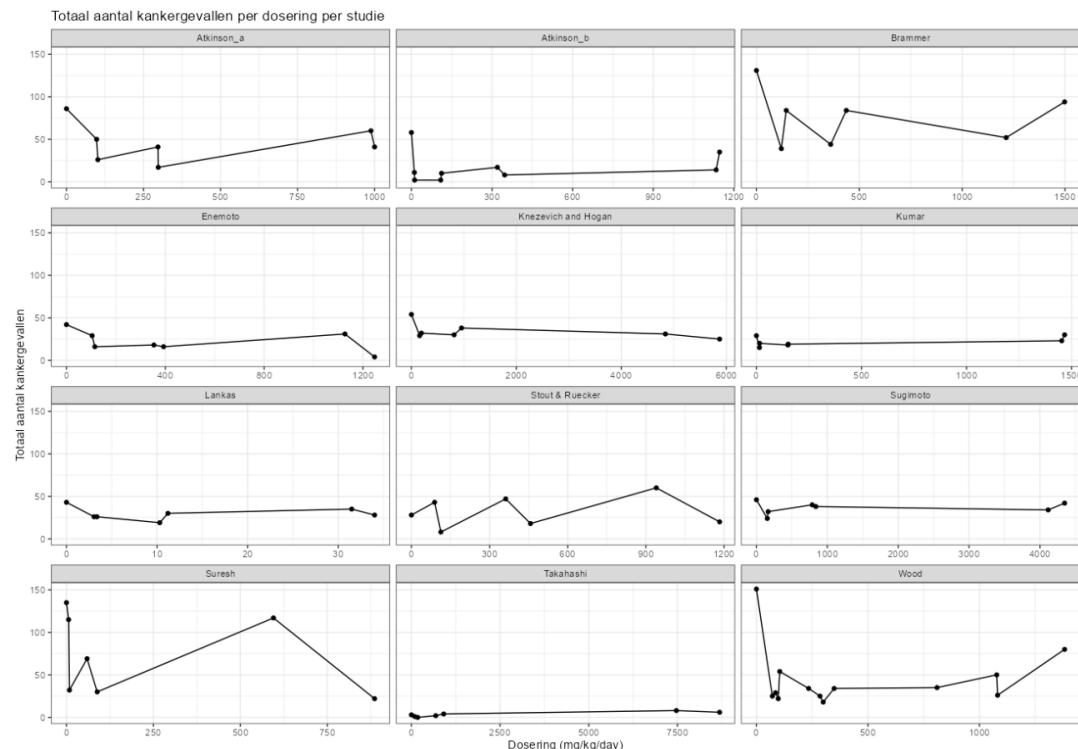
Een van de problemen met deze dataset is het aantal doublures. Als we kijken naar **Figuur 55** dan zien we dat er voor één studie een heleboel getallen zijn die wijzen op dezelfde dieren. Per geslacht en per dosering hebben we cellen met X aantal dieren en binnen die X aantal dieren zien we X aantal tumoren. We weten niet of sommige dieren meer tumoren hebben dan andere dieren. Al dit soort informatie is niet duidelijk op basis van de studie van Portier. Laten we wederom het meest belangrijke deel uit de methode-sectie erbij halen:

*Individual tumor counts for the individual studies are reanalyzed using the exact form of the Cochran-Armitage (C-A) linear trend test in proportions [37]. Reanalyses are conducted on all primary tumors where there are at least 3 tumors in all of the animals in a sex/species/strain combination (regardless of dosing). In addition, any tumor where a positive finding ( $p \leq 0.05$ , one-sided C-A trend test) is seen in at least one study is also evaluated, regardless of number of animals with the tumor, in all studies of the same sex/species/strain.*

*When adenomas and carcinomas are seen in the same tissue, a combined analysis of adenomas and carcinomas is also conducted. The minimum of three tumors is used since the exact version of the C-A test cannot detect tumors in studies of this size with less than at least 3 tumors. Additional file 2: Tables S1–S13 provide the tumor count data for all tumors with a significant trend test ( $p \leq 0.05$ ) in at least one study of the same sex/species/strain*

*along with the doses used (mg/kg/day) and the number of animals examined microscopically in each group. Pairwise comparisons between individual exposed groups and control are conducted using Fisher's exact test [37] and are provided for comparison with other reviews.*

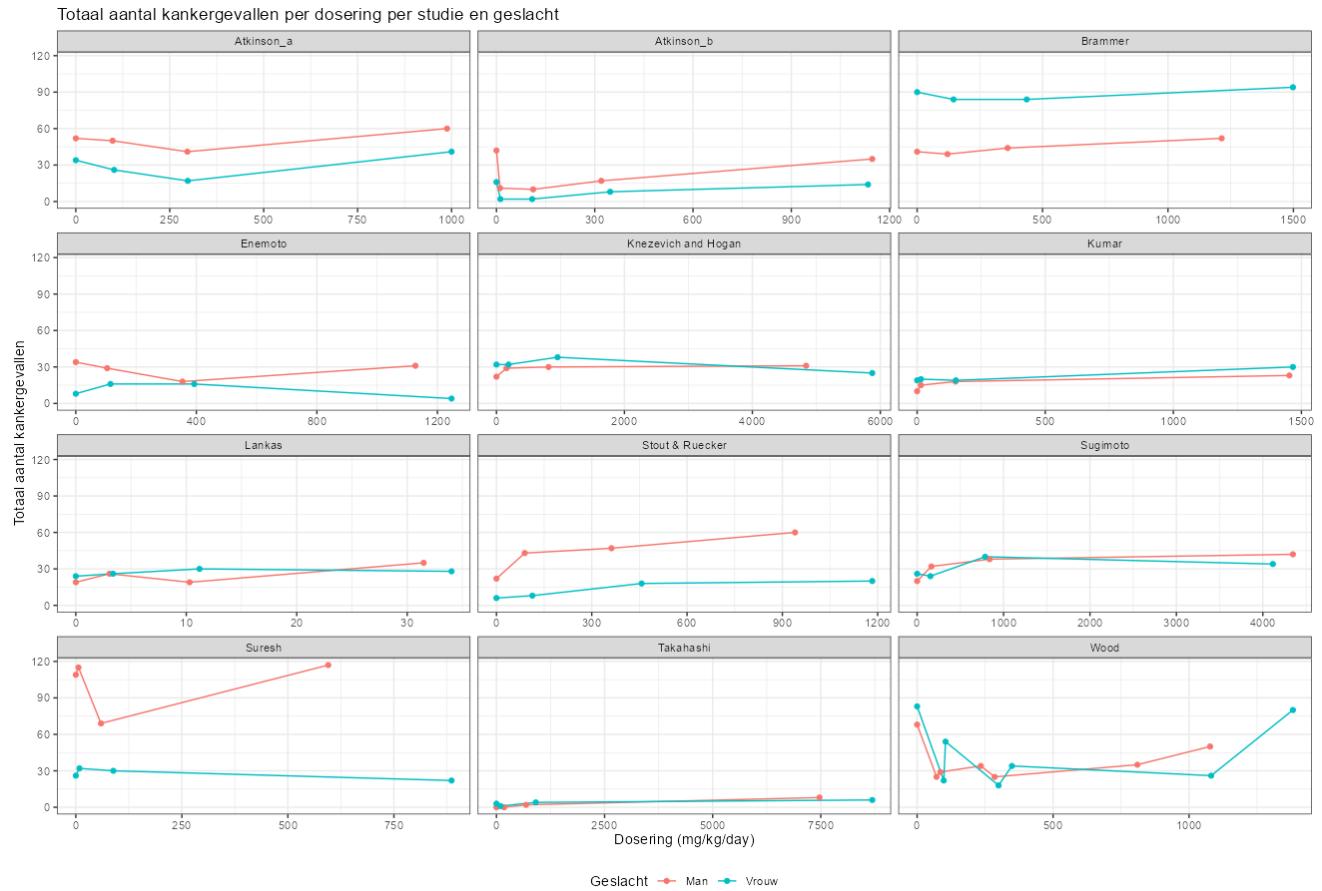
Hieruit blijkt geen eenduidig antwoord op de vraag of we meerdere tumoren zien bij meerdere dieren en/of meerdere tumoren bij één enkel dier. Wat we wel kunnen doen is het zekere voor het onzekere nemen door grafieken maken waarbij we het aantal tumoren tellen per dosering terwijl de groepsgrootte gelijk blijft. Omdat een studie niet van tevoren is ontworpen om een bepaalde tumorsoort wel of niet te ontdekken, betekent dit dat we voor een bepaalde dosering een X aantal dieren hebben waarbinnen X aantal tumoren worden gezien. Dit geeft een andere kijk op de zaak<sup>72</sup>. We kunnen dit gegeven nu op een aantal manieren visualiseren. Om te beginnen kunnen we het totaal aantal tumoren tellen per studie over tijd. Omdat de doseringen uitlopen kunnen we de studies apart tonen. We krijgen dan **Figuur 67**.



**Figuur 67.** Totaal aantal kankergevallen per dosering per studie.

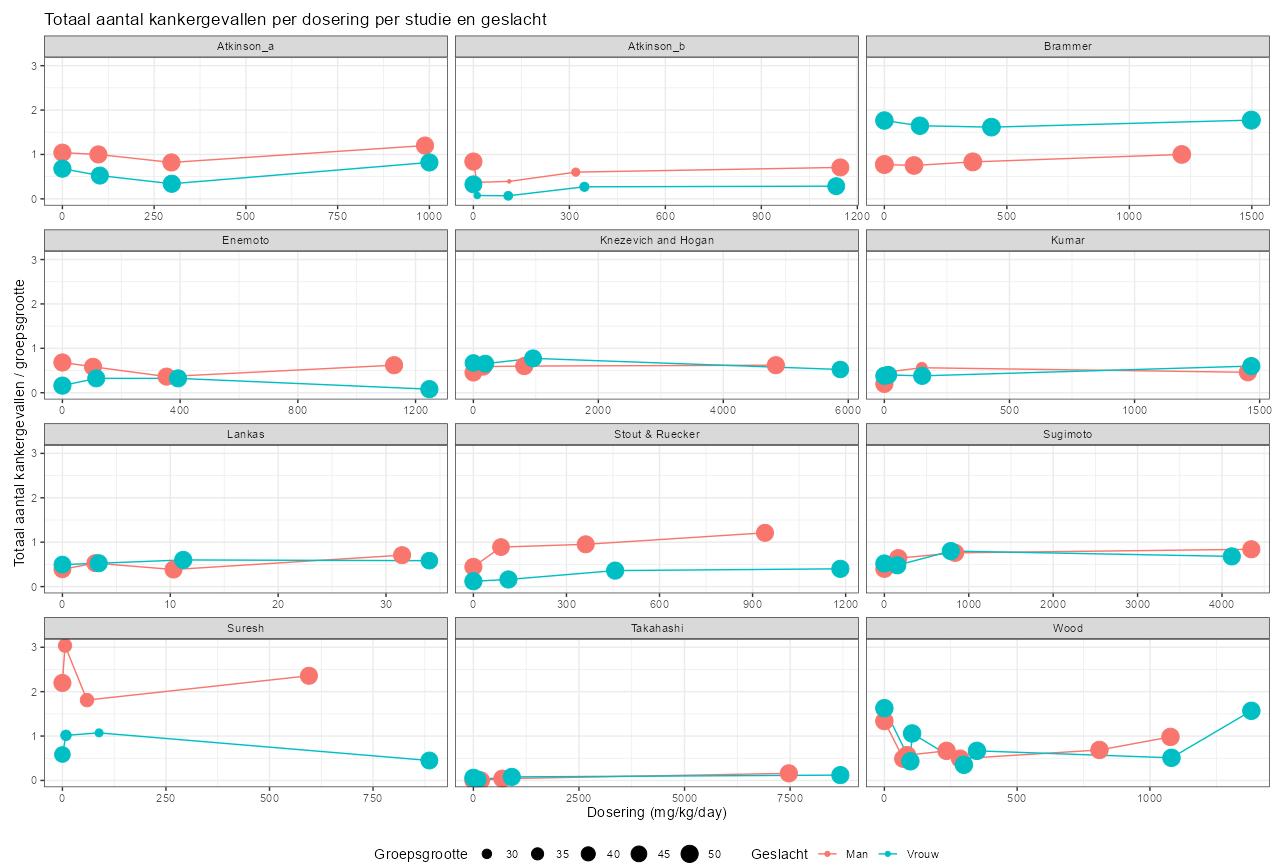
<sup>72</sup> Maar doet niets af aan het eerder commentaar dat tumoren die niet worden gezien ook niet worden meegeteld, wat niet correct is.

Als we het splitsen per geslacht krijgen we **Figuur 68**. Deze laatste figuur laat parallelle lijnen zien. Het lijkt er niet evident op dat er sprake is van een interactie tussen dosering en geslacht.



**Figuur 68.** Totaal aantal kankergevallen per studie, dosering en geslacht.

Wat we nu nog meer kunnen doen is deze aantallen in proportie zetten tegen het aantal geïncludeerde dieren (de deler) per dosering per studie. Het mag duidelijk zijn dat daarmee het totaal aantal kankergevallen, of tumorsoorten, groter kan zijn dan het aantal dieren. Dit komt omdat er mogelijke meerdere tumorsoorten in één enkel dier zitten. Dit alles stelt ons in staat om de invloed van de deler mee te nemen en het resultaat laat zich zien in **Figuur 69**: de groepsgrootte is in deze figuur opgenomen in zowel de y-as als de grootte van de bollen.



**Figuur 69.** Totaal aantal kankergevallen in verhouding tot groepsgrootte per studie, dosering en geslacht.

## Wat kunnen we hieruit afleiden?

We blijven keer op keer zien dat een groot deel van de tumorsoorten bij de nul-dosering plaatsvindt. Ook zien we dat studies verschillen in de soort tumoren én de hoeveelheid per tumorsoort. Dit alles maakt dat zowel het groeperen van studies en ook het analyseren van één enkele studie lastig is. Vergelijken we doseringen over alle tumorsoorten heen dan hebben we al snel te maken met een hele hoop berekeningen<sup>73</sup>.

Toch wordt het nu echt tijd om dieper in de berekeningen te duiken. Niet alleen omdat grafieken ons maar tot een bepaald niveau kunnen brengen, maar ook omdat er heel veel berekeningen zijn gedaan in de studie van Portier zelf. En het zijn juist die berekeningen waarom ik dit rapport nu opstel. Deze berekeningen moeten daarom echt het uitgangspunt vormen van dit rapport. De rest volgt daarna.

<sup>73</sup> Die vervolgens weer leiden tot een hoop p-waarden.

## Glyfosaat: eenzijdig en tweezijdig toetsen

---

*The choice of whether to use a one- or two-side test should be made at the design rather than the analysis stage. A two-sided statistical hypothesis test tests for a difference from the negative control (in a pair-wise comparison) in either direction. A one-sided comparison tests for a difference in only one pre-specified direction, but as a consequence has more power. In a carcinogenicity study, the expectation is often that the change will be an increase in tumours in the treated group so a onesided test may be considered more appropriate, although this can be controversial. If the treatment could also be protective (i.e., reduce tumour incidence or delay it) then a two-sided comparison may be more appropriate. Regulatory authorities may have specific opinions. For instance, the US EPA (2005) notes that either "a two-tailed test or a one-tailed test may be used".<sup>74</sup>*

---

Bovenstaande quote komt van de OECD richtlijn waarnaar werd gerefereerd in de BNNVARA/Zembla reportage. In deze quote staat dat éénzijdig testen een controversiële stap is, maar laat dit door Geert de Snoo en collega's nu juist als de meest noodzakelijk stap worden gezien: een tweezijdige toets kijkt naar twee richtingen (positief of negatief effect) terwijl een eenzijdige toets maar naar één richting kijkt (positief óf negatief). Volgens de Snoo **moeten** we wel eenzijdig testen, want een beschermend effect van glyfosaat is haast onmogelijk. Zodra we dit doen zullen de resultaten zien wat ze moeten aantonen.

Toch is dit niet geheel wat er met één-of tweezijdig toetsen wordt bedoeld. Het gaat er bij één-of tweezijdig toetsen niet om of een interventie beschermt of niet, maar juist of het mogelijk is dat een effect één of twee richtingen kan hebben. Het is dus vooral een toets die wordt ingezet wanneer het signaal helemaal niet zo duidelijk is.

Hoewel het nu wellicht klinkt alsof ik twee eenzelfde zaken verschillend behandel is dit niet het geval: in de statistiek draait het namelijk om de schatting van een effect. Deze schattingen kennen vaak geen harde afkapwaarden<sup>75</sup> en men zal dus rekening moeten

---

<sup>74</sup> [https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453\\_9789264221475-en](https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453_9789264221475-en)

<sup>75</sup> Een uitzondering zou het aantal graden Kelvin zijn, maar aantal graden Celsius kan zowel positief als negatief zijn.

houden met positieve én negatieve waarden (zie het voorbeeld van de lengte van Nederlandse mannen en vrouwen). Om dit in ogenschouw te nemen is er voor gekozen om de afkapwaarde ook meer stringent te maken: je wil als onderzoeker zeker weten dat een effect ook een signaal is én niet gedeeltelijk ruis. Een schatting moet dus enige waarde kennen. Dit alles staat los van de problemen van het frequentistische statistisch testen. Het gaat er dus gewoonweg om dat we van tevoren openhouden dat een schatting beide kanten op kan gaan puur en alleen omdat dit fysisch mogelijk is.

In het geval van glyfosaat betekent dit niets anders dan dat in de nul-dosering meer kanker wordt gevonden dan in de doseringen MET glyfosaat. De relatie is dan zo variabel dat het gemiddelde negatief is (ook al is de relatie te variabel om dat te zeggen). Dit wil dan niet direct zeggen dat glyfosaat beschermd, maar wel dat er meer bewijs nodig is om te zeggen dat het schadelijk is. En als we kijken naar de grafieken in het vorige hoofdstuk dan is dan is die stelling niet ver verwijderd van wat wij observeren. De proportie kanker is namelijk het hoogst in de nul-dosering, ook al zitten hier niet minder dieren vergeleken met de andere doseringen.

Het is nu tijd om het werk van Portier te gaan controleren. Voordat we ons weer buigen over het modelleren van een dose-response relatie gaan we eerst proberen de gerapporteerde p-waarden te repliceren. Daarmee herhaal ik de procedure zoals ik deze in **Tabel 16** en **Tabel 17** al liet zien.

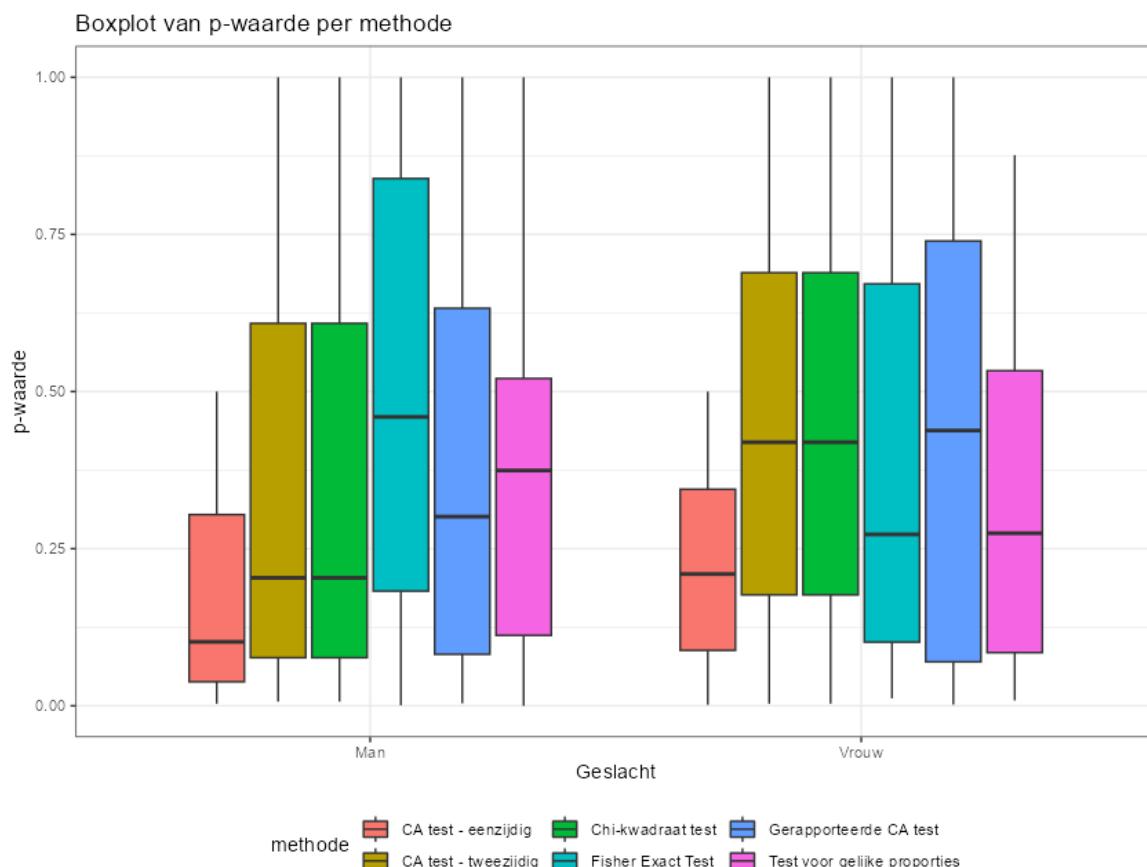
## Het werk van Portier controleren

Dit deel van het rapport is wellicht het meest belangrijke, maar ook procedureel gezien het meest eenvoudige. Het enige wat ik gedaan heb is de data, zoals gerapporteerd in Supplementary Material 2, inladen in een statistiekprogramma en dan per tumorsoort, per geslacht én per studie vier verschillende testen uitrekenen. Dit zijn niet de enige testen die je zou kunnen uitvoeren, maar het zijn wel de meest gangbare gezien de data. Van de CA-test heb ik zowel de eenzijdige als tweezijdige test toegevoegd. Het resultaat laat zich zien in **Tabel 18**. Ik heb per methode 183 toetsen uitgevoerd. Van 23 mogelijke toetsen was de data te summier om tot een zinvol antwoord te komen.

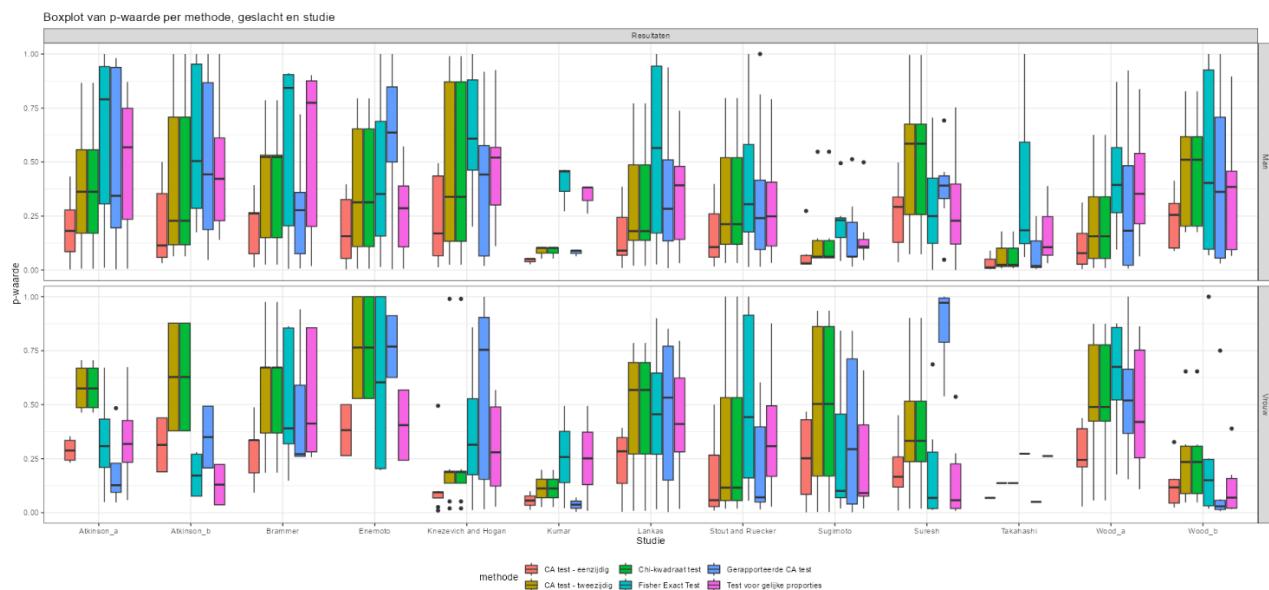
Methode	NA	$p$ -waarde $\leq 0.05$		$p$ -waarde $\leq 0.01$	
		Ja	Nee	Ja	Nee
CA test – eenzijdig	23	43	140	10	173
CA test – tweezijdig	23	21	162	7	176
Chi-kwadraat test - tweezijdig	23	21	162	7	176
Fisher Exact test - tweezijdig	23	22	161	4	179
Test voor gelijke proporties - tweezijdig	23	26	157	6	177
Gerapporteerde CA Test Portier - tweezijdig	23	34	149	12	171

**Tabel 18.** Aantal significante resultaten, op basis van een andere grenswaarde. Elke methode heeft 183 toetsen uitgevoerd per grenswaarde. Onderaan staat het aantal statistisch significante resultaten zoals gerapporteerd door Portier. NA: *not applicable* oftewel de toets kan niet worden uitgevoerd.

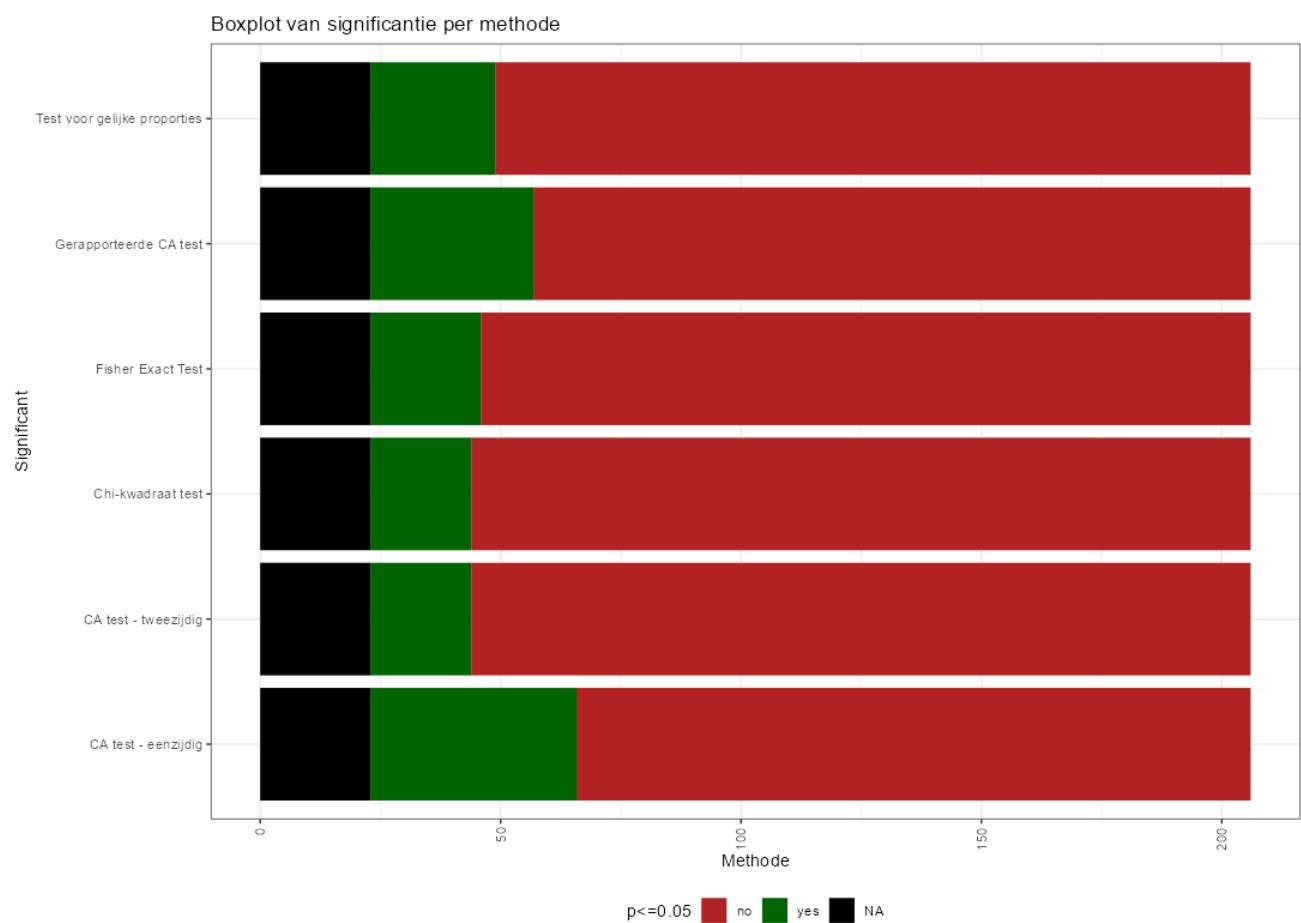
Belangrijk is hoe zeer de verschillende testen van elkaar verschillen. Dit kan ik grafisch laten via **Figuur 70 t/m Figuur 74**.



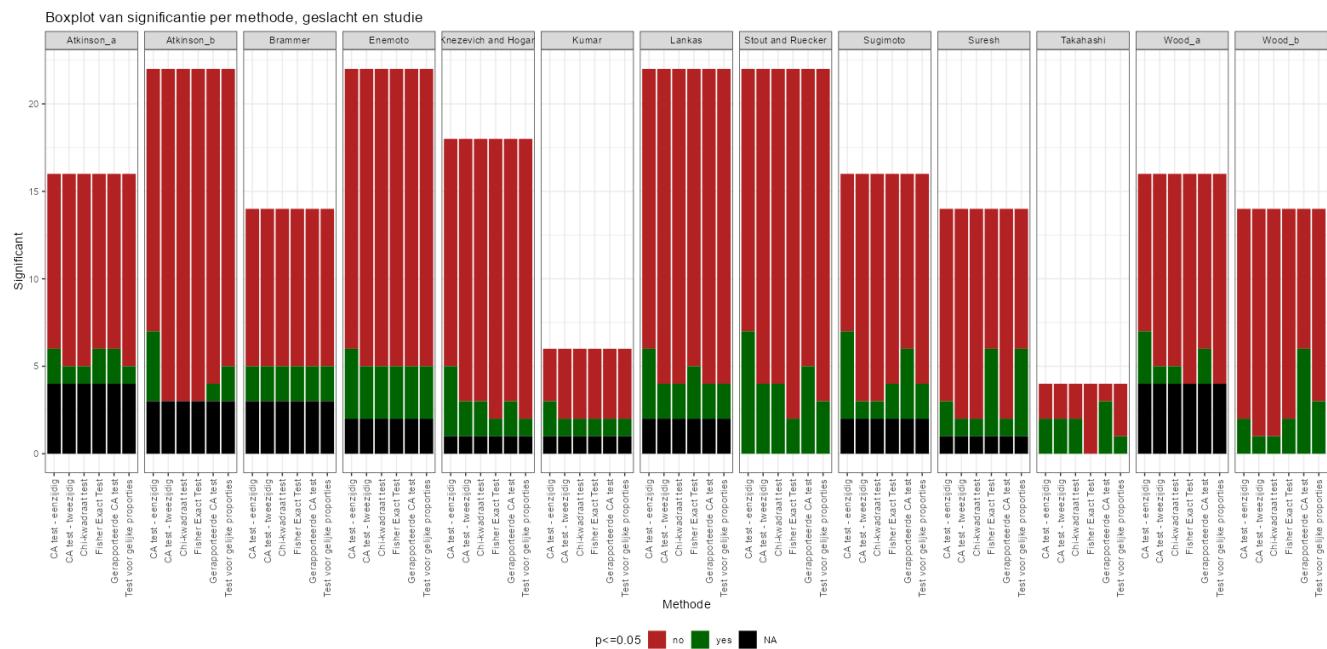
**Figuur 70.** Boxplot die de verdeling van  $p$ -waarden laat zien per methode én geslacht.



**Figuur 71.** Boxplot die de verdeling van p-waarden laat zien per methode, geslacht én studie.



**Figuur 72.** Boxplot die per methode het aantal significante resultaten laat zien voor  $p \leq 0.05$ . Daarmee is het een visuele representatie van **Tabel 18**. De eenzijdige CA test zoals door mij uitgevoerd laat het grootste aantal significante resultaten zien. Daarna komen de resultaten van Portier: het aantal ligt tussen de eenzijdige en tweezijdige CA test.



Figuur 73. Boxplot die per methode en per studie het aantal significante resultaten laat zien voor  $p \leq 0.05$ .



Figuur 74. Heatmap die het aantal statistisch significant bevindingen laat zien per tumorsoort én per methode. De heatmap is zo gemaakt dat een rij alleen wordt aangemaakt als er ook maar één statistisch significant resultaat is per tumorsoort. Er is uiteraard naar meer tumorsoorten gekeken. Wat opvalt is dat de gerapporteerde Portier resultaten bijna altijd de meest significante verschillen laat zien per tumorsoort.

Wat in elk van deze figuren zichtbaar is, direct, is dat het aantal statistisch significante resultaten helemaal niet zo groot is. Dat hebben we al eerder benoemd. Ook blijkt dat elk van de gekozen methoden, los van de eenzijdige CA-toets, minder statistisch significante resultaten laat zien dat Portier zelf. In afwachting van zijn antwoord op mijn vragen kunnen we niets anders dan dit nu voor lief nemen.

Wat we wel tot ons mogen nemen is dat ook ik statistisch significante resultaten krijg als ik de methoden van Portier overneem. Het is dus niet alsof ik geen resultaten zie. Zo gek mag dat ook niet klinken, want de gebruikte methoden komen rechtstreeks uit de OECD rapportages. Zo is de CA-test een meer dan gangbare test voor beoordelen van mogelijk carcinogene bijwerken. Ook het rapporteren per tumorsoort is gangbaar. Toch kan ik, vanuit een statistisch oogpunt, niet meegaan op de manier waarop dit door Portier nu is gedaan, maar daarover later meer. Zover zijn we namelijk nog niet.

Alvorens we het probleem van meerdere testen gaan bespreken is het wellicht verstandiger om het eerst te hebben over de data die we **niet** zien, maar die wel relevant genoeg zijn om mee te nemen. Ik spreek dan niet dat we een aantal statistische toetsen zijn vergeten, maar eerder dat er in sommige gevallen wel tumorsoorten zijn meegenomen waarvoor geen observaties zijn. In andere gevallen is dat dan weer niet het geval en deze keuzes zijn onnavolgbaar. Ik zal nu uitleggen waarom het uiterst belangrijk is om sowieso mee te nemen wat er niet is geobserveerd.

## Wat als we nu meenemen wat we niet zagen?

---

*Statistical analysis of a long-term study should be performed for each tumor type separately. The incidence of benign and malignant lesions of the same cell type, usually within a single tissue or organ, are considered separately and are then combined when scientifically defensible (McConnell et al., 1986).<sup>76</sup>*

---

<sup>76</sup> [https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453\\_9789264221475-en](https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453_9789264221475-en)

Bovenstaand citaat uit de OECD documentatie is eigenlijk volstrekt: het is standaard om voor elke tumorsoort een aparte analyse uit te voeren. Dit is ook de reden, zo denk ik, waarom er door Portier zoveel statistische testen zijn uitgevoerd:

*Individual tumor counts for the individual studies are reanalyzed using the exact form of the Cochran-Armitage (C-A) linear trend test in proportions [37]. Reanalyses are conducted on all primary tumors where there are at least 3 tumors in all of the animals in a sex. Tables S1–S13 provide the tumor count data for all tumors with a significant trend test ( $p \leq 0.05$ ) in at least one study of the same sex/species/strain along with the doses used (mg/kg/day) and the number of animals examined microscopically in each group. Pairwise comparisons between individual exposed groups and control are conducted using Fisher's exact test [37] and are provided for comparison with other reviews.*

Uit bovenstaande denk ik dat de belangrijkste zin deze is: "*tumor count data for all tumors with a significant trend test in at least one study of the same sex/species/strain*". Het lijkt er dus op dat als ook maar één enkele studie een positieve trend test laat zien in één specifieke tumor/geslacht/soort/genotype klasse, dit voor de rest van de studies ook gerapporteerd wordt. Dit zou betekenen dat als er één studie is die een positieve test laat zien voor bijvoorbeeld tumor X/geslacht Y/Soort Q én genotype Z we getallen zullen zien voor alle andere studies ongeacht of de tumor wel of niet gezien is. We zouden daarmee een behoorlijke reeks rijen moeten zien met nul observaties.

Laten we dit eens testen door een eerste significante test tot ons te nemen.

Gemakshalve zal ik **Figuur 55** hier nog een keer tonen en de bovenste rij nemen: *Kidney Adenoma (original pathology)* bij mannelijke CD-1 muizen. Deze is namelijk positief op basis van de Trend Test. Als ik de tekst goed begrijp dan zou ik nu voor elke studie die CD1 muizen heeft geïncludeerd eenzelfde rij moeten zien. Ook als er geen kanker is geobserveerd.

**Table S1.** Tumors of interest in male and female CD-1 mice from the 24-month feeding study of Knezevich and Hogan (1983) [11] – Study A

Tumor	Doses (mg/kg/day) or Tumor Incidence <sup>1</sup>			Trend Test p-value
Males	0	157	814	4841
Kidney Adenomas (original pathology)	0/49	0/49	1/50	3/50
Kidney Adenomas <sup>2</sup>	1/49	0/49	0/50	1/50
Kidney Carcinomas <sup>2</sup>	0/49	0/49	1/50	2/50
Kidney Adenomas and Carcinomas <sup>2</sup>	1/49	0/49	1/50	3/50
Malignant Lymphomas	2/49	5/49	4/50	2/50
Hemangiosarcomas <sup>3</sup>	0/49	0/49	1/50	0/50
Alveolar-Bronchiolar Adenomas	5/48	9/50	9/50	9/50
Alveolar-Bronchiolar Carcinomas	4/48	3/50	2/50	1/50
Alveolar-Bronchiolar Adenomas and Carcinomas	9/48	12/50	11/50	10/50
Females	0	190	955	5874
Hemangiomas	0/49	1/49	1/50	0/50
Harderian Gland Adenomas	2/45	0/48	1/49	0/44
Harderian Gland Carcinomas	0/45	0/48	0/49	0/44
Harderian Gland Adenomas and Carcinomas	2/45	0/48	1/49	0/44
Alveolar-Bronchiolar Adenomas	10/49	9/50	10/49	1/50
Alveolar-Bronchiolar Carcinomas	1/49	3/50	4/49	4/50
Alveolar-Bronchiolar Adenomas and Carcinomas	11/49	12/50	14/49	5/50
Spleen Composite Lymphosarcoma	1/50	1/48	1/49	5/49
Malignant Lymphomas	5/49	6/49	6/49	10/49

1 – Doses are given in the rows marked “Males” and “Females”, tumor counts appear on the rows with the individual tumors; 2 – tumor counts obtained from EPA]; \* 0.01<p≤0.05 for Fisher’s Exact Test; \*\* p≤0.01 for Fisher’s Exact Test

Een direct zoektocht in de data laat zien dat alleen voor de studie van Knezevich & Hogan de tumorsoort *Kidney Adenoma (original pathology)* wordt genoemd. Mijn interpretatie is dus niet correct. Laat ik het nog een keer proberen, maar dan bij de vrouwen. Ik neem nu de tumorsoort *Spleen Composite Lymphosarcoma*, maar het resultaat blijft hetzelfde. Alleen deze studie rapporteert deze tumorsoort.

Een vlugge analyse laat zien dat over 13 studies heen we 48 verschillende tumorsoorten vinden. Als we dit opsplitsen per geslacht dan zien we 33 unieke tumorsoorten bij de vrouwen en 25 bij de mannen<sup>77</sup> - dit is veel meer dan de 9 tumorsoorten die we zien bij de mannen én de vrouwen in de Knezevich & Hogan studie (die tezamen 14 unieke tumorsoorten tonen).

In afwachting van een antwoord gegeven door Portier moeten we de data nemen zoals deze zijn, wat betekent dat we niet in elke studie dezelfde tumorsoort zien. Dit heeft wel implicaties voor de analyses en dat kunnen we wellicht het best aantonen als we een

<sup>77</sup> Dat dit tezamen geen 48 is komt natuurlijk omdat ze elkaar niet uitsluiten.

tumoroort nemen die wel wordt gerapporteerd, maar in één enkele studie geen enkele incidentie kent.

Bij de vrouwen zien we dat voor *Harderian Gland Carcinomas* er géén incidenties zijn, maar de tumor wordt wel genoemd. Een statistische analyse is dan niet mogelijk, maar de rij staat er wel. Als ik zoek naar welke studies deze rij nog meer hebben meegenomen dan vind ik vier studies. In drie van die studies is er geen enkele incidentie wat betekent dat deze tumoroort wel wordt gerapporteerd maar niet is geobserveerd in die studies. In lijn met eerdere uitspraken van Portier zien we dat de *Harderian Gland Carcinomas* alleen in CD-1 muizen wordt gerapporteerd. Het lijkt dus te kloppen dat een tumoroort die ergens wel voorkomt (in de studie van Wood welteverstaan) ook wordt gerapporteerd in andere studies: ook als deze tumor niet gezien wordt. Dit zorgt voor een meer eerlijke vergelijking.

Maar we blijven wel met een vraag zitten. Waarom zien we dit wel voor *Harderian Gland Carcinomas*, maar niet bij *Kidney Adenoma (original pathology)*? Er zijn namelijk vijf studies met mannelijke CD-1 muizen: Atkinson (1993)<sup>78</sup>, Knezevich & Hogan, Sugimoto, Takahashi en Wood. Deze discrepanties zijn niet verklaarbaar door alleen maar te kijken naar de data.

Een kleine zoektocht laat zien dat voor 16 tumoroorten er rijen werden toegevoegd aan 10 studies waarbij er geen incidentie is<sup>79</sup>. Als we de tweede uit de lijst nemen (*Hemangiomas*) dan zien we dat deze is opgenomen in vier studies die wederom alleen CD-1 vrouwelijke muizen hebben geïncludeerd. Een diepere kijk op de tumoroort aan het einde van de lijst (*Pituitary Carcinomas*) toont drie studies<sup>80</sup>. De studie van Brammer komt hier twee keer naar voren: één keer voor vrouwelijke Wistar ratten en één keer voor mannelijke. Dit is dus een studie die zelf geen tumor heeft gezien in deze vorm, maar wel heeft gerapporteerd. Dat is netjes.

Als we dus zoeken naar tumoroorten die zijn gerapporteerd in de tabellen van Portier, maar geen incidentie tonen dan zie ik inderdaad heel netjes dat een tumoroort die wordt gezien in één studie in één segment in de combinatie geslacht/soort/tumor/genotype.

---

<sup>78</sup> Er zijn twee Atkinson studies geïncludeerd.

<sup>79</sup> Harderian Gland Carcinomas, Hemangiomas, Harderian Gland Adenomas, Harderian Gland Adenomas and Carcinomas Carcinomas, Kidney Adenomas, Kidney Adenomas and Carcinomas, Hemangiosarcomas, Thyroid Follicular-Cell Carcinomas, Skin Keratoacanthomas, Pancreas Islet Cell Carcinomas, Thyroid C-cell Carcinomas, Adrenal Cortical Adenoma, Adrenal Cortical Carcinoma, Hepatocellular Carcinomas, Pituitary Carcinomas

<sup>80</sup> Brammer, Suresh, Wood

Maar als we kijken naar de tumorsoorten die wel gezien worden, dan is het niet zo dat we die overal zien. De studie van Knezevich & Hogan laat dit al zien, omdat de tumorsoort *Kidney Adenomas (original pathology)* niet overal wordt getoond: in de vijf unieke studies die kijken naar mannelijke CD-1 muizen zien we maar één studie die kijkt naar deze tumorsoort, namelijk die van Knezevich & Hogan.

Toch voelt het alsof ik ergens iets mis. Om er zeker van te zijn dat ik hier geen punt ga maken van iets wat geen punt is, moet ik er zeker van zijn dat er studies zijn waarin tumorsoorten niet worden meegenomen die ergens anders wel worden gerapporteerd. Wat ik kan doen is voor één bepaalde groep (bijvoorbeeld mannelijke CD-1 mannelijke muizen) laten zien hoe vaak een tumorsoort wordt gerapporteerd. Ik weet dat er vijf studies zijn met deze muizen en dus zou ik elke tumorsoort vijf keer moeten zien terugkomen: alleen dan weet ik dat een studie telkens de tumor meeneemt ongeacht de incidentie.

Een snelle berekening laat zien dat in deze vijf studies negen unieke tumoren worden getoond. Deze tumoren worden echter niet in alle vijf de studies gezien (**Tabel 19**).

Tumorsoort	Hoeveel studies	Géén incidentie
Alveolar-Bronchiolar Adenomas	4	Nee
Alveolar-Bronchiolar Adenomas and Carcinomas	4	Nee
Alveolar-Bronchiolar Carcinomas	4	Nee
Hemangiosarcomas	4	Ja
Kidney Adenomas	5	Ja
Kidney Adenomas (original pathology)	1	Nee
Kidney Adenomas and Carcinomas	5	Ja
Kidney Carcinomas	5	Ja
Malignant Lymphomas	4	Nee

**Tabel 19.** Aantal studies met mannelijke CD-1 muizen per tumorsoort. Wat deze tabel laat zien is dat in vijf studie waarin mannelijke CD-1 muizen werden onderzocht er in totaal 9 unieke tumorsoorten werden gezien, maar niet elke studie heeft bericht over elke tumorsoort. Dit zou wel moeten.

Om er zeker van te zijn wil ik kijken naar de bovenste rij: *Alveolar-Bronchiolar Adenomas*. Deze wordt inderdaad gerapporteerd in Atkinson (1993), Knezevich & Hogan, Sugimoto en Wood, en ontbreekt daarmee dus in Takahashi. In die studie had ik nu dus verwacht dat ik deze tumorsoort wel zou zien, maar dan met geen incidentie. In de andere studies zien wel incidentie.

Nu is de vraag of er in deze negen tumorsoorten rijen zijn met géén incidentie. Dit laat ik ook zien in **Tabel 19** en om eerlijk te zijn kan ik er geen touw aan vastknopen. Wat ik

eigenlijk had verwacht te zien is dat elke tumorsoort in elke studie zichtbaar zou zijn. Maar als ik per studie kijk dan zie ik dat dit niet het geval is (**Tabel 20**).

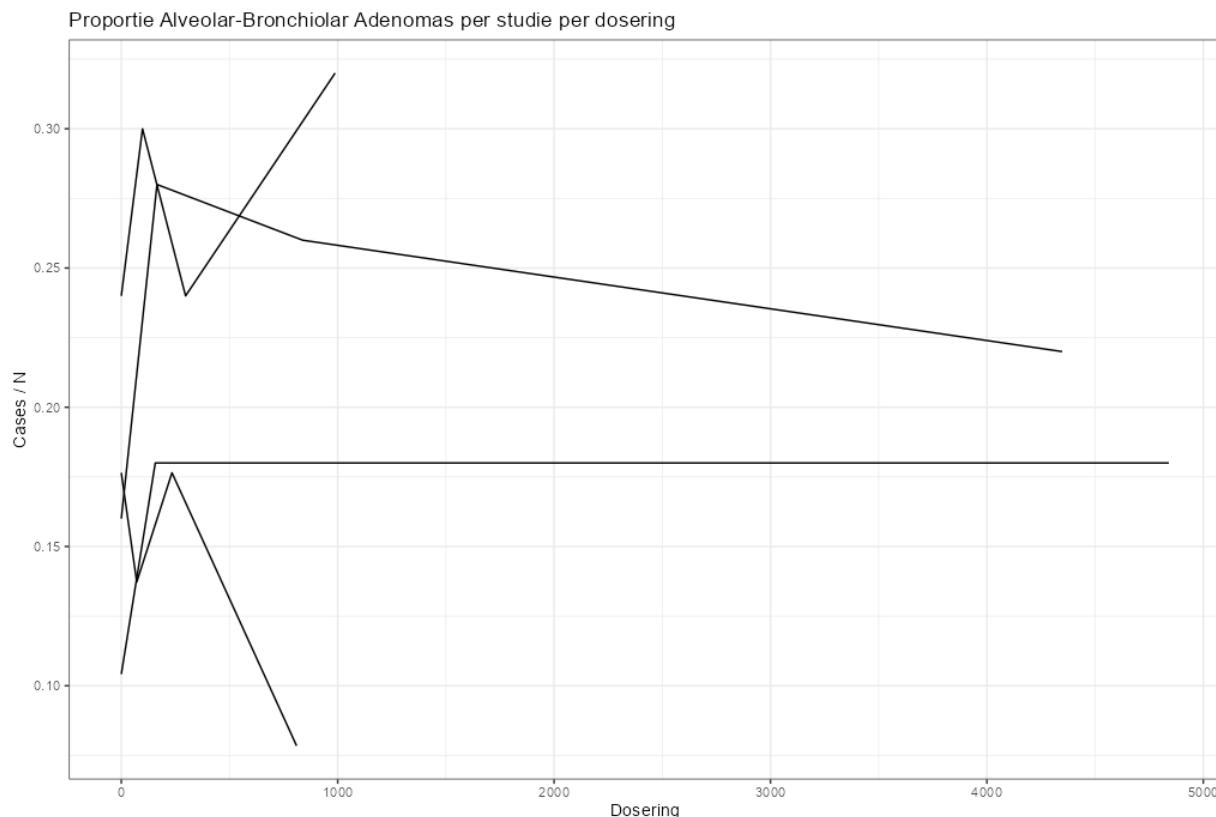
<b>Studie</b>	<b>Aantal tumoren</b>
Atkinson (1993)	8
Knezevich & Hogan	9
Sugimoto	8
Takahashi	3
Wood	8

**Tabel 20.** Aantal tumoren per studie met mannelijke CD-1 muizen.

In totaal werden negen unieke tumorsoorten gemeld.

Hoewel ik nu voor elke combinatie van geslacht/soort/tumor/genotype een beoordeling zou kunnen maken, is deze exercitie al voldoende om te zeggen dat er tumoren ontbreken: namelijk de tumoren waarbij er geen incidentie is. De tumoren die wel zonder incidentie worden genoemd hebben allemaal een evenknie waarbij er wél incidentie wordt vermeld.

Laten we gemakshalve daarom blijven bij de mannelijke CD-1 muizen. We weten dat er negen unieke tumorsoorten zijn en we verwachten dus dat elke tumorsoort gerapporteerd wordt in elk van de vijf studies. Dat is hier niet het geval. Daarmee ontbreken dus in totaal 8 rijen aan ‘data’: één rij bij Atkinson (1993), één rij bij Sugimoto, vijf rijen bij Takahashi en één rij bij Wood. Zouden wij die rijen toevoegen dan zouden de resultaten er hoogstwaarschijnlijk anders uitzien. Hoe anders kunnen we voor dit voorbeeld gewoon uitrekenen. Voor *Alveolar-Bronchiolar Adenomas* missen we exact één rij. De data die we wel hebben ziet er als volgt uit (**Figuur 75**).



**Figuur 75.** Dose-response relatie voor vier studies met data voor *Alveolar-Bronchiolar Adenomas*. Een vijfde studie ontbreekt. Zouden we deze studie toevoegen dan zou de grafiek er niet anders uitzien. Daar zit dus niet het verschil.

Als ik in deze figuur een rij zou toevoegen met géén incidentie heb dan zou die lijn niet zichtbaar zijn. Er zou ook geen impact zijn<sup>81</sup>. Dat kan dus beter en wat we kunnen doen is een nu-dosering afzetten tegen een niet-nul-dosering. Daarmee voegen we dus alle doseringen samen en tellen alleen de incidentie in die groep. We kunnen dan toevoegen wat we missen. In tabelvorm zou het er als volgt uitzien (**Tabel 21**):

Tumoroort	Aantal studies	Nul-dosering			Niet-nul-dosering			
		Cases	N	Ratio	Cases	N	Ratio	
Alveolar-Bronchiolar Adenomas	4	34	199	0.171	128	603	0.212	<b>0.041</b>
Alveolar-Bronchiolar Adenomas	5	34	249 <sup>82</sup>	0.136	128	753 <sup>83</sup>	0.170	<b>0.034</b>
<b>Verschil</b>				<b>0.035</b>			<b>0.042</b>	<b>0.007</b>

**Tabel 21.** Aantal studies met mannelijke CD-1 muizen met data over Alveolar-Bronchiolar Adenomas. De bovenste rij is wat we kunnen opmaken uit de studie van Portier. De onderste rij als we de studie zouden toevoegen met geen incidentie. Dit verandert de ratios.

<sup>81</sup> Dat komt natuurlijk omdat we de studies niet samenvoegen. Als we dit wel zouden doen dan zou het wel gewicht krijgen.

<sup>82</sup> 50 muizen toegevoegd omdat de gemiddelde nul-dosering een N had van 50.

<sup>83</sup> 150 muizen toegevoegd omdat de gemiddelde studie vier doseringen had met een N=50 bij elk. Dat betekent dat we gemiddeld genomen drie groepen met elk N=50 moeten toevoegen.

Wat direct opvalt is dat de toevoeging van één rij met geen incidentie de ratio met ongeveer eenzelfde verschil vermindert. Strikt genomen verandert er niks in de vergelijking: deze was 0.041 en zou dan hypothetisch veranderen in 0.034 – een verschil van 0.007. Eigenlijk is dit niet noemenswaardig. Met eenzelfde verschil gaan de absolute waarden omlaag.

Maar wat al snel vergeten wordt is dat we hier spreken van één enkele toevoeging.

Hoe anders ziet het eruit als we kijken naar *Kidney Adenomas (original pathology)*. Het toevoegen van vier studies met nul incidentie laat het verschil in ratio's nagenoeg halveren (**Tabel 22**). Toegegeven: een stijging van 0.0268 was al niet veel (het gaat om ongeveer 3%), maar de daling naar 1% maakt het resultaat niet noemenswaardig. Dit alles zijn dus wel zaken om rekening mee te houden.

Tumornoort	Aantal studies	Nul-dosering			Niet-nul-dosering			
		Cases	N	Ratio	Cases	N	Ratio	
Kidney Adenomas (original pathology)	1	0	49	0	4	49	0.0268	<b>0.0268</b>
Kidney Adenomas (original pathology)	5	0	295	0	4	295	0.0136	<b>0.0136</b>
<b>Verschil</b>				<b>0</b>			<b>0.0132</b>	

**Tabel 22.** Aantal studies met mannelijke CD-1 muizen die data hebben gerapporteerd over *Kidney Adenomas (original pathology)*. De bovenste rij is wat we kunnen opmaken uit de studie van Portier. De onderste rij als we de studie zouden toevoegen die niks rapporteert over deze tumor.

Wat ik nu kan doen, is voor elke groep laten zien hoeveel unieke tumoren er zijn en in hoeveel studies deze gemeld worden. Dit laat zien waar er rijken ontbreken en wat de impact van die ontbrekende rijken is op de ratio's zoals deze nu worden gerapporteerd. Om dit te doen moeten we eerst alle unieke combinaties uit de matrix geslacht/soort/genotype halen (**Tabel 23**).

Geslacht	Ras	Soort			
		CD-1	Sprague-Dawley	Swiss-Albino	Wistar
Man	Muizen	144		16	
Vrouw	Muizen	132		8	
Man	Ratten		272		96
Vrouw	Ratten		102		72

**Tabel 23.** Aantal rijken met informatie per Geslacht / Ras / Soort.

We hebben in feite dus acht combinaties en voor die acht combinaties kunnen we achterhalen hoeveel unieke tumoren we hebben én hoeveel studies er zijn. Uit **Tabel 24**

kunnen we nog niet heel veel afleiden, alleen maar dat het aantal unieke tumoren verschilt per combinatie én dat we verwachten een X-aantal unieke tumoren te zien in een Y-aantal studies gegeven de combinatie.

Geslacht	Ras	Soort	Aantal unieke tumoren	Aantal studies
Man	Muizen	CD-1	9	5
Vrouw	Muizen	CD-1	9	5
Man	Muizen	Swiss-Albino	4	1
Vrouw	Muizen	Swiss-Albino	2	1
Man	Ratten	Sprague-Dawley	21	4
Vrouw	Ratten	Sprague-Dawley	9	4
Man	Ratten	Wistar	8	3
Vrouw	Ratten	Wistar	6	3

**Tabel 24.** Het aantal unieke tumoren per combinatie geslacht/ras/soort én het aantal studies waarin we deze combinatie zien. Het nut van deze getallen is dat we kunnen achterhalen hoe vaak een combinatie alle mogelijke tumorsoorten rapporteert. Bijvoorbeeld: in de vijf studies met mannelijke CD-1 muizen verwachten we dat elke studie getallen rapporteert over elk van de negen individuele tumoren.

Belangrijker is om per combinatie geslacht/ras/soort te achterhalen hoeveel studies er zijn en hoeveel tumoren elke studie rapporteert (**Tabel 25**). Hieruit kunnen we direct ontlenen welke studie bepaalde tumoren wel of niet heeft gerapporteerd.

Studie	Geslacht	Soort	Aantal unieke tumoren		Aantal missende tumoren
			per combinatie	in studie	
Atkinson 1993a	Man	CD-1	9	8	1
Atkinson 1993a	Vrouw	CD-1	9	7	2
Knezevich & Hogan	Man	CD-1	9	9	0
Knezevich & Hogan	Vrouw	CD-1	9	9	0
Sugimoto	Man	CD-1	9	8	1
Sugimoto	Vrouw	CD-1	9	8	1
Takahashi	Man	CD-1	9	3	6
Takahashi	Vrouw	CD-1	9	1	8
Wood 2009a	Man	CD-1	9	8	1
Wood 2009a	Vrouw	CD-1	9	8	1
Kumar	Man	Swiss-Albino	4	4	0
Kumar	Vrouw	Swiss-Albino	2	2	0
Atkinson 1993 b	Man	Sprague-Dawley	21	16	5
Atkinson 1993 b	Vrouw	Sprague-Dawley	9	6	3
Enemoto	Man	Sprague-Dawley	21	16	5
Enemoto	Vrouw	Sprague-Dawley	9	6	3

Lankas	Man	Sprague-Dawley	21	16	5
Lankas	Vrouw	Sprague-Dawley	9	6	3
Stout & Ruecker	Man	Sprague-Dawley	21	16	5
Stout & Ruecker	Vrouw	Sprague-Dawley	9	6	3
Brammer	Man	Wistar	8	8	0
Brammer	Vrouw	Wistar	6	6	0
Suresh	Man	Wistar	8	8	0
Suresh	Vrouw	Wistar	6	6	0
Wood 2009 b	Man	Wistar	8	8	0
Wood 2009 b	Vrouw	Wistar	6	6	0

**Tabel 25.** Aantal unieke tumoren per combinatie én in elke studie. De eerste rij laat zien dat Atkinsons 1993a mannelijke CD-1 muizen includeert. In die combinatie zijn negen unieke tumoren gerapporteerd over vijf studies heen. In de studie van Atkinson 1993a vinden we maar acht tumoren. Er ontbreekt er dus één. De studie van Stout & Ruecker rapporteert in de mannelijke Sprague-Dawley ratten 16 tumoren terwijl in die combinatie 21 unieke tumoren zijn gerapporteerd. Er ontbreken er dus zes.

We zien dus hier hoeveel rijen we per combinatie ‘missen’. Om de combinatie van dit gemis op een transparante manier te tonen, is het wellicht niet gek om per combinatie op te halen hoeveel significante verschillen er gevonden zijn. Voor elk significant verschil kunnen we dan achterhalen wat er met deze bevinding gebeurt als we de ‘missende’ rijen toevoegen.

We hebben in totaal 34 significante resultaten<sup>84</sup>. In elke studie valt minstens één statistisch significant verschil te vinden en het gros van die statistische bevindingen vinden we in de studie van Stout & Ruecker ( $n=5$ ). Deze studie heeft Sprague-Dawley ratten gebruikt en we weten dat bij deze studie 5 tumorsoorten niet zijn meegenomen bij de mannen en drie tumorsoorten bij de vrouwen. Dit maakt deze studie een mooie casus om te laten zien wat er gebeurt als we de tumorsoorten toevoegen die we niet hebben gezien, maar wel zouden moeten meetellen. Hieronder staat de tabel van Stout & Ruecker zoals aangegeven in Supplementary Material 2 (**Tabel 26**). We zien inderdaad vijf significante resultaten: drie bij de mannen en twee bij de vrouwen.

We kunnen nu een aantal dingen doen. Ten eerste kunnen we de rijen van de tumoren toevoegen waarvan we weten dat deze in deze combinatie voorkomt bij andere studies, maar niet in deze studie. Voor de analyse van deze studie zou het niets uitmaken, maar voor de analyse van de tumorsoort in zijn geheel wel. Daarmee zou de ratio significant / niet significant steeds meer in het voordeel van niet significant vallen.

<sup>84</sup> Dat wil zeggen, resultaten waarvan de p-waarde kleiner of gelijk is aan 0.05

Tumor	Doses (mg/kg/day) or Tumor Incidence <sup>1</sup>				Trend Test p-value
Males	0	89	362	940	
Testicular Interstitial Cell Tumors	2/50	0/50	3/50	2/50	0.296
Pancreas Islet Cell Adenomas	1/48	8/47*	5/50	7/49*	0.147
Pancreas Islet Cell Carcinomas	1/48	0/47	0/50	0/49	1.000
Pancreas Islet Cell Adenomas or Carcinomas	2/48	8/47*	5/50	7/49*	0.206
Thyroid C-cell Adenomas	0/50	4/48	8/48**	5/50*	0.089
Thyroid C-cell Carcinomas	0/50	2/48	0/48	1/50	0.442
Thyroid C-cell Adenomas and Carcinomas	0/50	6/50*	8/50**	6/50*	0.097
Thyroid Follicular-cell Adenomas	2/50	1/48	3/48	2/50	0.408
Thyroid Follicular-cell Carcinomas	0/50	0/48	0/48	1/50	0.255
Thyroid Follicular-cell Adenoma and Carcinoma	2/50	1/48	3/48	3/50	0.232
<b>Hepatocellular Adenomas</b>	<b>3/50</b>	<b>2/50</b>	<b>3/50</b>	<b>8/50</b>	<b>0.015</b>
Hepatocellular Carcinomas	3/50	2/50	1/50	2/50	0.637
<b>Hepatocellular Adenomas and Carcinomas</b>	<b>6/50</b>	<b>4/50</b>	<b>4/50</b>	<b>10/50</b>	<b>0.050</b>
Kidney Adenomas	0/50	2/50	0/50	0/50	0.813
<b>Skin Keratoacanthomas</b>	<b>0/49</b>	<b>3/46</b>	<b>4/50</b>	<b>5/48</b>	<b>0.042</b>
Skin Basal Cell Tumors	0/49	0/46	0/50	1/48	0.249
Females	0	113	457	1183	
<b>Thyroid C-cell Adenomas</b>	<b>2/50</b>	<b>2/50</b>	<b>6/50</b>	<b>6/50</b>	<b>0.049</b>
Thyroid C-cell Carcinomas	0/50	0/50	1/50	0/50	0.500
Thyroid C-cell Adenomas and Carcinomas	2/50	2/50	7/50	6/50	0.052
Adrenal Cortical Adenoma	1/50	2/50	2/50	1/50	0.603
<b>Adrenal Cortical Carcinoma</b>	<b>0/50</b>	<b>0/50</b>	<b>0/50</b>	<b>3/50</b>	<b>0.015</b>
Adrenal Cortical Adenoma and Carcinoma	1/50	2/50	2/50	4/50	0.090

**Tabel 26.** De resultaten van Stout & Ruecker zoals gerapporteerd in de Supplementary Material 2 file van Pointer. Zichtbaar zijn de 5 significante vergelijkingen (dikgedrukt): drie bij de mannen en twee bij de vrouwen.

Eigenlijk moet de focus dus niet op één enkele studie liggen, maar op de combinatie van studies. We krijgen daarmee een herhaling van zetten zoals gezien in **Tabel 21** en **Tabel 22**. Wat ik nu exact kan doen, is kijken welke tumorsoorten er ontbreken bij de Sprague-Dawley ratten.

We weten dat bij de mannelijke Sprague-Dawley ratten er 21 unieke tumoren zijn en bij de vrouwelijke Sprague-Dawley ratten zijn dat er negen. Ook weten we dat er voor elke combinatie vier studies gedaan zijn. Wat ik nu kan doen is die tumorsoorten selecteren

waarin de discrepantie tussen wat daadwerkelijk is gerapporteerd én wat men had moeten rapporteren het grootst is. Dit is bij de tumorsoorten die maar één enkele keer genoemd worden: *Liver Carcinomas*, *Liver Neoplastic Nodules*, *Liver Nodules and Carcinomas*, *Skin Epithelioma* en *Thyroid Follicular-Cell Adenomas and Carnomas*.

Tumorsoort	Aantal studies	Nul-dosering			Niet-nul-dosering		
		Cases	N	Ratio	Cases	N	Ratio
<i>Liver Carcinomas</i>	1	0	50	0	3	150	0.02
<i>Liver Carcinomas</i>	4	0	200	0	3	600	0.005
<i>Liver Neoplastic Nodules</i>	1	3	50	0.06	7	150	0.0467
<i>Liver Neoplastic Nodules</i>	4	3	200	0.015	7	600	0.0117
<i>Liver Nodules and Carcinomas</i>	1	3	50	0.06	10	150	0.0667
<i>Liver Nodules and Carcinomas</i>	4	3	200	0.015	10	600	0.0167
<i>Skin Epithelioma</i>	1	1	50	0.02	7	115	0.0609
<i>Skin Epithelioma</i>	4	1	200	0.005	7	460 <sup>85</sup>	0.0152
<i>Thyroid Follicular-Cell Adenomas and Carnomas</i>	1	1	47	0.0213	10	147	0.0680
<i>Thyroid Follicular-Cell Adenomas and Carnomas</i>	4	1	188	0.0053	10	588	0.0170

**Tabel 27.** Per tumorsoort de discrepantie tussen daadwerkelijk gerapporteerd en wat men had moeten rapporteren. Dit is alleen voor de Sprague-Dawley ratten.

Als we nu kijken in welke van deze studies de analyse door Portier significant is bevonden dan zien alleen bij *Skin Epithelioma (Keratoacanthomas)* een p-waarde  $\leq 0.05$  (0.047 in Atkinson 1993b). Hoewel het eigenlijk geen berekening meer behoeft om aan te tonen dat een toevoeging van rijen met nul incidentie de p-waarde laat stijgen, zullen we de exercitie toch voltooien. Om een gedegen simulatie te doen, moeten we de dosering toevoegen waarvan het resultaat zichtbaar is in **Tabel 28**. De gerapporteerde p-waarde wordt nu niet meer behaald: ongeacht of we éénzijdig of tweezijdig toetsen.

<sup>85</sup>  $(115/3)*3*3$  oftewel  $115*4$

Studie		Dosering					P-waarde
Atkinson 1993b		0	11	112	320	1147	
	Cases	1	2	0	0	5	0.047
	N	50	25	19	21	50	
Atkinson 1993b + drie studies met nul incidentie	Cases	1	2	0	0	5	0.1909 <sup>86</sup> 0.0954 <sup>87</sup>
	N	200	100	76	84	200	

**Tabel 28.** De p-waarde zoals gerapporteerd in Pointer voor Skin Epithelioma (Keratoacanthomas) in Atkinson 1993b.

Daaronder staat de p-waarde met de CA-test als we de resterende drie studies met nul incidentie toevoegen. De door mij berekende p-waarde verandert dan. Hierbij moet wel gesteld worden dat ik de gerapporteerde p-waarde nooit heb weten te repliceren.

Ik wil het voor nu bij deze voorbeelden laten. De reden is niet luiheid, maar eerder gepaste voorzichtigheid omdat we nu weten dat een groot deel van de studies niet alle data rapporteert. In feite wordt alleen gerapporteerd wat wordt geobserveerd. Wanneer er wél een tumorsoort wordt toegevoegd zonder incidentie is het niet direct duidelijk waarom. Ik heb dus eigenlijk geen idee waarom er soms wel nul-incidentie wordt toegevoegd en waarom op andere momenten weer niet. De constatering dat dit het geval zou eigenlijk voldoende moeten zijn om de bevindingen van Porties met gepaste voorzichtigheid te behandelen. Voordat we terugkeren bij dit onderdeel is het goed om stil te staan bij een veel bekender probleem: het probleem van meerdere testen<sup>88</sup>.

<sup>86</sup> Tweezijdige CA-test

<sup>87</sup> Éénzijdige CA-test

<sup>88</sup> Een aantal zaken die ik hier zal noemen heb ik ook genoemd in de controle van de Portier resultaten. Om dit rapport meer leesbaar te houden zullen er zo en dan doublures leesbaar zijn.

## Het probleem van meerdere testen

---

*The evaluation of any one animal cancer study involves a large number of statistical tests that could lead to false positives. To evaluate this issue, the probability that all of the results in any sex/species/strain could be due to false positive results is calculated. Overall, a total of 496 evaluations are done for these 13 studies including the few evaluations done against historical controls. There are 41 evaluations at 37 tumor/site combinations with a trend test  $p \leq 0.05$ ; the probability that all of these are due to false positives is 0.001. Similarly, looking at the evaluations resulting in  $p \leq 0.01$ , the probability that all of the findings are due to false positives is < 0.001. The strongest evidence is for male CD-1 mice, the probability of seeing 11 positive findings at  $p \leq 0.05$  and 8 at  $p \leq 0.01$  are both below 0.001. (see Additional file 2: Table S14).*

---

Bovenstaande quote is afkomstig uit de studie van Portier waarin hij dus zelf al het probleem van meerdere testen adresseert. Zelf geeft hij aan 496 testen te hebben uitgevoerd waarvan er 41 significant zijn met  $p \leq 0.05$ . Vervolgens benoemt hij iets wat ik als bijzonder zou bestempelen. Portier zegt namelijk dat dat de kans dat deze allemaal vals positief zijn gelijk is aan 0.001. Dit betekent dat als we de regels van de frequentistische statistiek volgen de nulhypothese zou moeten zijn dat elke bevinding vals positief is. De kans dat dit het geval is, is 0.001 wat daarmee zo klein is, dat de nulhypothese niet houdbaar is. Dit maakt dat de nulhypothese wordt verworpen. Wat we trouwens niet weten is welke test niet vals positief is. Dit is probleem uit de frequentistische statistiek die we vaak tegenkomen: de binaire aart van de hypothesetoetsing maakt dat alles over één kam wordt geschorst. Wordt de nulhypothese verworpen dan moet alles in de alternatieve hypothese waar zijn. Wordt de nulhypothese niet verworpen dan moet alles in de alternatieve hypothese wel onwaar zijn.

Wat Portier hier eigenlijk doet is twee keer dezelfde fout maken. Niet alleen door heel vaak te testen op een statistisch significant verschil zonder de theoretische  $\alpha$ -waarde aan te passen, maar ook door het aantal gevonden p-waarden samen te voegen in een nieuwe statistisch test en daar dan weer een p-waarde voor uit te rekenen.

Laten we eerst eens kijken hoe er volgens de OECD richtlijnen moet worden omgegaan met p-waarden verkregen uit meerdere testen:

*There is also the methodological problem of the use of a multiple testing procedure where one hypothesis test is used to choose another test which can complicate quantifying the true probability values associated with various comparisons<sup>89</sup>*

Wie snel zoekt in de statistische literatuur vindt andere interessante uitgangspunten:

*One possible strategy to deal with this problem (employed, e.g., in the early NCI bioassays) is to use a Bonferroni-type multiple comparisons adjustment. A discussion of Bonferroni adjustments has been given by Mantel (16).<sup>90</sup>*

Nu weten we dat we met 206 testen te maken hebben. Althans, als we naar de data kijken. Dat zijn 290 testen minder dan gerapporteerd door Portier. Nu kan het heel goed zijn dat er in het artikel nog een hoop analyses zijn die maar summier zijn gerapporteerd, of die niet direct zichtbaar zijn in de tabellen. Voor nu dat doet er eigenlijk niet toe, want 206 p-waarden zijn een hele hoop p-waarden.

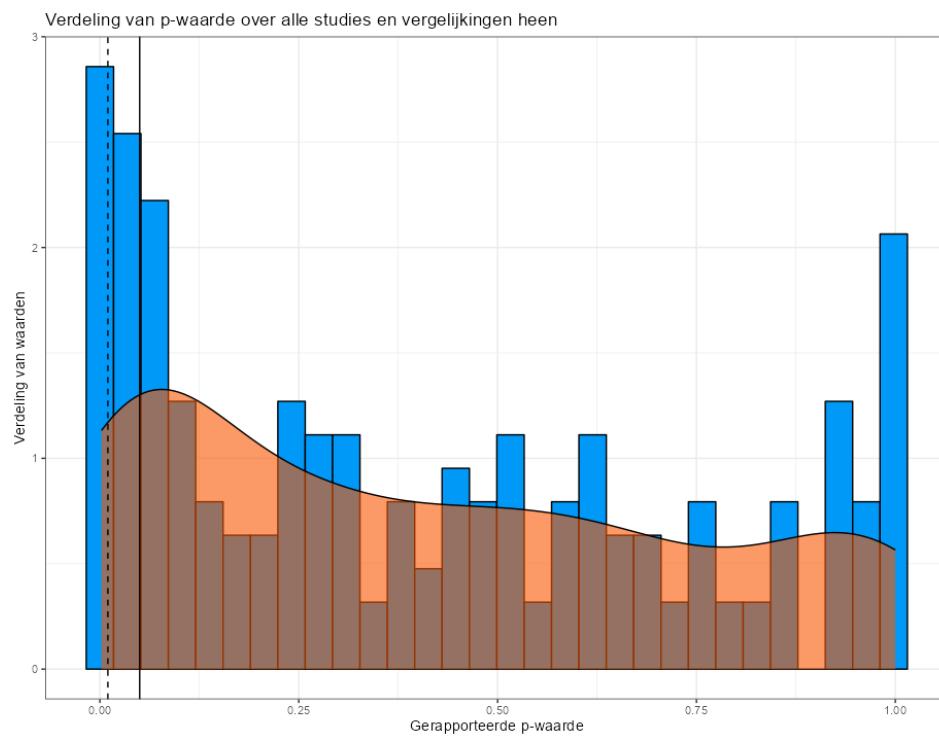
Een goed begin is het tonen van de verdeling van deze p-waarden **Figuur 76**. Op basis van die gegevens kan ik uitrekenen hoeveel p-waarden  $\leq 0.05$  en hoeveel er kleiner zijn dan  $\leq 0.01$ <sup>91</sup>: dat zijn er 34 en 12 respectievelijk. Van 23 vergelijkingen kon geen p-waarde worden berekend: in deze groepen was er geen kanker en/of waren de cellen te klein om betekenisvolle analyses uit te voeren. Als ik dan per studie kijk kan ik **Figuur 77** maken.

---

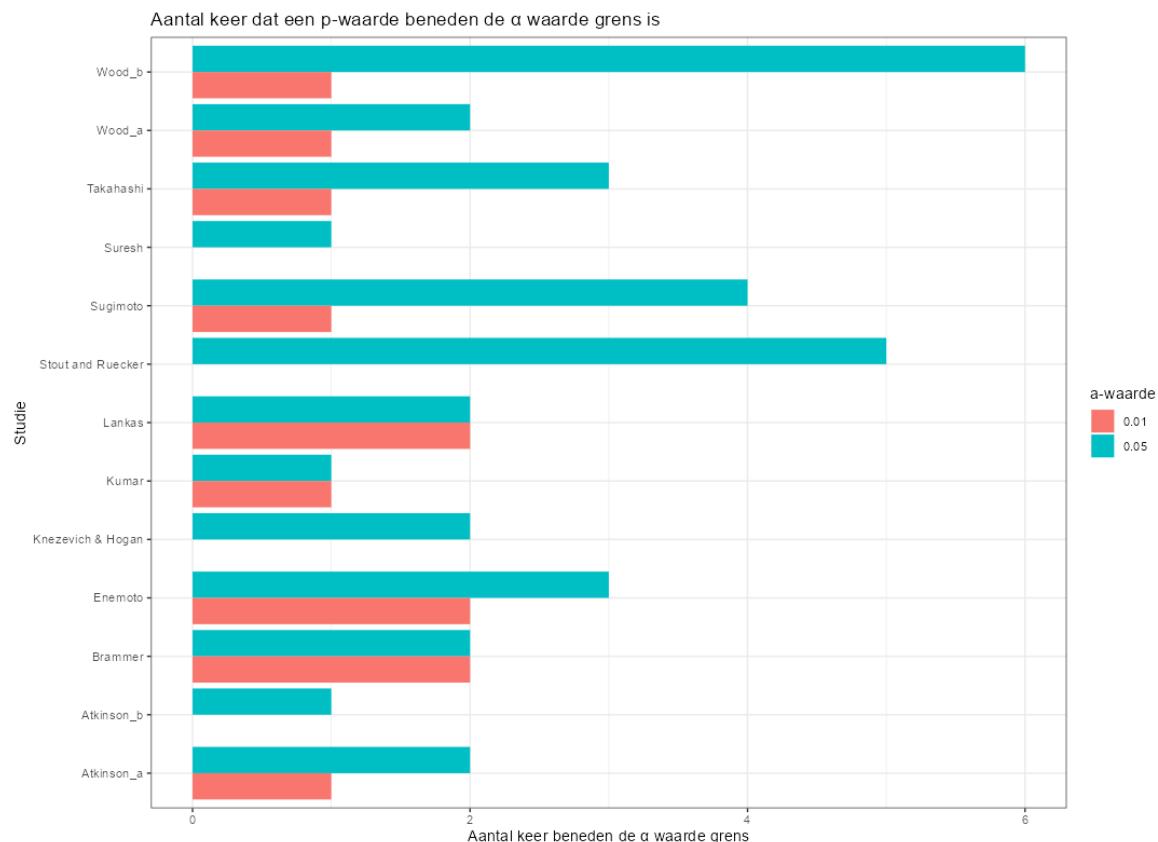
<sup>89</sup> [https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453\\_9789264221475-en](https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453_9789264221475-en)

<sup>90</sup> Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. Haseman. doi: 10.1289/ehp.8458385.

<sup>91</sup> Deze bevinding is niet wederzijds uitsluitend: een p waarde  $\leq 0.01$  zal ook  $\leq 0.05$ , maar niet omgekeerd. Ik kan de getallen dus niet bij elkaar optellen.



**Figuur 76.** Verdeling van de gerapporteerde p-waarden. De stippellijn links is de grens van statistisch significantie op 0.05.

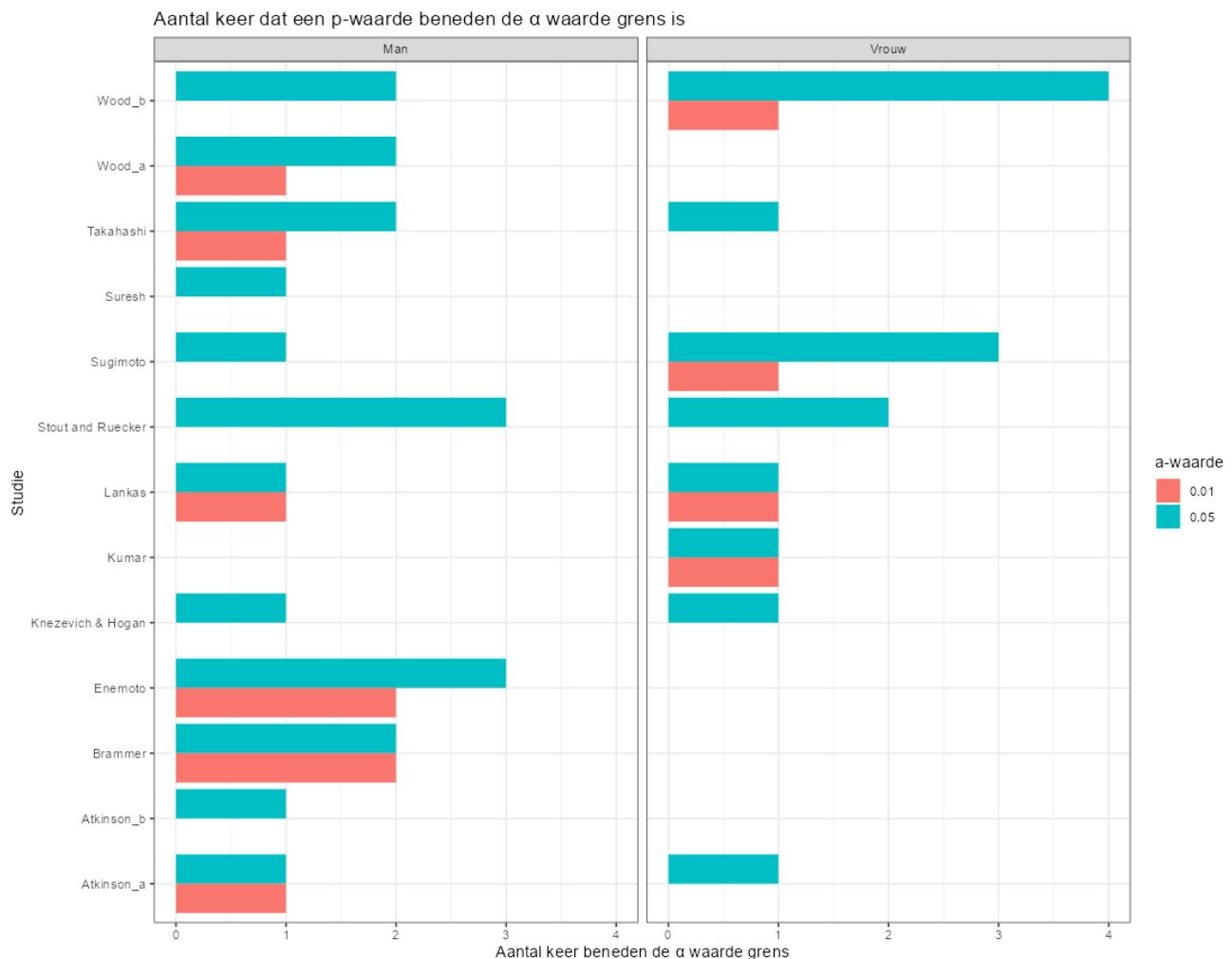


**Figuur 77.** Per studie het aantal toetsen dat onder een gespecificeerde grenswaarde viel. Dit zijn resultaten zoals door Portier gerapporteerd.

We weten dat we 13 studies hebben en we tellen in elke studie minstens één statistisch significante  $p$ -waarde. Toch kent niet elke studie kent evenveel significante resultaten. Dit zou niet mogen verbazen, omdat studies en dieren verschillen. We zullen dus altijd te maken hebben met variantie. Vervolgens kunnen we kijken hoeveel significante verschillen we vinden per geslacht (**Tabel 29**) en per geslacht én studie (**Figuur 78**).

Geslacht	$P \leq 0.05$	$P \leq 0.01$
Man	20	8
Vrouw	14	4

**Tabel 29.** Aantal significante resultaten per grenswaarde en geslacht.



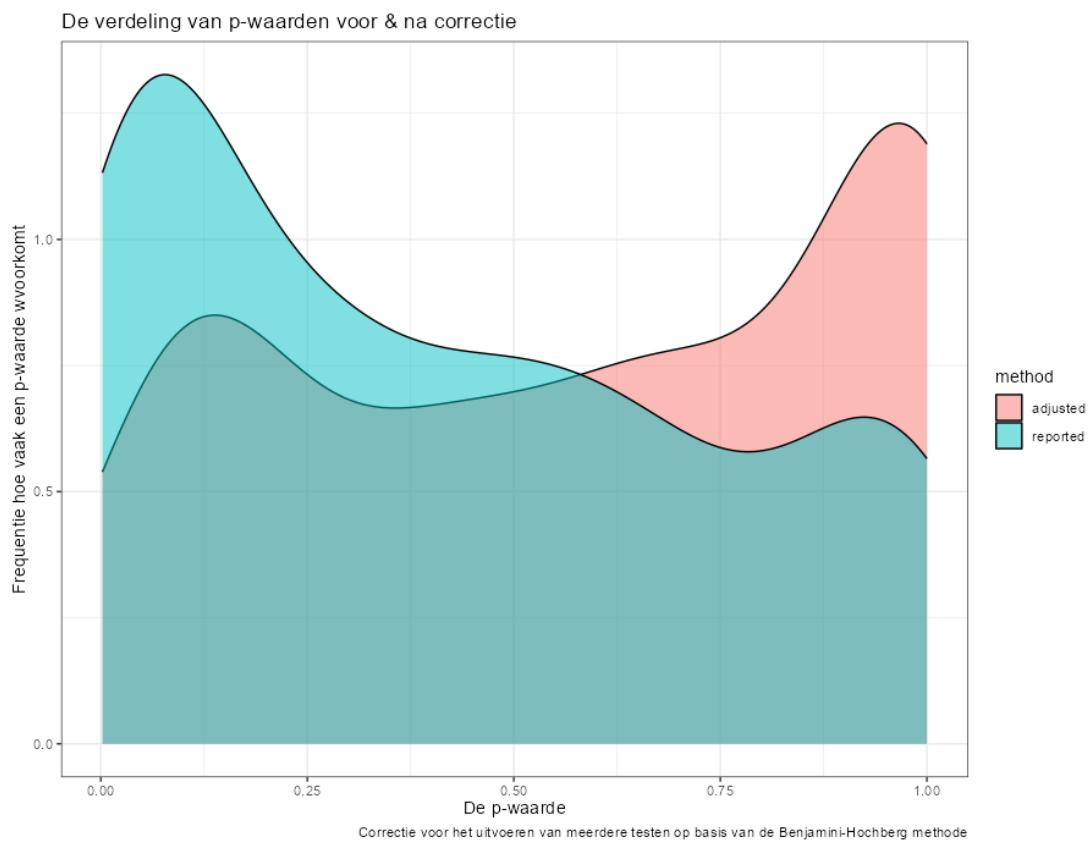
**Figuur 78.** Per studie en geslacht het aantal toetsen dat onder een gespecificeerde grenswaarde viel. Dit zijn resultaten zoals door Portier gerapporteerd.

Het wordt nu tijd om deze bevindingen te corrigeren voor een toename in het aantal vals positieven en wat ik ga doen is eigenlijk heel basaal: ik neem het aantal p-waarden (alle 206) en corrigeer deze dan voor het aantal uitgevoerde testen (206 dus). Dat kan ik op verschillende manieren doen met als resultaat **Tabel 30**:

Aantal statistische testen	Correctiemethode	$p \leq 0.05$	$p \leq 0.01$
206	Geen	34	12
206	Bejamini-Hochberg	0	0
206	Holm	0	0
206	FDR	0	0
206	Bonferroni	0	0

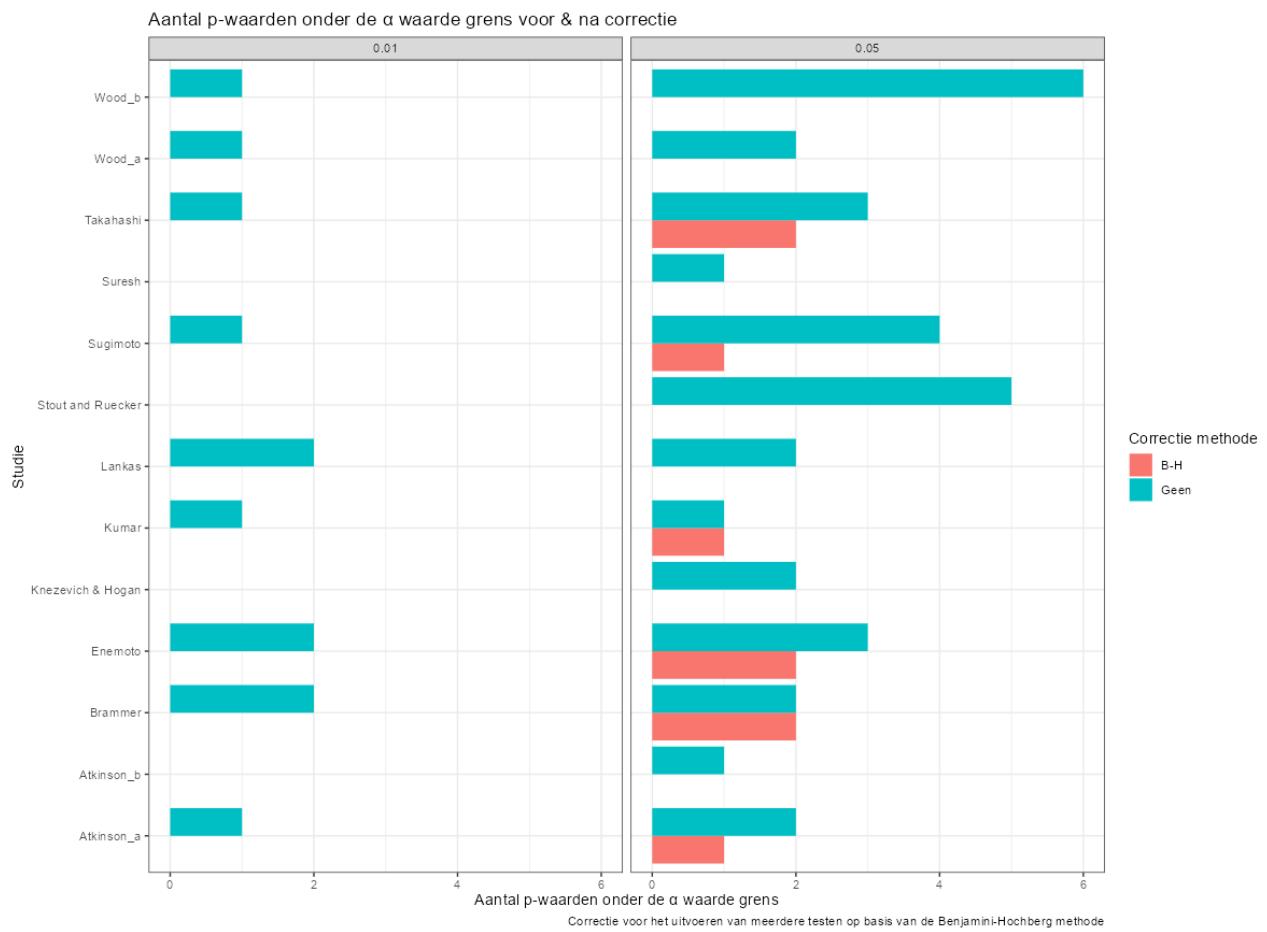
**Tabel 30.** Aantal statistische testen wel of niet gecorrigeerd per grenswaarde.

Als we dezelfde procedure uitvoeren, maar dan per studie (waardoor de correctiefactor omlaag gaat) zien we de volgende verdeling in **Figuur 79**:



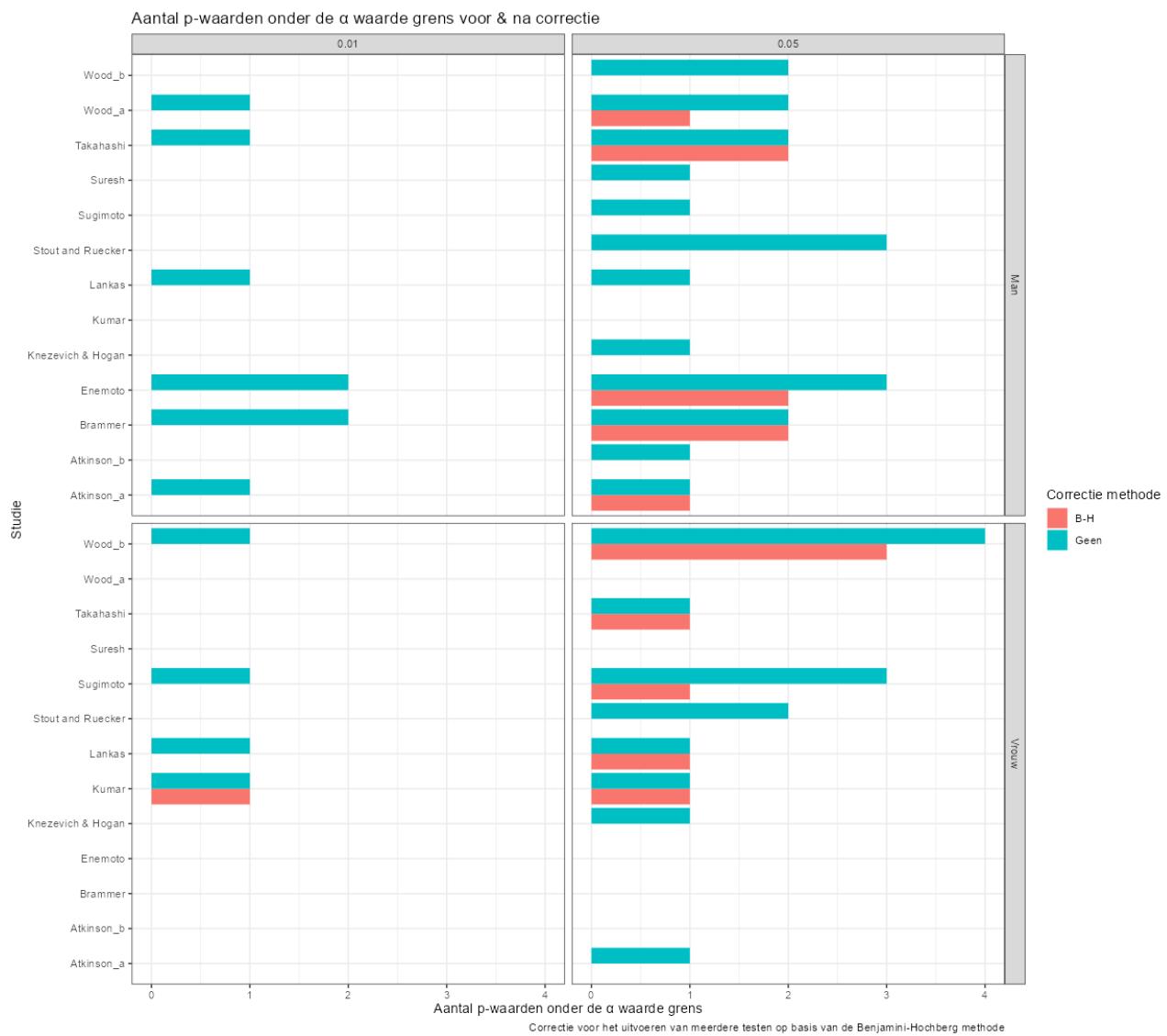
**Figuur 79.** Verdeling van het aantal gerapporteerde p-waarden met (adjusted) of zonder (reported) correctie.

Als we wederom een staafdiagram maken (**Figuur 80**) zien we dat geen enkele  $p$  waarde  $\leq 0.01$ . Als we de grenswaarde van  $\alpha$  optrekken zien we ineens wel weer statistisch significante resultaten, maar toch een stuk minder dan zonder correcties. We gaan van 34 naar zes statistisch significante resultaten.



**Figuur 80.** Per studie en geslacht het aantal toetsen dat onder een gespecificeerde grenswaarde viel met of zonder correctie methode. Dit zijn resultaten zoals door Portier gerapporteerd.

Laten we nog één keer de data opsplitsen en dat doen we door per studie en per geslacht te kijken naar de resultaten. Als wie die splitsing maken dan corrigeren we puur en alleen voor het aantal testen die gedaan zijn in de matrix geslacht\*studie: dit is ook exact de manier waarop de data door Portier wordt getoond. Wat volgt is het resultaat zoals zichtbaar in **Figuur 81**. We zien hier dat de correctie-methode een stuk minder aanpast en dat is geheel in lijn met de theorie: het aantal gemaakte testen is per categorie ook kleiner geworden.



**Figuur 81.** Per studie en grenswaarde het aantal toetsen dat onder een gespecifieerde grenswaarde viel met of zonder correctie methode. Dit zijn resultaten zoals door Portier gerapporteerd.

Als we uiteindelijk de data zo zouden opsplitsen zodat we nog één test per groep overhouden dan verdwijnt de correctiefactor. Met deze exercitie is hopelijk duidelijk geworden dat een correctie op meerdere testen niet zomaar mag worden overgeslagen. De analyse van Portier dat het aantal statistische significantie bevindingen groter is dan kans past niet in de frequentistische statistiek: we zullen deze berekening bewaren voor het stuk over de Bayesiaanse statistiek.

## Wat kunnen we hieruit afleiden?

Dit stuk vormt eigenlijk de kern van dit rapport. Het is hier waarin we getracht hebben om het werk van Portier te herhalen. Dat is ons niet gelukt. Ook zien we niet direct in hoe het verkleinen van de drempelwaarde (door naar eenzijdig testen over te stappen) helpt in het meer significant krijgen van de analyses.

Een ander belangrijk punt wat we hebben doorlopen is het gemis van nul-waarden. In sommige studies staan wel degelijk nul-waarden beschreven voor tumorsoorten wat betekent dat de afwezigheid van incidentie van een specifieke tumorsoort genoemd wordt als die andere studies wel is gezien. Dat is netjes omdat een samensmelting van gegevens een meer betrouwbare analyse maakt van de ratio wel/geen tumor. We hebben ook gezien dat het toevoegen van die ontbrekende gegevens (de nul-incidentie) de ratio's doet veranderen waardoor een significant effect vaak verdwijnt. Omdat het hier om een nogal theoretische exercitie gaat heb ik dit niet voor elke tumorsoort gedaan of voor elke combinatie van ras en geslacht. Toch laten de voorbeelden die we wel hebben doorlopen al voldoende doorschemeren dat het niet meetellen van nul-incidentie een effect moet hebben op de latere analyses in Portier's werk.

De meest ingrijpende exercitie die ik heb toegepast is een simpele correcte voor het uitvoeren van meerdere testen. Als we de gangbare literatuur volgen in het corrigeren voor vals-positieven blijkt dat geen bevinding meer statistisch significant is. Dit is een belangrijke bevinding om tot ons te nemen, omdat het ook laat zien hoe kwetsbaar de frequentistische statistiek is én hoe belangrijk het is dat we ons blijven bedenken op welke aannames de resultaten zijn gebaseerd.

## Glyfosaat: Dose-Response analyses

We hebben in de voorgaande hoofdstukken gezien hoe lastig het is om het werk van Portier na te bootsen: het lukt mij niet om exact dezelfde bevindingen te krijgen. Ook hebben we gezien dat het uitvoeren van zoveel verschillende testen problematisch is. Hier moet voor worden gecorigeerd, ook al doet Portier dat zelf niet. We zagen dat na het toepassen van correctie-methoden het gros van de bevindingen niet meer statistisch significant was.

Om niet te verzanden in een welles-nietes discussie is het wellicht raadzaam om dit rapport uit te breiden met verdere dose-response analyses. Dit betekent dat we gaan kijken of het aantal kankergevallen per dosering een richting laat zien die statistisch significant afwijkt van nul. Een klein voorproefje had ik al reeds eerder getoond, maar er zijn meerdere methoden om een dose-response analyse uit te voeren. Ik zal er een aantal laten zien.

Voordat ik hiermee aan de slag ga, wil ik eerst een korte tussenstop maken en het gebruik van historische controlegroepen adresseren. Dit onderdeel werd ook al belicht in de BNNVARA/Zembla rapportage. Een historische controlegroep betreft een groep die niet is geobserveerd in de desbetreffende studie, maar wel wordt meegenomen als vergelijkingsmateriaal. Een voorbeeld zou zijn als de controlegroep uit studie A wordt gebruikt als controlegroep voor studie B, C én D. Dit is evident problematisch omdat de omstandigheden waarin gegevens uit studie A zijn genomen anders is dan studie B, C of D. Soms heeft een studie geen keus, maar als het even kan getuigd het van goed gebruik om een eigen controlegroep mee te nemen. Dit wordt door de OECD benadrukt:

*Historical control data can help interpret results in a number of situations (see GD 35). In any discussion about historical control data, it should be stressed that the concurrent control group is always the most important consideration in the testing for increased tumour rates<sup>92</sup>.*

Zoals ik al zei zijn er meerdere manieren om dose-response analyses te doen. Dit komt omdat het allemaal onderdelen zijn van zogenaamde regressieanalyses waarbij de te voorspellen variabele( (hier kanker) kan worden afgezet tegen mogelijk verklarende variabelen (dosering, geslacht, soort dier etc.). Bij een meer ‘traditionele dose-response

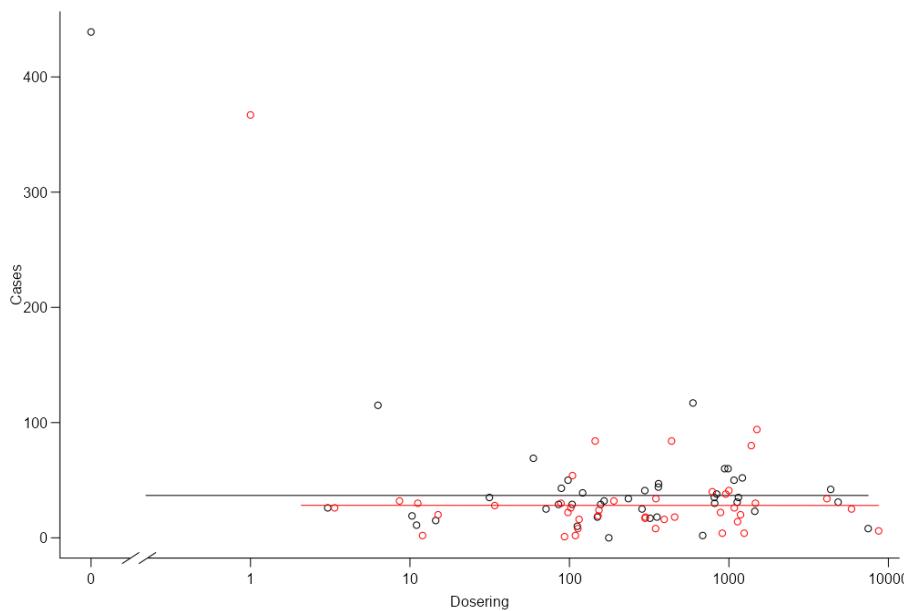
---

<sup>92</sup> [https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453\\_9789264221475-en](https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453_9789264221475-en)

analyse’ wordt er gebruik gemaakt van non-lineaire formules, terwijl de meer ‘traditionele analyses’ gebruik maken van regressie technieken. Hoewel de non-lineaire vorm lastiger te bepalen is, kent deze een aantal voordelen boven op de lineaire methodiek die het de moeite waard maakt om als eerste te proberen<sup>93</sup>.

## Non-lineaire dose-response analyse

Zoals we al eerder zagen was het niet bepaald makkelijk om een dose-response analyse uit te voeren (**Figuur 59** en **Figuur 60**). Sterker nog, in mijn statistiekprogramma kwam ik niet eens tot een wiskundige oplossing die stabiel genoeg was om ook onzekerheid te schatten. Ik kreeg daarom ook geen betrouwbaarheidsinterval. Het uitproberen van verschillende statistische verdelingen voegt weinig toe en echt verbazen mag dit alles niet. Gooien we de data namelijk allemaal op één grote hoop dan zien we dat het gros van de kankergevallen zichtbaar is in de controlegroep (nul-dosering) en daarna afzwakt (**Figuur 82**). Toch zou het te makkelijk zijn om het hier nu bij te laten. Daarom maken we een uitstap naar de lineaire methoden.



**Figuur 82.** Dose-response relatie op basis van alle data. Gekeken is naar het aantal gerapporteerde tumorgevallen, los van het aantal geïncludeerde dieren. Zoals te zien valt is het haast onmogelijk om een dose-response curve te maken.

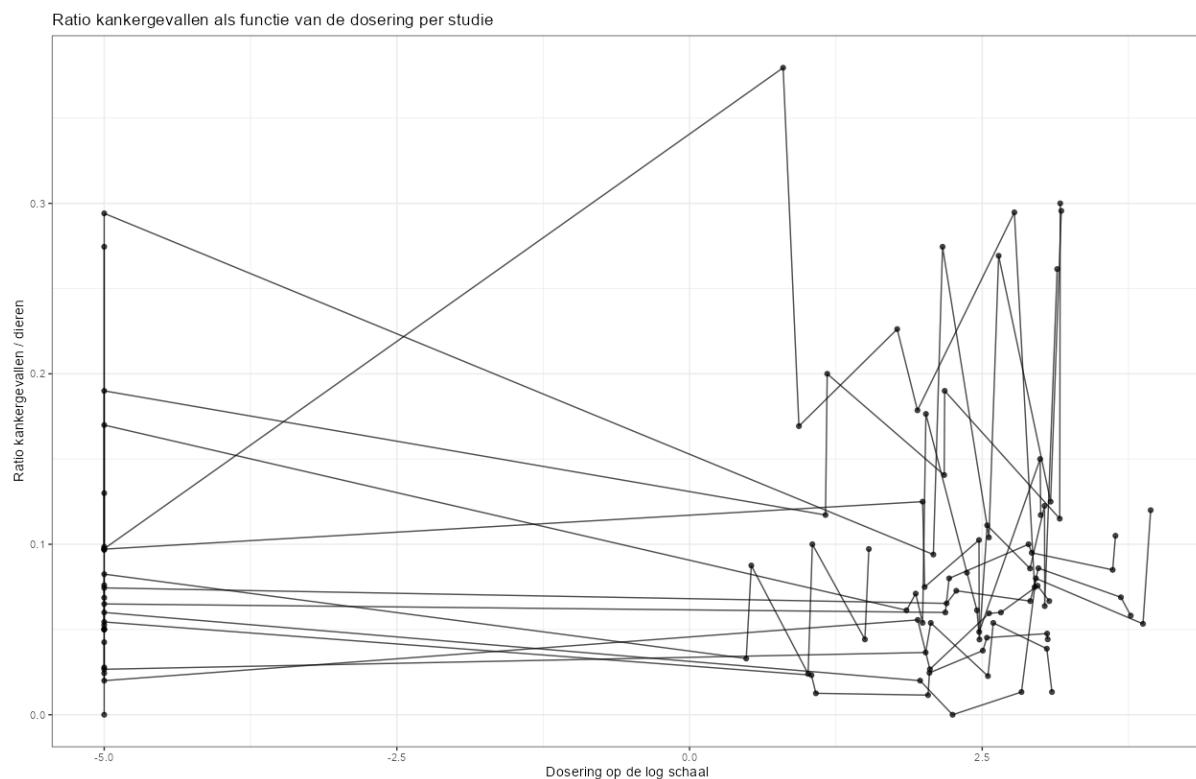
<sup>93</sup> Let wel: een lineaire methode kan wel degelijk non-lineaire voorspellingen maken. Het verschil zit hem in de verhoudingen tussen de variabelen.

## Lineaire dose-response analyse

Het voordeel van dit type analyse is dat het wiskundig makkelijker is om tot een stabiele oplossing te komen. Een analytisch resultaat is vaak eenvoudiger te behalen<sup>94</sup>

Voordat we aan de slag gaan wil ik de data nog één keer grafisch weergeven. Dit doe ik omdat het soort analyse nu ook iets anders is. Stel dat ik nu voor elke studie, ongeacht het type kancersoort, de ratio kanker/geen kanker laat zien voor elke dosering. Dan krijgen we

**Figuur 83.**



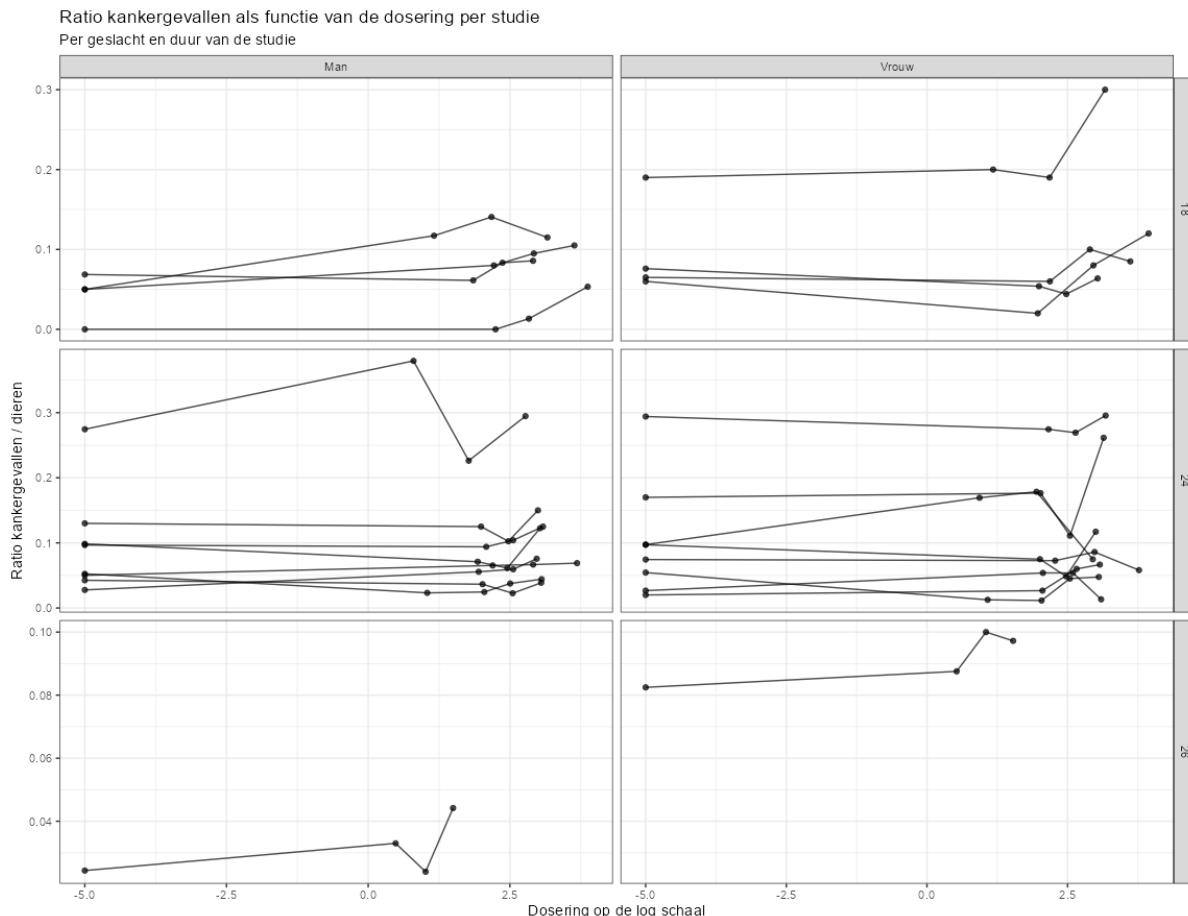
**Figuur 83.** Dose-response curve met op de x-as de dosering op de log schaal. Elke lijn is een studie. Op de Y-as is de ratio van het aantal kankergevallen per het aantal dieren per dosering.

Als we deze gegevens opdelen in geslacht en de lengteduur van de studie dan krijgen we

**Figuur 84.** Nu beginnen we in sommige cellen een iets andere lijn te zien en een stijging van kanker als gevolg van de dosering kunnen we niet direct uitsluiten. Een stijging (of daling) in sommige cellen is echter niet voldoende om beroep te doen op de uitspraak dat een bevinding groter is 'dan kans dat' (cq. statistische significantie). Laten we daarom de data

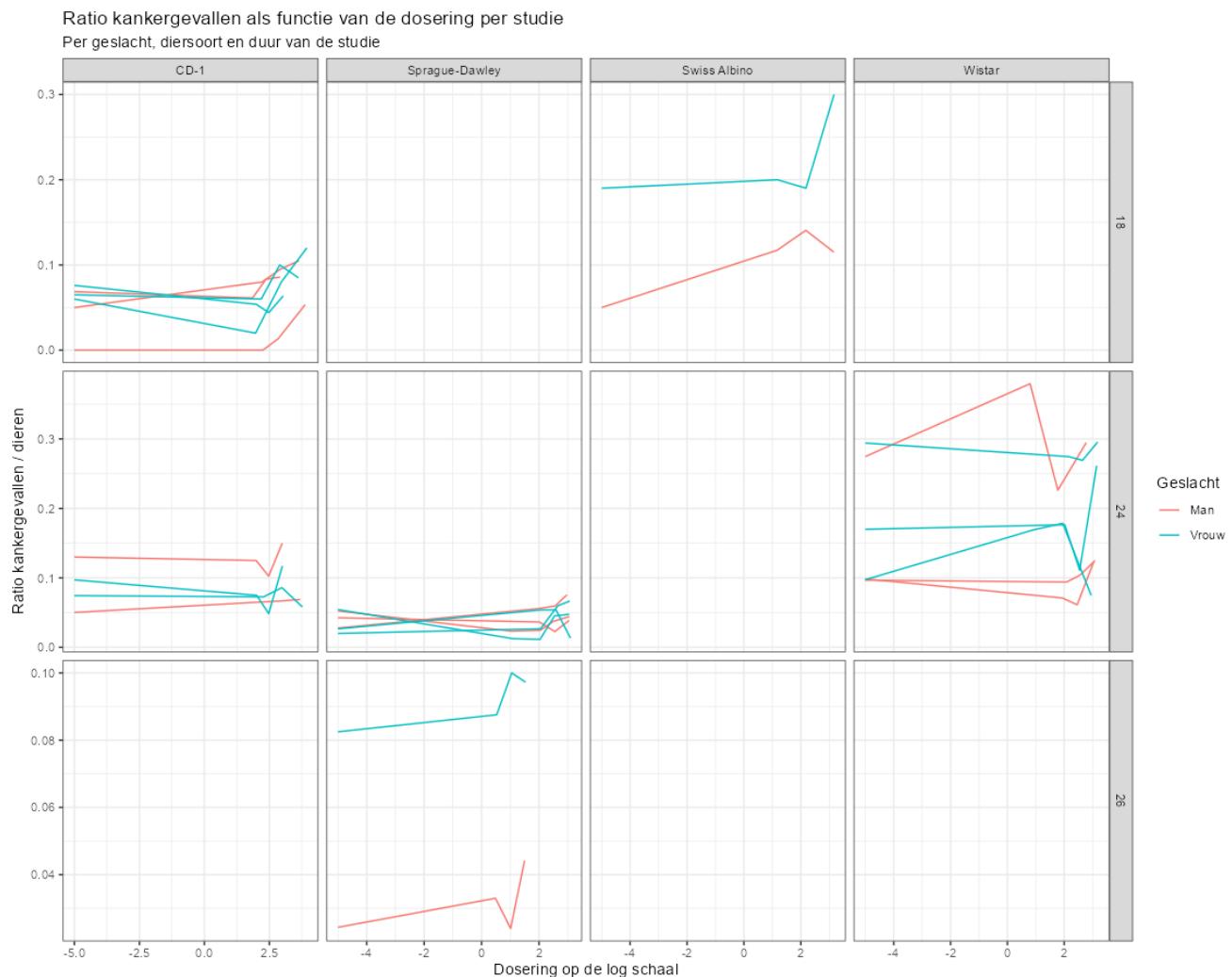
<sup>94</sup> Tenzij de relatie echt non-lineair is. Dan is de non-lineaire manier de beste en makkelijkste manier met de meest voor de hand liggende interpretaties. De grafische weergave van de studies zoals door Portier gerapporteerd laat heel duidelijk zien waarom dit niet het geval is.

verder opsplitsen en nu ook soort dier meenemen (**Figuur 85**). Wat we dan zien is dat sommige lijnen wel degelijk een dose-response curve laten zien. Maar dit is zeker niet altijd het geval. Er is dus sprake van variatie. Ook is er sprake van de onzekerheid per puntschatting. Dit wordt vaak vergeten bij het zien van figuren zoals hier beneden.



**Figuur 84.** Dose-response curve met op de x-as de dosering op de log schaal. Elke lijn is een studie. Op de Y-as is de ratio van het aantal kankergevallen per het aantal dieren zoals geïncludeerd per dosering. Een verder onderverdeling is gemaakt per geslacht (kolom) en de duur van de studie (rijen).

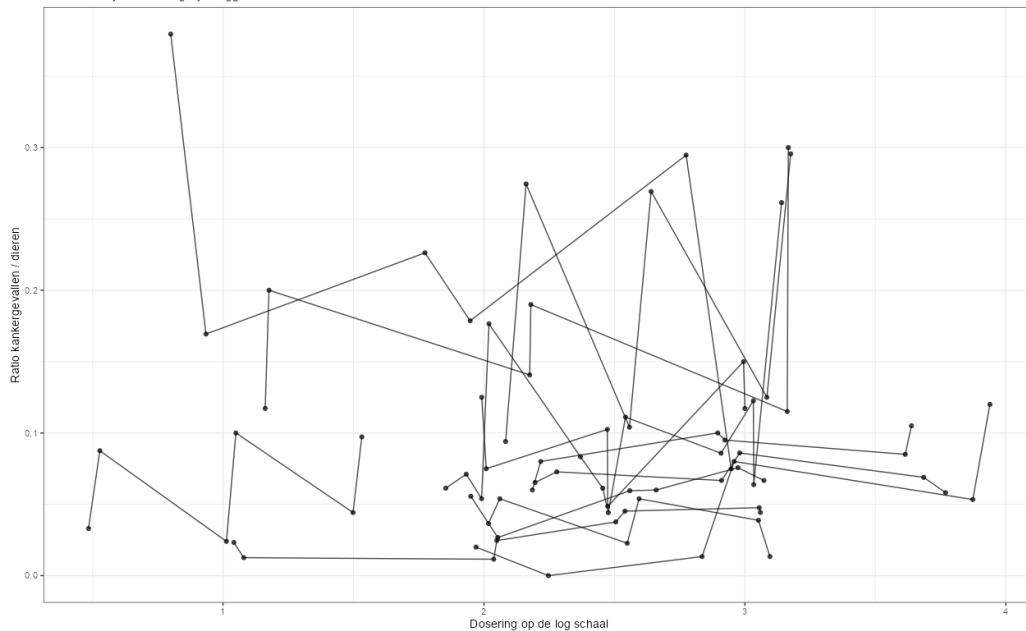
Wat we nu zouden kunnen doen is het weglaten van de controlegroep (**Figuur 85**, **Figuur 86** en **Figuur 87**). Eigenlijk is dit 'not-done' omdat het de natuurlijke correctiefactor verwijdert, maar het helpt wel om het effect van de dosering wat beter te bekijken. Omdat het gros van de kankergevallen nu wegvalt zouden we op zijn minst stijgende lijnen mogen verwachten: een hogere dosering zou in het geval van carcinogeniteit ergens moeten leiden tot een stijgende ratio.



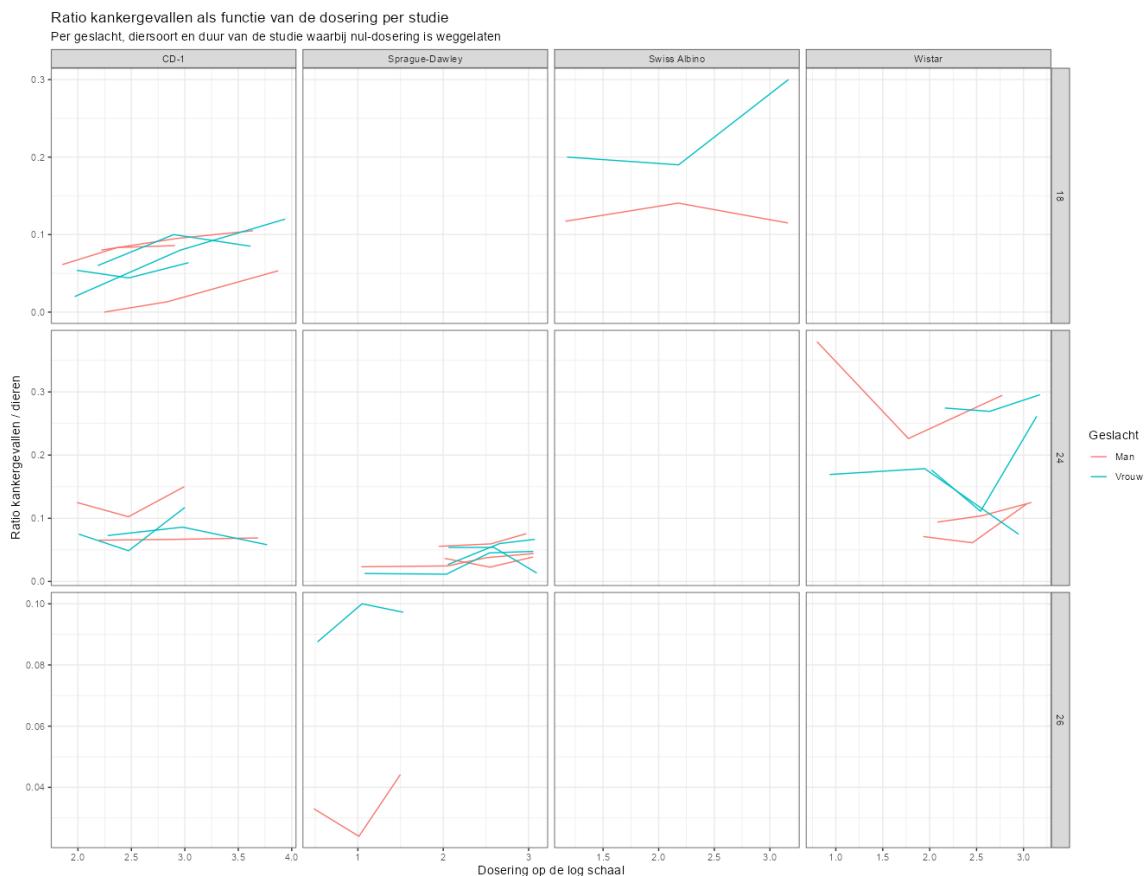
**Figuur 85.** Ratio kankergevallen als functie van de dosering per studie (log schaal). Nu opgedeeld per geslacht, soort en duur van de studie. Opvallend is het aantal rechte lijnen. Ook kunnen we een aantal flinke dose-response relaties zien.

Wat opvalt aan **Figuur 85**, **Figuur 86** en **Figuur 87** is dat een verdere opsplitsing gecombineerd met het weghalen van de nul-dosering soms wel degelijk een stijgende lijn laat zien. Nu is het opsplitsen van de data niet gek, en soms zelfs nodig, maar het weglaten van de nul-dosering betekent effectief het weglaten van de controlegroep. Dit betekent eigenlijk ook dat de grafieken weinig waard zijn: de controlegroep zet namelijk de andere groepen in perspectief en zonder dat perspectief kun je geen vergelijking maken. Dat de controlegroep zoveel kankergevallen laat zien wordt verder niet verklaart en is ook een van de redenen waarom ik zoveel verschillende analyses heb losgelaten op de gerapporteerde data van Portier. Het ziet er eigenlijk uit alsof het niet kan.

Ratio kankergevallen als functie van de dosering per studie  
Resultaten bij nul-dosering zijn weggelaten.



**Figuur 86.** Ratio kankergevallen als functie van de dosering per studie. Elke lijn is een studie. Deze keer is de nul-dosering weggelaten.



**Figuur 87.** Ratio kankergevallen als functie van de dosering per studie. Elke lijn is een studie. Deze keer is de nul-dosering weggelaten. Opsplitsing per soort, duur van de studie en geslacht.

## Linear Mixed Model

Een dose-response relatie maken is meer dan alleen grafieken tonen. Elke dose-response relatie is feitelijk een model waarin de dosering gekoppeld wordt aan het aantal tumorgevallen. Dit getal is de ‘y-variabele’. De rest zijn de ‘x-variabelen’.

De dosering zelf is in dit geval lastig te modelleren, maar wat misschien nog lastiger is, is de ‘juiste’ keuze voor de y-variabele. Zo kan ik kiezen tussen het aantal tumorgevallen per tumorsoort, of per studie, of per dosering. Ik kan er ook voor kiezen om de data te modelleren als ratio zoals ik al liet zien in **Figuur 83**. Dan heb ik nog de mogelijkheid om de data te splitsen in twee groepen: de nul-dosering én de niet-nul-dosering. Deze keuzevrijheid betekent echter niet dat er meerdere juiste manieren zijn. Dit is vaak het moeilijke in modelleren: je bent op zoek naar een formule die de data zoals geobserveerd het best kan nabootsen op een zodanige manier dat de statistische eigenschappen van die formule acceptabel zijn.

Ik zal dus een aantal modellen moeten maken en deze met elkaar vergelijken. Een mooi soort model om mee te beginnen is een [Linear Mixed Model \(LMM\)](#). Een LMM is ingesteld op het bepalen van de variatie tussen groepen met herhaalde metingen. In dit geval dus de hoeveelheid variatie in de dose-response relatie tussen studies. Hoewel elke studie een aparte groep dieren heeft per dosering, en er dus geen herhaalde metingen zijn per dier, is het wel zo dat de metingen plaatsvinden in dezelfde studie. Dat betekent dat er studie-specifieke effecten zijn die moeten worden meegenomen. Een LMM model doet dit op een geraffineerde wijze door variatie op te delen in onderdelen (dit moeten we zelf specificeren) en dan uit te rekenen welk onderdeel van de dataset welke contributie maakt aan de totale variatie zoals geobserveerd.

In **Figuur 88** kunnen we het resultaat zien van een LMM die per studie, geslacht en soort kijkt hoeveel dieren er zijn meegenomen per dosering en wat het totaal aantal kankergevallen is geweest zoals geobserveerd. Dit is waarom de y-variable ‘*sum Cases*’ heet: dit betreft het opgetelde aantal kankergevallen. Als ik de data zou visualiseren zou deze er exact zo uitzien als in **Figuur 83**. Alleen de y-as zou anders zijn.

De reden om ‘*sum Cases*’ te nemen is omdat verdeling van die variabele voldoende spreiding heeft waarmee ik kan rekenen. Om te zien of het model voldoet aan de statistische eisen van een LMM moet ik eerst de analyse uitvoeren. Één van die assumpties is dat de restwaarden normaal verdeeld zijn én dat de mate van variantie niet afhankelijk mag zijn van

de voorspelde waarde. Het is een onterechte aanname dat je deze assumpties met de ruwe data kunt testen.

<i>Predictors</i>	sum Cases		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	-18.95	-82.52 – 44.62	0.555
Dosering log	0.37	-1.05 – 1.78	0.609
Geslacht [Vrouw]	<b>-7.05</b>	<b>-13.64 – -0.46</b>	<b>0.036</b>
Duur	2.39	-0.69 – 5.46	0.127
Soort [Sprague-Dawley]	-13.96	-32.76 – 4.85	0.144
Soort [Swiss Albino]	-1.36	-25.44 – 22.72	0.911
Soort [Wistar]	23.86	4.99 – 42.73	<b>0.014</b>
Dosering log * Geslacht [Vrouw]	-0.13	-2.12 – 1.87	0.898
<b>Random Effects</b>			
$\sigma^2$	282.76		
$\tau_{00}$ Studie	75.71		
ICC	0.21		
N Studie	13		
Observations	106		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.415 / 0.538		

**Figuur 88.** Resultaten van een LMM. Geslacht en soort lijken statistisch significant te zijn.

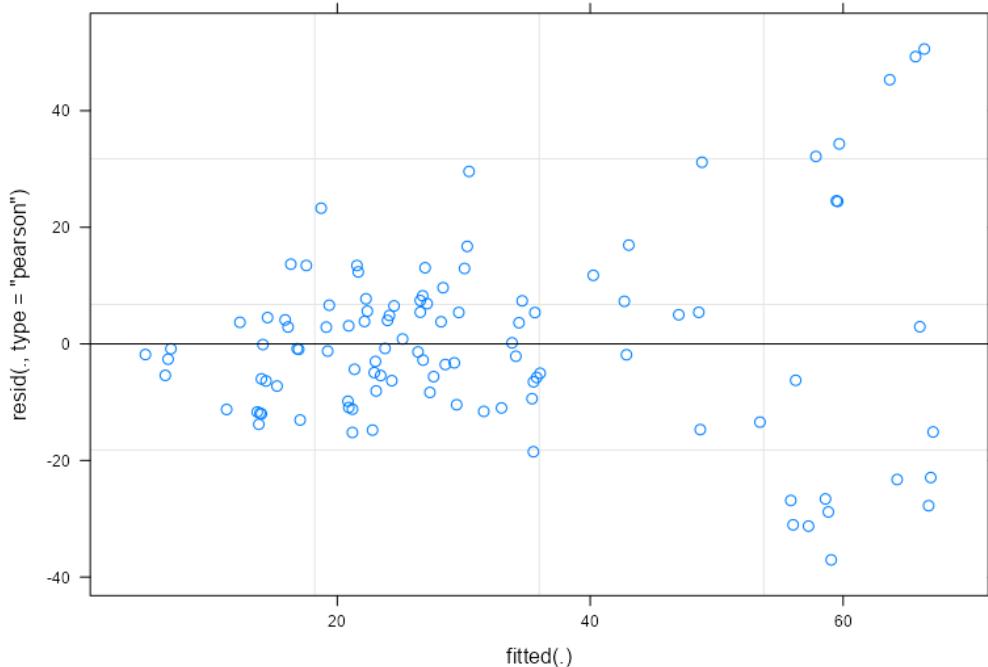
**Figuur 88** laat de resultaten zien van een LMM waarbij de dosering op de log schaal wordt gekoppeld aan het opgeteld aantal tumoren per geslacht, soort en studie. In de linker kolom staan de *Predictors*: dit zijn de variabelen die ik probeer te koppelen aan de hoeveelheid geobserveerde tumorsoorten. Zo is te zien dat de dosering op de log schaal niet statistisch significant is. Dat wil zeggen dat de relatie tussen dosering en het aantal tumorsoorten niet statistisch significant afwijkt van een horizontale lijn. Dit is de lijn van ‘geen effect’.

In feite toets dit model dus of de relatie tussen dosering en het aantal kankergevallen een richtingscoëfficiënt heeft die statistisch significant verschillend is van nul. Dat lijkt hier niet het geval te zijn, maar we mogen niet zomaar de problemen uit de frequentistische statistiek negeren. Dit betekent dat het niet kunnen verwerpen van de nulhypothese betekent dat deze hypothese klopt.

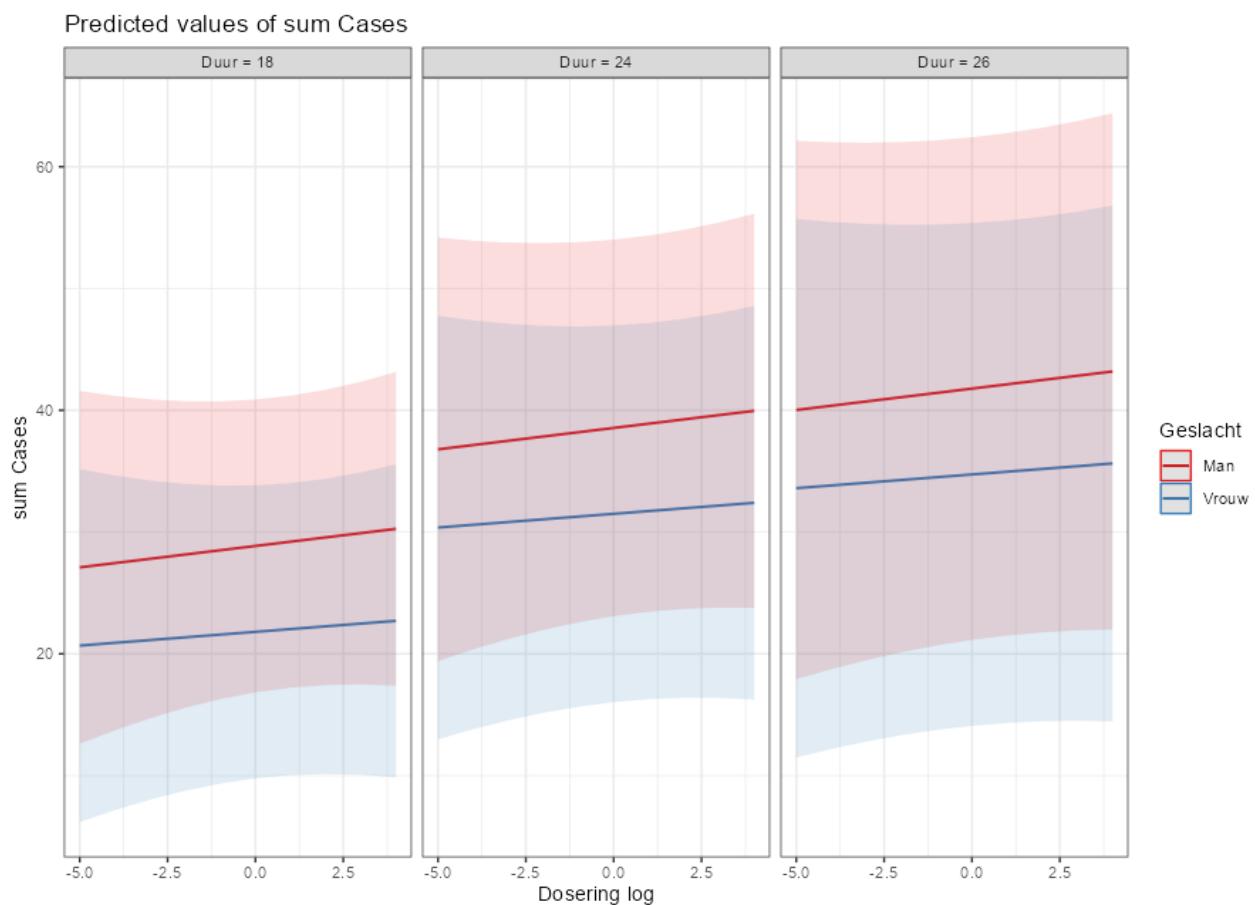
Uit de tabel blijkt verder dat geslacht en soort wel statistisch significant zijn. Dit vereist wat interpreteerwerk: vrouwen laten in totaal minder tumoren zien. Bij soort gaat het om de vergelijking tussen Wistar ratten én de referentiesoort wat in dit geval de CD-1 muizen betreft. Het resultaat oppert dat we bij de Wistar ratten meer tumoren zien dan bij de CD-1 muizen.

Ik heb trouwens expres gekozen voor de interactie *Geslacht\*Dosering* omdat Portier zijn data expliciet per geslacht opdeelt. Het toevoegen van nog meer interacties zou de data onnodig veel opdelen met als gevolg dat de cellen te klein worden om betekenisvolle analyses op uit te voeren. Uiteraard is de lezer vrij om die analyse zelf uit te voren met de data en codes vanuit de [GitHub pagina](#).

In **Figuur 89** zien we een vaak voorkomende grafiek die laat zien of de restwaarden (het verschil tussen voorspelling en observatie) normaal én gelijk verdeeld zijn. Daar lijkt het niet helemaal op. Waarschijnlijk dat een andere y-variabele (eentje die rekening houdt met het aantal observaties) een betere manier is om een dose-response relatie te bouwen. Ondanks dat het niet perfect is, is het ook weer niet zo slecht dat we helemaal niks kunnen met dit model. Daarom is het toch zinvol om te kijken welke voorspellingen dit model produceert (**Figuur 90**).



**Figuur 89.** Verdeling van restwaarden afkomstig van een LMM model. Hier wordt grotendeels aan voldaan, maar zien we op het einde tot een waaier-effect. Mogelijkerwijs is het opgeteld aantal kankergevallen niet de beste y-variabele om dosering aan te koppelen.

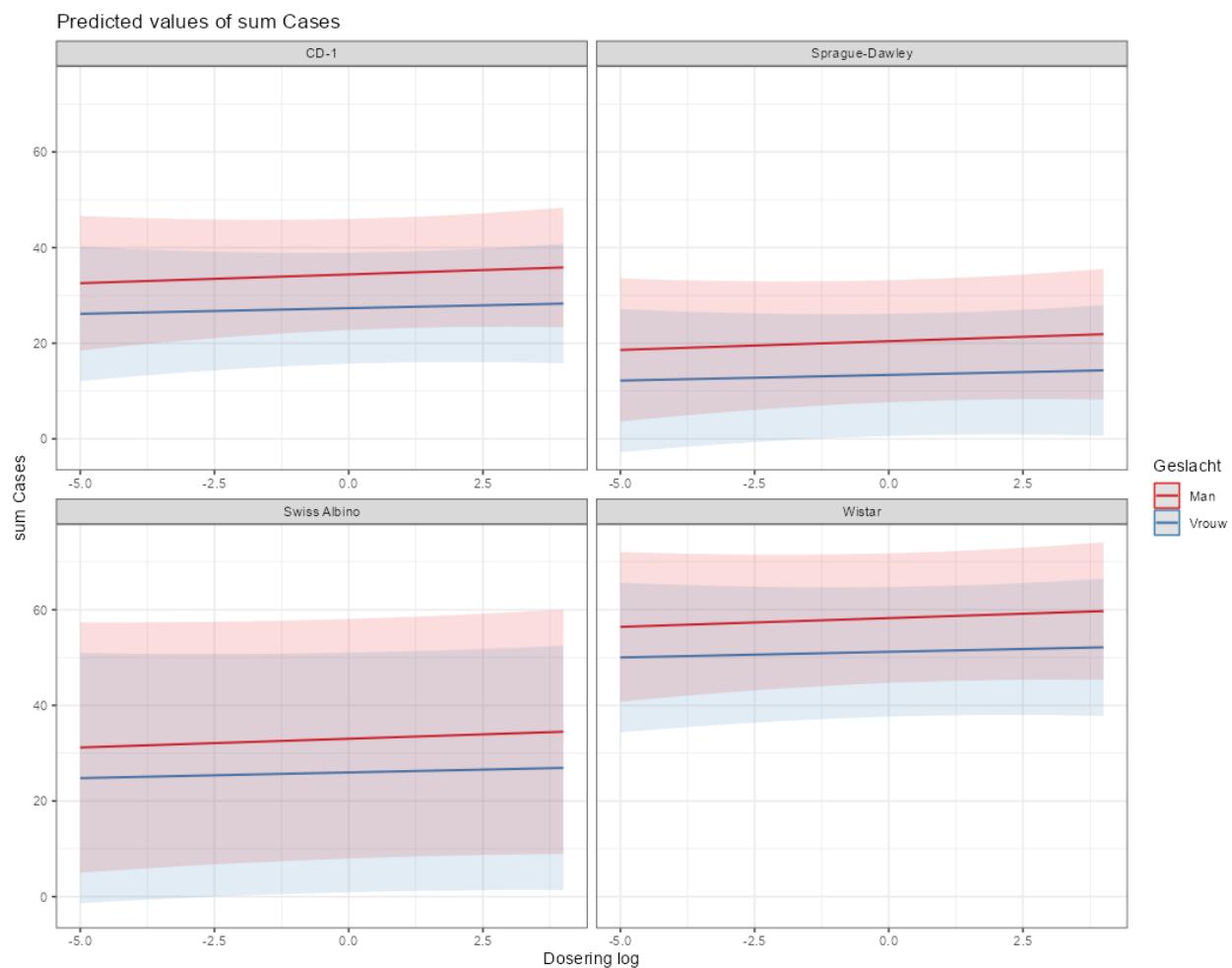


**Figuur 90.** Voorspellingen afkomstig van een LMM model met als y-variabele het opgeteld aantal kankergevallen (*sum Cases*). De lijn laat de voorspelde waarde zien als functie van de dosering (log schaal), geslacht en de duur van de studie. De lijnen laten de gemiddelde voorspelling zien, de band eromheen is de onzekerheid.

Bovenstaand **Figuur 90** laat zien dat we door de onzekerheidsband ook een perfecte horizontale lijn kunnen trekken. Wanneer die mogelijkheid er is, is het haast onmogelijk om in de frequentistische statistiek de nulhypothese ( $\text{coëfficiënt} = 0$ ) te verwerpen. In **Figuur 91** zien we de relatie tussen dosering en opgeteld aantal kankergevallen per soort.

In het voorgaande model hebben we relatie tussen dosering en het aantal kankergevallen lineair gehouden (**Figuur 94**). In het volgende model tracht ik de relatie te modelleren met een non-lineaire relatie die lineair is in de parameters. Dit worden ook wel *natural splines* genoemd. Het resultaat is te zien in **Figuur 92**. In dit model verdwijnt de significantie van *Geslacht*, maar blijft die van de Wistar ratten overeind. Het model zelf is niet bepaald verklarend dus het is maar de vraag hoeveel waarde we hieraan moeten

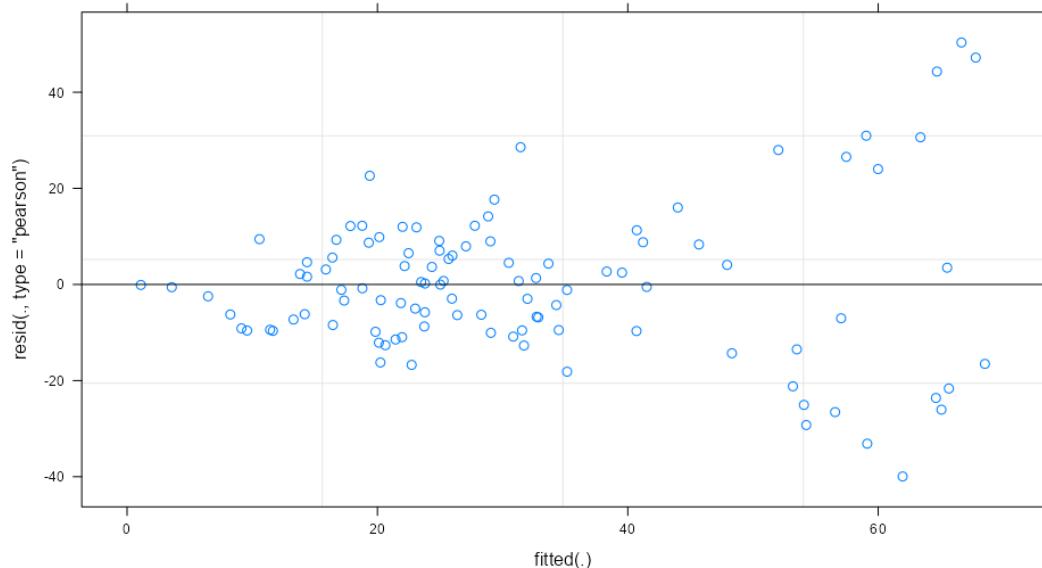
hechten. **Figuur 93** laat een grotere waaier zien dan **Figuur 89**. Het model laat duidelijk te wensen over wat zichtbaar wordt in **Figuur 95**.



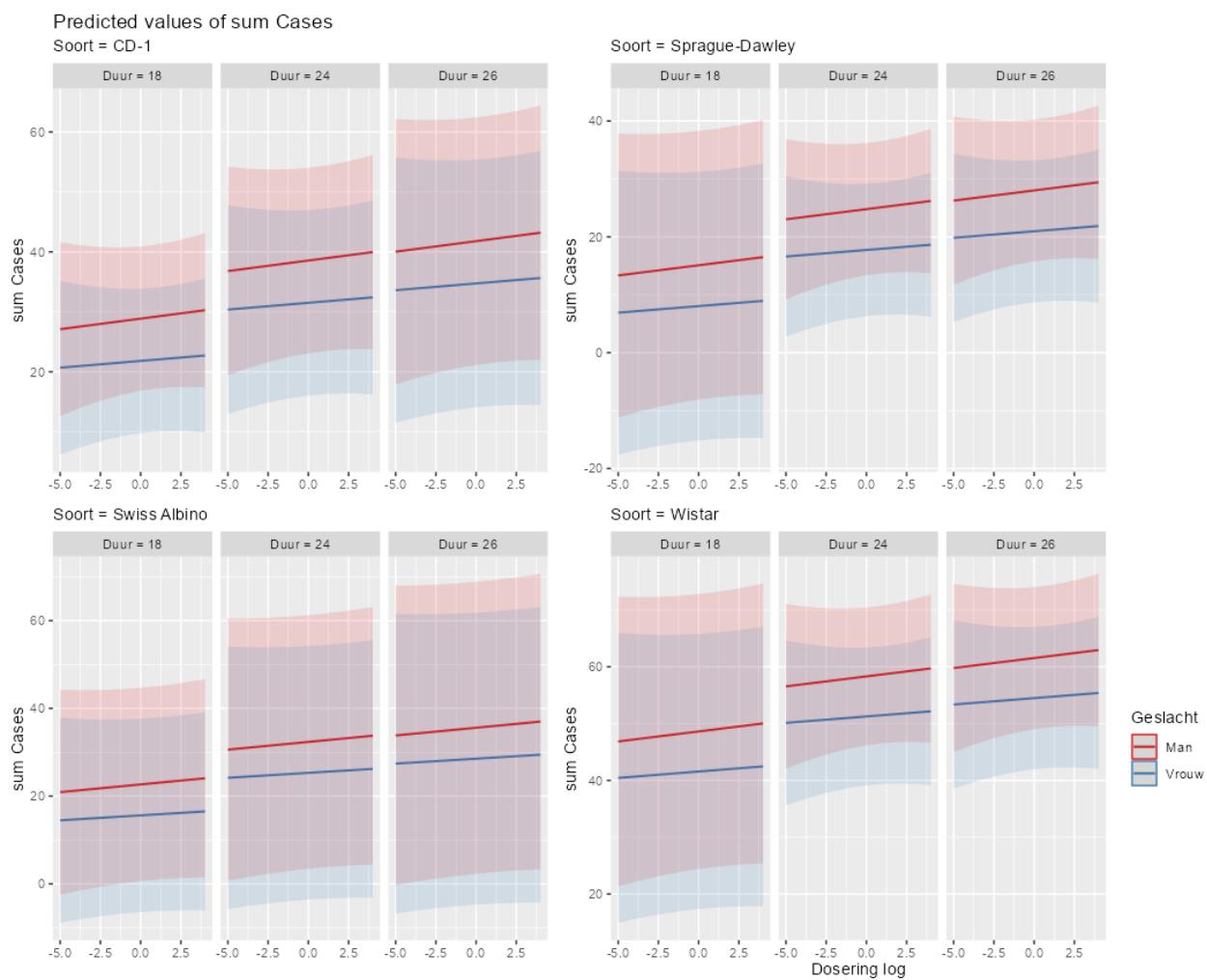
**Figuur 91.** Voorspellingen afkomstig van een LMM model met als y-variabele het opgeteld aantal kankergevallen (*sum Cases*). De lijn laat de voorspelde waarde zien als functie van de dosering (log schaal), geslacht en soort.

Predictors	sum Cases		
	Estimates	CI	p
(Intercept)	-23.63	-91.26 – 43.99	0.489
Dosering log [1st degree]	-4.18	-29.56 – 21.21	0.745
Dosering log [2nd degree]	13.89	-29.11 – 56.89	0.523
Dosering log [3rd degree]	8.33	-17.97 – 34.63	0.531
Geslacht [Vrouw]	-5.57	-18.70 – 7.56	0.402
Duur	2.49	-0.75 – 5.74	0.131
Soort [Sprague-Dawley]	-12.65	-32.55 – 7.24	0.210
Soort [Swiss Albino]	0.28	-25.25 – 25.81	0.983
Soort [Wistar]	24.78	4.88 – 44.68	<b>0.015</b>
Dosering log [1st degree]	7.71	-26.41 – 41.83	0.655
* Geslacht [Vrouw]			
Dosering log [2nd degree]	-18.67	-76.88 – 39.55	0.526
* Geslacht [Vrouw]			
Dosering log [3rd degree]	9.84	-24.37 – 44.05	0.569
* Geslacht [Vrouw]			
<b>Random Effects</b>			
$\sigma^2$	284.12		
$\tau_{00}$ Studie	87.32		
ICC	0.24		
N Studie	13		
Observations	106		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.414 / 0.551		

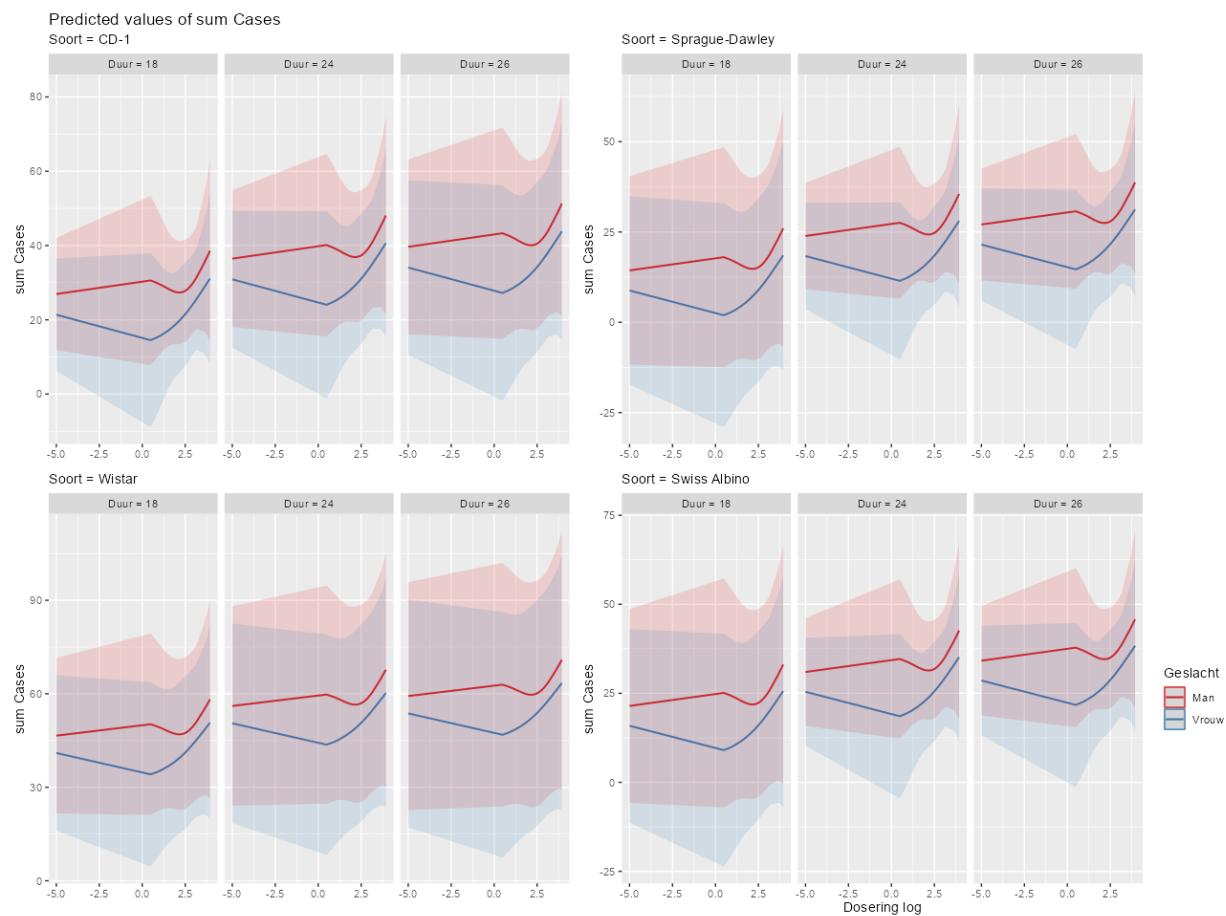
**Figuur 92.** Resultaat van een LMM, gelijk aan **Figuur 88**, maar dan met een non-lineaire relatie tussen dosering en het opgeteld aantal kankergevallen.



**Figuur 93.** Restwaarden van een LMM met natural splines. De waaier op het einde is te groot, wat maakt dat het model ook te veel te wensen overlaat.



**Figuur 94.** Voorspellingen vanuit het model per geslacht, soort en duur. De relatie is duidelijk lineair gemodelleerd, maar de nulhypothese ( $\text{coëfficiënt}=0$ ) kan niet worden vervangen. Daarvoor zijn de onzekerheidsbanden te groot.



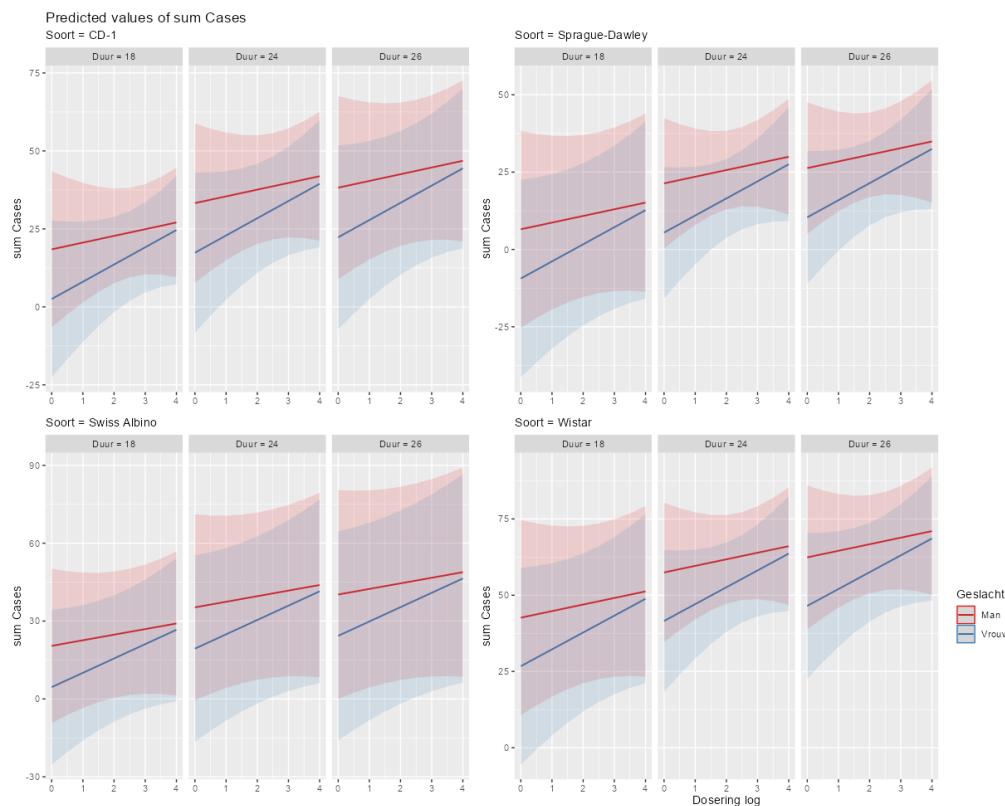
**Figuur 95.** Een LMM met natural splines die gek gedrag laat zien. Dit is geen behulpzaam model.

De vorige modellen laten geen relatie zien tussen dosering en het aantal kankergevallen. Een manier om dit wel te laten zien is door de nul-dosering weg te halen en daarna te herhalen wat hierboven al gedaan is. De resultaten van het model met een lineaire relatie kunnen we zien in **Figuur 96** en **Figuur 97**. Het model met een non-lineaire relatie (de *natural splines*) zien we in **Figuur 98** en **Figuur 99**. Beide modellen tonen inderdaad een positieve richtingscoëfficiënt, maar de onzekerheid is te groot om beslissend te zijn.

Predictors	sum Cases		
	Estimates	CI	p
(Intercept)	-26.04	-103.05 – 50.97	0.502
Dosering log	2.15	-5.52 – 9.83	0.578
Geslacht [Vrouw]	-15.91	-40.20 – 8.38	0.196
Duur	2.47	-1.06 – 6.00	0.167
Soort [Sprague-Dawley]	-11.92	-33.58 – 9.74	0.276
Soort [Swiss Albino]	2.00	-25.86 – 29.86	0.886
Soort [Wistar]	24.18	2.55 – 45.80	<b>0.029</b>
Dosering log * Geslacht [Vrouw]	3.37	-6.27 – 13.02	0.488

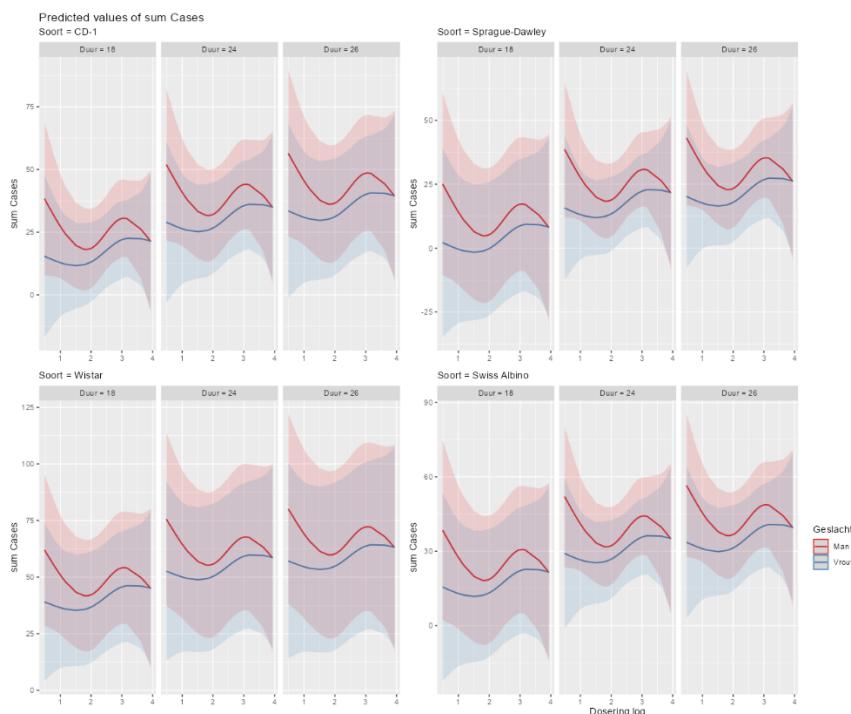
  

Random Effects	
$\sigma^2$	277.48
$\tau_{00}$ Studie	97.64
ICC	0.26
N Studie	13
Observations	80
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.401 / 0.557

**Figuur 96.** Resultaat van een LMM, gelijk aan **Figuur 88**, maar dan zonder de nul-dosering.**Figuur 97.** Voorspellingen vanuit het model per geslacht, soort en duur. De nul-dosering is verwijderd. De relatie heeft nu een meer duidelijke richtingscoëfficiënt, maar toont nog steeds grote onzekerheden rondom deze schattingen. Dit maakt dat ook in dit model de nulhypothese niet verworpen kan worden.

Predictors	sum Cases		
	Estimates	CI	P
(Intercept)	-2.30	-79.96 – 75.36	0.953
Dosering log [1st degree]	7.58	-15.54 – 30.70	0.515
Dosering log [2nd degree]	-37.58	-97.62 – 22.46	0.216
Dosering log [3rd degree]	-0.36	-28.35 – 27.63	0.980
Geslacht [Vrouw]	-22.95	-56.23 – 10.33	0.173
Duur	2.26	-1.16 – 5.68	0.192
Soort [Sprague-Dawley]	-13.26	-34.24 – 7.72	0.211
Soort [Swiss Albino]	0.15	-26.85 – 27.15	0.991
Soort [Wistar]	23.65	2.70 – 44.60	<b>0.028</b>
Dosering log [1st degree]	3.82	-26.61 – 34.25	0.803
* Geslacht [Vrouw]			
Dosering log [2nd degree]	38.44	-39.84 – 116.73	0.330
* Geslacht [Vrouw]			
Dosering log [3rd degree]	10.47	-24.97 – 45.90	0.557
* Geslacht [Vrouw]			
<b>Random Effects</b>			
$\sigma^2$	281.47		
$\tau_{00}$ Studie	86.89		
ICC	0.24		
N Studie	13		
Observations	80		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.416 / 0.554		

**Figuur 98.** Resultaat van een LMM, gelijk aan **Figuur 92**, maar dan zonder de nul-dosering toegevoegd.



**Figuur 99.** Non-lineaire relatie op basis van natural splines vanuit een model zonder nul-dosering. In dit model valt heel duidelijk te zien dat de relatie tussen dosering en opgeteld aantal kankergevallen gekke patronen begint te vertonen. Deze patronen zijn niet echt, maar eerder het gevolg van grote onderlinge variatie gecombineerd met een model wat aan [overfitting](#) doet.

De laatste pagina's waren wellicht wat ruw voor de ogen met zoveel grafieken en tabellen, maar beiden zijn belangrijk genoeg om te laten zien. Net als Portier modelleren we de dose-response relatie nu ook met een regressie waarbij we gebruik maken van een algoritme wat uitstekend werk doet in het modelleren van variantie.

Geen van de modellen weet een significante relatie te vinden tussen de dosering en het aantal kankergevallen. Dat wil niet zeggen dat dit het einde is van de analyse. Zo kunnen we de y-variabele aanpassen en aan de slag gaan met zogenaamde Generalized Linear Mixed Models (GLMMs). Daarnaast kunnen we ook nog aan de slag met Bayesiaanse analyses. Laten we eerst maar kijken naar de GLMMs.

## Generalized Linear Mixed Model

---

*The C-A trend test belongs to the general class of logistic regression models. To evaluate the consistency of a tumor finding across multiple studies using the same sexspecies- strain combinations, logistic regression with individual background responses and dose trends are fit to the pooled data using maximum likelihood estimation.*

---

Bovenstaand stuk tekst komt uit de studie van Portier. Het is een beschrijving van zijn manier om een dose-response analyse uit te voeren door alle studies met dezelfde diersoort te combineren. Een meer gedetailleerde beschrijving volgt in **Figuur 100**.

Deze analyse werd dus gedaan per diersoort, maar wat niet direct zichtbaar is, is welke y-variabele is gehanteerd. Omdat er gebruik is gemaakt van logistische regressie ga ik uit van de standaard notatie. Dit betekent dat de y-variabele een ratio is van het aantal kankergevallen ten op zichtte van het aantal dieren dat is meegenomen in de studie. Dit zou ook het dichtst in de buurt komen van de originele CA-test.

$$p = \frac{e^{\alpha_i + \beta \cdot dose}}{1 + e^{\alpha_i + \beta \cdot dose}} \quad (1)$$

where  $p$  is the probability of having a tumor,  $\alpha_i$  is a parameter associated with the background tumor response (dose = 0) for study  $i$  and  $\beta$  is a parameter associated with a change in the tumor response per unit dose (slope). A common positive trend is seen in the pooled analysis when the null hypothesis that the slope is 0 ( $H_0: \beta = 0$ ) is rejected (statistical  $p$ -value  $\leq 0.05$  using a likelihood-ratio test) in favor of the alternative that the slope is greater than 0 ( $H_A: \beta > 0$ ). The heterogeneity of slopes (all studies have different slopes vs all studies have a common slope) is tested using the model:

$$p = \frac{e^{\alpha_i + \beta_i \cdot dose}}{1 + e^{\alpha_i + \beta_i \cdot dose}} \quad (2)$$

where  $p$  and  $\alpha_i$  are as in equation (1) and  $\beta_i$  is a parameter associated with the slope for study  $i$ . Heterogeneity is seen in the pooled analysis when the null hypothesis that the slopes are equal ( $H_0: \beta_1 = \beta_2 = \beta_3 = \dots$ ) is rejected (statistical  $p$ -value  $\leq 0.05$  using a likelihood-ratio test) in favor of the alternative that at least one of the slopes is different.

**Figuur 100.** Knipsel uit de studie van Portier (pagina 4).

Om toch enige duiding te krijgen kunnen we verder kijken in de tekst waar we het volgende lezen:

*For CD-1 mice, there are studies of 18 months (3) and 24 months (2) so analyses are conducted separately for 18 month studies and 24 month studies and then a combined analysis is performed. In SD rats, one study had 26 months of exposure and the remaining 3 had 24 months of exposure so similar grouped analyses are conducted.*

Het lijkt erop dat de analyses per diersoort zijn gedaan en per duur van de studie. Omdat we te maken hebben met een klein aantal studies zal dit de onzekerheid van een analyse alleen maar vergroten. Het is lastig om hier een keuze te maken: combineren we de studies en accepteren we extra risico met het introduceren van extra variantie, of voeren we een groot aantal sub-analyses uit en accepteren we het extra risico op extra onzekerheid én vals positieven? Ik zal in mijn eigen analyses de resultaten van Portier (voor zover ik deze kan duiden uit de studie) meenemen om te laten zien waar de verschillen liggen (of de

overeenkomsten). Maar ik zal niet starten met de manier waarop het door Portier is gedaan. Dat is omdat ik de studies eerst wil analyseren op de manier waarop ik het zou doen. Vervolgens haal ik het werk van Portier aan om te vergelijken, maar niet om te repliceren. Ik kan namelijk niet exact zien wat er gedaan is.

De term Generalized Linear Mixed Model, of GLMM, is een type model wat gelijk is aan een LMM in de manier waarop de data wordt geïnterpreteerd<sup>95</sup>. Dit betekent dat ook een GLMM eerst kijkt naar de verschillende variantiecomponenten in een dataset. Het grote verschil met een LMM is dat de assumptie van normaliteit mag worden losgelaten. Dat betekent dat een  $y$ -variabele in een GLMM niet hoeft te bestaan uit een continue reeks van getallen, waarvan de restwaarden normaal verdeeld zijn, maar ook discreet mag zijn (afwezigheid van decimalen). Het tellen van het aantal kankergevallen wordt gezien als een discrete variabele. Een ratio is een waarde die continu lijkt, maar eigenschappen kent die anders zijn dan de eigenschappen van bijvoorbeeld een temperatuurmeter. Dit is omdat bij een ratio de variantie afhankelijk is van de waarde zelf. Een ratio is dus een waarde die met een GLMM wordt gemodelleerd.

We kunnen en mogen dus niet zomaar een LMM gebruiken voor elk soort data. De reden dat het enigszins werkzaam was in het vorige voorbeeld is het gevolg van het optellen van het aantal tumorsoorten. Toch was dit geen perfecte toepassing, omdat er geen negatieve getallen kunnen voorkomen. Dat betekent dat we werken met een telling die een harde ondergrens heeft én deze ondergrens ook laat zien. LMM modellen kunnen dan tegen problemen aanlopen. Met een GLMM analyse hoeven we ons daar eigenlijk geen zorgen meer om te maken. Laten we beginnen met de meest gebruikte vorm van een GLMM: het binomiaal model. Deze werkt dus met een binomiale verdeling.

## Binomiaal model

Wie nog nooit van een binomiaal model gehoord heeft, hoeft alleen maar te denken aan het opgooien van een munt. In **Figuur 41** liet ik al zien dat het opgooien van een munt een exercitie is die moet aantonen dat een munt wel of niet zuiver is. Dat wil zeggen dat de kans

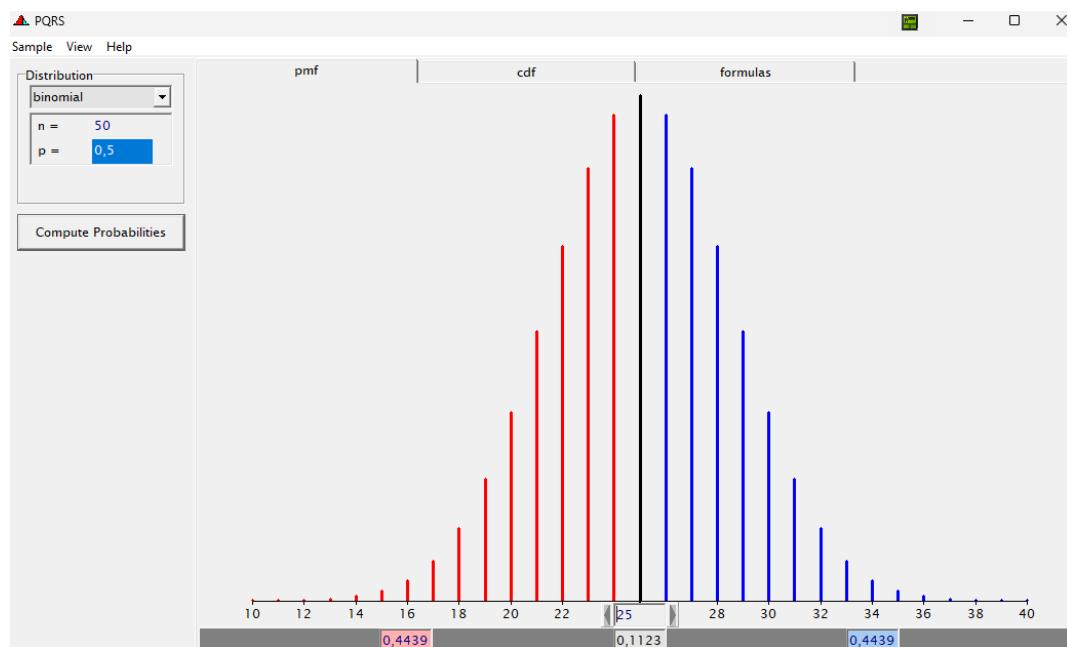
---

<sup>95</sup> Theoretisch gezien, maar wiskundig gezien zijn er heel wat verschillen in de manier waarop de berekening tot stand komt. Een GLMM is vele malen complexer om te berekenen en ook de wiskunde is nog niet helemaal af. Een perfect introductieboek is het werk van Walter Stroup, Marina Pthukhina en Julie Garai (<https://www.amazon.nl/-/en/Walter-W-Stroup/dp/1498755569>).

op kop of munt ongeveer 50% is. In hetzelfde figuur was goed zichtbaar dat de empirische observatie voor het aantonen van de correctheid van het 50%-model aardig wat handelingen met zich meebrengt. Uiteindelijk is een oneindig aantal momenten nodig om exact op 50% uit te komen.

Een binomiaal model verschilt van een normaal model<sup>96</sup> in de hoeveelheid parameters die we kunnen bepalen: dit hebben we ook al een aantal keren besproken in het voorbeeld over de lengte van Nederlandse mannen en vrouwen. Bij een normaalverdeling heb je twee parameters: gemiddelde én spreiding. Bij een binomiaal model kunnen we nog steeds spreken van gemiddelde en spreiding, maar in dit geval is de spreiding afhankelijk van het gemiddelde en het gemiddelde is geen echt gemiddelde maar een ratio (beide zijn verwachtingen).

Een makkelijke manier om dit te laten zien is door het binomiale model toe te passen. Stel, we zeggen dat de kans op kop of munt daadwerkelijk 50% is. En we doen 50 observaties. Hoe vaak kop en hoe vaak munt mogen we dan verwachten?<sup>97</sup>



**Figuur 101.** De verdeling van mogelijke uitkomsten bij het 50 keer opgooien van een munt afkomstig uit een binomiaal verdeling met kans 50%.

<sup>96</sup> Ik bedoel hier uiteraard een model dat gebruik maakt van de normaalverdeling.

<sup>97</sup> Zowel Figuur 41 als eigen observaties laten eenduidig zien dat 25 keer kop of 25 keer munt geen gegeven is bij het 50 keer opgooien van een munt. We mogen dus ook niet verwachten dat een binomiaal model zomaar als antwoord ‘25’ uitspuugt.

**Figuur 101** laat zien wat de verdeling van mogelijkheden is bij het 50 keer opgooien van een munt. **Figuur 102** laat zien wat de achterliggende formules zijn.

```
P(X=x) = (nCx) p^x (1-p)^{n-x} for x = 0, 1, ..., n
P(X = 25) = 0,112275172659
Expectation = np = 25
Variance = np(1 - p) = 12,5
Standard deviation = 3,535533905933
Moment generating function M(t) = (1 - p + pe^t)^n
```

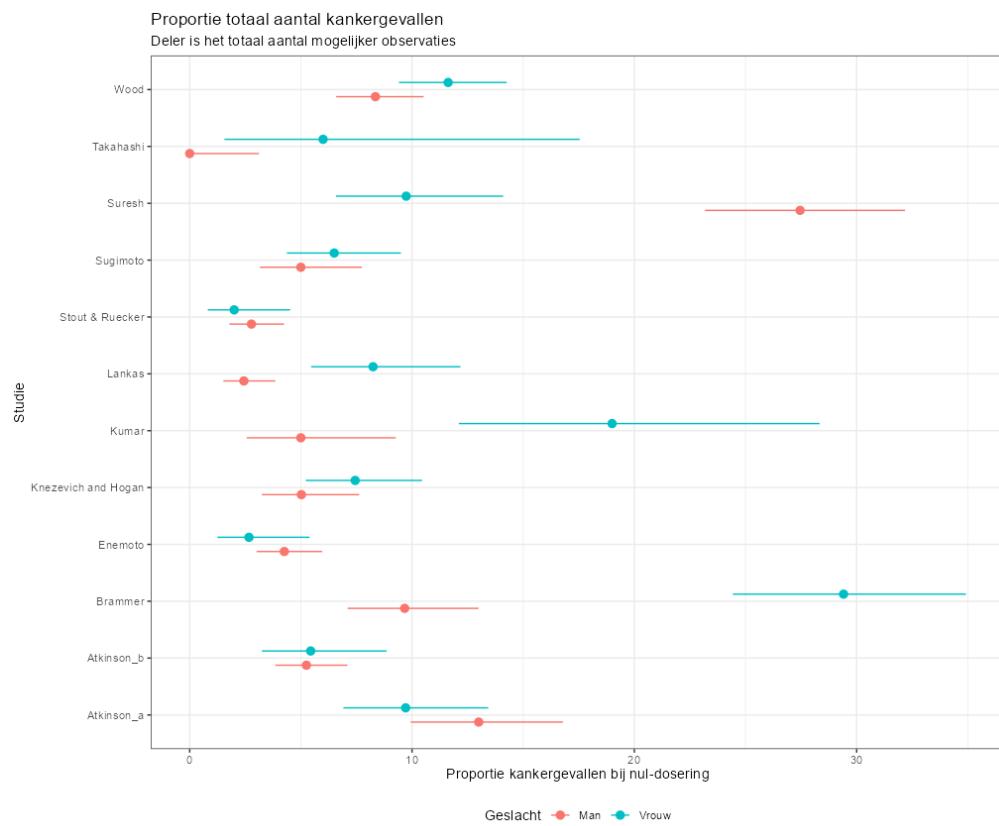
The distribution of the total number of successes in a series of n independent Bernoulli trials

**Figuur 102.** Formules die horen bij een binomiaal verdeling. De uitkomsten zijn gebaseerd op 50 observaties en een verwachte kans van 50%.

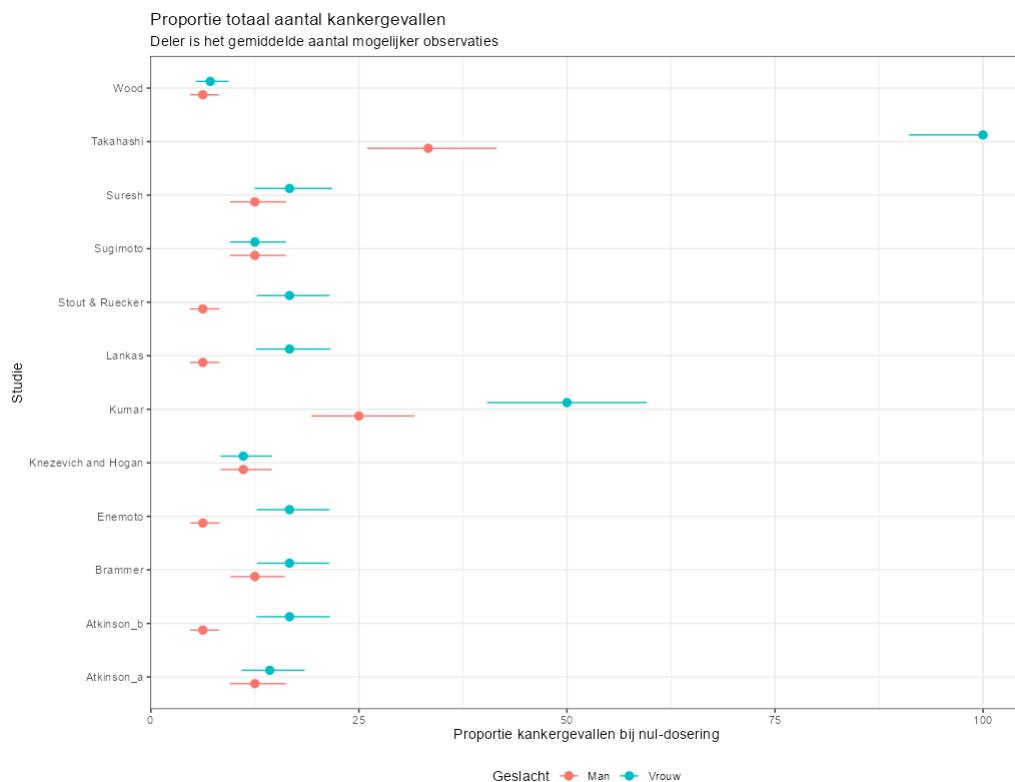
Wat opvalt is dat het verwacht aantal keren kop of munt inderdaad 25 is, maar dat de variantie een afgeleide is van de voorgestelde kans (50%) en het aantal keer munt opgooien. We hebben hier te maken met een deterministische formule. De lezer snapt nu wellicht direct waarom de normaalverdeling vaak zo aantrekkelijk is: het onafhankelijk kunnen schatten van de variantie brengt extra vrijheid mee. In dit geval maakt het allemaal weinig uit: een gemiddelde toepassen op een ratio leidt tot gekke modellering.

Een GLMM is net als een LMM een mooie manier om met variantie te werken. **Figuur 103** laat zien wat de proportiekankergevallen bij de nul-dosering is, per studie en per geslacht, met onzekerheidsinterval<sup>98</sup>. Dit is ongeacht het soort kanker, omdat dit de grafiek alleen maar zou verwateren. Wat direct opvalt is het grote onderscheid tussen proporties in zowel één studie en ook tussen studies. Omdat een proportie afhankelijk is van een deler kunnen we deze aanpassen. Dat is wat ik laat zien in **Figuur 104** nadat ik de deler transformeer tot het gemiddelde aantal mogelijke observaties. Wat dan opvalt is dat de spreiding tussen studies een stuk kleiner is, maar eigenlijk is dit geen juiste weergave van de data. Een proportie van het aantal kankergevallen, zeker als we kijken naar de nul-dosering, is gewoonweg het aantal kankergevallen delen door het aantal observaties. Dat is **Figuur 103.**

<sup>98</sup> Ik kan niet vaak genoeg noemen hoe onzekerheid wordt vermeden om een relatie scherper voor te stellen dan deze op basis van de data mag worden aangenomen.



**Figuur 103.** Proportie kankergevallen bij nul-dosering waarbij de deler het totaal aantal mogelijke observaties is.



**Figuur 104.** Proportie kankergevallen bij nul-dosering waarbij de deler het gemiddelde aantal mogelijke observaties is.

Laten we nu eens gaan rekenen. Om te beginnen start ik een model waarbij ik alle data combineer, en variabelen zoals dosering, geslacht, soort en duur van de studie meeneem. In feite is het model gelijk aan de vorige modellen met als enige verschil dat ik nu modelleer met een binomiale verdeling. Het resultaat zien we in **Figuur 105**<sup>99</sup> waarbij er wel degelijk een verschil lijkt te zijn in geslacht en tussen soorten.

cbind(Cases,N-Cases)			
Predictors	Odds Ratios	CI	p
(Intercept)	0.08	0.06 – 0.11	<0.001
Dosering log	1.02	0.99 – 1.05	0.170
Geslacht [Vrouw]	1.25	1.03 – 1.51	0.026
Duur [18]	0.70	0.46 – 1.05	0.084
Duur [26]	1.43	0.86 – 2.39	0.166
Soort [Sprague-Dawley]	0.44	0.30 – 0.66	<0.001
Soort [Swiss Albino]	2.82	1.66 – 4.77	<0.001
Soort [Wistar]	2.19	1.48 – 3.25	<0.001
<b>Random Effects</b>			
$\sigma^2$	3.50		
$\tau_{00}$ ID	0.21		
$\tau_{00}$ Studie	0.02		
ICC	0.01		
N Studie	13		
N ID	106		
Observations	106		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.101 / 0.106		

**Figuur 105.** Resultaat van een GLMM met een binomiale verdeling.

Alvorens we de resultaten helemaal tot ons nemen is het verstandig om stil te staan bij het onderdeel Random Effects: dit is het gedeelte van het model waarin wordt gekeken welke factoren bijdragen aan de variatie. Zo zagen we in **Figuur 103** al dat de proportie

<sup>99</sup> De cbind(Cases, N-Cases) is een lelijke notatie die ik makkelijk had kunnen veranderen. Ik heb dat bewust niet gedaan omdat ik de lezer wil laten zien hoe de ‘y-variabele’ tot stand is gekomen.

kankergevallen bij de nul-dosering verschilt over de studies. Dat betekent dat het facet *Studie* een behoorlijke bijdrage lijkt te leveren aan de variatie, ook al kan het best zijn dat *Studie* zelf niet de juiste naam is voor die bijdrage<sup>100</sup>. In de uitkomst van dit model lijkt *Studie* helemaal niet zoveel bij te dragen en dit komt hoogstwaarschijnlijk door een extra toevoeging die ik heb gedaan. Ik zal dit nu proberen uit te leggen.

In een GLMM model wordt de variantie vaak verkeerd geschat. Dit komt door de beperkte vrijheidsgraden van de gehanteerde verdelingen. Wanneer variantie als kleiner wordt geschat dan deze daadwerkelijk is, wordt de kans op een vals positieve groter. We noemen dit in het Engels *underdispersion* en dit is het directe gevolg van het niet meer zelf kunnen bepalen van de variantie van het theoretisch model. In de binomiaalverdeling is de variantie namelijk afhankelijk van het aantal observaties vermenigvuldigd met de kans op succes (hier tumorsoort). Wanneer de variantie wordt overschat noemen we dat *overdispersion*.

Om dit enigszins te omzeilen heb ik een aparte kolom aangemaakt zodat het model wel in staat is om variatie te bepalen. Dit doe ik door elke rij een unieke code te geven ('ID') en vervolgens de variatie tussen de rijen uitreken. Dit maakt dat de variantie die anders onder *Studie* zou vallen nu (gedeeltelijk) naar *ID* gaat. Het resultaat zonder *ID* is zichtbaar in **Figuur 106**. We zien duidelijk dat studie niet alle variantie verklaart.

---

<sup>100</sup> Één van de belangrijkste onderdelen van het modelleren is het interpreteren van de variabelen en de relaties tussen variabelen in een model. Als in een model blijkt dat het facet ‘studie’ behoorlijk variantie verklarend is, dan is het nog maar de vraag wat maakt dat wij dit vinden tussen studies. Welk onderdeel, of welke onderdelen tussen studies, maakt dat er variatie is? Dat is vaak helemaal niet zo makkelijk om te zien.

<i>Predictors</i>	cbind(Cases,N-Cases)		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.08	0.06 – 0.12	<0.001
Dosering log	1.02	1.01 – 1.03	0.004
Geslacht [Vrouw]	1.22	1.13 – 1.31	<0.001
Duur [18]	0.67	0.43 – 1.07	0.092
Duur [26]	1.18	0.67 – 2.08	0.573
Soort [Sprague-Dawley]	0.49	0.31 – 0.76	0.002
Soort [Swiss Albino]	2.78	1.55 – 4.99	0.001
Soort [Wistar]	2.30	1.48 – 3.57	<0.001
<b>Random Effects</b>			
$\sigma^2$	3.29		
$\tau_{00}$ Studie	0.06		
ICC	0.02		
N Studie	13		
Observations	106		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.103 / 0.118		

**Figuur 106.** Resultaat van een GLMM met een binomiale verdeling zonder de correctie

Het kan trouwens goed zijn dat ik geen goed model heb gemaakt. Een vlugge blik op **Figuur 103** laat namelijk zien dat er verschillen zijn tussen geslacht in een studie. Misschien moeten we het facet *Geslacht* in het model wel helemaal anders bepalen en namelijk als variantiecomponent en niet als individuele voorspeller voor kankergevallen.

Het resultaat zien we in **Figuur 107**. We hebben nu meerdere *Random Effects*, waarbij er zowel een ‘Studie’ als een ‘ID’ component is, maar ook een *Studie:GeslachtVrouw* component. Laatstgenoemde kijkt naar de variatie tussen geslachten in een studie. Het probleem met deze toevoeging is echter tweevoudig:

1. Variatie berekenen tussen twee mogelijkheden kent te weinig afwijkingen om een goede schatter te zijn.

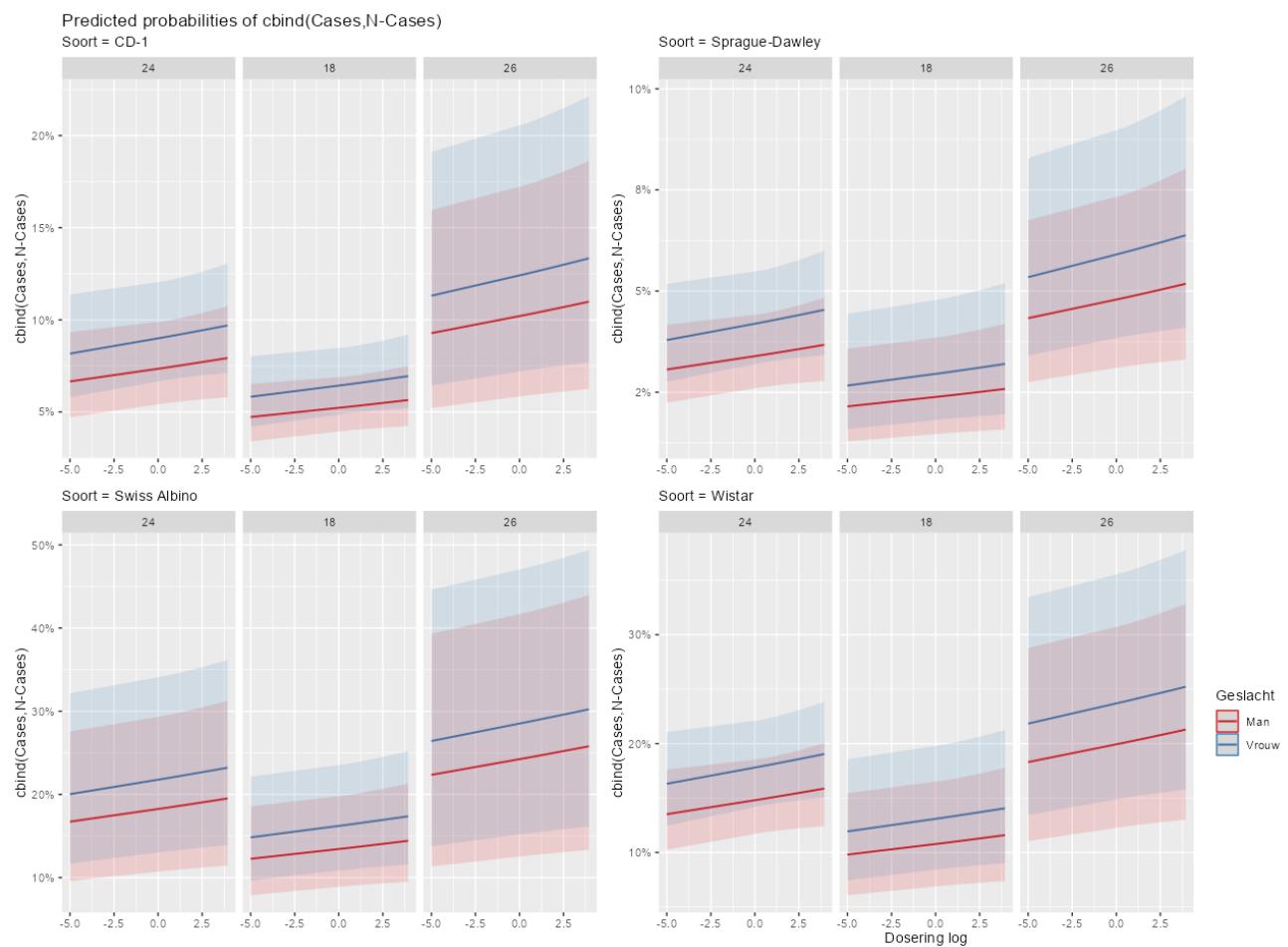
2. De correlatie tussen studies is nu nagenoeg 1 ( $p_{01}$  Studie). Dit is altijd een teken dat het model meer parameters heeft dan zinvol is wat weer kan leiden tot problemen in het uitrekenen van een stabiele oplossing.

chind(Cases,N-Cases)			
Predictors	Odds Ratios	CI	p
(Intercept)	0.09	0.07 – 0.11	<0.001
Dosering log	1.02	1.00 – 1.04	0.039
Duur [18]	0.79	0.58 – 1.08	0.135
Duur [26]	1.86	1.21 – 2.85	0.005
Soort [Sprague-Dawley]	0.46	0.34 – 0.63	<0.001
Soort [Swiss Albino]	3.16	2.08 – 4.80	<0.001
Soort [Wistar]	2.43	1.80 – 3.28	<0.001
<b>Random Effects</b>			
$\sigma^2$	3.35		
$\tau_{00}$ ID	0.06		
$\tau_{00}$ Studie	0.32		
$\tau_{11}$ Studie.GeslachtVrouw	0.56		
$\rho_{01}$ Studie	-0.98		
ICC	0.09		
N Studie	13		
N ID	106		
Observations	106		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.100 / 0.179		

**Figuur 107.** Resultaat van een GLMM met een binomiale verdeling waarin een extra variatiecomponent is opgenomen - *Studie:GeslachtVrouw*.

Wat duidelijk mag zijn is waarom Portier heeft gekozen om de analyses per soort en geslacht op te delen. Het is gewoonweg eenvoudiger om het zo te doen. Maar, zoals gezegd, wil ik eerst mijn eigen model de revue laten passen en dus ga ik aan de slag met het model uit

**Figuur 106.** Omdat ik de variabelen soort, geslacht en duur heb toegevoegd kan ik uit het model de dose-response voor elk van deze combinaties tonen. Het resultaat is **Figuur 108**.



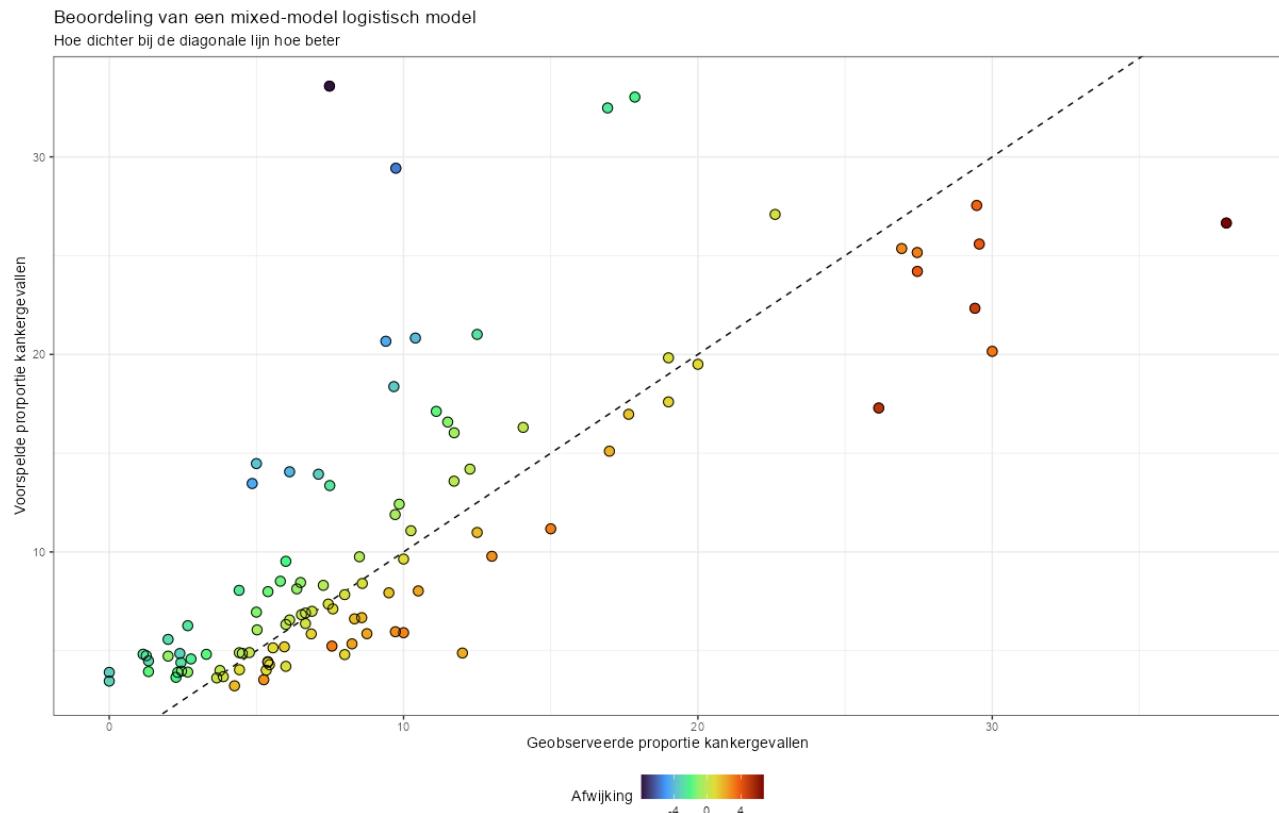
**Figuur 108.** De voorspellingen uit het GLMM model zoals weergegeven in **Figuur 106**.

Wat wederom opvalt is dat ik door elk van deze facetten een rechte lijn kan trekken. Dat mag niet verbazen: het model gaf zelf al aan dat een relatie tussen dosering en de y-variabele niet significant was (lees: de variatie op de richtingscoëfficiënt was té groot om een signaal op te pikken).

Een interpretatie van het model en haar uitkomsten heeft weinig zin als er geen beoordeling is van het model zelf. Omdat een afbeelding die de restwaarden toont niet meer zo eenvoudig te interpreteren is in een binomiaal model moet ik iets anders verzinnen<sup>101</sup>. Wat ik wel kan doen is tonen of dat wat het model voorspelt gelijk is wat het model observeert. Onder de assumptie van Maximum Likelihood zou dit (hoewel zeker niet perfect) een duiding kunnen zijn van de adequaatheid van het model. Het resultaat van deze

<sup>101</sup> De restwaarden hoeven niet meer normaal verdeeld te zijn, noch is het nodig dat de variatie overal gelijk is. Steker nog: omdat de variatie een functie is van het gemiddelde is deze per definitie overal anders.

exercitie, die ik hier nu grafisch weergeef<sup>102</sup>, is te zien in **Figuur 109**. Wat opvalt is dat de algemene tendens wel wordt gevolgd, maar dat er wel degelijk verschillen zitten tussen de geobserveerde proportie en de voorspelde proportie. Dit wil zeggen dat conclusies trekken uit het huidige model niet de beste basis heeft. Eigenlijk moeten we verder gaan kijken.



**Figuur 109.** De beoordeling van het GLMM model met binomiale verdeling. Goed te zien is de afstand tussen wat het model voorspeld aan proportie kankergevallen en wat daadwerkelijk is geobserveerd.

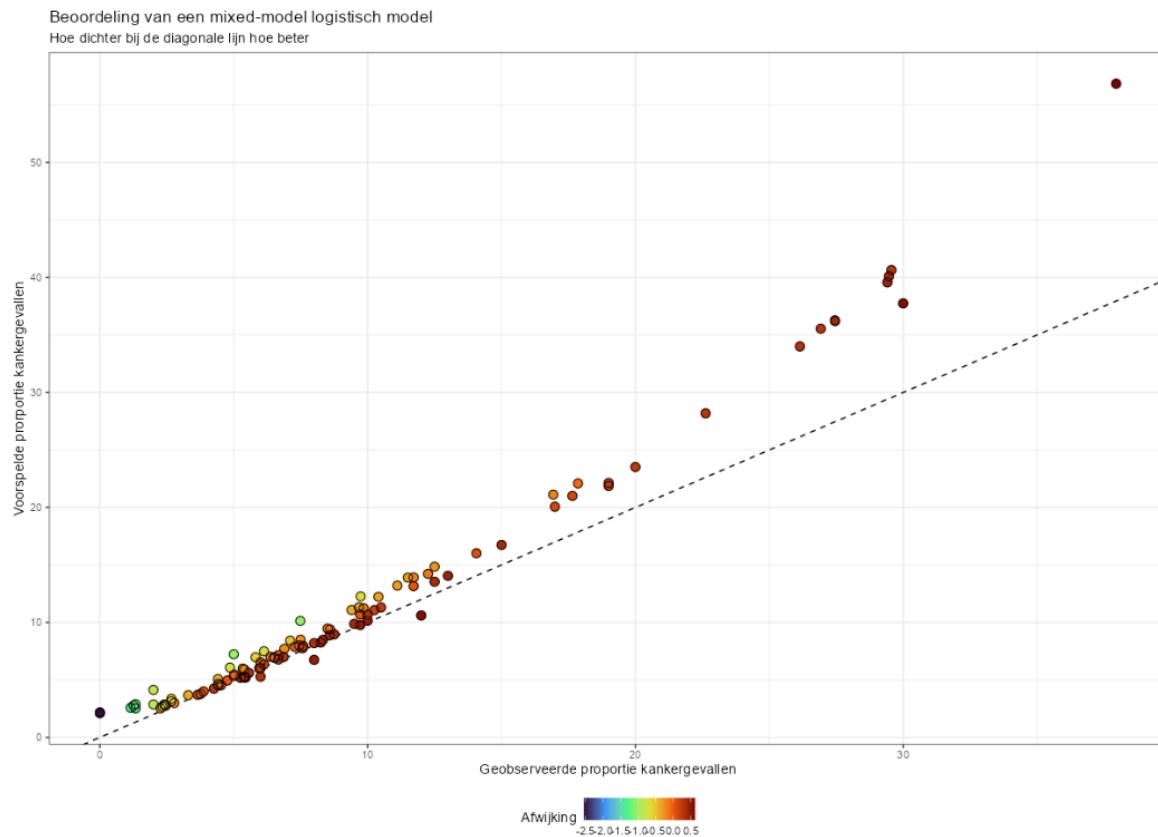
We kunnen, net als bij de eerdere LMM modellen, proberen om weer met *splines* te werken. Met *splines* werken is altijd een beetje ‘gevaarlijk’ door hun sporadisch gedrag bij weinig datapunten<sup>103</sup>

Laten we de toch de proef op de som nemen en kijken wat een model met *splines* ons laat zien. Voordat we de resultaten tonen is het wellicht beter om eerst de beoordeling te laten zien. Dan weten we ook of het de moeite waarde is om überhaupt naar de resultaten te kijken. Het resultaat van een GLMM model met *splines* zien we in **Figuur 110**.

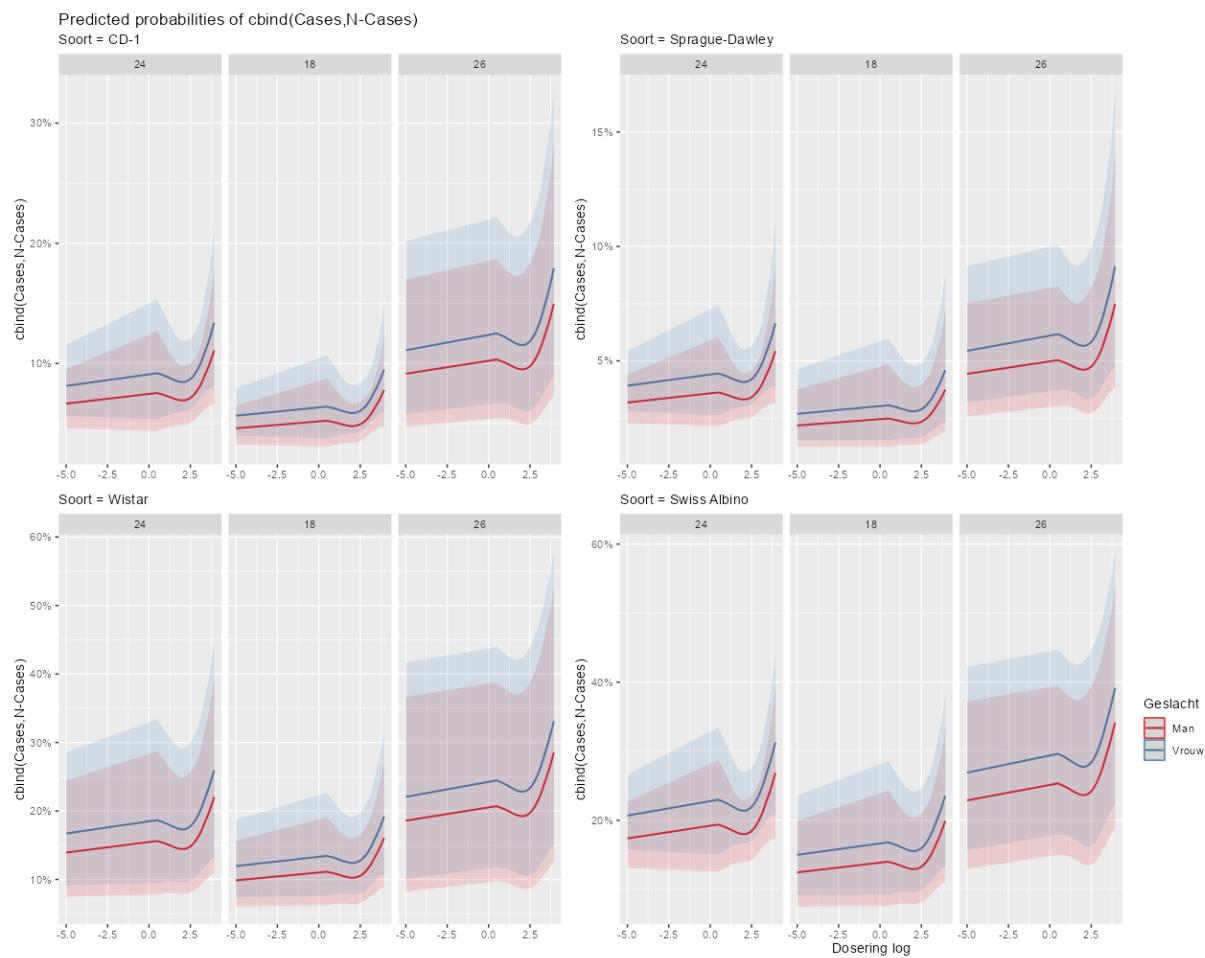
<sup>102</sup> Er zijn officiële waarden die bepalen hoe geschikt een model is ten opzichte van de data en ten opzichte van de andere modellen. Voor nu is het beter om dit grafisch te doen.

<sup>103</sup> Meest studies hebben maar 4 doseringen.

Hoewel het lijkt alsof dit model het veel beter doet is hier eigenlijk sprake van overfitting: het model zit zo dicht op de geobserveerde waarden dat het geen zinvol model meer is. Een betere manier om dit te laten zien is door middel van **Figuur 111**.



**Figuur 110.** De beoordeling van het GLMM model met binomiale verdeling en met splines. Dit model laat een vele sterkere correlatie zien tussen geobserveerde en voorspelde proporties kankergevallen, maar neigt sterk naar *overfitting*.



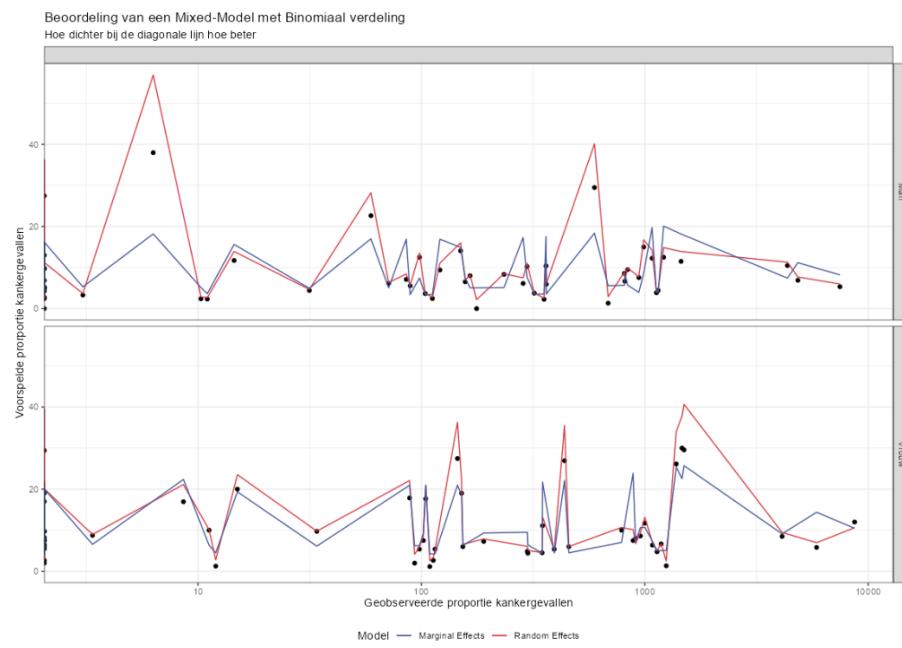
**Figuur 111.** De voorspelde waarden afkomstig van het GLMM model met binomiale verdeling en met splines.

De grote knik op het einde is niet logisch. Wie toch twijfelt of dit model niet gewoon een goed model is kunnen we meenemen naar **Figuur 112**. Dit model toont, per geslacht, de geobserveerde proporties. De rode en blauwe lijn zijn beiden afkomstig van het model uit **Figuur 110**. De blauwe lijn toont het gemiddelde, of marginale, model. De rode lijn toont de GLMM met alle conditionele effecten. Dit betekent dat we nu ook de *random effects* meenemen. Wat zichtbaar is, is dat beide modellen verschillen waarbij het rode model wel heel dicht tegen de geobserveerde punten aanligt. Dit is een teken dat het model aan *overfitting* doet.

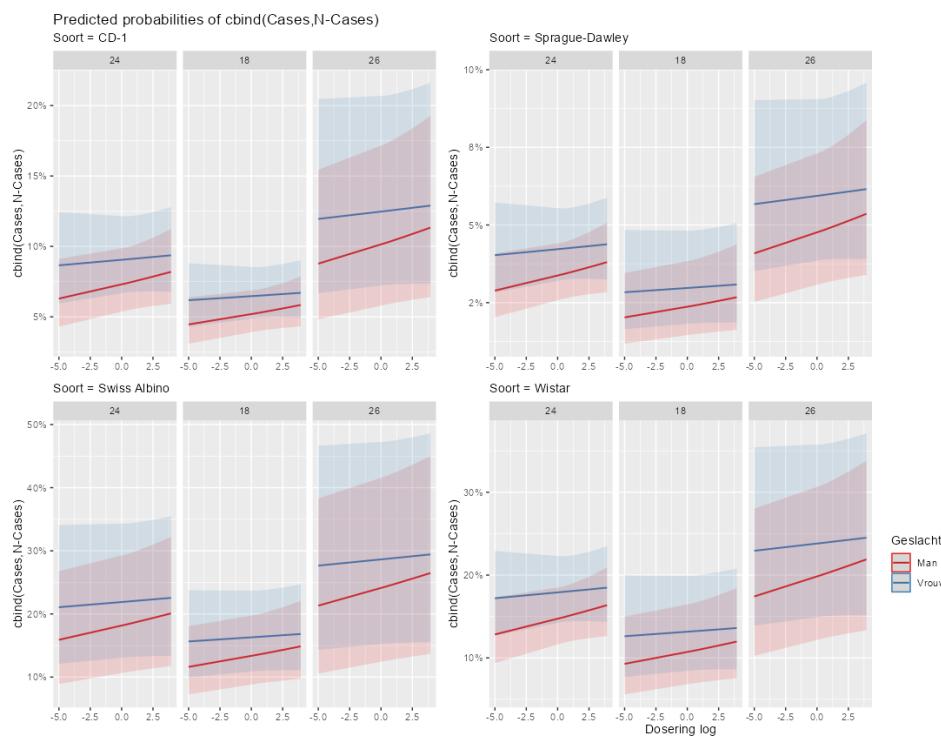
Ook als we dit model wel zouden hanteren, dan nog zouden we, wederom, geen relatie vinden tussen dosering en de proportie kanker. Maar wie een kijk neemt in de studie van Portier ziet wel degelijk dikgedrukte letters die aangeven dat voor de CD-1 muizen er kancersoorten zijn die wel degelijk doseringsafhankelijk zijn. Ons rest daarom nog één model

en dat is het model met de interactie tussen dosering en geslacht. Het resultaat zien we in

**Figuur 113.**

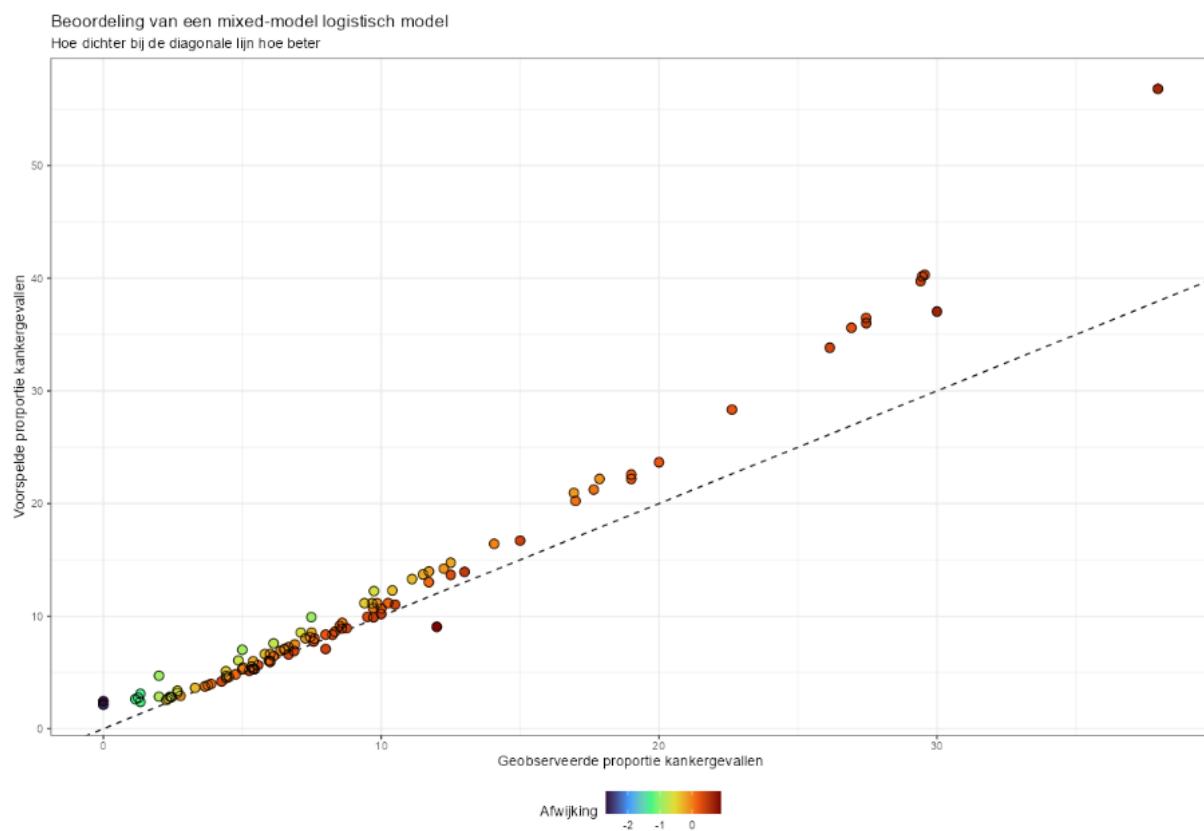


**Figuur 112.** De beoordeling van het GLMM model met binomiale verdeling. De blauwe lijn toont het gemiddelde, of marginale, model. De rode lijn toont de daadwerkelijke GLMM met alle conditionele effecten.



**Figuur 113.** Voorspelde proporties aan kanker voor een GLMM model met binomiale verdeling en een interactie tussen dosering en geslacht.

De onzekerheidsbanden zijn zo groot dat er geen significant effect kan zijn tussen de groepen op basis van de dosering. Het model zelf steekt niet eens zo slecht in elkaar (**Figuur 114**). Alleen op het einde gaan de voorspellingen mis omdat de voorspelde proportie groter is dan de geobserveerde proportie. Omdat het model grotere aantallen voorspeld dan geobserveerd zou het model eerder moeten concluderen dat er een relatie is tussen dosering en kanker. Desondanks is de dosering niet statistisch significant (**Figuur 115**). Wat ons nu rest is dus proberen om, wederom, het werk van Portier te repliceren.



**Figuur 114.** De correlatie tussen geobserveerde en voorspelde proporties kankergevallen uit een GLMM model met interactie tussen dosering en geslacht.

<i>Predictors</i>	cbind(Cases,N-Cases)		
	Odds Ratios	CI	p
(Intercept)	0.08	0.06 – 0.11	<0.001
Dosering log	1.03	0.99 – 1.08	0.135
Geslacht [Vrouw]	1.26	1.04 – 1.54	0.020
Duur [18]	0.69	0.46 – 1.05	0.083
Duur [26]	1.43	0.86 – 2.39	0.167
Soort [Sprague-Dawley]	0.44	0.30 – 0.66	<0.001
Soort [Swiss Albino]	2.82	1.66 – 4.78	<0.001
Soort [Wistar]	2.19	1.48 – 3.26	<0.001
Dosering log * Geslacht [Vrouw]	0.98	0.92 – 1.04	0.466
<b>Random Effects</b>			
$\sigma^2$	3.49		
$\tau_{00\text{ ID}}$	0.20		
$\tau_{00\text{ Studie}}$	0.02		
ICC	0.01		
N <sub>Studie</sub>	13		
N <sub>ID</sub>	106		
Observations	106		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.101 / 0.106		

**Figuur 115.** Samenvatting van een GLMM model met binomiale verdeling en een interactie tussen dosering en geslacht.

### Replicatie van het Logistisch model van Portier

In de studie van Portier worden verschillen grafieken getoond die per soort laten zien wat de resultaten zijn. Ik zal beginnen met de CD-1 muizen zoals getoond in **Figuur 116**. De kolom waar we nu op moeten focussen is de *Common Trend*: dit is het resultaat van een logistisch model met alle studies per soort, geslacht en tumorsoort. Ik zal beginnen met *Kidney Adenomas* bij mannelijke CD-1 muizen.

**Table 3** P-values for the Cochran-Armitage trend test and pooled logistic regression analysis for tumors with at least one significant trend test ( $p \leq 0.05$ ) or Fisher's exact test ( $p \leq 0.05$ ) in male and female CD-1 mice

Tumor	Individual study p-values for trend <sup>a</sup>					Common Trend	Heterogeneity Test
	A	B	C	D	E		
Males							
Kidney Adenomas	0.442 (0.138) <sup>d</sup>	0.938	0.062 ( <b>0.009</b> ) <sup>d</sup>	---	<b>0.019</b> <b>0.006</b>	0.268	
Kidney Carcinomas	0.063 (< <b>0.001</b> ) <sup>d</sup>	0.938	---	---	0.250 <b>0.031</b>	0.546	
Kidney Adenomas and Carcinomas	0.065 ( <b>0.008</b> ) <sup>d</sup>	0.981	0.062 ( <b>0.009</b> ) <sup>d</sup>	---	<b>0.005</b> < <b>0.001</b>	0.106	
Malignant Lymphomas	0.754	0.087	<b>0.016</b>	<b>0.007</b>	ND <sup>c</sup>	0.093	0.007
Hemangiosarcomas	0.505		<b>0.004</b> 0.062 ( <b>0.005</b> ) <sup>d</sup>	---	ND <sup>c</sup> <b>0.033</b>	0.007	
Alveolar-Bronchiolar Adenomas	0.294	0.231	0.513	0.924	ND <sup>c</sup>	0.384	0.409
Alveolar-Bronchiolar Carcinomas	0.918	0.456	0.148	<b>0.028</b>	ND <sup>c</sup>	0.407	0.083
Alveolar-Bronchiolar Adenomas and Carcinomas	0.576	0.231	0.294	0.336	ND <sup>c</sup>	0.346	0.826
Females	A	B	C	D	E		
Hemangiomas	0.631	---	<b>0.002</b>	0.438	ND <sup>c</sup> <b>0.031</b>	0.155	
Harderian Gland Adenomas	0.877	ND <sup>c</sup>	<b>0.040</b>	0.155	ND <sup>c</sup>	0.155	0.052
Harderian Gland Carcinomas	---	ND <sup>c</sup>	---	1.00	ND <sup>c</sup>	0.500	1.00
Harderian Gland Adenomas and Carcinomas	0.877	ND <sup>c</sup>	<b>0.040</b>	0.372	ND <sup>c</sup>	0.184	0.110
Alveolar-Bronchiolar Adenomas	0.999	0.144	0.800	0.656	ND <sup>c</sup>	0.996	0.211
Alveolar-Bronchiolar Carcinomas	0.183	0.110	0.623	0.601	ND <sup>c</sup>	0.268	0.544
Alveolar-Bronchiolar Adenomas and Carcinomas	0.985		<b>0.048</b> 0.842	0.688	ND <sup>c</sup>	0.982	0.241
Malignant Lymphomas	0.070 <sup>e</sup>	0.484	0.294	0.353	<b>0.050</b> <b>0.012</b>	0.995	

<sup>a</sup> – Study A is Knezevich and Hogan [11] (Additional file 2: Table S1), Study B is Atkinson et al. [12] (Additional file 2: Table S2), Study C is Sugimoto [13] (Additional file 2: Table S3), Study D is Wood [14] (Additional file 2: Table S4), Study E is Takahashi [15] (Additional file 2: Table S5); <sup>b</sup> – three dashes “---” indicates all tumor counts are zero; <sup>c</sup> – ND indicates there is no data available for this tumor in this study; <sup>d</sup> – using historical control data (see text for details) and Tarone's test; <sup>e</sup> – Spleen composite lymphosarcomas (malignant lymphomas) are also significantly increased in female mice in this study (see Additional file 2: Table S1)

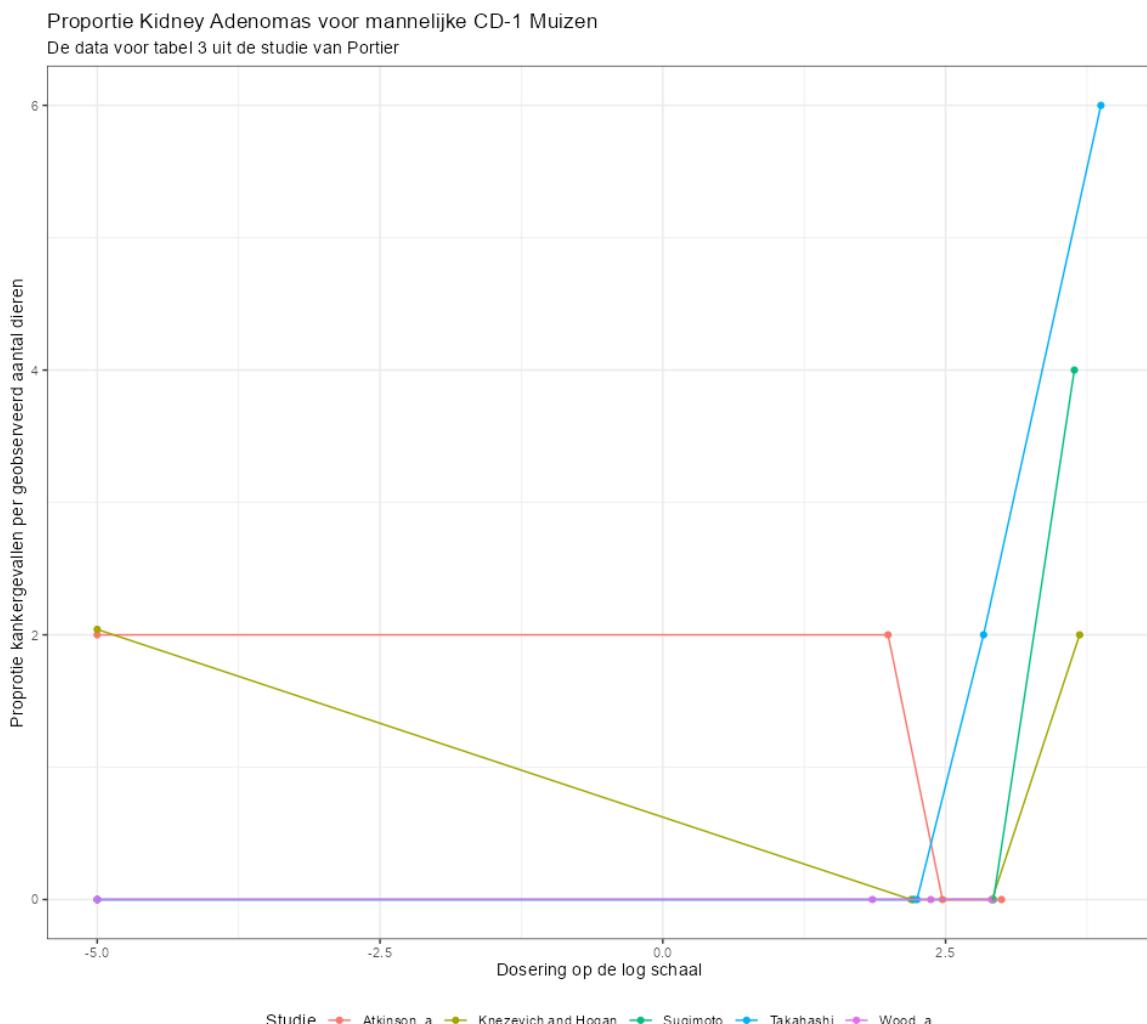
**Figuur 116.** Tabel 3 uit de Portier studie waarin voor CD-1 muizen studies worden gecombineerd per tumorsoort en geslacht.

Hier zijn inderdaad vijf studies geïncludeerd. Het is nu de bedoeling dat ik een resultaat vind, in de dose-response, die lijkt op het resultaat van **Figuur 116**. Ik zal wederom gebruik maken van een GLMM.

Ik krijg direct een foutmelding dat er te weinig data is om een stabiele GLMM uit te voeren. Toch heeft de computer er geen problemen mee om mij het onvoltooide werk te tonen wat zichtbaar is in **Figuur 117**. We zien direct dat er geen *Random Effects* component is voor ‘Studie’. Met deze gegevens is het handig om de data grafisch weer te geven (**Figuur 118**). Vaak zien we dan wat er scheelt en wat nu opvalt is dat er voor Wood geen data is.

cbind(Cases,N-Cases)			
Predictors	Odds Ratios	CI	p
(Intercept)	0.01	0.00 – 0.02	<0.001
Dosering log	1.11	0.88 – 1.39	0.393
<b>Random Effects</b>			
$\sigma^2$	3.29		
$\tau_{\text{Studie}}$	0.00		
N Studie	5		
Observations	20		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.036 / NA		

**Figuur 117.** Resultaten van een incorrect GLMM model.



**Figuur 118.** De studies zoals in de data weergegeven voor mannelijke CD-1 muizen en *Kidney Adenomas*. Dit is de data die past bij de allereerste rij uit **Figuur 116**.

Wat ook opvalt is dat het einde van de log schaal wel degelijk een stijging laat zien. Het interpreteren van deze plot wordt bemoeilijkt door het ontbreken van de variantie. Daardoor lijken de proporties significanter dan ze wellicht zijn. Zo liet het Logistisch model uit **Figuur 117** geen verschil zien tussen de proportie kanker en de dosering, maar dat model is niet betrouwbaar. De vraag is nu hoe we dit model wel betrouwbaar kunnen maken.

Laten we om te beginnen de data van Wood weghalen. Als we dit doen behouden we echter hetzelfde probleem. Dit is dus niet de oplossing. Maar wellicht werkt het beter als we data lenen uit een andere studie? Dat zou betekenen dat we voor de Wood studie data moeten invullen. Eigenlijk is dat best gek want de afwezigheid van *Kidney Adenomas* in de gehele Wood studie is ook informatie! Dit hebben we al besproken. Toch zouden we kunnen zien of het helpt als ik incidentie ga toevoegen. Als ik dit namelijk doe voor elke dosering blijft de proportie over tijd voor die studie gelijk. Daarmee voeg ik wel data toe om mee te rekenen, maar installeer ik geen relatie die er niet is. Als ik wat getallen uitprobeer dan blijkt het getal twee te werken: dus twee kankergevallen per dosering. Het resultaat zien we in **Figuur 119**. We hebben nu wel degelijk variantie tussen de studies<sup>104</sup>. Maar nog steeds geen significant resultaat.

cbind(Cases,N-Cases)			
Predictors	Odds Ratios	CI	p
(Intercept)	0.02	0.01 – 0.03	<0.001
Dosering log	1.04	0.90 – 1.21	0.593
<b>Random Effects</b>			
$\sigma^2$	3.29		
$\tau_{00}$ Studie	0.13		
ICC	0.04		
N Studie	5		
Observations	20		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.006 / 0.045		

**Figuur 119.** Resultaten van een model waarin ik data heb toegevoegd.  
Het gaat hier om mannelijke CD-1 muizen en de mogelijke relatie tussen dosering en *Kidney Adenomas*.

<sup>104</sup> Omdat het toevoegen van twee cases per dosering afwijkt van de één à drie cases die we vinden bij de andere studies. Hiermee lossen we het probleem van lege cellen op én het probleem van te weinig variantie tussen de studies. Uiteraard is wat we hier niet doen niet geheel correct: de proportie kanker per dosering voor *Kidney Adenomas* in Wood\_a is gewoonweg nul!

Wie trouwens **Figuur 119** wil vergelijken met **Figuur 116** moet mentaal wat berekeningen doen. Dat komt omdat de resultaten in **Figuur 119** de Odds Ratios laat zien. Een Odds Ratio is in het kort een ratio van twee ratio's wat betekent dat het getal 1 een indicatie is van geen verschil.

Onderstaande tekst in het wit, met zwarte achtergrond, is de uitkomst zoals ik deze zie wanneer ik het statistiekprogramma opdracht geef om een model te maken. De dosering is op de logschaal en de richtingscoëfficiënt is 0.04. Nou weet ik niet of Portier ook de logschaal heeft gebruikt, maar het lijkt er op dat dit niet het geval is. Veel uitmaken mag het niet want de 0.04 is niet statistisch afwijkend van nul<sup>105</sup>. Dat maakt dat de p=0.006 die Portier vindt heel anders is dan de p=0.593 die ik heb gevonden<sup>106</sup>. Ik kan wederom niet duiden waarom dit is.

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
```

```
Family: binomial ( logit )
```

```
Formula: cbind(Cases, N - Cases) ~ Dosering_log + (1 | Studie)
```

```
Data: .
```

AIC	BIC	logLik	deviance	df.resid
55.4	58.3	-24.7	49.4	17

```
Scaled residuals:
```

Min	1Q	Median	3Q	Max
-0.9755	-0.8899	0.1193	0.5722	2.0276

```
Random effects:
```

Groups	Name	Variance	Std.Dev.
--------	------	----------	----------

Studie	(Intercept)	0.1335	0.3654
--------	-------------	--------	--------

Number of obs:	20	groups:	Studie, 5
----------------	----	---------	-----------

```
Fixed effects:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.10807	0.32422	-12.671	<0.0000000000000002 ***
Dosering_log	0.04059	0.07598	0.534	0.593

```
---
```

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

```
Correlation of Fixed Effects:
```

(Intr)	
--------	--

Dosering_lg	-0.301
-------------	--------

<sup>105</sup> De variatie rondom die 0.04 is 0.075 wat ongeveer twee keer zo groot is dus het mag niet verbazen

<sup>106</sup> Ik heb in deze analyse niet meer de correctie voor *overdispersion* gedaan omdat het model wederom niet tot een oplossing zou komen. Eigenlijk is het doen van een dergelijke analyse te breekbaar om zinvol uit te voeren.

Het is nu zeer aannemelijk dat ik voor geen enkele van de tumorsoorten waarden ga zien die vergelijkbaar zijn met de waarden van Portier. Wat ik wil kan doen is een tabel maken, op basis van de data zoals aangevoerd door Portier waarin ik zijn resultaten vergelijk met die van mij (**Tabel 31**). Ik kies gemakshalve alleen de resultaten die door Portier als significant zijn bestempeld. Dit doe ik omdat dit relevante gerapporteerde bevindingen zijn en ik wil graag zien of ik dit kan repliceren<sup>107</sup>. Ik zal steeds weer de studie van Wood voorzien van 2 Cases per dosering. Dit zou geen invloed mogen hebben op de gemiddelde richtingscoëfficiënt in de relatie tussen dosering en aantal kankergevallen (omdat in beide gevallen de relatie onderling nul is).

Geslacht	Tumor	Common Trend (Portier)	GLMM
Man	Kidney Adenomas	0.006	0.593
Man	Kidney Carcinomas	0.031	0.664
Man	Kidney Adenomas and Carcinomas	<0.001	0.417
Man	Hemangiosarcomas	0.033	0.228
Vrouw	Hemangiomas	0.031	0.151
Vrouw	Malignant Lymphomas	0.012	0.648

**Tabel 31.** De p-waarden voor mannelijke en vrouwelijke CD-1 muizen zoals gerapporteerd in Portier en zoals gevonden door het door mij gehanteerde GLMM model.

Zoals de te zien valt kan ik geen enkel van de significante resultaten repliceren. Het is verstandig om de andere tabellen, zoals aangedragen door Portier ook mee te nemen. Ik zal beginnen bij de resultaten zoals beschreven in **Figuur 120**. De resultaten zet ik in **Tabel 32** neer.

---

<sup>107</sup> Het kan heel goed zijn dat wat in de Portier studie als niet significant wordt bevonden door mij als wel significant wordt bevonden. Maar dat is niet het doel van dit rapport. Het doel is om te laten zien hoe keuzes van invloed zijn op het resultaat. In de statistiek bestaat er niet zo iets als een sluitende bevinding.

**Table 4** P-values for the Cochran-Armitage trend test and pooled logistic regression analysis for tumors with at least one significant trend test or Fisher's exact test ( $p \leq 0.05$ ) in male and female Sprague-Dawley rats

Tumor	Individual study p-values for trend <sup>a</sup>				Common Trend	Heterogeneity Test
Males	G	H	I	J		
Testicular Interstitial Cell Tumors	<b>0.009</b>	0.296	0.580	0.594	0.461	0.105
Pancreas Islet Cell Adenomas	0.512	0.147 ( <b>0.007</b> ) <sup>c</sup>	0.974	0.859	0.849	0.143
Pancreas Islet Cell Carcinomas	0.251	1.000	–	0.500	0.731	0.166
Pancreas Islet Cell Adenomas or Carcinomas	0.316	0.206	0.974	0.844	0.875	0.185
Thyroid C-cell Adenomas	0.743	0.089	0.278	0.631	0.210	0.532
Thyroid C-cell Carcinomas	0.505	0.442	0.495	0.565	0.322	0.898
Thyroid C-cell Adenomas and Carcinomas	0.748	0.097	0.197	0.642	0.175	0.526
Thyroid Follicular-cell Adenomas	0.122	0.408	0.067	0.966	0.464	0.055
Thyroid Follicular-cell Carcinomas	— <sup>b</sup>	0.255	0.443	1.000	0.448	0.137
Thyroid Follicular-cell Adenoma and Carcinoma	0.122	0.232	0.099	0.986	0.446	0.031
Hepatocellular Adenomas	0.471	<b>0.015</b>	0.325	0.500	<b>0.029</b>	0.664
Hepatocellular Carcinomas	0.062	0.637	0.760	0.642	0.803	0.269
Hepatocellular Adenomas and Carcinomas	0.173	<b>0.050</b>	0.480	0.690	0.144	0.428
Kidney Adenomas	0.938	0.813	1.000	<b>0.004</b>	<b>0.039</b>	0.002
Skin Keratoacanthomas	— <sup>b</sup>	<b>0.042</b>	<b>0.047</b>	<b>0.029</b>	< 0.001	0.998
Skin Basal Cell Tumors	0.251	0.249	1.000	<b>0.004</b>	< 0.001	0.009
Females	G	H	I	J		
Thyroid C-cell Adenomas	0.679	<b>0.049</b>	0.207	0.912	0.287	0.150
Thyroid C-cell Carcinomas	<b>0.003 (&lt; 0.001)</b> <sup>c</sup>	0.500	— <sup>b</sup>	— <sup>b</sup>	0.385	0.041
Thyroid C-cell Adenomas and Carcinomas	0.072 ( <b>0.037</b> ) <sup>c</sup>	0.052	0.207	0.912	0.275	0.071
Adrenal Cortical Adenoma	0.851	0.603	— <sup>b</sup>	0.626	0.713	0.750
Adrenal Cortical Carcinoma	0.386	<b>0.015</b>	0.493	— <sup>b</sup>	<b>0.031</b>	0.199
Adrenal Cortical Adenoma and Carcinoma	0.801	0.090	0.493	0.626	0.195	0.520

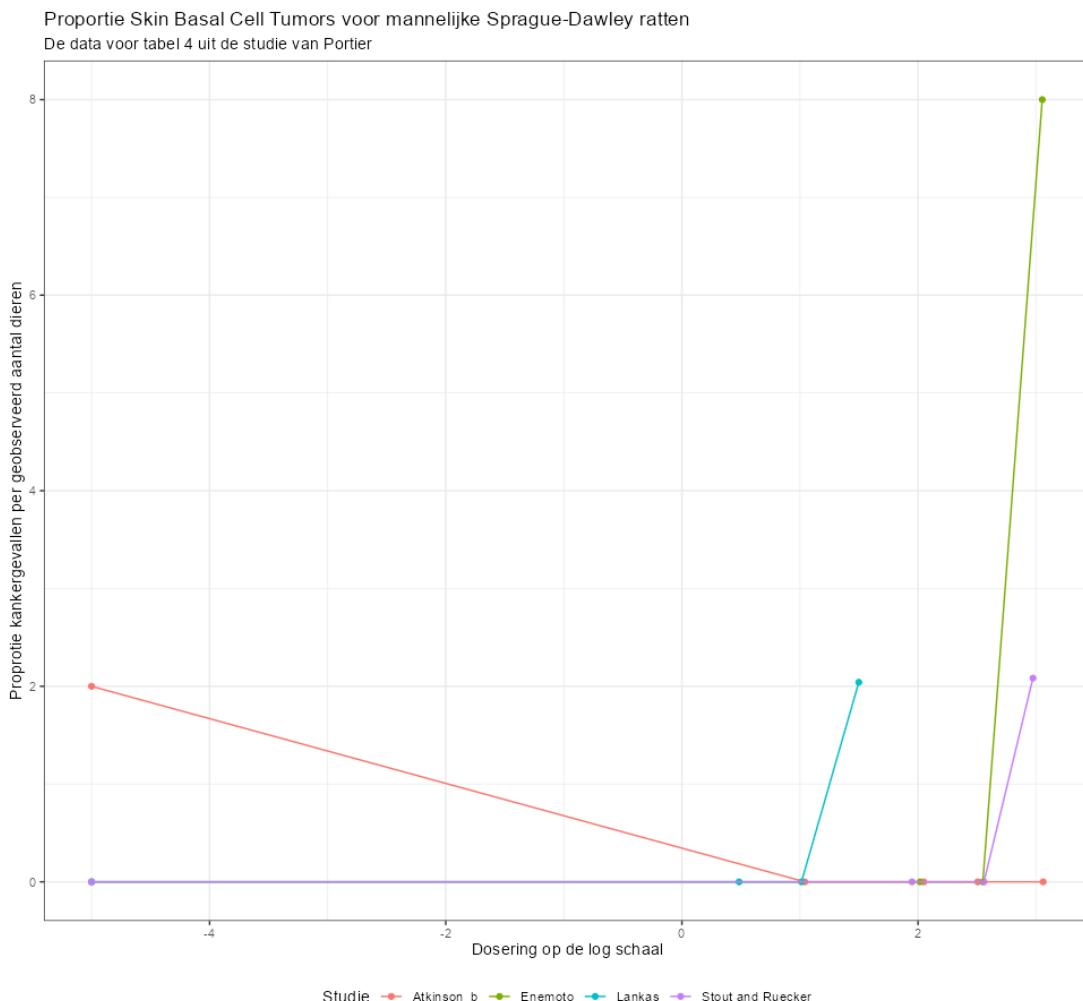
<sup>a</sup> – Study G is Lankas [17] (Additional file 2: Table S7), Study H is Stout and Ruecker [18] (Additional file 2: Table S8), Study I is Atkinson et al. [12] (Additional file 2: Table S9) and Study J is Enemoto [20] (Additional file 2: Table S10); <sup>b</sup> – three dashes “—” indicates all tumor counts are zero; <sup>c</sup> – using historical control data (see text for details) and Tarone’s test

**Figuur 120.** Tabel 4 uit de Portier studie.

Geslacht	Tumor	Common Trend (Portier)	GLMM
Man	Hepatocellular Adenomas	0.029	0.480
Man	Kidney Adenomas	0.039	0.743
Man	Skin Keratoacanthomas	<0.0001	<0.0001
Man	Skin Basal Cell Tumors	<0.0001	0.263
Vrouw	Adrenal Cortical Carcinoma	0.031	0.170

**Tabel 32.** De p-waarden voor mannelijke en vrouwelijke Sprague-Dawley ratten zoals gerapporteerd in Portier en zoals gevonden door het door mij gehanteerde GLMM model.

Wederom kan ik het gros van Portier niet repliceren. Maar het statistisch significante resultaat voor *Skin Keratoacanthomas* zie ik ook, dus is het zinvol om de ruwe data te tonen. Op deze manier krijgen we een idee hoe een statistisch significant effect er daadwerkelijk uitziet. **Figuur 121** laat zien dat op het einde van de dosering de proportie omhoog schiet.



**Figuur 121.** Relatie tussen dosering en de proportie Skin Basal Cell Tumors voor mannelijke Sprague-Dawley ratten.

Laten we de exercitie afmaken en aan de slag gaan met Tabel 5 uit de Portier studie (**Figuur 122**). De resultaten zien we in **Tabel 33**. Wederom kan ik geen enkele van de resultaten repliceren.

**Table 5** P-values for the Cochran-Armitage trend test and pooled logistic regression analysis for tumors with at least one significant trend test or Fisher's exact test ( $p \leq 0.05$ ) in male and female Wistar rats

Tumor	Individual study p-values for trend <sup>a</sup>			Common Trend	Homogeneity Test
	K	L	M		
Males					
Hepatocellular Adenomas	0.391	<b>0.008</b>	0.418	<b>0.048</b>	0.156
Hepatocellular Carcinomas	0.418	---	1.000	0.492	0.242
Hepatocellular Adenomas and Carcinomas	0.286	<b>0.008</b>	0.610	<b>0.029</b>	0.194
Pituitary Adenomas	0.376	0.277	<b>0.045</b>	0.057	0.664
Pituitary Carcinomas	0.692	---	1.000	0.771	0.956
Pituitary Adenomas and Carcinomas	0.454	0.277	0.059	0.073	0.700
Skin Keratoacanthomas	---	0.387	<b>0.030</b>	<b>0.032</b>	0.823
Adrenal Pheochromocytomas	<b>0.048</b>	0.721	0.306	0.273	0.210
Females					
Mammary Gland Adenomas	0.539	0.941	0.062	0.448	0.015
Mammary Gland Adenocarcinomas	1.000	0.271	<b>0.042</b>	0.071	0.008
Mammary Gland Adenomas and Adenocarcinomas	0.729	0.590	<b>0.007</b>	0.113	0.064
Pituitary Adenomas	0.967	0.261	<b>0.014</b>	0.105	0.023
Pituitary Carcinomas	1.000	–	0.750	0.748	0.491
Pituitary Adenomas and Carcinomas	0.976	0.261	<b>0.017</b>	0.129	0.019

<sup>a</sup> – Study J is Suresh [21] (Additional file 2: Table S11), Study K is Brammer [22] (Additional file 2: Table S12), and Study L is Wood et al. [14] (Additional file 2: Table S13); <sup>b</sup> – three dashes “—” indicates all tumor counts are zero

**Figuur 122.** Tabel 5 uit de Portier studie.

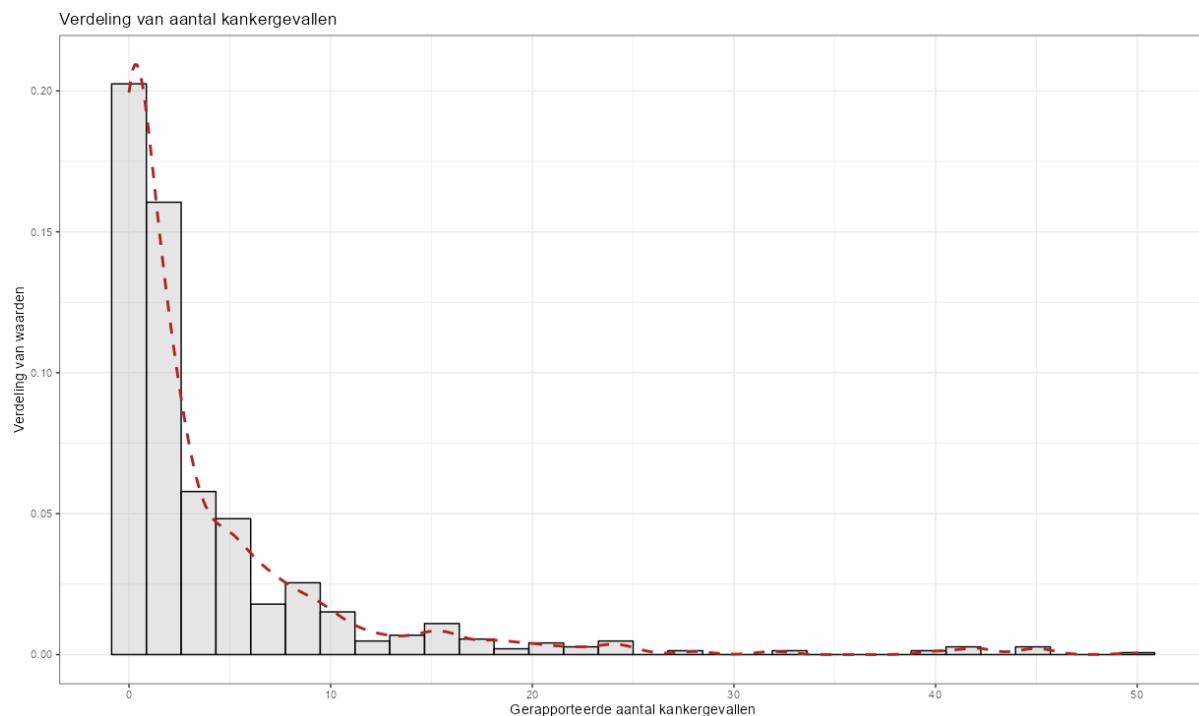
Geslacht	Tumor	Common Trend (Portier)	GLMM
Man	Hepatocellular Adenomas	0.048	0.066
Man	Hepatocellular Adenomas and Carcinomas	0.029	0.735
Man	Skin Keratoacanthomas	0.032	0.580

**Tabel 33.** De gevonden p-waarden door Portier en door het GLMM model wat ik heb gebruikt. Het gaat hier om mannelijke Wistar ratten.

Met deze bevindingen is het niet meer zinvol om hier al te veel tijd aan te besteden. Zouden we de positieve resultaten van Portier in een correctie voor meerdere testen zetten dan zou het gros ook niet meer statistisch significant zijn. Om toch nog een complete analyse te doen wil ik uitwijken naar een zogenaamd Poisson model.

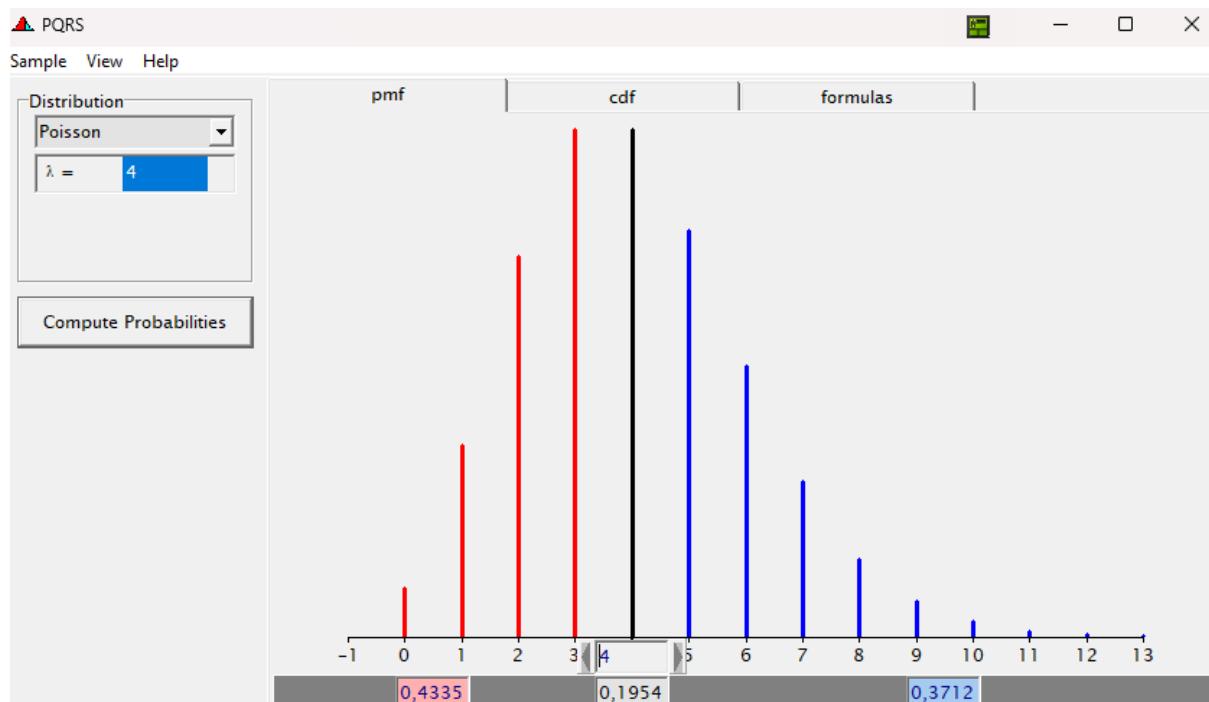
## Poisson model

Een Poisson model is een ander veelgebruikt model voor discrete data en wordt vaak gebruikt bij data die voortkomen uit tellingen. Een kenmerk van een Poisson model is de lange staart die we ook zien als we het aantal gerapporteerde kankergevallen als een frequentieverdeling uiteenzetten (**Figuur 123.**). Wat opvalt is dat het gros van de data een proportie van nul laat zien.



**Figuur 123.** Verdeling van het gerapporteerde aantal kankergevallen. Deze verdeling lijkt sterk op de traditionele verdeling die we zien bij Poisson verdelingen.

Voordat we gaan rekenen is het verstandig om een kleine introductie aan te bieden. De Poisson verdeling is namelijk net als de binomiale verdeling minder flexibel dan de normaalverdeling. In de Poisson verdeling is de variantie gelijk aan het gemiddelde, wat heel vaak zal leiden tot *overdispersion*. Ik kan dit heel makkelijk laten zien door het gemiddelde aantal gerapporteerde kankergevallen uit te rekenen en dan de frequentieverdeling te tonen door middel van een theoretisch Poisson model. Het gemiddelde is 4 en het resultaat zien we in **Figuur 124**. Dit lijkt niet op de data uit **Figuur 123**. We weten dus eigenlijk nu al dat een Poisson model niet zal voldoen, maar wellicht helpt het als we een model maken waarin we verschillende verklarende factoren opnemen (geslacht, dosering, soort) én rekening houden met *overdispersion* (zoals we dat gedaan hebben in een GLMM met binomiale verdeling).



**Figuur 124.** De frequentieverdeling bij een gemiddelde van vier vanuit een Poisson verdeling. In een Poisson verdeling staat de variantie gelijk aan het gemiddelde, dus vier.

Om deze analyse te kunnen doen moet ik ook hier weer data groeperen en dus, net als bij een LMM, moeten werken met het opgetelde aantal kankergevallen<sup>108</sup>. Het resultaat van het model zien we in **Figuur 125**.

Voordat we dieper ingaan op de uitkomsten van dit model wil ik nog vermelden dat ik verschillende correctiemethoden heb toegepast. Niet alleen heb ik wederom een kolom aangemaakt waarbij elke rij een ander ID krijgt, maar ik heb ook een zogenaamde *offset* toegepast. De offset is letterlijk het anker waartegen het aantal kankergevallen wordt afgezet net als de deler in een binomiale verdeling. Omdat per studie en per dosering het aantal geobserveerde gevallen kan verschillen, bepaal ik door middel van de *offset* dat hier rekening mee wordt gehouden. In dit model houd ik dus rekening met het over-of onderschatten van variantie én het feit dat de groepen per dosering per geslacht per studie niet even groot is.

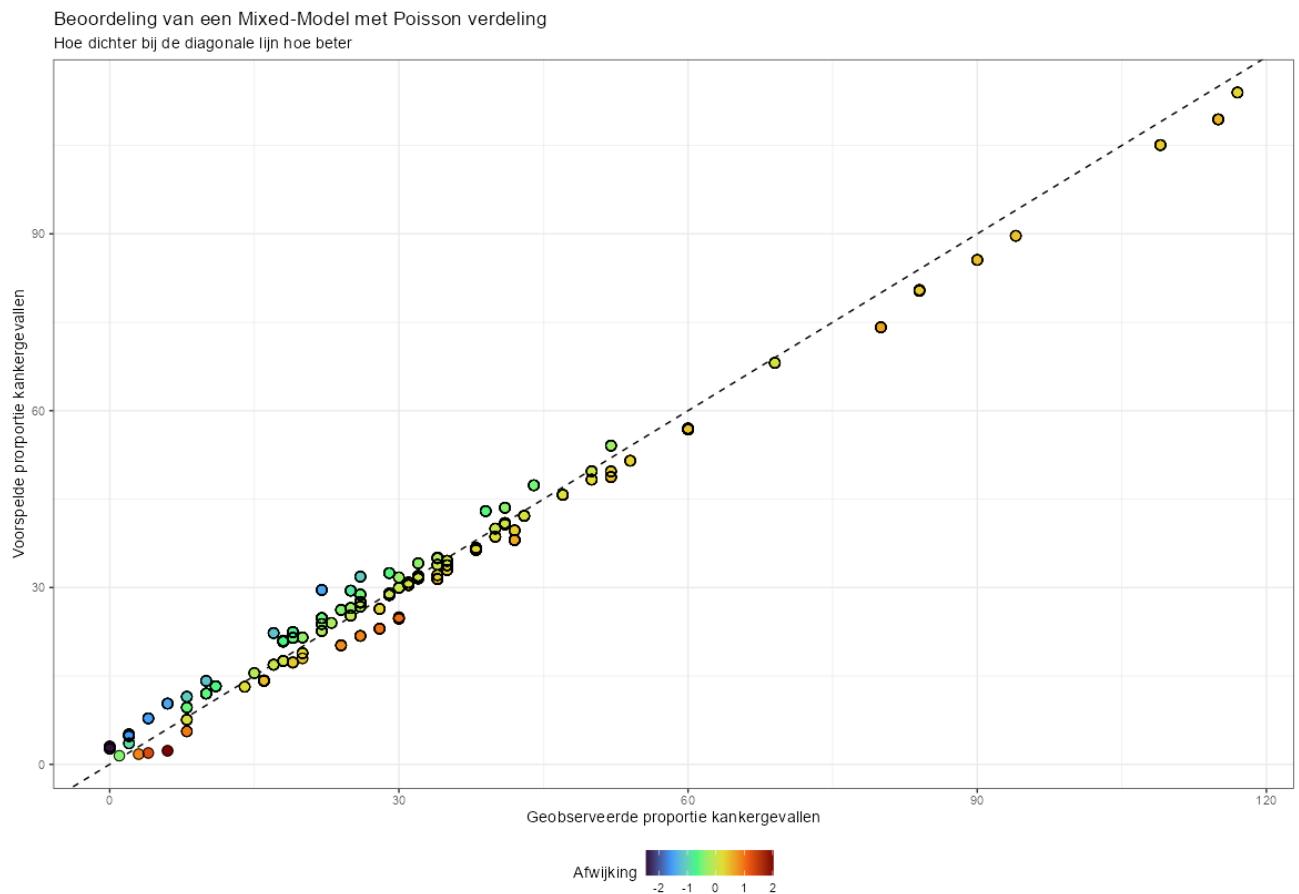
<sup>108</sup> Anders werk ik met een verdeling van gerapporteerde kankergevallen per kancersoort waarvan we weten dat deze eigenlijk niet compleet is. Ik werk daarom liever met de totale som. Op deze manier heb ik niks te maken met de rijen waarin een kancersoort wel wordt genoemd, maar er geen kanker is gezien. Het enige wat ik tot mij neem is de gerapporteerde kankergevallen per studie, per dosering, per geslacht.

Als we kijken naar de resultaten zien we voor het eerst een sterk significante relatie met de dosering. Dit vraagt dat we heel goed kijken naar de validiteit van het model, door mij nu afgebeeld in **Figuur 126** en **Figuur 127**<sup>109</sup>. Niet omdat ik iets zie wat ik niet wil zien, maar omdat ik iets zie wat ik steeds niet kon vinden.

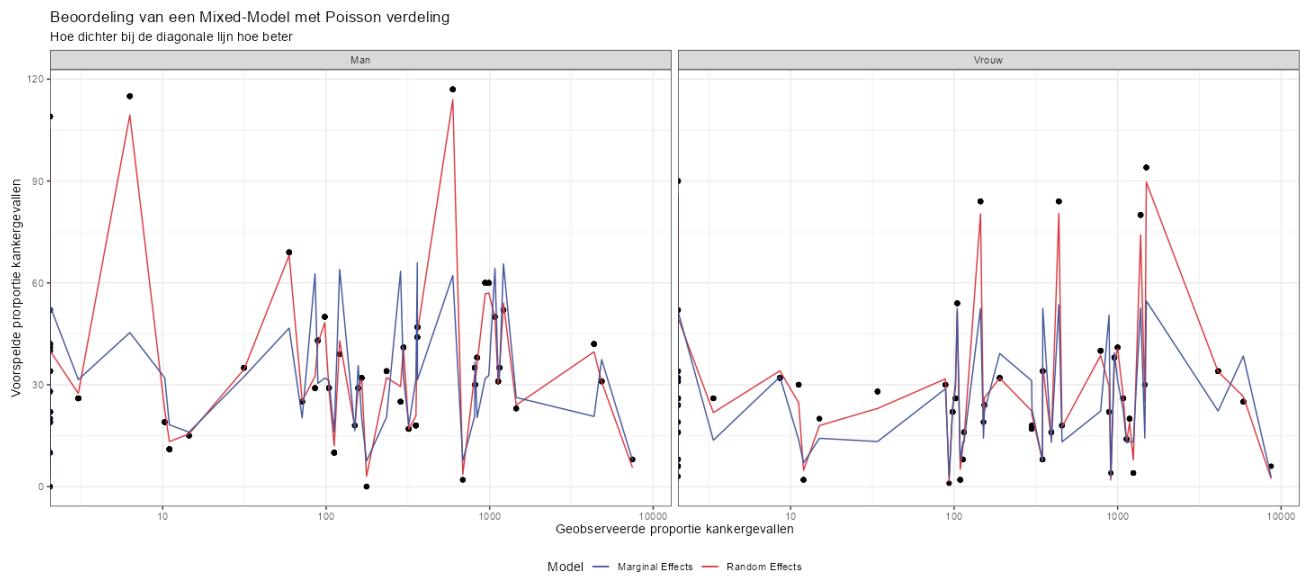
<i>Predictors</i>	sum Cases		
	<i>Incidence Rate Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.08	0.05 – 0.11	<0.001
Dosering log	1.02	1.01 – 1.03	<0.001
Geslacht [Vrouw]	1.16	1.10 – 1.23	<0.001
Duur [18]	0.62	0.37 – 1.03	0.067
Duur [26]	1.05	0.56 – 1.97	0.874
Soort [Sprague-Dawley]	0.49	0.30 – 0.81	0.005
Soort [Swiss Albino]	2.57	1.34 – 4.92	0.005
Soort [Wistar]	1.92	1.17 – 3.16	0.010
Dosering log * Geslacht [Vrouw]	0.98	0.96 – 1.00	0.014
<b>Random Effects</b>			
$\sigma^2$	2.83		
$\tau_{00\text{ ID}}$	0.11		
$\tau_{00\text{ Studie}}$	0.08		
ICC	0.03		
N <sub>Studie</sub>	13		
N <sub>ID</sub>	842		
Observations	842		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.094 / 0.117		

**Figuur 125.** Bevindingen van een Poisson model met als y-variabele het opgeteld aantal kankergevallen.

<sup>109</sup> Ik zeg dit natuurlijk niet omdat het resultaat mij niet bevalt, maar omdat ik voor het eerst een relatie heb gevonden die ik in geen ander model heb gevonden. Ook is het grafisch zelden zichtbaar geweest.

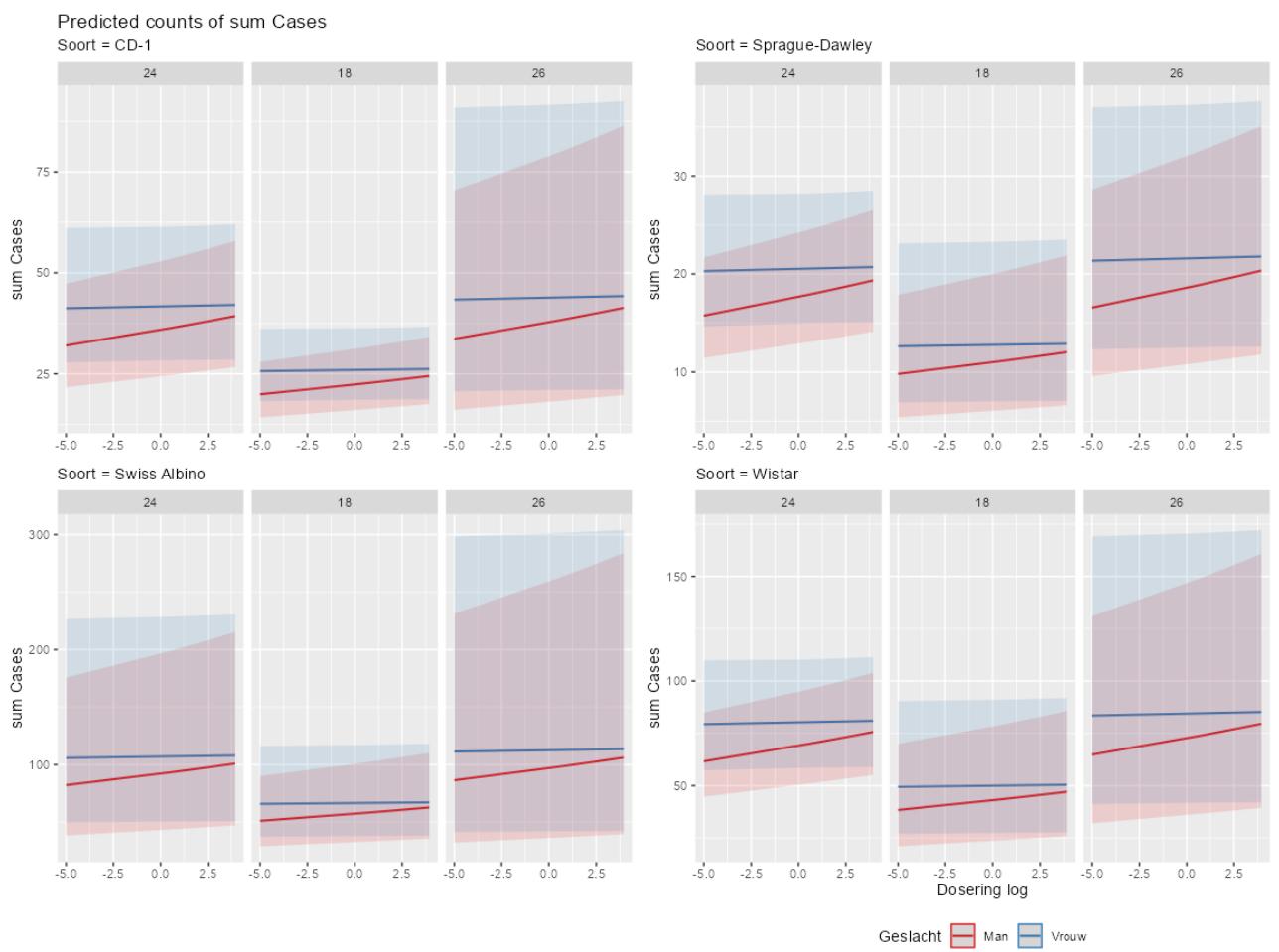


**Figuur 126.** Het voorspelde aantal kankergevallen afgezet tegen het geobserveerde aantal kankermodellen op basis van een Poisson model.



**Figuur 127.** Beoordeling van het Poisson model. De blauwe lijn toont het gemiddelde, of marginale, model. De rode lijn toont de daadwerkelijke GLMM met alle conditionele effecten.

Kijken we naar de voorspellingen van het model per dosering, soort en duur dan zien we dat er voor vrouwen helemaal geen relatie met dosering meer is (**Figuur 128**). Die van mannen stijgt, maar de onzekerheidsbanden zijn zo groot dat ik er makkelijk een rechte lijn door kan trekken. Hoewel de algemene trend in het model statistisch significant is, kan ik uit de voorspellingen maar moeilijk achterhalen waar die significantie op is gebaseerd. Het lijkt erop dat een relatie eerder bij mannen dan bij vrouwen te zien is.



**Figuur 128.** Voorspelde waarden uit het Poisson model met de interactie tussen geslacht en dosering.

Halen we er de ruwe resultaten bij vanuit het model dan zien we het volgende.

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)

[glmerMod]

Family: poisson ( log )

Formula:

sum\_Cases ~ Dosering\_log \* Geslacht + Duur + Soort + offset(log(sum\_N)) +  
(1 | Studie) + (1 | ID)

Data: df\_poiss

AIC	BIC	logLik	deviance	df.resid
6476.9	6529.0	-3227.5	6454.9	831

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.73493	-0.47245	0.03878	0.28290	2.44230

Random effects:

Groups	Name	Variance	Std.Dev.
--------	------	----------	----------

ID	(Intercept)	0.10754	0.3279
----	-------------	---------	--------

Studie	(Intercept)	0.07501	0.2739
--------	-------------	---------	--------

Number of obs: 842, groups: ID, 842; Studie, 13

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.572425	0.196838	-13.069	< 0.0000000000000002 ***
Dosering_log	0.022952	0.005176	4.434	0.000009243 ***
GeslachtVrouw	0.149574	0.028778	5.197	0.000000202 ***
Duur18	-0.473817	0.259076	-1.829	0.06742 .
Duur26	0.050595	0.319797	0.158	0.87429
SoortSprague-Dawley	-0.708846	0.253299	-2.798	0.00513 **
SoortSwiss Albino	0.943117	0.332213	2.839	0.00453 **
SoortWistar	0.654822	0.253596	2.582	0.00982 **
Dosering_log:GeslachtVrouw	-0.020713	0.008440	-2.454	0.01413 *

---

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

Correlation of Fixed Effects:

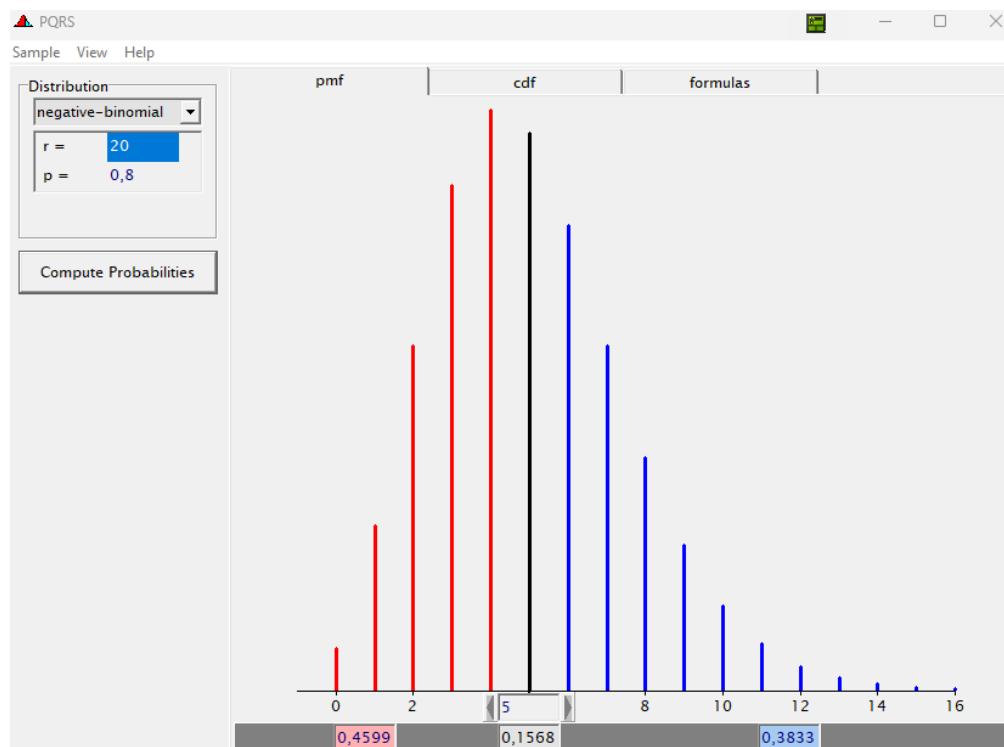
	(Intr)	Dsrng_	GslchV	Duur18	Duur26	SrtS-D	SrtSwA	SrtWst
Dosering_lg	-0.019							
GeslachtVrw	-0.069	0.102						
Duur18	-0.757	-0.005	0.009					
Duur26	0.000	0.013	-0.004	0.000				
SrtSprg-Dwl	-0.775	0.001	0.029	0.588	-0.316			
SrtSwssAlbn	0.000	0.006	0.004	-0.332	0.000	0.000		
SoortWistar	-0.773	0.004	0.006	0.587	0.000	0.600	0.000	
Dsrng_lg:GV	0.009	-0.608	-0.171	0.003	-0.003	0.001	-0.001	0.000

Wat opvalt is dat de dosering wel degelijk significant is, met een groei van 0.02 op de log schaal. Omdat de onzekerheid rondom die 0.02 zo klein is (0.005) is het statistisch significant verschillend van nul, maar een groot effect kunnen we het niet noemen. Dit zien we terug in **Figuur 128.**

## Negative binomial

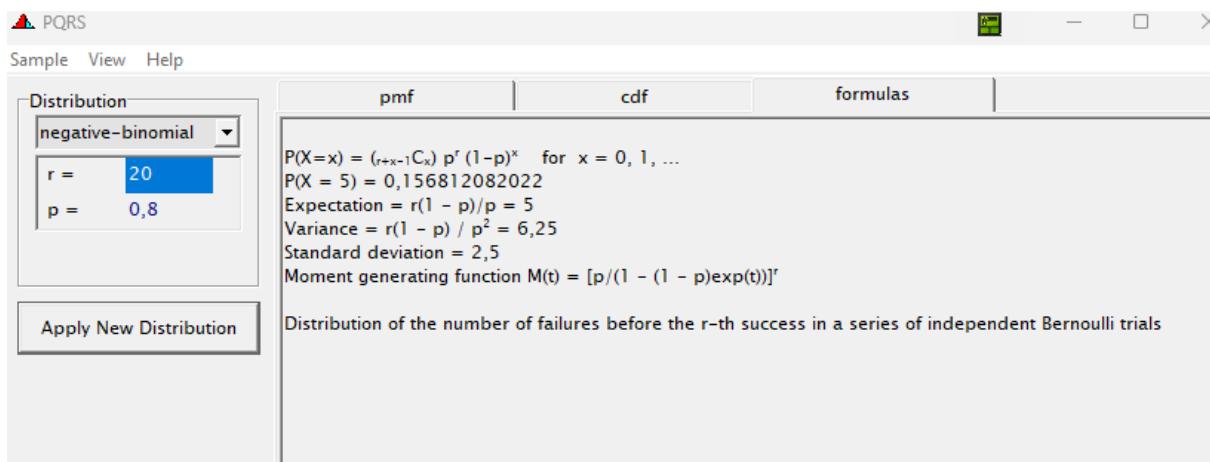
Een ander model wat vaak gebruikt wordt om tellingen te modelleren is het [negative binomial model](#), oftewel de negatieve binomiaal. Deze verdeling modelleert niet het aantal successen, zoals in de binomiale verdeling, maar juist de tijd waarin er een 'mislukking' optreedt. We kunnen bijvoorbeeld het gooien van een 6 op een dobbelsteen definiëren als een succes, en het gooien van een ander getal als een mislukking, en vragen hoeveel mislukte worpen er zullen zijn voordat we het derde succes zien ( $r = 3$ ). In zo'n geval zal de kansverdeling van het aantal mislukkingen dat optreedt een negatieve binomiale verdeling zijn.

Een alternatieve formulering is om het aantal totale proeven te modelleren (in plaats van het aantal mislukkingen). In feite is voor een bepaald (niet-willekeurig) aantal successen ( $r$ ) het aantal mislukkingen ( $n - r$ ) willekeurig omdat het aantal totale proeven ( $n$ ) willekeurig is. We kunnen bijvoorbeeld de negatieve binomiale verdeling gebruiken om het aantal dagen  $n$  (willekeurig) te modelleren dat een bepaalde machine werkt (gespecificeerd door  $r$ ) voordat hij kapot gaat (**Figuur 129**). Zie hier de relatie met het modelleren van het aantal kankergevallen.



**Figuur 129.** Voorbeeld van een verdeling op basis van de negatieve binomiale verdeling.

Hoewel de verdeling een andere interpretatie kent, is dit voor het modelleren niet per se van belang<sup>110</sup>. Belangrijker is het benoemen van datgene wat de negatieve binomiaal zo aantrekkelijk maakt in vergelijking met de Poisson verdeling: namelijk het hebben van twee onafhankelijke parameters. Dit betekent, theoretisch, dat we flexibeler zijn in de parameterisering van het gemiddelde én de variantie (**Figuur 130**).



**Figuur 130.** Formules die bij de negatieve binomiale verdeling horen plus de uitkomsten bij  $r=20$  en  $p=0.8$ .

Nu we weten dat de negatieve binomiale verdeling goed uit de voeten kan met discrete data, zoals tellingen, is het alleen maar logisch dat ik het toepas op de dezelfde data als waar ik de Poisson verdeling voor heb gebruikt. De resultaten van dat exacte model (wat zichtbaar is in de *Formula* regel van de output hieronder) toont een significante relatie voor de dosering op de log schaal. De grootte van het effect is ongeveer gelijk aan die in het Poisson model.

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod]
  Family: Negative Binomial(9.5238) ( log )
  Formula: sum_Cases ~ Dosering_log * Geslacht + Duur + Soort + offset(log(sum_N)) +
    (1 | Studie)
  Data: df_nb

  AIC   BIC  logLik deviance df.resid
  6460.6 6512.7 -3219.3  6438.6    831

  Scaled residuals:
    Min   1Q Median   3Q  Max
  -2.0597 -0.8303  0.0531  0.6732  3.1463

  Random effects:
  Groups Name        Variance Std.Dev.
  Studie (Intercept) 0.07466  0.2732
```

<sup>110</sup> De y-variabele bepaalt de interpretatie, net als de link functie van het model. Die is nog steeds op de log schaal.

Number of obs: 842, groups: Studie, 13

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.549403	0.196236	-12.992	< 0.0000000000000002 ***
Dosering_log	0.020845	0.005061	4.119	0.00003802162 ***
GeslachtVrouw	0.173461	0.028717	6.040	0.00000000154 ***
Duur18	-0.464255	0.258497	-1.796	0.07250 .
Duur26	0.068920	0.319049	0.216	0.82897
SoortSprague-Dawley	-0.691656	0.252598	-2.738	0.00618 **
SoortSwiss Albino	0.954321	0.331415	2.880	0.00398 **
SoortWistar	0.704820	0.252909	2.787	0.00532 **
Dosering_log:GeslachtVrouw	-0.017715	0.008305	-2.133	0.03291 *

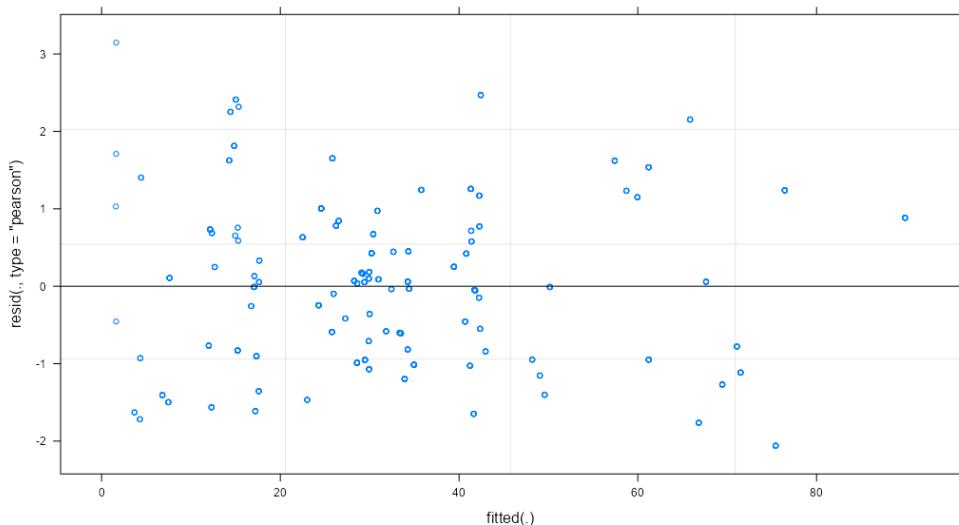
---

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

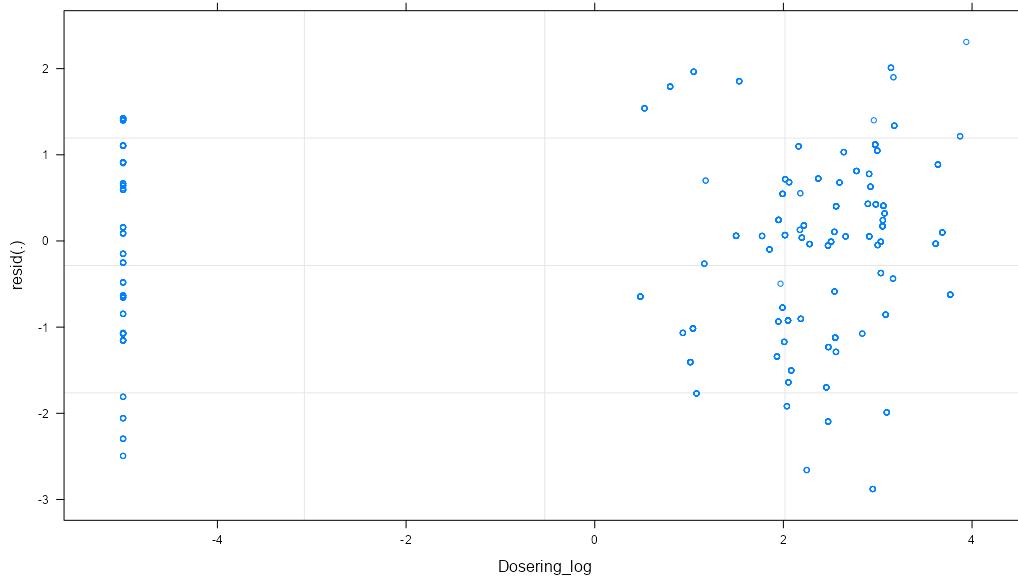
Correlation of Fixed Effects:

	(Intr)	Dsrng_GslchV	Duur18	Duur26	SrtS-D	SrtSwA	SrtWst
Dosering_lg	-0.017						
GeslachtVrw	-0.063	0.099					
Duur18	-0.756	-0.007	0.006				
Duur26	0.001	0.009	-0.014	0.000			
SrtSprg-Dwl	-0.775	0.002	0.026	0.588	-0.316		
SrtSwssAlbn	0.000	0.004	-0.002	-0.332	0.000	0.000	
SoortWistar	-0.773	0.006	-0.002	0.587	0.000	0.600	0.000
Dsrng_lg:GV	0.009	-0.602	-0.172	0.004	0.001	-0.001	-0.003

Nu we twee keer een model vinden wat, in tegenstelling tot andere modellen, wel een significante relatie vindt tussen dosering en het aantal kankergevallen is het zinvol om de modellen beter te bekijken. Hieronder staan twee figuren die een relatie teken tussen de restwaarden van het model en het aantal kankergevallen (**Figuur 131**) én de restwaarden van het model en de dosering (**Figuur 132**).



**Figuur 131.** Relatie tussen restwaarden van het model en het aantal kankergevallen.

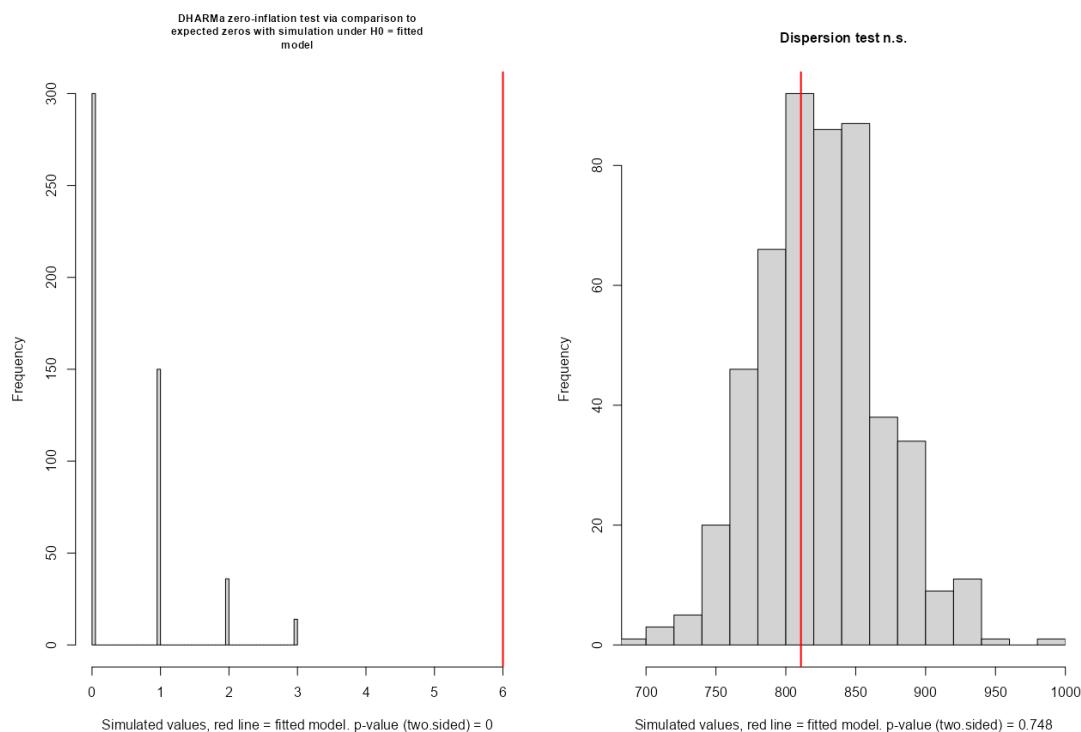


**Figuur 132.** Relatie tussen restwaarden van het model en de dosering (op de log schaal).

Dit ziet er eigenlijk prima uit en op basis van deze figuren is het niet logisch om de modellen te verwerpen. Toch is er iets wat mij niet lekker zit, namelijk de grote hoeveelheid nul-kankergevallen. Dit zagen we al in **Figuur 123** en we noemen dit fenomeen ook wel een zero-inflated verdeling waar ook weer specifieke modellen voor zijn. Samengevat betekent dit dat de modellen die ik tot nu heb gebruikt geen rekening houden met het mogelijke feit dat de data zoals we deze hebben geobserveerd tot stand is gekomen door niet één maar twee modellen: één model voor de nullen, en één model voor de rest.

We kunnen dit statistisch toetsen. Eigenlijk is dat niet meer nodig want **Figuur 123** is al behoorlijk evident, maar omdat we toch zo diep in de frequentistische statistiek zitten kunnen we het er net zo goed bijnemen. Door gebruik te maken van het eerdere negatieve binomiale model kunnen we een simulatie uitvoeren om zowel visueel (**Figuur 133**) als statistisch te toetsen op:

1. Meer nullen in de dataset dan verwacht
2. Meer variantie in de data dan verwacht.



**Figuur 133.** Simulatie op basis van negatief binomiaal model om te bezien of het model meer nullen heeft dan we voorspellen én of het model meer variantie heeft dan we meenemen.

De figuren, maar ook de gegevens hieronder, tonen overduidelijk dat het model goed overweg kan met de variantie, maar het aantal nullen stevig onderschat. Dit is een teken dat het model op een andere manier moet worden geanalyseerd. We kunnen dan later terugkomen om te zien of de resultaten van modellen die rekening houden met veel nullen gelijk zijn aan de modellen die hier geen rekening meer houden. We noemen deze modellen ook wel zero-inflated models of hurdle models. Hoewel beide modellen uit twee componenten bestaan, zijn ze **niet** gelijk aan elkaar. Het verschil tussen een zero-inflated model en een hurdle model is als volgt<sup>111</sup>:

1. Het zero-inflated model zal twee componenten modelleren - de kans op nul en de kans op X.
2. Het hurdle model zal ook twee componenten modelleren - de kans op nul en de kans op  $\neq 0$ .

<sup>111</sup> <https://medium.com/dev-genius/mixture-component-zero-inflated-and-hurdle-models-44c5e6fe5d7f>

In tegenstelling tot de zero-inflated modellen behandelen hurdle-modellen nul-kankergevallen en niet-nul-kankergevallen als twee volledig afzonderlijke categorieën, in plaats van de nul-kankergevallen te behandelen als een mengeling van structurele en steekproef waarden. De modellen ogen dus hetzelfde en tonen in bepaalde scenario's wellicht dezelfde resultaten, maar het zijn **niet** dezelfde modellen. Het is dus zinvol om naar beide modellen te kijken en we kunnen dit doen voor zowel de Poisson verdeling als de negatief binomiale verdeling.

```
DHARMA zero-inflation test via comparison to expected zeros with simulation under H0 = fitted model

data: simulationOutput
ratioObsSim = 11.364, p-value < 0.00000000000000022
alternative hypothesis: two.sided
```

```
DHARMA nonparametric dispersion test via mean deviance residual fitted vs. simulated-refitted

data: simulationOutput
dispersion = 0.98221, p-value = 0.748
alternative hypothesis: two.sided
```

## Zero-Inflated & hurdle models

Ondanks dat de modellen niet hetzelfde zijn zal ik ze wel tezamen behandelen. Dat doe ik omdat de procedure voor het toepassen van de modellen én het beoordelen ervan nagenoeg gelijk is. Uiteindelijk ben ik als volgt te werk gegaan:

1. Ik heb eerst gekozen voor een bepaalde verdeling: de Poisson of de negatieve binomiaal.
2. Ik heb vervolgens verschillende modellen gemaakt die verschillen in het aantal meegenomen parameters én de keuze om wel of niet de nullen te modelleren.

Het volgende stuk tekst is de output van drie modellen waarin ik zoek naar de relatie tussen dosering en kanker. Verder neem ik de variatie tussen de studies mee. Door de bocht genomen zijn deze drie modellen allemaal Poisson modellen, maar het eerste model is een gewoon Poisson model, het tweede model is zero-inflated én het derde model is een hurdle

model. Wat direct opvalt is dat het hurdle model de data beter beschrijft dan de andere twee modellen. Dat betekent dat het binair maken van de kansen (kans op 0 en kans op niet 0) beter past bij de data.

```
Data: df_nb
Models:
fit.p.TMB.1: sum_Cases ~ Dosering_log + (1 | Studie) + offset(log(sum_N)), zi=~0, disp=~1

fit.zip.TMB.1: sum_Cases ~ Dosering_log + (1 | Studie) + offset(log(sum_N)), zi=~1, disp=~1

fit.hnp.TMB.1: sum_Cases ~ Dosering_log + (1 | Studie) + offset(log(sum_N)), zi=~., disp=~1

      Df   AIC    BIC   logLik   deviance Chisq Chi Df Pr(>Chisq)
fit.p.TMB.1     3  8396.4 8410.6 -4195.2   8390.4
fit.zip.TMB.1   4  8398.4 8417.4 -4195.2   8390.4   0.000  1       1
fit.hnp.TMB.1   6  8373.6 8402.0 -4180.8   8361.6   28.795 2     0.0000005589 ***
---
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
```

Uiteindelijk kunnen we deze exercitie meerdere keren herhalen door telkens meer parameters toe te voegen. Wat dan blijkt is dat het hurdle model elke keer weer beter bij de data past. Wat dan rest is om al die hurdle modellen met elkaar te vergelijken om te bezien hoeveel parameters ons model moet hebben om het dichtst bij de data te staan<sup>112</sup>. De resultaten van die exercitie staan hier beneden en in **Figuur 134**.

---

<sup>112</sup> Waarbij we dus wel moeten oppassen voor overfitting én moeten oppassen dat we geen afscheid nemen van biologische parameters.

```
Data: df_nb
Models:
fit.hnp.TMB.0: sum_Cases ~ 1 + (1 | Studie) + offset(log(sum_N)), zi=~., disp=~1

fit.hnp.TMB.1: sum_Cases ~ Dosering_log + (1 | Studie) + offset(log(sum_N)), zi=~., disp=~1

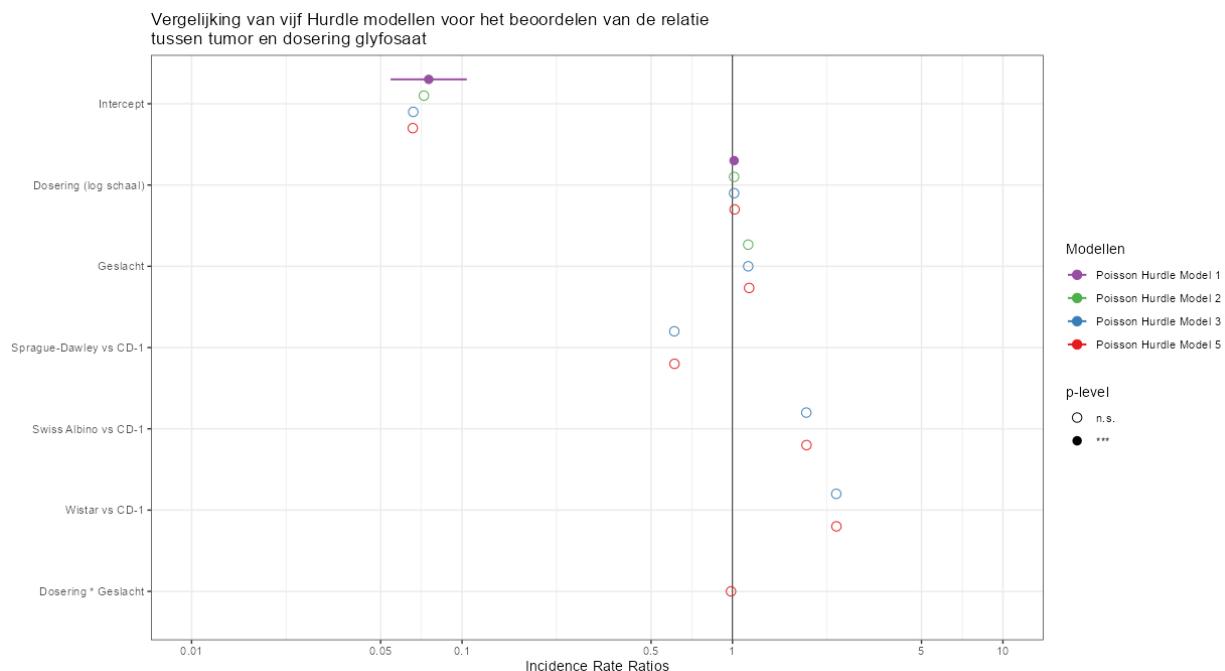
fit.hnp.TMB.2: sum_Cases ~ Dosering_log + Geslacht + (1 | Studie) + offset(log(sum_N)), zi=~., disp=~1

fit.hnp.TMB.3: sum_Cases ~ Dosering_log + Geslacht + Soort + (1 | Studie) + offset(log(sum_N)), zi=~., disp=~1

fit.hnp.TMB.5: sum_Cases ~ Dosering_log * Geslacht + Soort + (1 | Studie) + offset(log(sum_N)), zi=~., disp=~1

fit.hnp.TMB.4: sum_Cases ~ Dosering_log + Geslacht + Soort + Duur + (1 | Studie) + offset(log(sum_N)), zi=~., disp=~1
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi_Df	Pr(>Chisq)
fit.hnp.TMB.0	4	8430.4	8449.4	-4211.2	8422.4			
fit.hnp.TMB.1	6	8373.6	8402.0	-4180.8	8361.6	60.792	2	0.00000000000006299 ***
fit.hnp.TMB.2	8	8259.3	8297.2	-4121.6	8243.3	118.353	2	< 0.00000000000000022 ***
fit.hnp.TMB.3	14	8248.7	8315.0	-4110.3	8220.7	22.600	6	0.0009422 ***
fit.hnp.TMB.5	16	8242.4	8318.2	-4105.2	8210.4	10.242	2	0.0059701 **
fit.hnp.TMB.4	18					2		



**Figuur 134.** Schattingen voor elke parameter in een specifiek Poisson Hurdle model.

Uit de tabel kunnen we halen dat *Poisson Hurdle Model 5* het beste bij de data past. Dat is het model met de meeste parameters en in **Figuur 134** het model met de rode bolletjes. Zoals te zien is, is de schatting voor dosering niet statistisch significant. De horizontale as van het figuur laat echter geen coëfficiënten zien, maar spreek van zogenaamde Incidence Rate Ratios. Wat deze metriek tracht te laten zien is het verschil in incidentie gegeven een bepaald moment. Voor een variabele als dosering betekent dit in feite hoeveel meer tumorgevallen we mogen verwachten zodra de dosering stijgt. Dat is in dit geval dus verwaarloosbaar. Wel zien we duidelijke verschillen tussen soorten, maar die schattingen zijn te onbetrouwbaar om significant te noemen. De richting is wel in lijn met eerdere resultaten.

We kunnen deze exercitie nu herhalen voor de negatieve binomiale modellen. Ik volg exact dezelfde procedure en zal exact dezelfde manier van rapporteren hanteren. Uiteindelijk kunnen we het beste Poisson model met het beste negatieve binomiaal model vergelijken. Maar eerst de resultaten van elk negatief binomiaal model hier beneden. Het mag niet verbazen dat het hurdle model ook hier weer als beste bij de data past. Het beste model is deze keer *NB Hurdle Model 3*.

```
Data: df_nb
```

```
Models:
```

```
fit.hnnb.TMB.0: sum_Cases ~ 1 + (1|Studie) + offset(log(sum_N)), zi=~., disp=~1
```

```
fit.hnnb.TMB.1: sum_Cases ~ Dosering_log + (1|Studie) + offset(log(sum_N)), zi=~., disp=~1
```

```
fit.hnnb.TMB.2: sum_Cases ~ Dosering_log + Geslacht + (1|Studie) + offset(log(sum_N)), zi=~., disp=~1
```

```
fit.hnnb.TMB.3: sum_Cases ~ Dosering_log + Geslacht + Soort + (1|Studie) + offset(log(sum_N)), zi=~., disp=~1
```

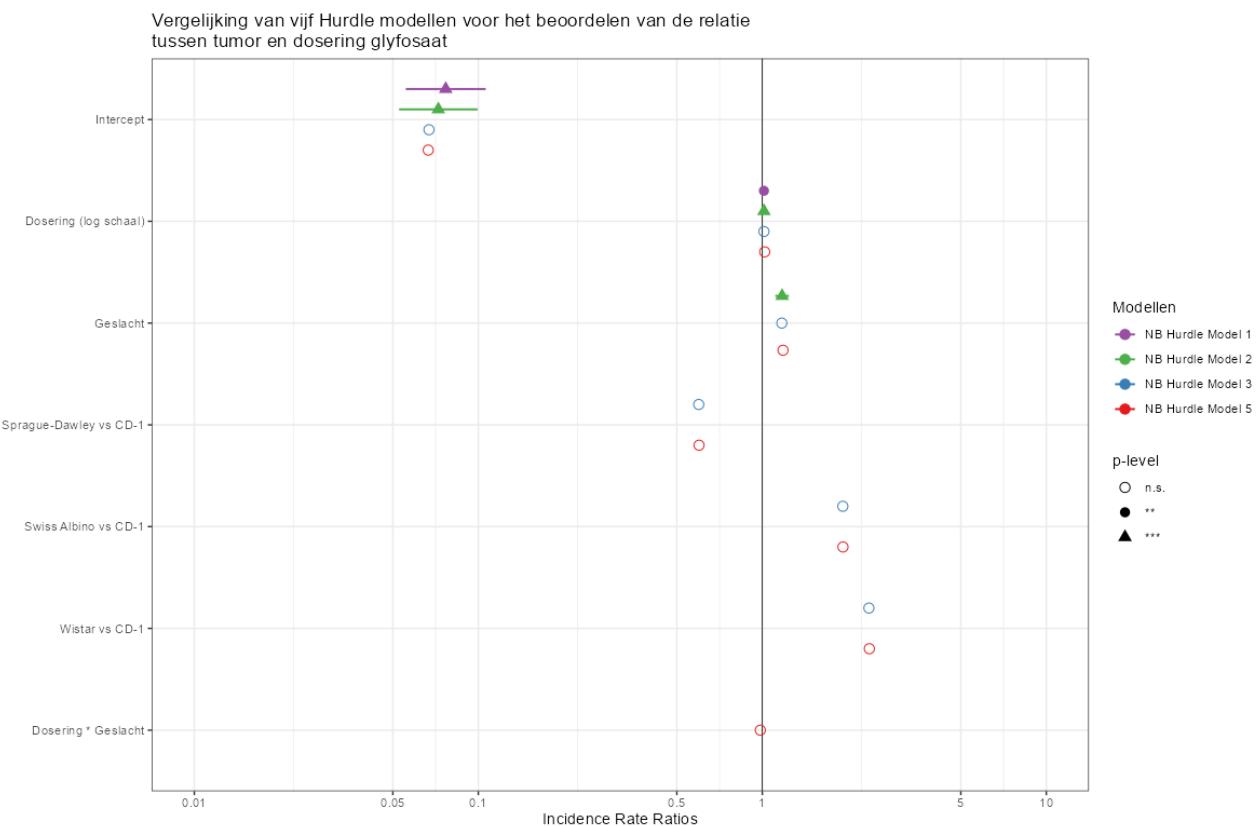
```
fit.hnnb.TMB.5: sum_Cases ~ Dosering_log * Geslacht + Soort + (1|Studie) + offset(log(sum_N)), zi=~., disp=~1
```

```
fit.hnnb.TMB.4: sum_Cases ~ Dosering_log + Geslacht + Soort + Duur + (1|Studie) + offset(log(sum_N)), zi=~., disp=~1
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi_Df	Pr(>Chisq)
fit.hnnb.TMB.0	5	6499.3	6522.9	-3244.6	6489.3			
fit.hnnb.TMB.1	7	6488.7	6521.9	-3237.4	6474.7	14.524	2	0.0007016 ***
fit.hnnb.TMB.2	9	6442.7	6485.3	-3212.3	6424.7	50.040	2	0.00000000001361 ***
fit.hnnb.TMB.3	15	6430.4	6501.5	-3200.2	6400.4	24.240	6	0.0004718 ***
fit.hnnb.TMB.5	17	6430.6	6511.1	-3198.3	6396.6	3.837	2	0.1468293
fit.hnnb.TMB.4	19						2	

---

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



**Figuur 135.** Schattingen voor elke parameter in een specifiek Poisson Hurdle model.

Bovenstaande figuur toont *NB Hurdle Model 3* in het blauw. Wat opvalt is dat in het blauwe model geen enkele van de parameters significant is. In de modellen met minder parameters is de dosering en het geslacht wel significant, maar zodra we soort toevoegen verdwijnt dit. Ook Portier kijkt specifiek per soort. Dit wil niet zeggen dat dosering geen relatie heeft met het aantal kankergevallen, maar wel dat het toevoegen van parameters de relaties in het model verschuift. Wie echter *NB Hurdle Model 3* met *NB Hurdle Model 1* en *NB Hurdle Model 2* vergelijkt ziet dat *NB Hurdle Model 3* echt beter past bij de geobserveerde data. We zouden dus een goede reden moeten hebben om dit model niet te kiezen. Vooralsnog ken ik die reden niet.

## Wat kunnen we hieruit concluderen?

Het lukt mij niet om een significante relatie te vinden tussen de dosering van glyphosate en het totaal aantal kankergevallen per dosering. In mijn analyses heb ik rekening gehouden met factoren zoals geslacht en soort, maar de grote hoeveel kanker bij de nul-dosering maakt het lastig om een statistisch significant resultaat te vinden. Een verdere splitsing per

tumornoort heb ik niet gedaan met de simpele reden dat een correctie op meerdere testen elk significant resultaat zou laten verdwijnen.

Wat mij nu nog rest is om de data op een Bayesiaanse manier te analyseren. Niet alleen stap ik dan uit de wereld van de statistisch significantie, maar het geeft mij ook meer vrijheden in het modelleren van de data doormiddel van zogenaamde *priors* en *hyperparameters*. Het is tijd om te stappen in de wereld van bewijs en het achterhalen welke hypothese het sterkst wordt ondersteund door de onderliggende data. Ik ben van mening dat dit uiteindelijk de enige juiste manier is om de data te analyseren.

## Glyfosaat: een Bayesiaanse analyse

---

*As mentioned previously there are different ‘schools of thought’ about the statistical analysis of data. Much of the work in toxicology has been carried out based upon a traditional frequentist approach particularly around the concept of hypothesis testing. While recognizing that alternative viewpoints exist and this is a controversial area, most of the emphasis in this document will be on the traditional approaches<sup>113</sup>.*

---

Dit citaat uit de OECD richtlijnen laat duidelijk zien hoe er over statistiek gedacht wordt in dit onderzoeks veld. Het komt er op neer dat men weet dat er verschillende paradigma’s zijn, maar dat er uiteindelijk toch wordt gekozen om gebruik te maken van de meer ‘traditionele’ aanpak. Helemaal verbazen hoeft het trouwens ook weer niet, want de frequentistische statistiek is het paradigma wat het meest wordt gedoceerd en het vaakst voorkomt. In die zin zou je zelfs met enige humor kunnen benoemen dat omdat iets heel vaak voorkomt het wel zo zal zijn. Precies in lijn met de frequentistische statistiek.

Toch wil ik verder gaan en nu, op het einde van dit rapport, naar de Bayesiaanse statistiek manevreren. We hebben al aardig wat modellen gemaakt, analyses gecontroleerd en voorbeelden tot ons genomen. Wat dit onderdeel bijzonder maakt is dat een groot deel van alles wat we tot nu toe hebben gedaan gaan loslaten. Daarentegen gaan we kennis maken met *priors*, *hyperparameters*, Monte Carlo simulaties en *likelihood ratios*. Tijd voor een korte introductie door middel van een simpel voorbeeld.

### Een simpele regressie als voorbeeld

Om dit rapport enigszins behapbaar te maken zal ik niet stil staan bij de exacte wiskundige formulering rondom de Bayesiaanse statistiek. Voor de geïnteresseerde lezer zijn er voldoende introductieboeken te verkrijgen in zowel het Nederlands<sup>114</sup> als in het Engels<sup>115</sup>.

---

<sup>113</sup> [https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453\\_9789264221475-en](https://www.oecd-ilibrary.org/environment/guidance-document-116-on-the-conduct-and-design-of-chronic-toxicity-and-carcinogenicity-studies-supporting-test-guidelines-451-452-and-453_9789264221475-en)

<sup>114</sup> Ronald Meester & Klaas Sloten. Kan dat geen toeval zijn? Een kritische blik op statistische bewijsvoering.

<sup>115</sup> Will Kurt. Bayesian Statistics The Fun Way. Understanding Statistics and Probability with Star Wars, LEGO, and Rubber Ducks.

Wat ik wel ga doen is een statistisch model (in dit geval een lineaire regressie zoals jullie die al kennen uit het hoofdstuk over de **Lineaire dose-response analyse**) tot me nemen en deze op een Bayesiaanse manier analyseren. Voordat ik dit doe zal ik uitleggen wat er exact veranderd in zowel de analyse als in de interpretatie van de onderzoeksresultaten.

In de analyse zelf betekent het dat we nu een stuk meer vrijheid krijgen om ons model op te stellen. We kunnen op de Bayesiaanse manier nog steeds een model specificeren waarin we bepalen welke variabelen we willen meenemen. In het model hieronder is dat *Dosering (op de log schaal)*, *Geslacht* en *Studie*. Dat verandert niet.

Wat wel verandert is het opstellen van de zogenaamde *prior*. De prior in de Bayesiaanse verdeling is de frequentieverdeling die past bij de kennis die je tot dan toe hebt over een variabele. Stel dat we een studie doen naar de lengte van jongens en meisjes in Nederland. En stel ook dat we al 1000 metingen hebben gedaan. Uit de meting blijkt dat voor beide groepen een normaalverdeling het beste bij de observaties past met gemiddelde X en standaard deviatie Y. Stel verder dat we nog eens 1000 mensen gaan onderzoeken. In dat geval vormt het eerste onderzoek hoogstwaarschijnlijk de prior voor het tweede onderzoek. Dit betekent dat het analyseren van de tweede set gegevens afhankelijk is van bevindingen uit de eerste set aan gegevens. Die afhankelijk is een kracht, maar wordt helaas nog te vaak gezien als een beperking omdat er zogenaamd naar een bepaalde bevinding toe kan worden gemodelleerd. Ik zal daarom ook laten zien waarom dit helemaal niet zo makkelijk is.

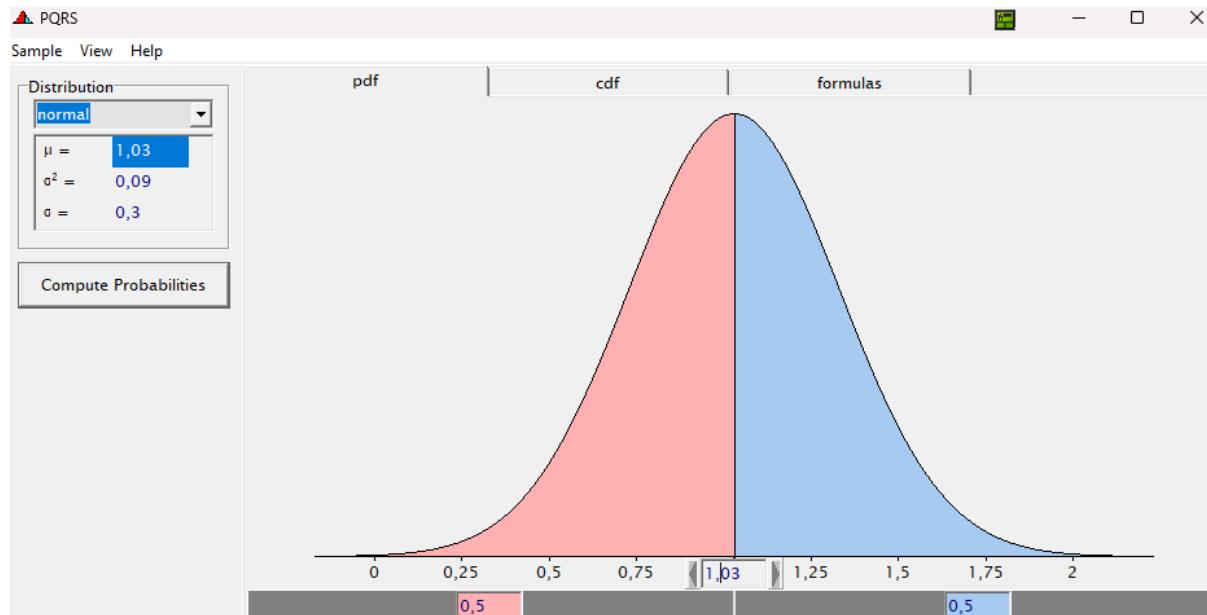
Wat betekent dit alles voor de regressie hier beneden? Ten eerste dat we niet zomaar met één druk op de knop de data kunnen inladen en de analyse kunnen uitvoeren. Een Bayesiaanse analyse verwacht dat je van tevoren aangeeft welke verdeling past bij de parameters die je wil gaan schatten. Belangrijker is dat je ook aangeeft hoe zeker je bent over die verdeling.

Stel nou dat we aannemen dat de relatie tussen Dosering (niet op de log schaal) en het totaal aantal tumorgevallen 1.03 is. Dit betekent dat met elke toename van de dosering het aantal kankergevallen met 1.03 toeneemt. Als de dosering dus van 0 naar 100 gaat, dan betekent dat een toename van  $100 * 1.03$  oftewel 13 kankergevallen meer<sup>116</sup>. Uiteraard is die

---

<sup>116</sup> Deze manier van redeneren is hoe een regressie werkt, maar eigenlijk laat het ook zien hoe breekbaar een dergelijke kijk naar de wereld is. De coëfficiënt 1.03 lijkt hier natuurlijk veel stabieler dan dat deze in werkelijk is (als het al een rechte lijn is).

1.03 een gemiddelde waarde en vergeten we hier de spreiding. Stel nou dat de standaard deviatie van die coëfficiënt 0.3 is en dat de schatting van die parameter past bij een normaalverdeling. We zouden dan **Figuur 136** kunnen maken.



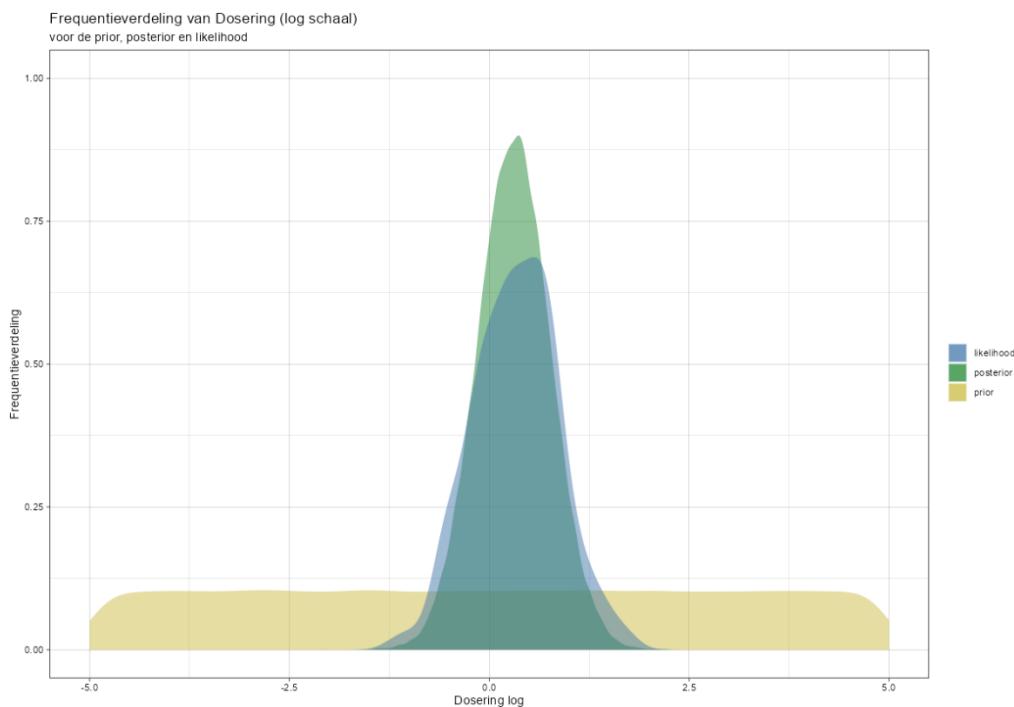
**Figuur 136.** Voorbeeld van een normaalverdeling met gemiddelde 1.03 en standaard deviatie 0.3 die zou kunnen gelden als prior voor de parameter Dosering.

De lezer kan nu terecht opmaken dat de schatting van de parameter *Dosering* eenzelfde verdeling volgt (namelijk de normaalverdeling) als de schatting voor lengte n de Nederlandse bevolking. We hebben dus wederom te maken met een kansverdeling.

Wie zich nu wellicht toch een beetje verloren voelt kunnen we geruststellen, want in het gehele rapport zijn de we kansverdelingen nooit uit het oog verloren. Zo hebben we in het hoofdstuk bij de dose-response analyses keer op keer schattingen gemaakt over de invloed van *Geslacht* en *Dosering*. Deze variabelen noemen we in de Bayesiaanse statistiek ook wel hyperparameters. Het zijn dus kansverdelingen en die kansverdelingen gebruiken we om op een wiskundige manier te duiden hoeveel informatie we al verzameld hebben én hoeveel informatie we uiteindelijk overhouden na het verzamelen van (nieuwe) gegevens. Het gaat dus boven alles over het vergaren van informatie en bepalen van zekerheden. Het duiden van parameters door middel van verdelingen maakt dit inzichtelijk. Dit hebben we al

vanaf de start gezien toen we het nog hadden over de verdeling van lengtes in Nederlandse mannen en vrouwen.

De verdeling uit **Figuur 136** zou dus kunnen volgen nadat we data hebben geanalyseerd door middel van de Bayesiaanse statistiek. In dat geval noemen we de verdeling een *posterior distribution*: een verdeling van nieuwe kennis die volgt op basis van de prior verdeling (oude kennis) én de analyse van de geobserveerde gegevens (nieuwe gegevens). Het is denk ik nu tijd om te laten zien hoe dat ongeveer werkt en wat we daarvoor nodig hebben is een aanname, de *prior distribution* en een set aan gegevens. We kunnen dan, door de *stelling van Bayes* te volgen, de *posterior distribution* berekenen<sup>117</sup>. Het resultaat zien we in **Figuur 137**. Wat hopelijk direct opvalt is dat de groene en blauwe kansverdeling haast gelijk zijn aan elkaar. Dat komt door de gele kansverdeling die heel vlak is en dus weinig informatie geeft. Het gros van de informatie komt dus van de geobserveerde waarden. Daarmee is ook de relatie met de frequentistische statistiek gelegd: de Bayesiaanse statistiek en de frequentistische statistiek tonen haast identieke resultaten bij zeer onzeker priors. Dit worden ook wel zwakke priors genoemd, vanwege hun beperkte bijdrage.

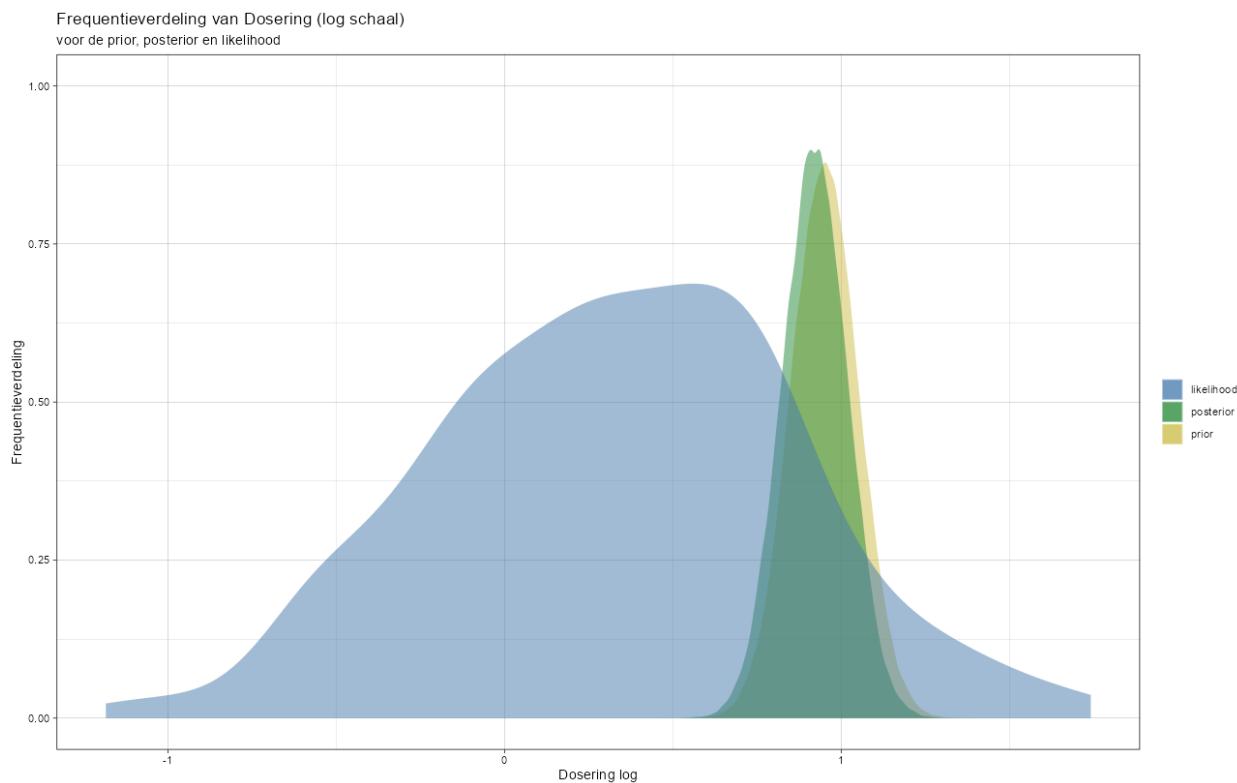


**Figuur 137.** Frequentieverdeling van de prior, likelihood en posterior van de Dosering parameter.

<sup>117</sup> Ik blijf hier de Engelse benaming volgen, maar wat ik eigenlijk bedoel is de kansverdeling vooraf (*prior probability*) en de kansverdeling achteraf (*posterior probability*). We noemen dit ook wel de kansverdeling op basis van oude kennis (*prior probability*) en de kansverdeling op basis van nieuwe gegevens (*posterior probability*).

De beste manier om deze grafiek te lezen is door te bedenken dat de *posterior* een functie is van de *prior* vermenigvuldig met de *likelihood*. Oftewel, nieuwe kennis is een functie van oude kennis vermenigvuldigd met nieuwe gegevens. En omdat ik niet met zekerheid durf te stellen hoe de verdeling van de coëfficiënt voor *Dosering* er precies uitziet, hangt deze grotendeels af van de berekeningen op basis van de observaties. Dit is geheel in lijn met hoe wij mensen kennis tot ons nemen: als wij geen mening hebben over een onderwerp (of hoogst onzeker zijn) dan is elke vorm van kennis op basis van nieuwe gegevens welkom. Zouden wij wel een sterke mening hebben, dan zouden wij ons maar moeilijk laten bewegen ook als de nieuwe gegevens overweldigend zijn. Ook dit kunnen we grafisch weergeven.

Stel nou dat we overtuigd zijn dat de relatie tussen glyfosaat en kankergevallen negatief is, met een coëfficiënt van 0.95 en standaard deviatie 0.1. We laten voor nu in het midden of we onze overtuiging baseren op daadwerkelijke kennis, of dat het gewoonweg een wens is, maar wat ik wel kan doen is tonen wat het effect is op de uitkomst ná het zien van nieuwe data. Het resultaat is zichtbaar in **Figuur 138**. Wat we nu kunnen is dat de prior en de posterior (oude en nieuwe kennis) heel dicht bij elkaar liggen. De verdeling van de geobserveerde waarden rondom *Dosering* ziet er dan een stuk meer onzeker uit. Dit plaatje is het gevolg van een zeer sterke prior. Deze is zo sterkt dat nieuwe gegevens niet direct leiden tot nieuwe kennis.



**Figuur 138.** Frequentieverdeling van de prior, likelihood en posterior van de Dosering parameter.

De oplettende lezer zal kunnen aandragen dat een hele sterke prior (gebaseerd op data, kennis, of gewoonweg vooringenomenheid) zal leiden tot een marginale verschuiving van de posterior kansverdeling. Met andere woorden: als ik wil dat glyfosaat **niet** kankerverwekkend is dan zal dit ook nooit blijken ongeacht de gegevens uit de studie. Het enige wat ik hoef te doen is een extreem sterke prior te bepalen.

Toch zal het niet zo'n vaart lopen. Ten eerste laten **Figuur 137** en **Figuur 138** zien hoe de onderlinge relatie eruit ziet. Dit maakt dat iemand die een hele sterke prior stelt deze ook moet kunnen verdedigen. Anders spreken we inderdaad van beïnvloeding.

Het mooie van de Bayesiaanse methodiek is dat we nu een wiskundige manier tot ons kunnen nemen om informatie te combineren. Oude kennis, gecombineerd met nieuwe gegevens, brengt nieuwe kennis. De bewijskracht veranderd daarmee (of niet). Bij het schatten van de *hyperparameters* is er nog het bijkomende voordeel dat het stellen van een prior maakt dat de posterior ook een duidelijke vorm krijgt. We hoeven namelijk niet alles aan de nieuwe geobserveerde waarden over te laten. Laten we met deze kennis over parameterschattingen, verdelingen, oude én nieuwe kennis eens aan de slag gaan met de regressie zoals al eerder benoemd.

Ik zal gemakshalve de output laten zien vanuit de frequentistische statistiek alvorens de output vanuit de Bayesiaanse statistiek te tonen<sup>118</sup>. Wat direct opvalt is dat ik deze keer de variabele *Geslacht* heb toegevoegd.

```
Linear mixed model fit by REML ['lmerMod']
Formula: sum_Cases ~ Dosering_log * Geslacht + (1 | Studie)
Data: df_lmer

REML criterion at convergence: 912.9

Scaled residuals:
    Min   1Q Median   3Q   Max
-2.1806 -0.4904 -0.0993  0.3170  3.0320

Random effects:
  Groups      Name        Variance Std.Dev.
  Studie (Intercept) 299.2     17.30
  Residual           282.7     16.81
Number of obs: 106, groups: Studie, 13

Fixed effects:
                     Estimate Std. Error t value
(Intercept)       35.4129  5.3401   6.632
Dosering_log       0.3722  0.7148   0.521
GeslachtVrouw     -7.0515  3.3182  -2.125
Dosering_log:GeslachtVrouw -0.1268  1.0042  -0.126

Correlation of Fixed Effects:
            (Intr) Dosering_ Dsrng_ GslchV
Dosering_Ig -0.075
GeslachtVrw -0.310
Dsrng_Ig:GV  0.053  -0.708  -0.176
```

De frequentistische output (boven) wijkt iets af van de Bayesiaanse output (beneden). Dit is de invloed van de verschillende priors.

---

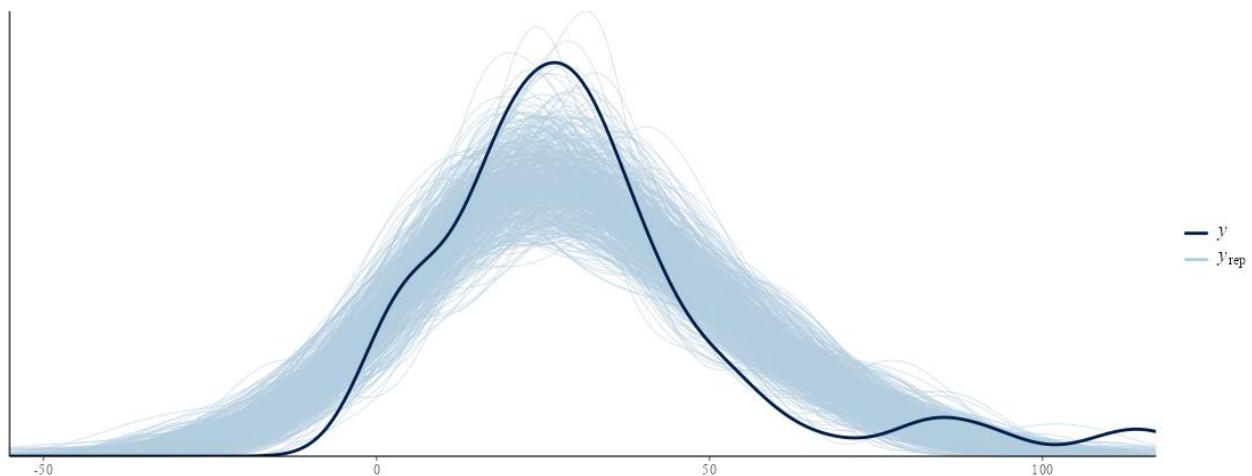
<sup>118</sup> Eigenlijk is het een beetje gek dat ik keer op keer spreek alsof iets vanuit de frequentistische statistiek komt of vanuit de Bayesiaanse statistiek. Waar het eigenlijk om gaat is dat ik de data op twee verschillende manier behandel en daarom ook met twee verschillende interpretaties ga komen. We kunnen hier echt spreken van een paradigma shift.

Family: gaussian							
Links: mu = identity; sigma = identity							
Formula: sum_Cases ~ Dosering_log * Geslacht + (1   Studie)							
Data: df_Lmer (Number of observations: 106)							
Draws: 4 chains, each with iter = 40000; warmup = 3000; thin = 1; total post-warmup draws = 148000							
Group-Level Effects:							
~Studie (Number of levels: 13)							
sd(Intercept)	Estimate	Est.Error	I-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
	16.60	1.68	13.58	20.08	1.00	3895	7706
Population-Level Effects:							
Intercept	Estimate	Est.Error	I-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
5.62	1.14	2.87	7.23	1.00	6294	11580	
Dosering_log	0.47	0.63	-0.77	1.72	1.00	6348	11482
GeslachtVrouw	-3.23	1.45	-4.94	0.37	1.00	7431	11019
Dosering_log:GeslachtVrouw	-0.33	0.89	-2.07	1.43	1.00	6444	11455
Family Specific Parameters:							
sigma	Estimate	Est.Error	I-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
	15.04	0.95	13.30	17.04	1.00	8901	14693
Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS and Tail_ESS are effective sample size measures, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).							

De Bayesiaanse output verschilt verder nog meer. Zo zien we ook intervallen<sup>119</sup> en achter elke parameter staan nog extra kolommen die iets zeggen over de kwaliteit van elke schatting. Door de manier waarop ik tot een antwoord kom (via [Monte Carlo simulatie](#)) heb ik te maken met processen die ontvankelijk zijn voor storingen. Ik wil er zeker van zijn dat het algoritme een oplossing vindt die ligt in een globaal minimum, en **niet** een lokaal minimum<sup>120</sup>. Een simpele manier om dit te beoordelen zien we in **Figuur 139**. Deze figuur laat de verdeling van de y-variabele zien vanuit de data (dikke lijn) én de trekkingen vanuit de *posterior probability* (dunne lijnen). Het komt niet helemaal overeen.

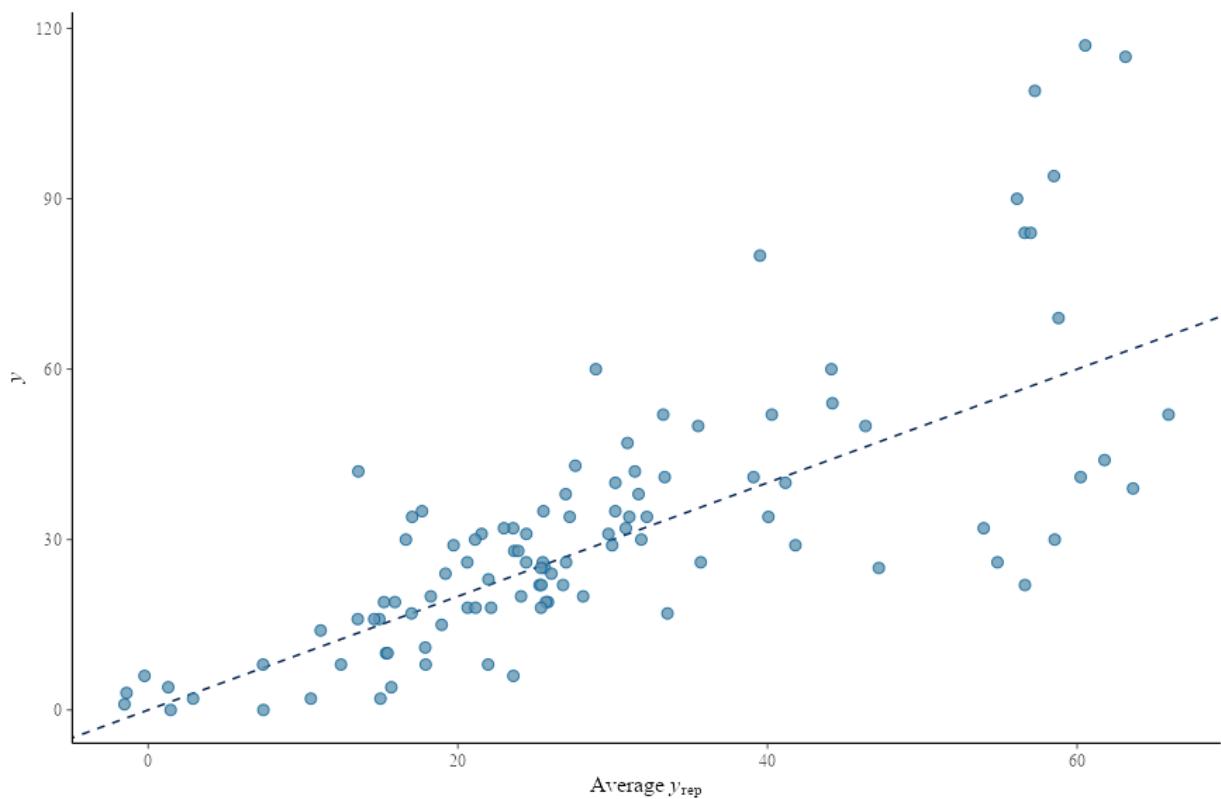
<sup>119</sup> We spreken nu niet meer van betrouwbaarheidsintervallen, maar daarover later meer.

<sup>120</sup> <https://nl.mathworks.com/help/optim/ug/local-vs-global-optima.html>

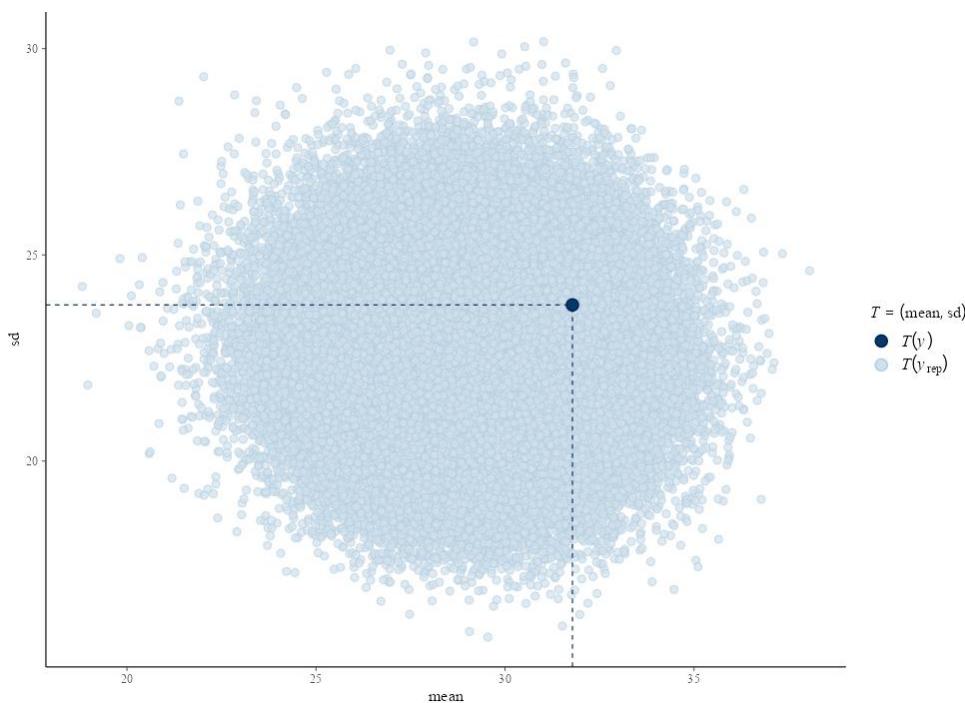


**Figuur 139.** Posterior check van het model. De donkere lijn is de geobserveerde waarde – de dunnen lijnen zijn trekken uit de verdelingen afkomstig van het model.

We zien die afwijking ook in **Figuur 140** en **Figuur 141**. Dit alles hoeft trouwens niet noodzakelijkerwijs slecht te zijn!

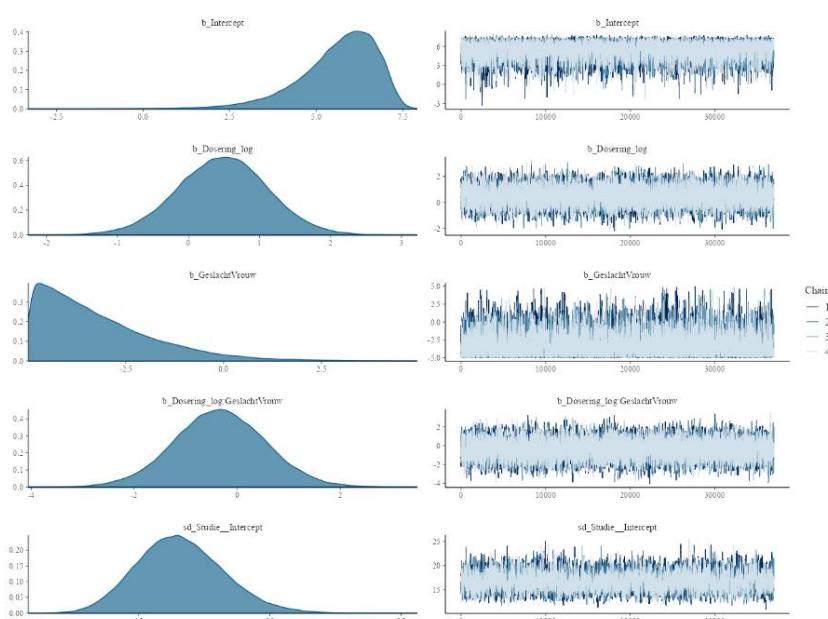


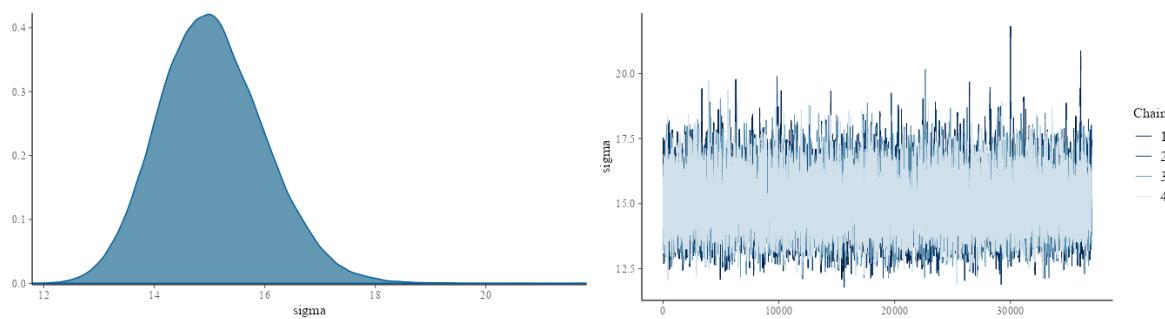
**Figuur 140.** Posterior check van het model waarbij zichtbaar wordt waar de afwijking tussen observatie en modelvoorspelling het grootst is.



**Figuur 141.** Posterior check van het model waarbij zichtbaar wordt waar de afwijking tussen observatie en modelvoorspelling het grootst is.

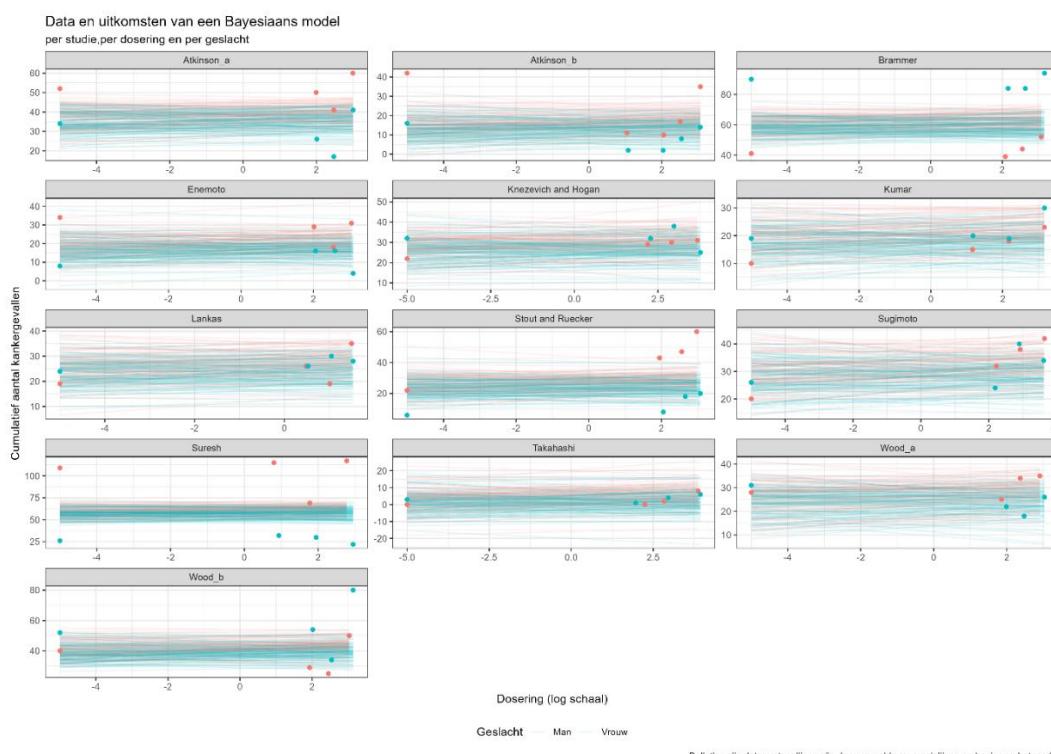
We kunnen ook kijken naar de schattingen van elke parameter in dit model. We hebben er zes (**Figuur 142**). Wat opvalt is dat ze allemaal een andere verdeling hebben (figuren links). De figuren rechts laat zien of de schattingen van de parameters voldoende betrouwbaar is – wat je wil zien is een wat chaotisch patroon, maar niet té chaotisch. Wat mij opvalt is dat de schatting van de parameter Geslacht tegen de ondergrens aanligt. Dat is vaak problematisch.





**Figuur 142.** De posterior verdeling van elke parameter in het model (links). Rechts zien we of de verdelingen, zoals deze tot stand zijn gekomen, op een manier is die zekerheid biedt aan de betrouwbaarheid van die verdelingen.

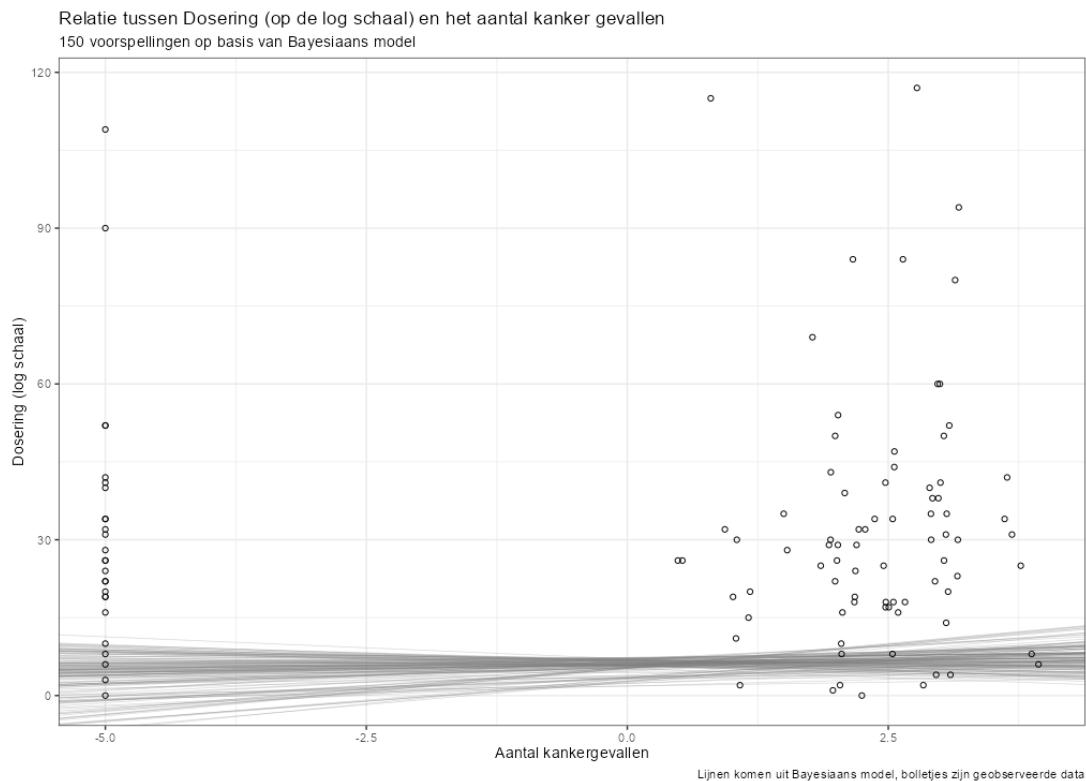
Wat we nu verder kunnen doen is de schatting van de Bayesiaanse statistiek afzetten tegen de geobserveerde data. Ik ben niet geïnteresseerd in een verdere vergelijking tussen de frequentistische statistiek en de Bayesiaanse statistiek, maar ik heb wel interesse in de vaardigheid van het Bayesiaanse model om schattingen te maken op basis van de prior verdelingen én de data. Dit kunnen we toetsen door trekkingen te doen uit de posterior verdelingen en dan voorspellingen te maken. Deze voorspellingen kunnen we vervolgens afzetten tegen de geobserveerde waarden (**Figuur 143**).



**Figuur 143.** Voorspellingen (lijnen) en geobserveerde waarden in de relatie tussen dosering en het aantal kankergevallen.

Het is trouwens niet nodig dat de voorspellingen (de lijnen) dicht tegen de geobserveerde data (de bolletjes) aanligt. Dat is een fout die vaak wordt gemaakt. In de Bayesiaanse statistiek gaat het niet om het schatten van de frequentie om te bepalen of een bepaalde waarde vaker of minder vaak voorkomt dan voorheen gedacht, maar gaat het bovenal over de hoe zeker je kunt zijn op basis van de informatie die vorhanden is. Het kan dus heel goed zijn dat een set nieuwe observaties leidt tot nieuwe kennis, maar dat hoeft niet te betekenen dat die nieuwe kennis voornamelijk beschreven kan worden vanuit de nieuwe observaties. Dit is waarom de prior zo belangrijk is<sup>121</sup>. Dit is ook waarom het zo belangrijk is om uitgevoerde stappen stuk voor stuk te beoordelen op nut én accuraatheid.

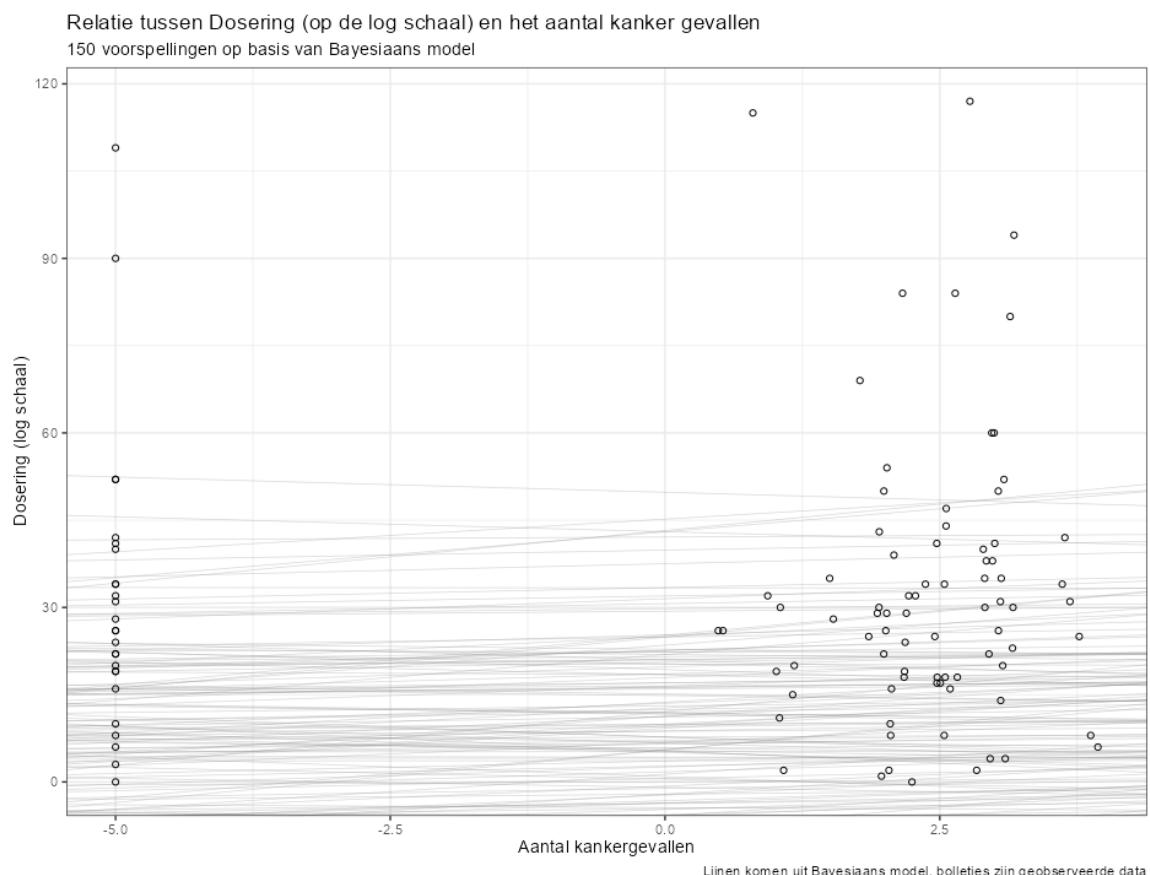
Toch helpt het wellicht om bovenstaande grafiek op te bouwen. Zo kunnen we eerst de voorspellingen uit het Bayesiaanse model afzetten tegen de geobserveerde gegevens (**Figuur 144**). Wie dit resultaat ziet zal denken dat het model het niet goed doet, maar dat is niet helemaal waar.



**Figuur 144.** Voorspellingen (lijnen) en geobserveerde waarden in de relatie tussen dosering en het aantal kankergallen. In dit voorbeeld neem ik alleen de marginale (gemiddelde) parameter waarden mee. Ik neem in dit voorbeeld niet de correctie voor studie mee.

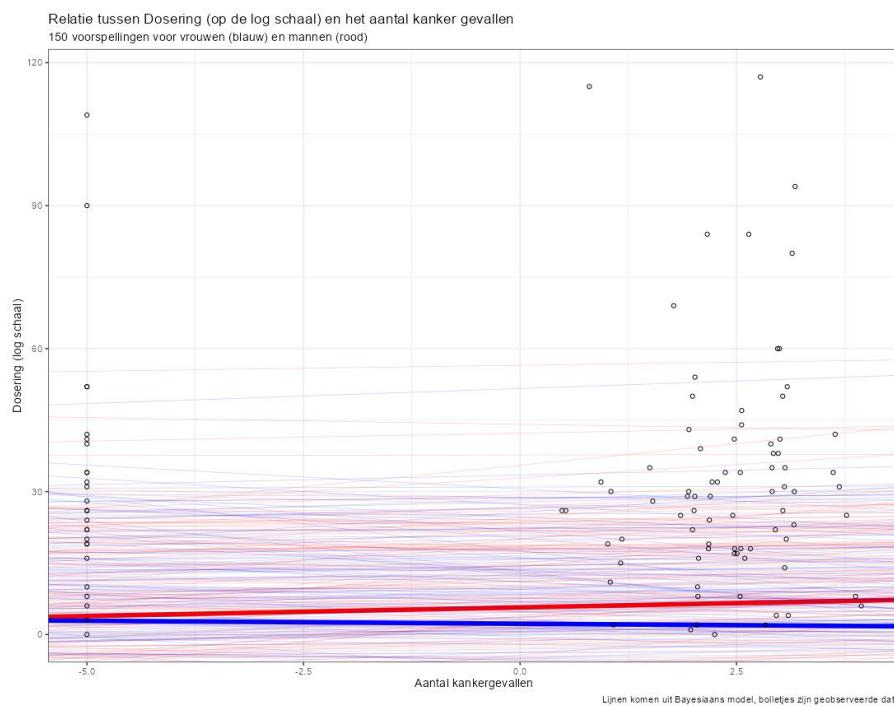
<sup>121</sup> In deze blog post beschreef ik dat een afwijking tussen voorspelling en geobserveerde data heel normaal is (<https://blog.devgenuis.io/analyzing-height-weight-5ba7449d6431>).

Wat er mist is een essentieel onderdeel, namelijk *Studie*. Dat onderdeel kan ik toevoegen en dan krijg ik **Figuur 145**. De eerste indruk is om te zeggen dat dit er ‘beter’ uitziet, maar wat we eigenlijk zien zijn voorspellingen die gebaseerd zijn op meer informatie uit het model. We komen daarmee ook een stap dichter bij **Figuur 143**. Door de variabele *Geslacht* toe te voegen krijgen we **Figuur 146**. In deze figuur zit nog meer informatie en wat direct opvalt is dat de gemiddelde lijn voor de relatie tussen *Dosering* en aantal kankergevallen veel scherper is dan de variatie rondom die lijnen doen vermoeden<sup>122</sup>. Wat ook opvalt uit de figuren is dat het model niet in staat is om de extremen te modelleren. Dit zien we ook in **Figuur 147**.

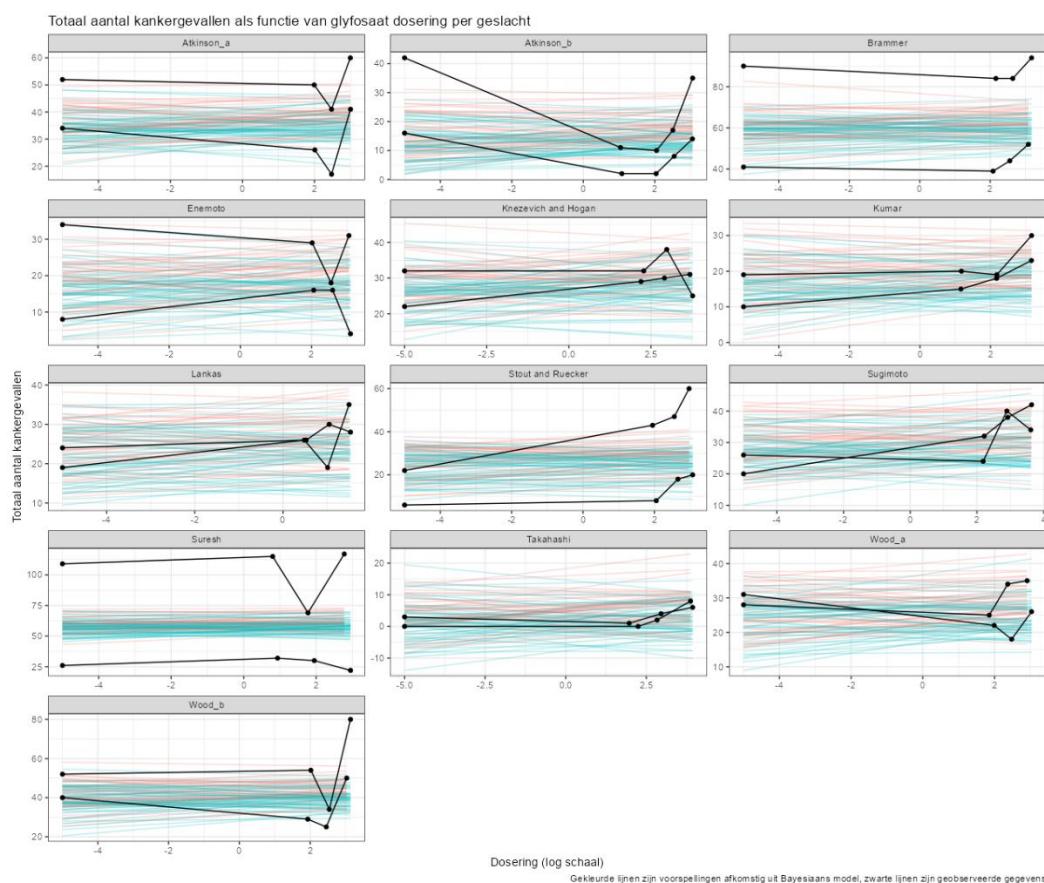


**Figuur 145.** Voorspellingen en geobserveerde waarden waarbij de correctie voor studie wel wordt meegenomen. Duidelijk zichtbaar, in vergelijking met **Figuur 144**, is de toename in spreiding vanuit de schattingen.

<sup>122</sup> Of, met andere woorden, de relatie tussen Dosering en kanker per geslacht kent meer onzekerheid dan op het eerste zicht aanwezig lijkt als je alleen naar de gemiddelde relatie kijkt.

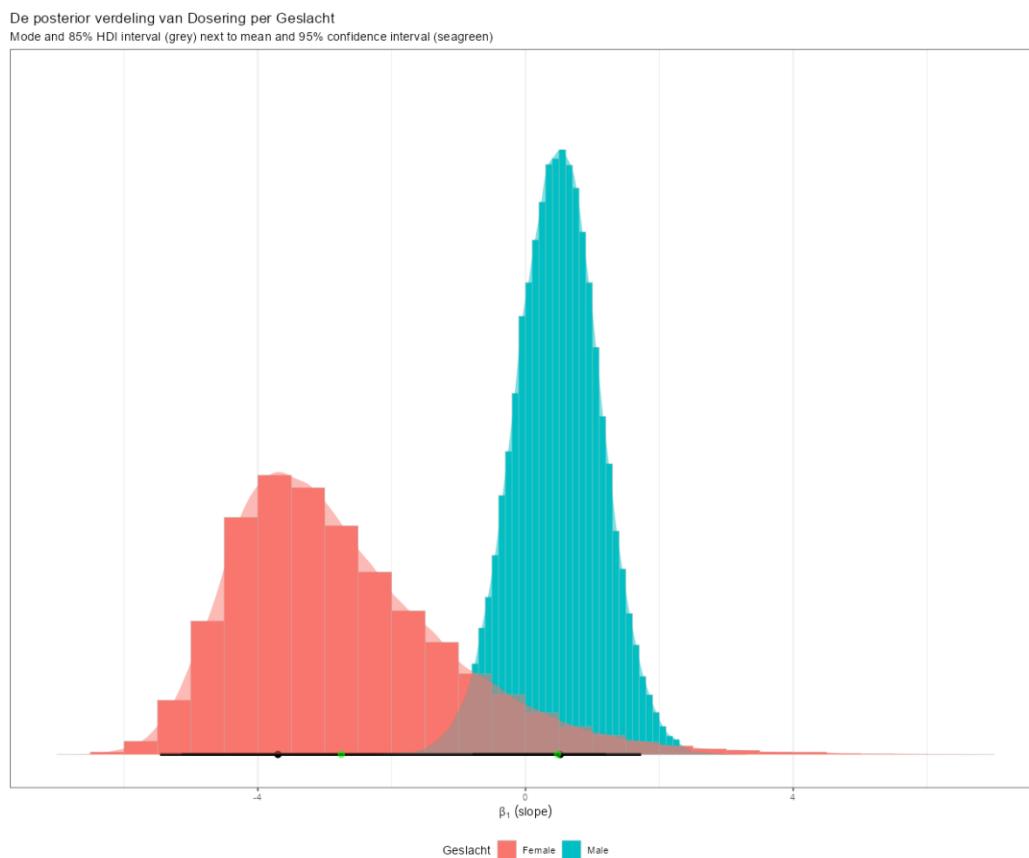


**Figuur 146.** Voorspellingen voor mannen en vrouwen voor de relatie tussen dosering en aantal kankergevallen.



**Figuur 147.** Relatie tussen dosering en aantal kankergevallen, per studie en per geslacht waarbij de gekleurde lijnen de voorspellingen zijn per geslacht en de zwarte lijnen en bolletjes de geobserveerde waarden zijn. Het gaat hier om totale kankergevallen per dosering per geslacht en per studie.

Deze modellen zijn dus niet de modellen die we willen gebruiken voor het eindresultaat, maar lenen zich wel voor de noodzakelijke uitleg rondom het Bayesiaanse principe (**Figuur 137, Figuur 138 en Figuur 148**). Maar we zijn er nog niet en wat rest is de introductie van de likelihood ratio<sup>123</sup> als vervanger van de statistische significantie waarover we al zo vaak hebben gesproken.



**Figuur 148.** De posterior verdeling van de parameter Dosering voor mannen en vrouwen.

## De likelihood ratio

De likelihood ratio (LR) geeft, in zijn meest eenvoudige beschrijving, weer hoe sterk de bewijslast voor hypothese 1 is ten opzichte van hypothese 2, gegeven de oude kennis die we al hadden en gegeven de nieuwe data die we hebben verzameld. De uitkomst is, zoals de naam al doet vermoeden, een ratio. Hoe hoger het getal, hoe sterker het bewijs in het voordeel voor de hypothese in de noemer. Dit is dus een hele andere benadering dan de statistische significantie die veel meer binair is in haar aanpak. De LR, daarentegen, is een

<sup>123</sup> Wordt ook wel Bayes factor genoemd.

continue schaal. Uiteraard ligt het in de aard van de mens om ook deze schaal op te delen in classificaties en zo een antwoord te hebben op de vraag: “hoe sterk moet de bewijslast zijn om relevant te zijn?”. We zouden kunnen zeggen dat elke toename vanaf één al telt. Zo geeft een LR van twee aan dat het bewijs tweemaal sterker is voor hypothese 1 (noemer) dan voor hypothese 2 (deler). Voor de meeste mensen is dit niet genoeg en daarom wordt vaak de volgende schaal gehanteerd (**Tabel 34**). Zoals je kunt zien neemt de schaal grote sprongen in de classificatie, maar wordt er eigenlijk niet aangegeven waarop die classificatie nou precies is gebaseerd. **Tabel 34** is daarom een indicatie, maar zeker niet leidend.

Likelihood Ratio	Kracht van het bewijs in voordeel van hypothese 1
<1	Negatief (in voordeel van hypothese 2)
1 tot 3.2	Zwak
3.2 tot 10	Substantieel
10 tot 31.6	Sterk
31.6 tot 100	Zeer sterk
>100	Sluitend

**Tabel 34.** Classificatie van de Likelihood Ratio (LR) in termen van bewijskracht.

We kunnen trouwens de LR uitrekenen voor verschillende modellen én voor parameters in één model. Omdat het model waaruit een parameter wordt geschat belangrijker is dan een bepaalde parameter is het beter om eerst modellen met elkaar te vergelijken. Daarvoor kunnen we vijf modellen nemen:

1. Een model zonder verklarende factoren.
2. Een model met *Dosering* als verklarende factor.
3. Een model met *Geslacht* als verklarende factor.
4. Een model met *Dosering* en *Geslacht* als verklarende factoren.
5. Een model met *Dosering*, *Geslacht* en de interactie tussen beiden als verklarende factoren.

Dit zou ons een idee kunnen geven van de kracht en het nut van de LR. We houden gemakshalve de prior verdelingen voor elke parameter hetzelfde en eindigen vervolgens met

een matrix (**Tabel 35**). De sterkste LR vinden we voor model 3 (hypothese 1) in vergelijk met model 5 (hypothese 2). Dat betekent dat dosering geen rol speelt<sup>124</sup>.

		Noemer				
		Model 1	Model 2	Model 3	Model 4	Model 5
Deler	Model 1	1	0.142	3.37	0.485	0.114
	Model 2	7.05	1	23.73	3.41	0.805
	Model 3	0.297	0.042	1	0.144	0.034
	Model 4	2.06	0.293	6.95	1	0.236
	Model 5	8.75	1.24	29.47	4.24	1

**Tabel 35.** Vergelijking tussen vijf modellen op basis van de Likelihood Ratio (LR). Model 3 is het model wat het best wordt verklaard gegeven de data en de priors.

Dit betekent ook dat we geen LR kunnen uitrekenen voor de parameter *Dosering*. Dat voorbeeld wil ik toch graag geven dus we gaan aan de slag met model 2: dit is het makkelijkste model om te gebruiken als voorbeeld. De volgende drie hypotheses lijken het meest zinvol om uit te rekenen<sup>125</sup>:

1. De coëfficiënt voor de *Dosering* is exact 0.
2. De coëfficiënt voor de *Dosering* is < 0.
3. De coëfficiënt voor de *Dosering* is > 0.

Hypothese 1	Hypothese 2	Likelihood Ratio
0	0.3	7.02
< 0	0.3	0.33
> 0	0.3	3.01

**Tabel 36.** Bewijs voor de hypothese dat de coëfficiënt voor Dosering exact nul is, kleiner dan nul of groter dan nul. De grootste bewijskracht ligt voor de hypothese dat de coëfficiënt exact 0 is, gegeven de priors en de data.

Het meest sterke bewijs is voor de hypothese dat de coëfficiënt voor *Dosering* nul is. Voor de oplettende lezer is hopelijk duidelijk dat de eerste hypothese toets een tweezijdige toets

<sup>124</sup> We hebben al eerder gezien dat een model als dit, dat geen rekening houdt met de vele tumorgevallen bij de nul-dosering, eigenlijk geen goed model is. Dus deze bevinding is niet sluitend, maar moet eerder worden gezien als uitleg in een voorbeeld.

<sup>125</sup> Feitelijk kunnen we elke hypothese uitrekenen die er is.

was. De laatste twee toetsen zijn beiden eenzijdig. Dit volgt natuurlijk uit de richting die ik aan hypothese 1 heb gegeven (kleiner dan óf groter dan nul).

## Bayesiaanse analyse van het hurdle model

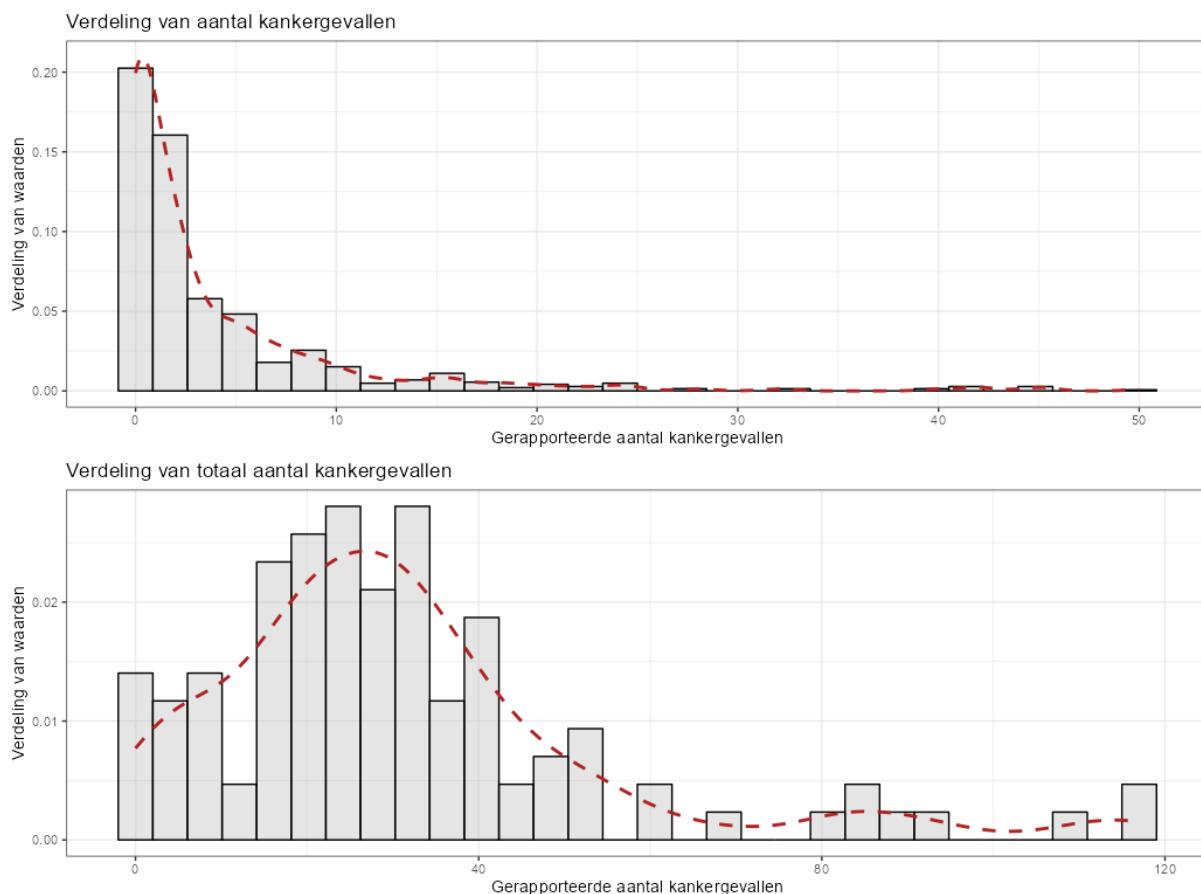
We hebben al een aantal keren gezien dat de verdeling van het totale aantal tumorgevallen een verdeling heeft die wordt gekenmerkt door een groot aantal nullen (**Figuur 123**). Deze verdeling is problematisch en kan ook niet zomaar worden genegeerd. Als we echter kijken naar de verdeling uit **Figuur 139** dan lijkt het wel weer mee te vallen, hoewel **Figuur 140** en **Figuur 141** duidelijk aangeven dat er een verschil zit tussen de verdeling zoals geobserveerd en de verdeling zoals gesimuleerd vanuit de posterior. Dat geeft altijd stof tot nadenken. Laten we daarom, gemakshalve, beginnen met de Bayesiaanse analyse van het hurdle model. Voordat ik hiermee begin is het essentieel om een aantal zaken vast te stellen:

1. Als we een verdeling maken van het aantal kanker gevallen dan krijgen we **Figuur 123**.
2. Als we een verdeling maken van het totaal aantal kanker gevallen per studie, geslacht en dosering én dit afzetten tegen **Figuur 123** dan krijgen we **Figuur 149**. Dan wordt duidelijk dat de scheve verdeling met veel nullen ineens minder scheef is, hoewel er nog steeds aardig wat nullen zijn. Dit verklaar wellicht waarom de hurdle modellen in het verleden het beter deden dan de andere modellen. Toch is het allemaal verre van perfect.
3. We modelleren al een hele tijd op het totaal aantal kankergevallen per studie, geslacht, dosering en verder. Dit blijf ik doen.

Het gemakkelijkste model om te maken is een model zonder verklarende factoren. Dat geeft ons een idee of het toepassen van een dergelijk model met al haar aannames wel past bij de data zoals geobserveerd. Gemakshalve doe ik ook geen enkele uitspraak over de prior verdelingen. Dit alles maakt dat dit Bayesiaanse model heel sterk lijkt op een model afkomstig uit de frequentistische statistiek<sup>126</sup>.

---

<sup>126</sup> In afwezigheid van een gedefinieerde prior wordt vaak automatisch gekozen voor een zeer zwakke prior wat betekent dat we eigenlijk geen enkele aannname doen over de verdeling. De posterior wordt op die manier haast enkel bepaald door



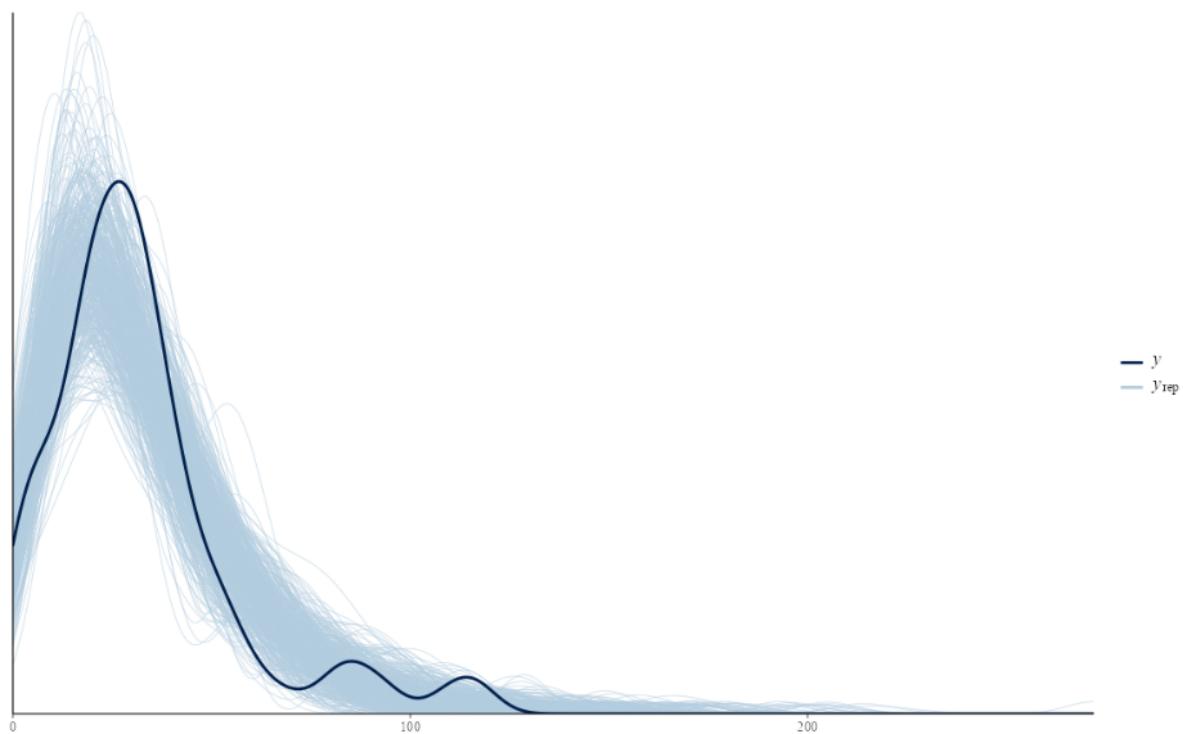
**Figuur 149.** Verdeling van het aantal kanker gevallen (boven, en verdeling van het totaal aantal kanker gevallen per studie, geslacht en dosering (onder).

De uitkomst van het hurdle model zonder verklarende factoren valt te zien in **Figuur 150** en

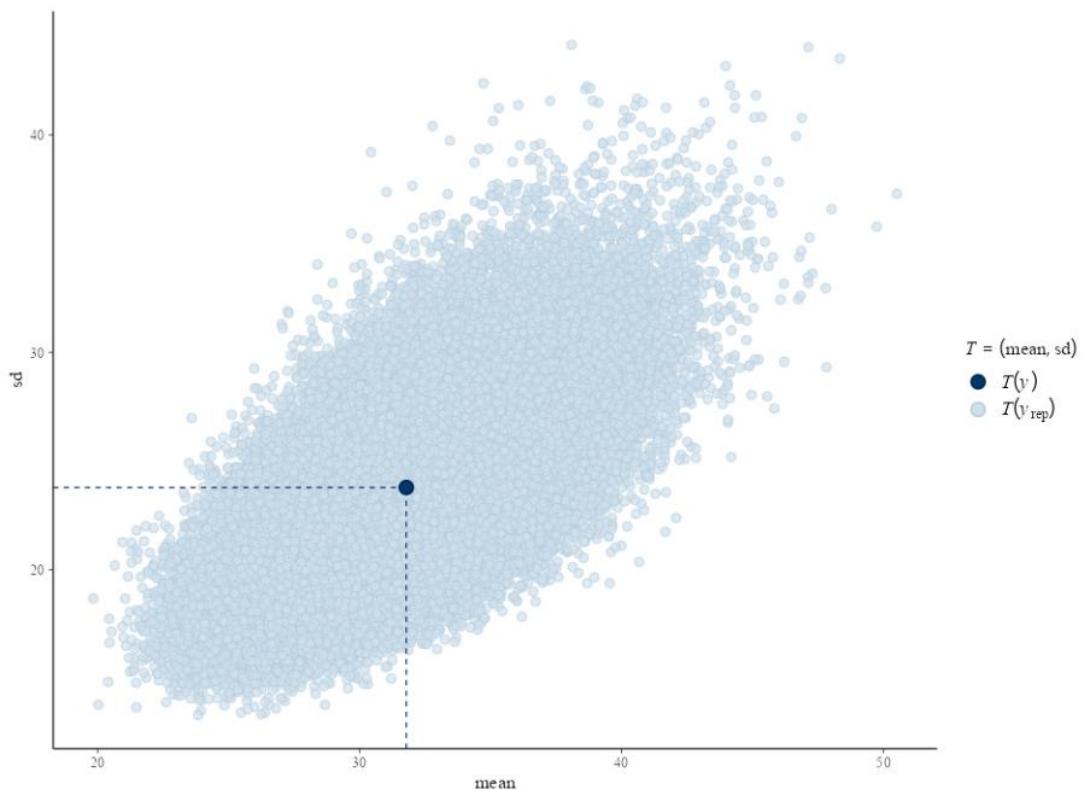
**Figuur 151.** Samengevat zien we het volgende waarbij vooral de *hu\_Intercept* parameter interessant is:

Population-Level Effects:							
	Estimate	Est.Error	I-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	3.48	0.07	3.34	3.62	1.00	140494	107126
hu_Intercept	-3.73	0.64	-5.16	-2.65	1.00	130879	84259
Family Specific Parameters:							
	Estimate	Est.Error	I-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
shape	2.02	0.31	1.47	2.66	1.00	134217	105957

de geobserveerde gegevens. Dit is dezelfde procedure als die wordt toegepast in de frequentistische statistiek waarin alleen datgene telt wat geobserveerd kan worden.



**Figuur 150.** De verdeling van de posterior voor een leeg hurdle model (lichte lijnen), en de verdeling van de geobserveerde tumorgevallen (donkere lijn).



**Figuur 151.** De verdeling van de posterior voor een leeg hurdle model (lichte bollen), en de verdeling van de geobserveerde tumorgevallen (donkere bol).

Een hurdle model modelleert eigenlijk twee onderdelen: de kans op nul en de kans op  $\neq 0$ .

De  $hu$  parameter is de kans op 0 en die kunnen we dus afleiden uit het model: -3.73.

Omgerekend in kansen wordt dit 0.02 of 2%. Dit is in lijn met een simpele berekening die laat zien hoe vaak we nul-incidentie zien. Dat is namelijk 2 van de 106 tellingen en daarmee, ongeveer, 2%. Als we het zo bekijken lijkt het alsof de nul helemaal geen probleem is.

Wellicht lastiger te modelleren zijn de twee hobbels vooraf de 100 en vlak daarna (**Figuur 150**). Toch blijkt het hurdle model elke keer beter te passen wanneer vergeleken met een standaard model (**Figuur 139**) met meer verklarende factoren. Wat we kunnen doen is dit model uitbreiden met parameters en dan zien of onze waardering voor dit type model blijft staan. Deze keer worden het 7 verschillende modellen:

1. Een model zonder verklarende factoren.
2. Een model met *Dosering* als verklarende factor.
3. Een model met *Geslacht* als verklarende factor.
4. Een model met *Dosering* en *Geslacht* als verklarende factoren.
5. Een model met *Dosering*, *Geslacht* en de interactie tussen beiden als verklarende factoren.
6. Een model met *Dosering*, *Geslacht* en *Soort* als verklarende factoren.
7. Een model met *Dosering*, *Geslacht*, de interactie tussen beiden en *Soort* als verklarende factoren.

		Noemer						
		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Deler	Model 1	1	0.046	28.42	1.25	0.090	46.45	3.36
	Model 2	21.68	1	616.15	27.10	1.95	1010	72.95
	Model 3	0.035	0.002	1	0.044	0.003	1.63	0.118
	Model 4	0.800	0.037	22.74	1	0.072	37.16	2.69
	Model 5	11.13	0.513	316.15	13.9	1	516.77	37.43
	Model 6	0.022	0.0009	0.612	0.027	0.002	1	0.072
	Model 7	0.297	0.014	8.45	0.372	0.027	13.81	1

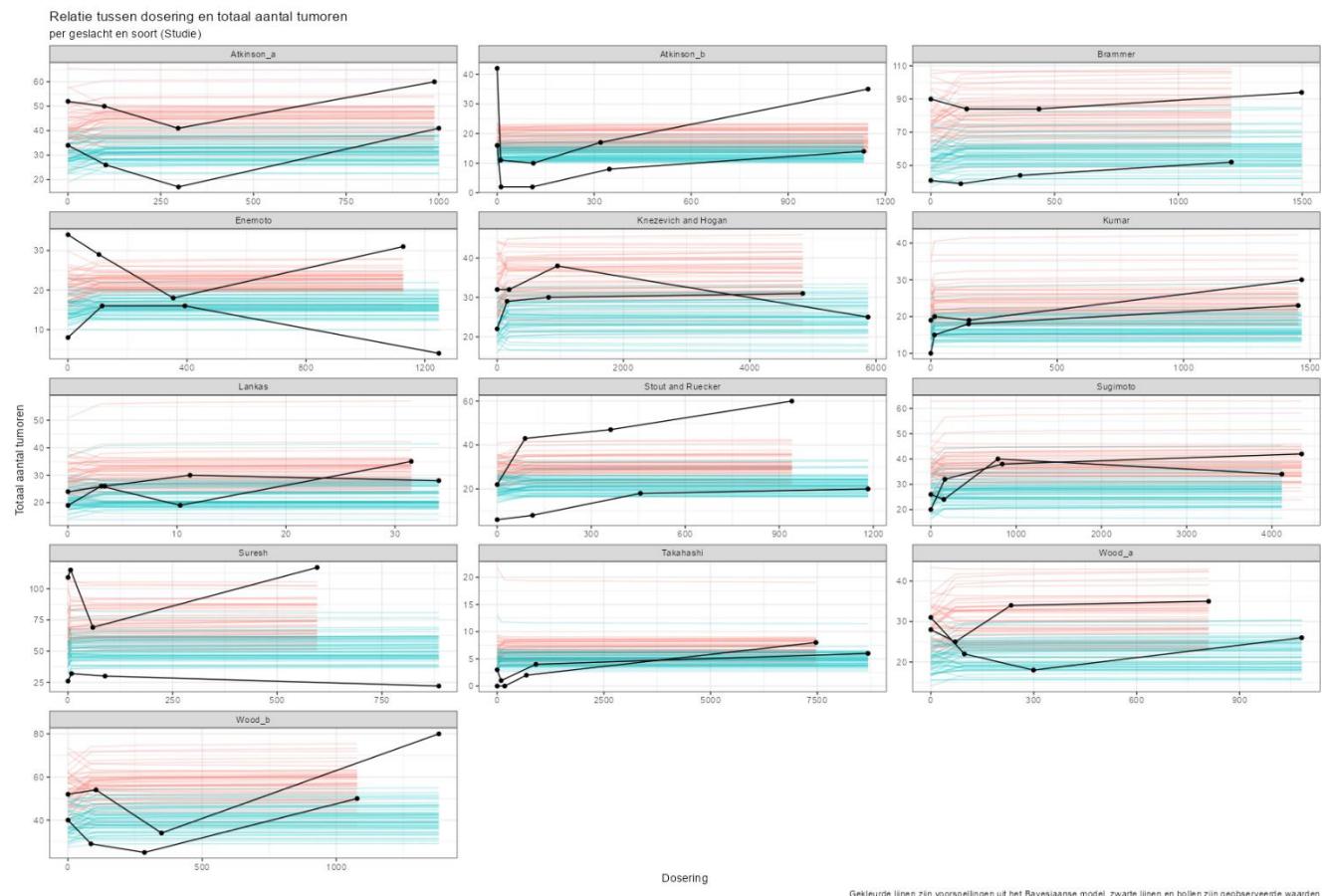
**Tabel 37.** De Likelihood Ratio (LR) matrix voor 7 modellen. Voor Model 6 is er het meeste bewijs, hoewel Model 6 en Model 3 haast evenveel bewijs krijgen toegekend. Dat betekent dat op basis van de data het maar lastig blijft wat de rol van de dosering is.

De resultaten uit **Tabel 37** laten zien dat model 6 het model is met het sterkste bewijs ten op zichtte van de andere modellen, met uitzondering van model 3. En als we model 3 afzetten tegen alle andere modellen dan is de bewijslast voor model 3 het sterkst, behalve wanneer vergeleken met model 6. Het is daarom het meest raadzaam om model 6 te kiezen als model waarvoor de bewijslast het sterkst is en op basis van dit model de LR uit te rekenen voor dosering (algemeen). Het resultaat zien we in **Tabel 38** wat erop neerkomt dat er geen bewijs is om aan te nemen dat Dosering een rol speelt.

Geslacht	Hypothese 1	Hypothese 2	Likelihood Ratio
Beiden	0	0.1	N/A
Beiden	< 0	0.1	0.39
Beiden	> 0	0.1	2.6

**Tabel 38.** Bewijs voor de hypothese dat de coëfficiënt voor Dosering exact nul is, kleiner dan nul of groter dan nul. De grootste bewijskracht ligt voor de hypothese dat de coëfficiënt groter dan 0 is, maar het bewijs is zwak.

Als we het model tot ons nemen en de voorspellingen afzetten tegen de geobserveerde waarden dan zien we **Figuur 152**. Ik heb gemakshalve de doseringswaarden op de originele schaal gelaten wat helpt in het interpreteren van de data. Wat direct opvalt is dat het model moeite heeft met het modelleren van de onderlinge relatie en dat komt omdat de relaties tussen dosering en tumorgevallen zich niet makkelijk laat vangen door één enkele functie. Wat verder ook opvalt is dat de relatie helemaal niet zo strak is wat de zwarte lijnen doen vermoeden. Het lijkt er haast op alsof het aantal kankergevallen afneemt bij het verhogen van de dosis en dan toeneemt bij grotere hoeveelheden van de dosis.



**Figuur 152.** Grafiek die relatie laat zien tussen dosering en aantal tumorgevallen, per geslacht en studie, op de originele schaal. De geobserveerde data (zwarte lijnen) lijken meer zeker dan de modelvoorspellingen aangeven.

## Modelleren van de verandering: vóóraf vs. áchteraf

Het modelleerwerk met de hurdle modellen laat zien hoe lastig het is om een dose-response model te maken. Eigenlijk hebben we dit door het gehele rapport al gezien: er valt maar lastig een functie te bouwen. Wat dan rest, is om de data binair te maken en te spreken over de controlegroep én de behandelgroep. Eigenlijk kunnen we dan ook spreken over een vóóraf en een áchteraf.

Deze analyses kunnen op verschillende manieren verwerkt worden. Zo kunnen we, zoals al genoemd, de dosering opsplitsen in twee groepen en dan het verschil uitrekenen. Belangrijk is dat we niet vergeten dat elke groep zijn eigen controle is, maar omdat alle controle groepen bij dosering nul beginnen hoef ik niet te corrigeren voor de startwaarde<sup>127</sup>.

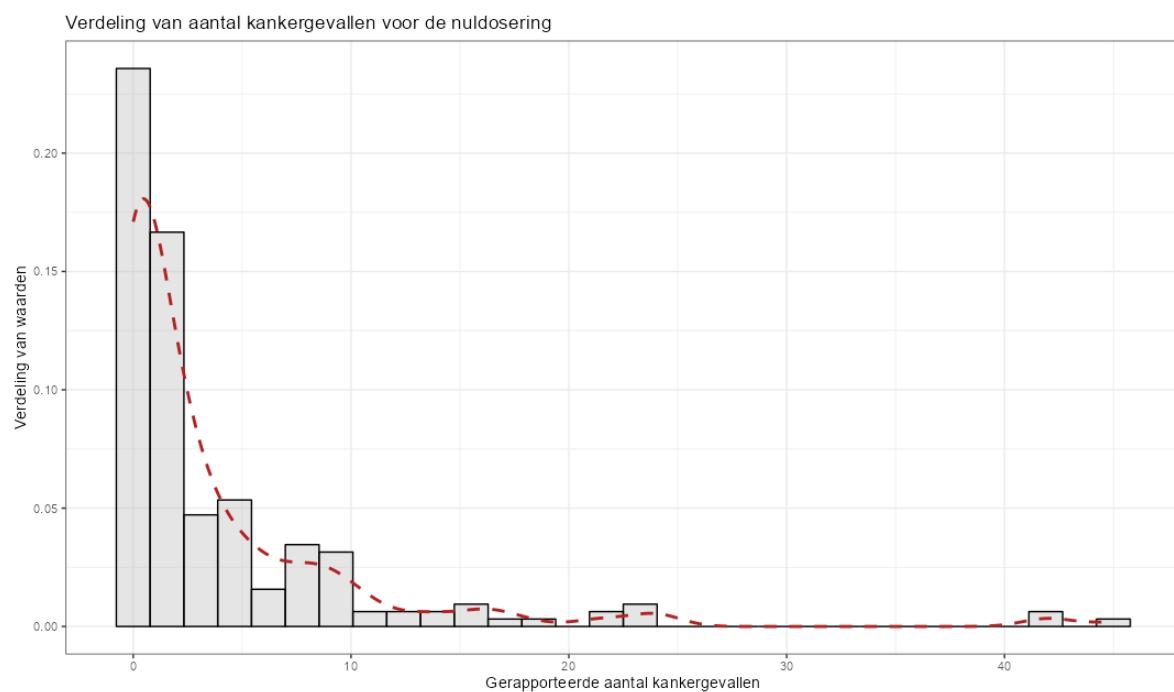
<sup>127</sup> Als ik nou een model had gemaakt dat rekening had gehouden met het type dieet om gewichtsverlies te modelleren is het wel belangrijk om het startgewicht mee te nemen: het startgewicht heeft namelijk vaak invloed op de mate van afvallen.

Wel moet ik rekening houden dat metingen gedaan worden in dezelfde omgeving. Om een dergelijke analyse uit te voeren ga ik de volgende stappen nemen:

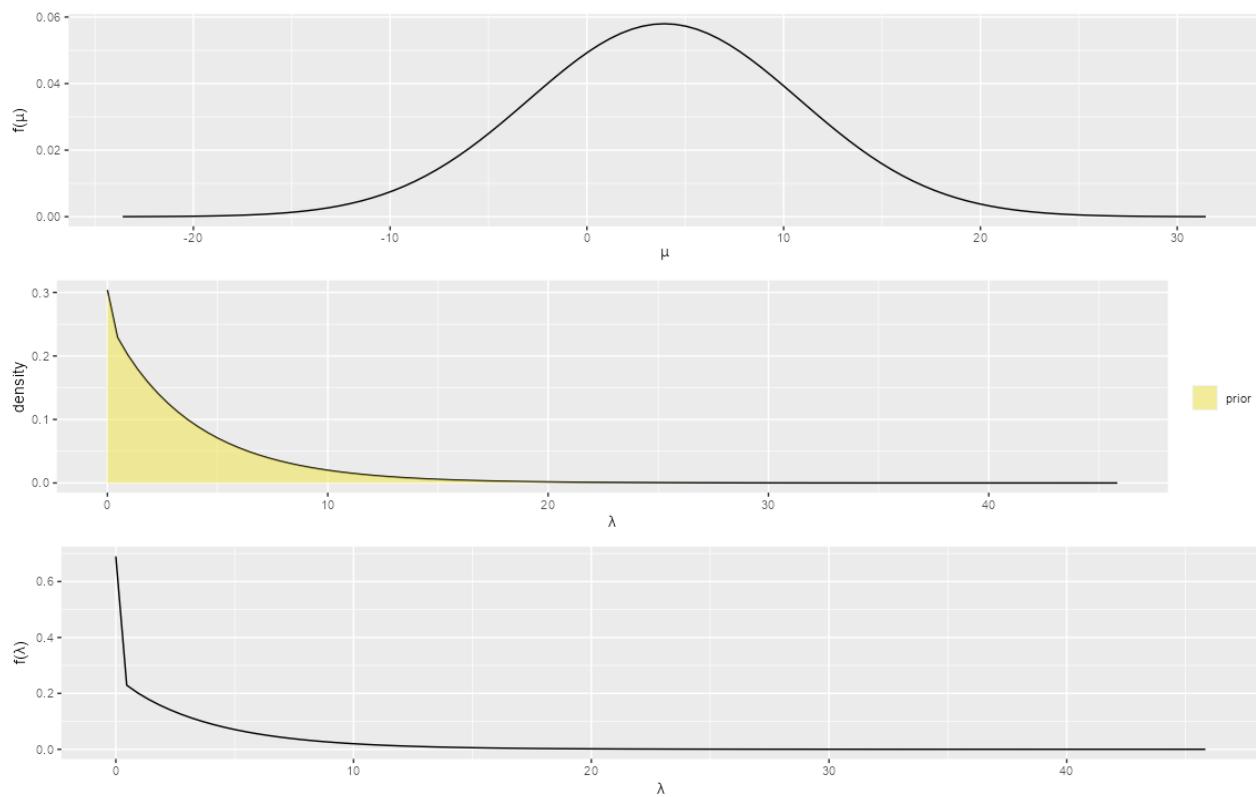
1. Ik splits de data in twee groepen: controle én behandeling. Dit zijn de doseringen vóór af (de nul-meting) en áchteraf.
2. Het aantal tumorgevallen tel ik bij elkaar op en deel ik door het aantal dieren. Ik werk dus met een binomiale verdeling. De aantallen zullen altijd op zijn minst per studie worden berekend.
3. De specifieke ratio (tumoraantallen / aantal dieren) die ik zal gebruiken hangt af van het specifieke model dat ik wil maken. Als ik alleen wil kijken naar het verschil tussen de nul-dosering en het verschil tussen de behandeling, dan tel ik alle cellen zoals weergegeven in *Supplementary Material 2* van Pointer. Dat betekent dat als er twee geslachten zijn, vijf tumorsoorten per geslacht met vier doseringen (nul-dosering en 3 glyfosaatdoseringen) ik voor de nul-dosering 10 groepen optel en voor de behandeling 30 groepen. Elke doseringsgroep is uniek, maar de tumoraantallen zijn per dezelfde groep dieren. Daarmee heb ik per studie één totaal aantal kankergevallen per nul-dosering én één totaal aantal kankergevallen per behandeling per studie.
4. Als ik wil kijken naar de invloed van behandeling én geslacht, dan deel ik het per studie, groep en dosering op. Wil ik soort toevoegen dan deel ik nog verder op. Op die manier kan ik uitrekenen wat de invloeden zijn per variabele.
5. Ik reken de hyperparameters uit voor dosering, geslacht én soort op basis van de priors (de nulmeting) en de data. Dan krijg ik de posterior verdeling.
6. Ik kan vanuit de prior, data en posterior de Likelihood Ratio (LR) voor de modellen uitrekenen.
7. Vanuit het model met de hoogste LR reken ik uit wat de mate van bewijs is voor de coëfficiënt van de dosering. Dit is hier het verschil tussen controle en behandeling.
8. Ik voer analyses uit met een zwakke prior én met een sterke prior.

Laten we, voordat we beginnen, even kijken naar de verdelingen. Een verdeling van het aantal tumorgevallen voor de nul-dosering identificeren zien we in **Figuur 153**. Deze doet sterk denken aan **Figuur 149** waar alle doseringen worden meegenomen. Het gemiddelde van de verdeling voor de nul-dosering is 3.93 en de standaard deviatie is 6.88, maar het is

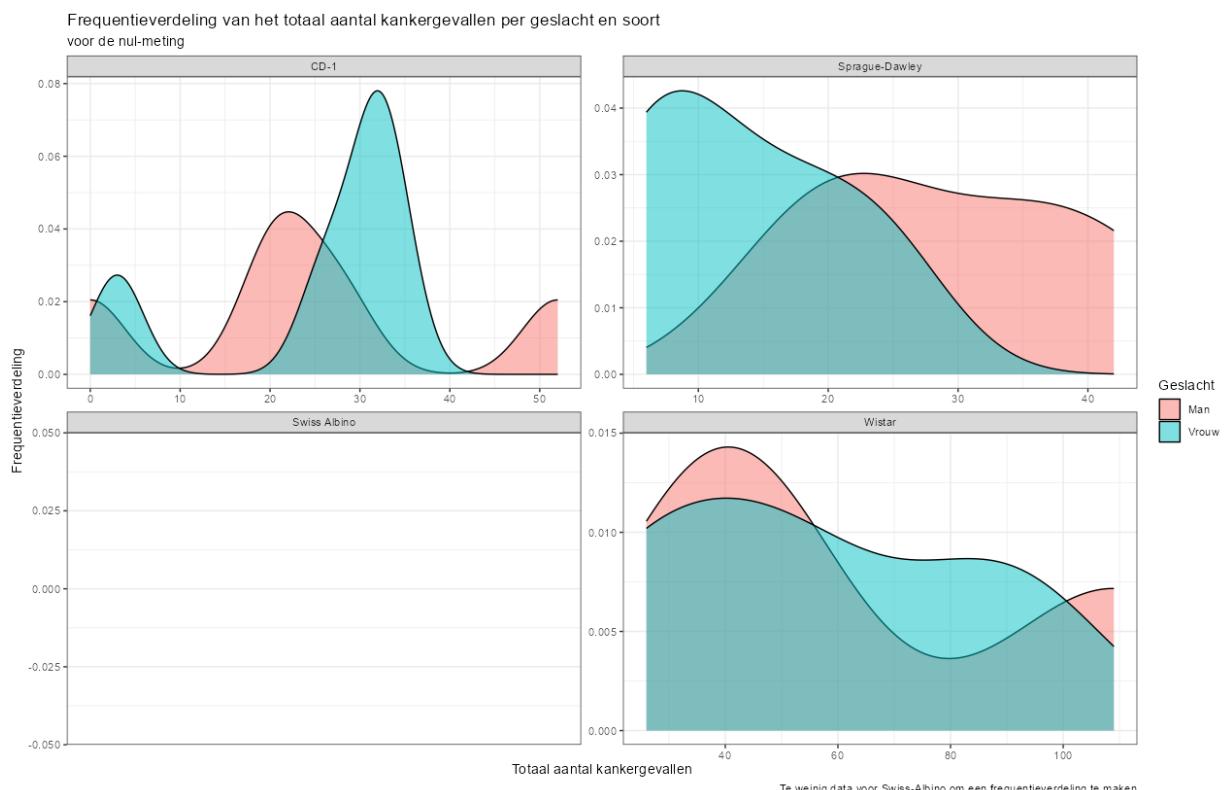
overduidelijk dat dit geen normaalverdeling is. De verdeling is behoorlijk scheef. Dat kan ik laten zien door drie verschillende delingen te nemen en de data te simuleren op basis van de geobserveerde gegevens en de eigenschappen van die delingen (**Figuur 154**). De laatste twee delingen lijken sterker op de geobserveerde data en lijken daarmee zinvolle priors voor het aantal tumorgevallen vooraf. Wel is het zo dan ik dan alle tumorgevallen meeneem, zonder een splitsing te maken per geslacht of soort. Doe ik dat wel, dan zijn de frequentieverdelingen niet meer heel soepel (**Figuur 155**). Ik moet dus goed oppassen hoe ver ik de data splits om een zinvolle analyse te maken.



**Figuur 153.** Verdeling van het aantal kankergevallen bij de nul-dosering.



**Figuur 154.** Drie verdelingen op basis van de geobserveerde data. Van boven naar beneden: de normaalverdeling, de gamma verdeling en de gamma-poisson verdeling. De gamma-poisson heet ook wel de negatieve binomiale verdeling en die hebben we al veelvuldig voorbij zien komen.

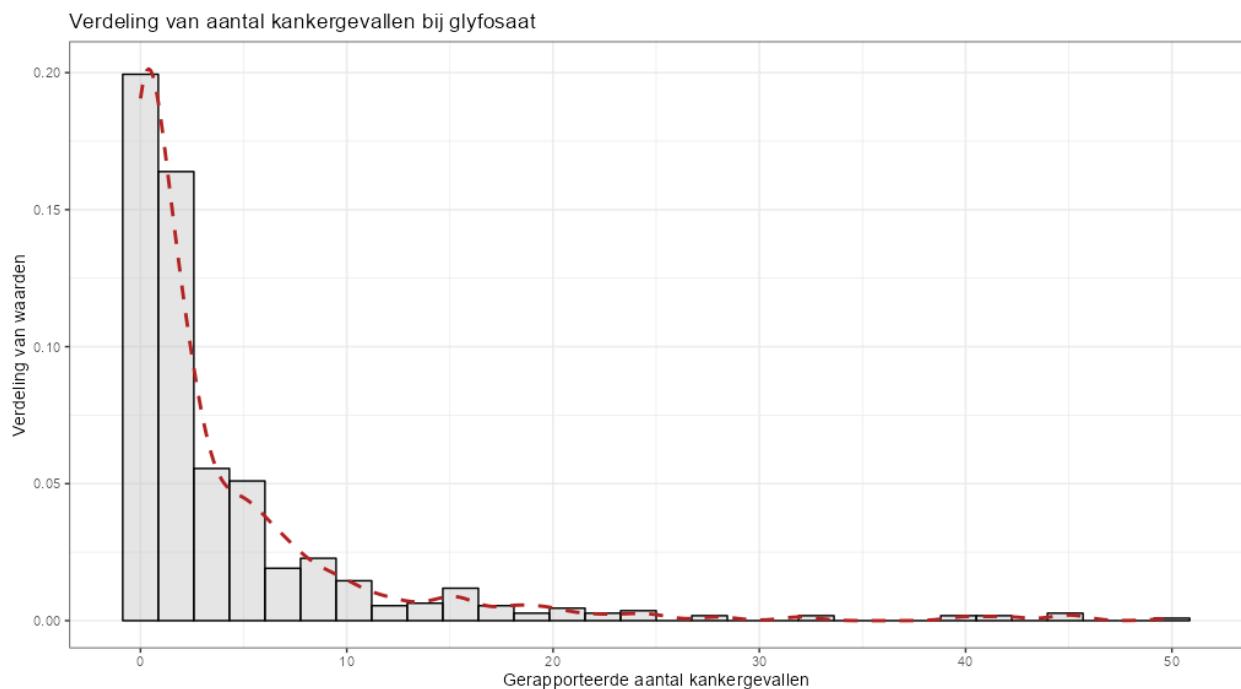


**Figuur 155.** Verdeling van het aantal kankergevallen per geslacht en soort voor de nul-verdeling (controlegroep).

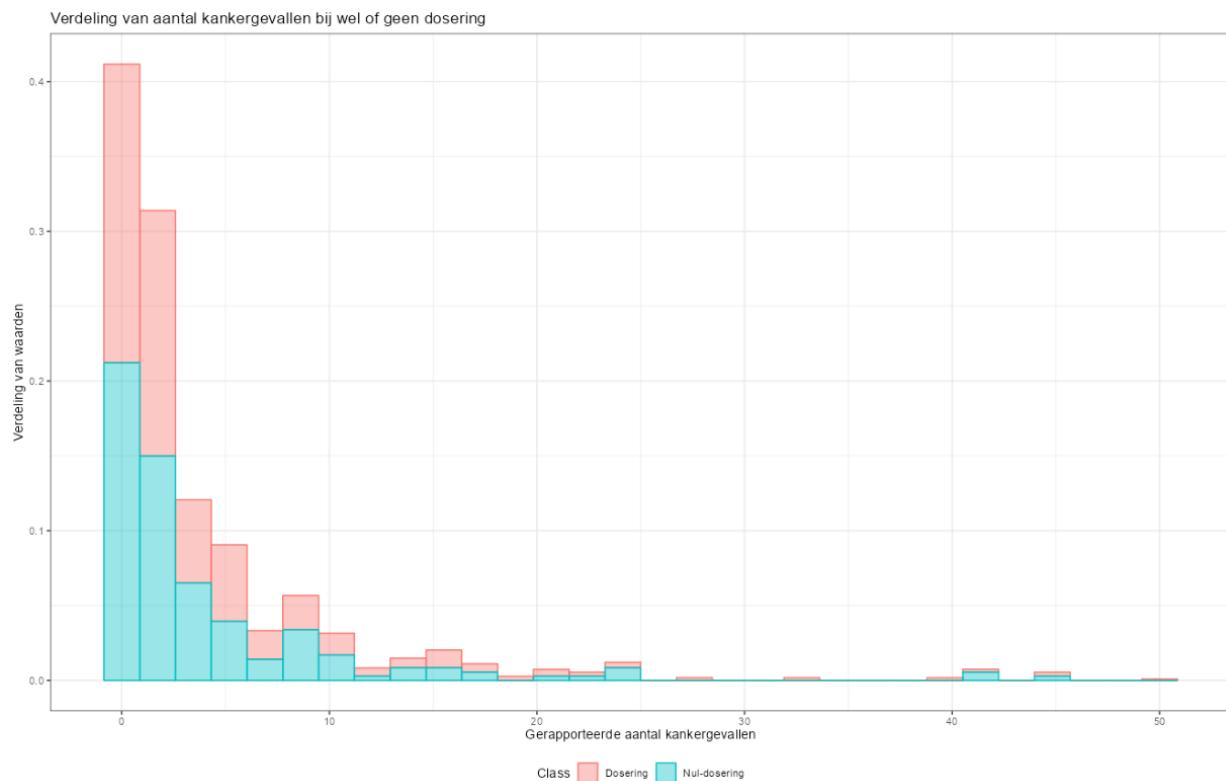
We kunnen deze exercitie herhalen voor het aantal tumorgevallen na het toedienen van glyfosaat (de behandelgroep). Laten we gemakshalve de data tonen over alle doseringen heen. We zien eigenlijk eenzelfde figuur (**Figuur 156**) als in **Figuur 153**. Een snelle berekening geeft een gemiddelde van 4.02 en een standaard deviatie van 6.88. Dat is haast hetzelfde als bij de nul-verdeling, maar we hebben wel met veel meer tellingen te maken (2563).

Vergelijken we **Figuur 156** met **Figuur 153** dan zien we **Figuur 157**. Dat lijkt nagenoeg op hetzelfde. Wanneer we uitgaan van de ratio van het totaal aantal kankergevallen per studie dan krijgen we **Figuur 158**. Deze benadering is te hoog-over en daarmee niet bruikbaar. Gemakshalve kunnen we een figuur maken zoals **Figuur 155** waarin ik laat zien wat de verdeling tumorgevallen is per geslacht en soort (**Figuur 159**).

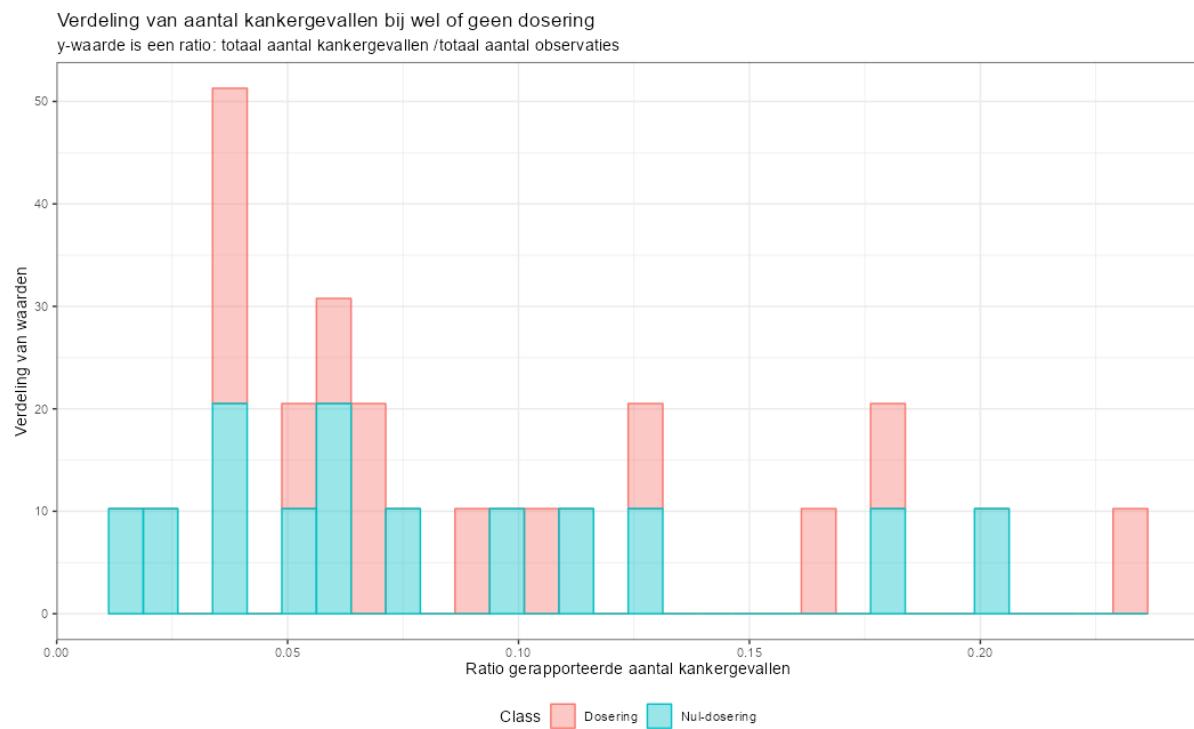
Nu heb ik al eerder geschreven dat ik deze keer met ratio's wil werken en niet per se met het aantal kankergevallen. De figuren tot nu deden dat vaak wel, maar dat is niet gepast voor de analyse die ik voor ogen heb. De reden dat ik deze figuren nogmaals wilde maken is om te laten zien hoe lastig het werken is met aantallen (de noemer) als je niet de deler meeneemt. Ook wil ik laten zien hoe breekbaar modelleren is.



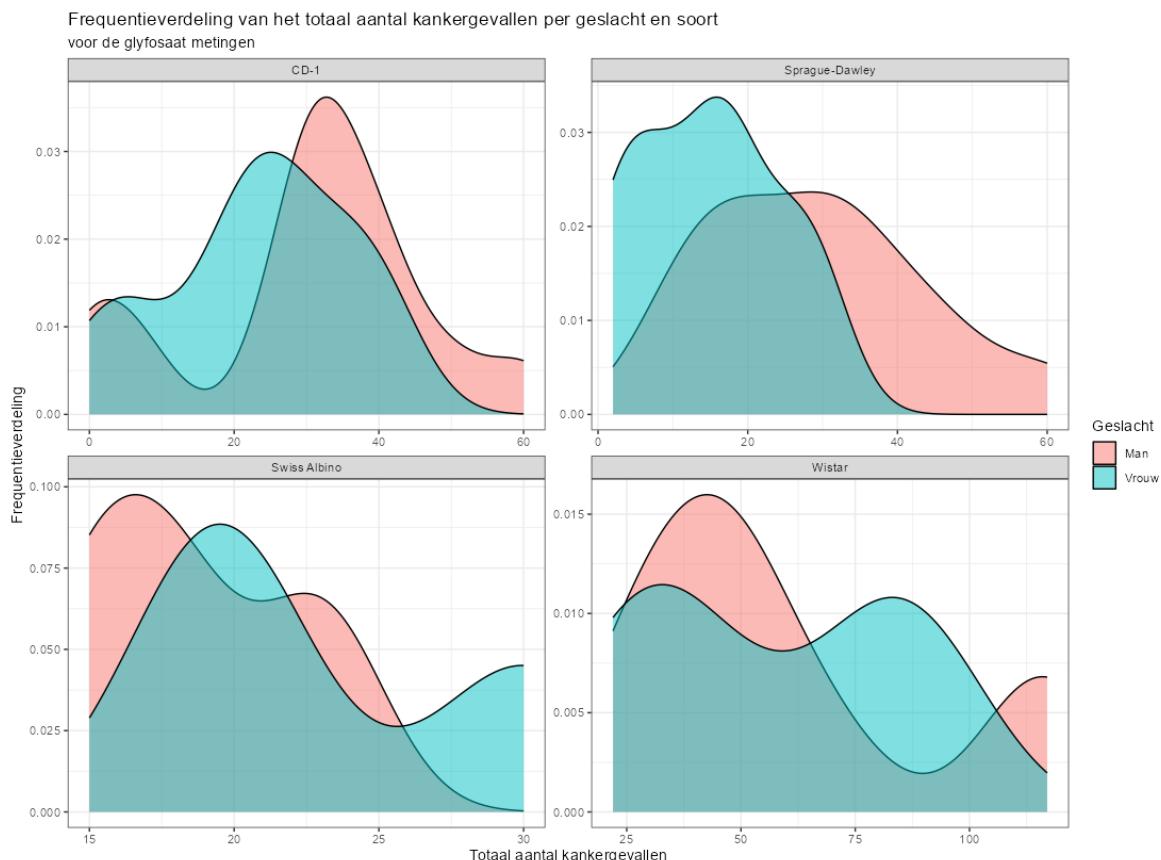
**Figuur 156.** Verdeling van aantal kankergevallen over alle doseringen heen.



**Figuur 157.** Verdeling van het aantal kankergevallen per nul-dosering en dosering waarbij dosering is samengevoegd tot één groep.



**Figuur 158.** Verdeling van het aantal kankergevallen bij wel of geen dosering waarbij de y-waarde (hier de x-as) de ratio is van het totaal aantal kankergevallen gedeeld door het totaal aantal observaties.



**Figuur 159.** Verdeling van het aantal kankergevallen per geslacht en soort over de glyfosaatverdelingen heen.

## De binomiaalverdeling

De binomiaalverdeling loont zich vaak uitstekend voor dit soort analyses, maar was bij eerder gebruik niet handig voor het analyseren van een dose-response relatie. Ik heb het vermoeden dat het deze keer beter zal werken en om te beginnen zal ik een binomiaal model maken zonder dat ik variabelen opneem (model 1) én een binomiaal model waarin ik wel een verschil maak tussen groepen (model 2). De data die ik gebruik hiervoor zijn als volgt:

Studie <chr>	Group <chr>	Cases <dbl>	N <int>	N_mean <dbl>	ratio_sum <dbl>	ratio_mean <dbl>
1 Atkinson_a	Control	86	750	50	0.115	1.72
2 Atkinson_a	Treatment	235	2247	49.9	0.105	4.71
3 Atkinson_b	Control	58	1094	49.7	0.0530	1.17
4 Atkinson_b	Treatment	99	2927	33.3	0.0338	2.98
5 Brammer	Control	131	730	52.1	0.179	2.51
6 Brammer	Treatment	397	2190	52.1	0.181	7.61
7 Enemoto	Control	42	1099	50.0	0.0382	0.841

8 Enemoto	Treatment 114	3282	49.7	0.0347	2.29
9 Knezevich and Hogan	Control 54	868	48.2	0.0622	1.12
10 Knezevich and Hogan	Treatment 185	2656	49.2	0.0697	3.76
11 Kumar	Control 29	300	50	0.0967	0.58
12 Kumar	Treatment 125	756	42	0.165	2.98
13 Lankas	Control 43	1071	48.7	0.0401	0.883
14 Lankas	Treatment 164	3257	49.3	0.0504	3.32
15 Stout and Ruecker	Control 28	1092	49.6	0.0256	0.564
16 Stout and Ruecker	Treatment 196	3256	49.3	0.0602	3.97
17 Sugimoto	Control 46	800	50	0.0575	0.92
18 Sugimoto	Treatment 210	2400	50	0.0875	4.2
19 Suresh	Control 135	664	47.4	0.203	2.85
20 Suresh	Treatment 385	1656	39.4	0.232	9.76
21 Takahashi	Control 3	200	50	0.015	0.06
22 Takahashi	Treatment 21	600	50	0.035	0.42
23 Wood_a	Control 59	816	51	0.0723	1.16
24 Wood_a	Treatment 160	2448	51	0.0654	3.14
25 Wood_b	Control 92	712	50.9	0.129	1.81
26 Wood_b	Treatment 272	2142	51	0.127	5.33

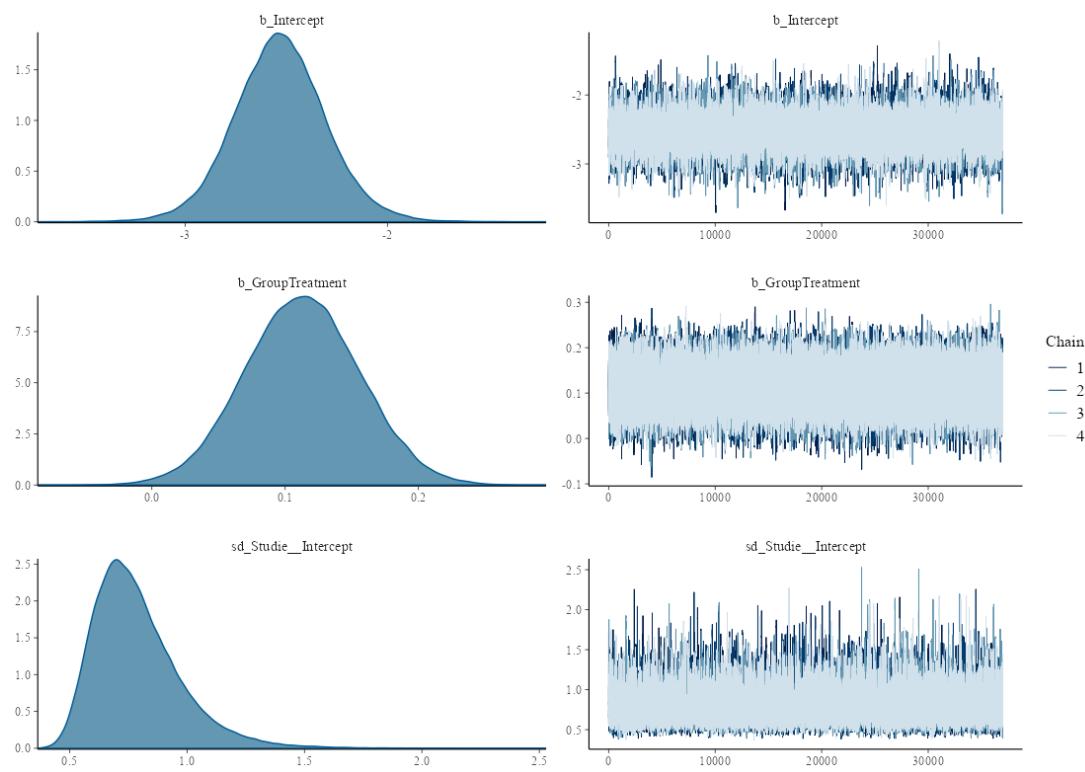
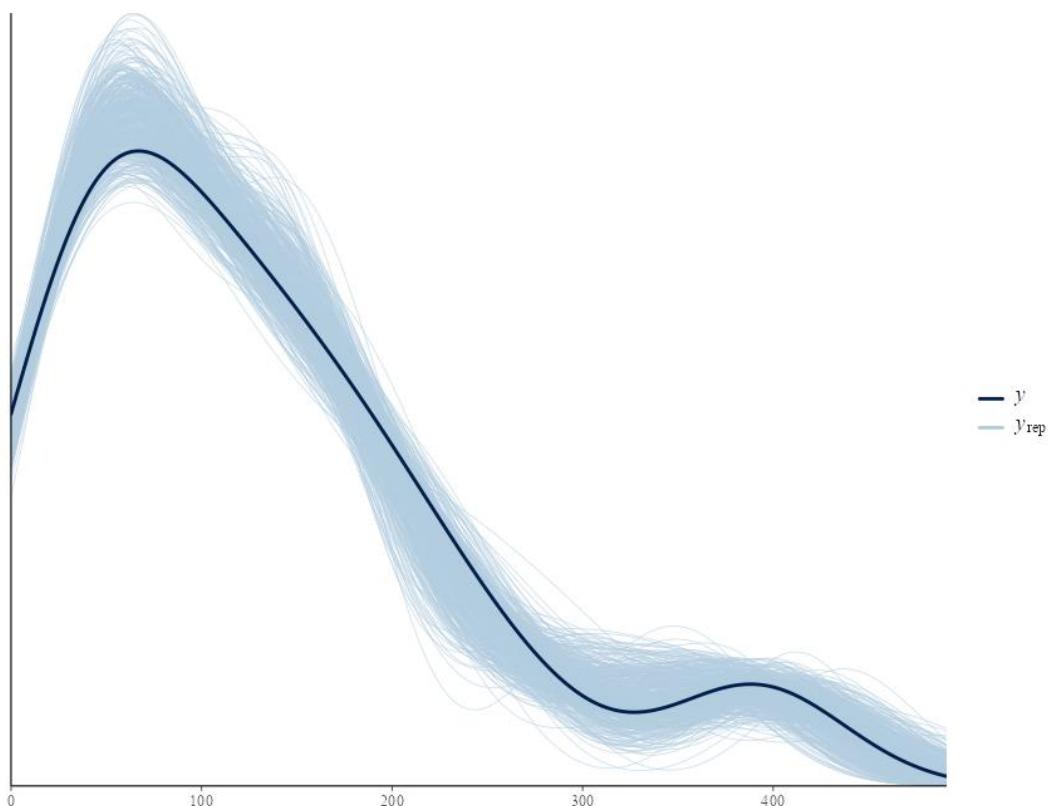
De resultaten van beide modellen staan in **Tabel 39** en tonen dat model 2 een grotere LR heeft gegeven de data en de priors. Nu is een LR van ongeveer 4 volgens **Tabel 34** ‘substantieel’, maar het is ook niet overweldigend. Toch geeft het een indicatie om naar de behandeling te kijken.

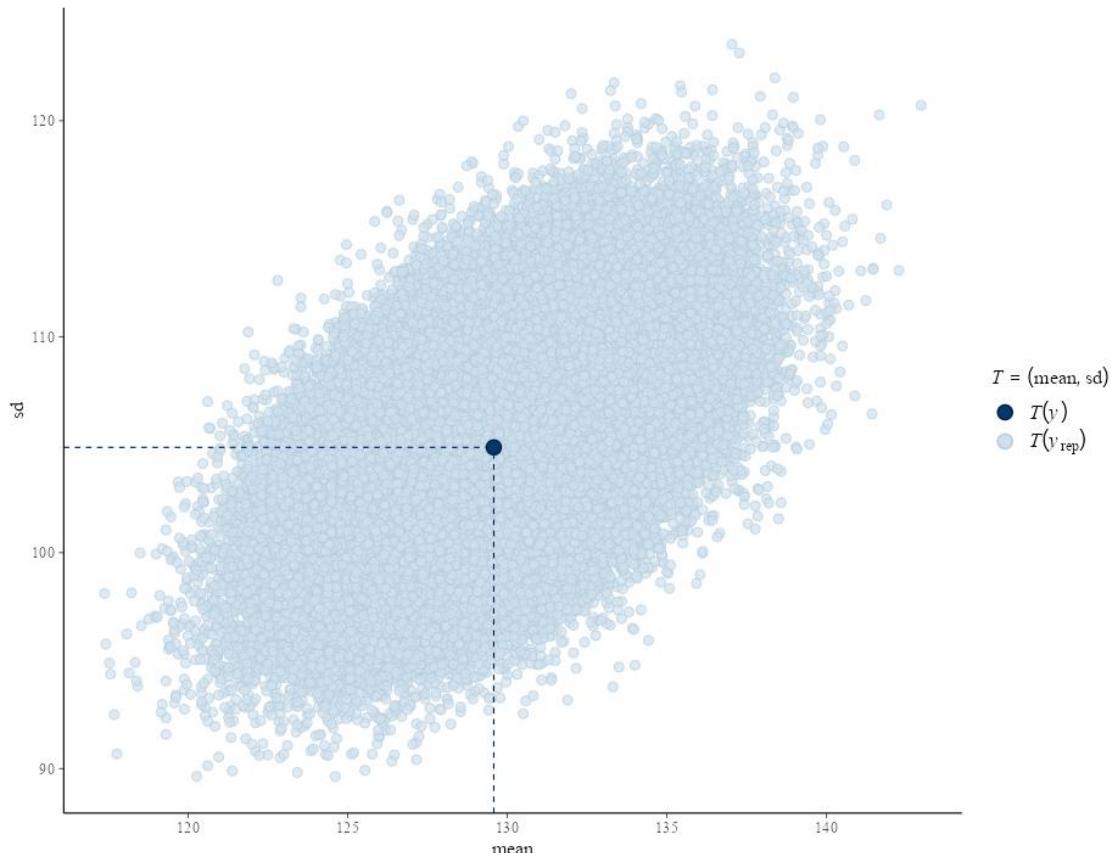
		Noemer	
		Model 1	Model 2
Deler	Model 1	1	0.286
	Model 2	3.5	1

**Tabel 39.** De LR voor model 1 en model 2.

Wat ook belangrijk is om te bezien is hoe betrouwbaar de schattingen van de parameters zijn. Dat zien we in **Figuur 160** en het ziet er naar uit dat de schattingen betrouwbaar zijn. Ook zien we in **Figuur 161** en **Figuur 162** dat de trekkingen uit de posterior verdeling een stuk meer overeenkomt dan de geobserveerde verdeling<sup>128</sup>.

<sup>128</sup> Nogmaals, dit hoeft niet exact hetzelfde te zijn, verre van, maar een grote afwijking behoeft ook een goede uitleg.

**Figuur 160.** Betrouwbaarheid van elke hyperparameter in model 2.**Figuur 161.** Posterior verdeling ten opzichte van geobserveerde verdeling (donkere lijn).



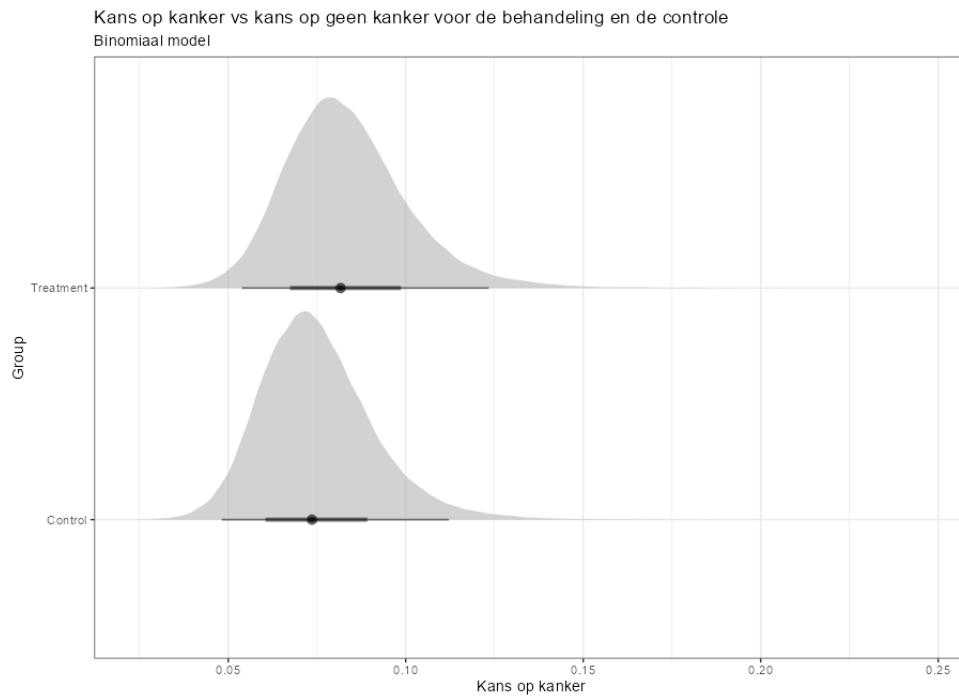
**Figuur 162.** Eenzelfde figuur als **Figuur 161** maar dan platgeslagen in twee variabelen (gemiddelde en standaard deviatie).

Nu we weten dat model 2 een hogere LR heeft dan model 1 én de posterior trekkingen betrouwbaar zijn, kunnen we kijken naar de invloed van behandeling op de onderliggende kans om kanker te krijgen. Dit is de kans van de controlegroep oftewel de nulmeting.

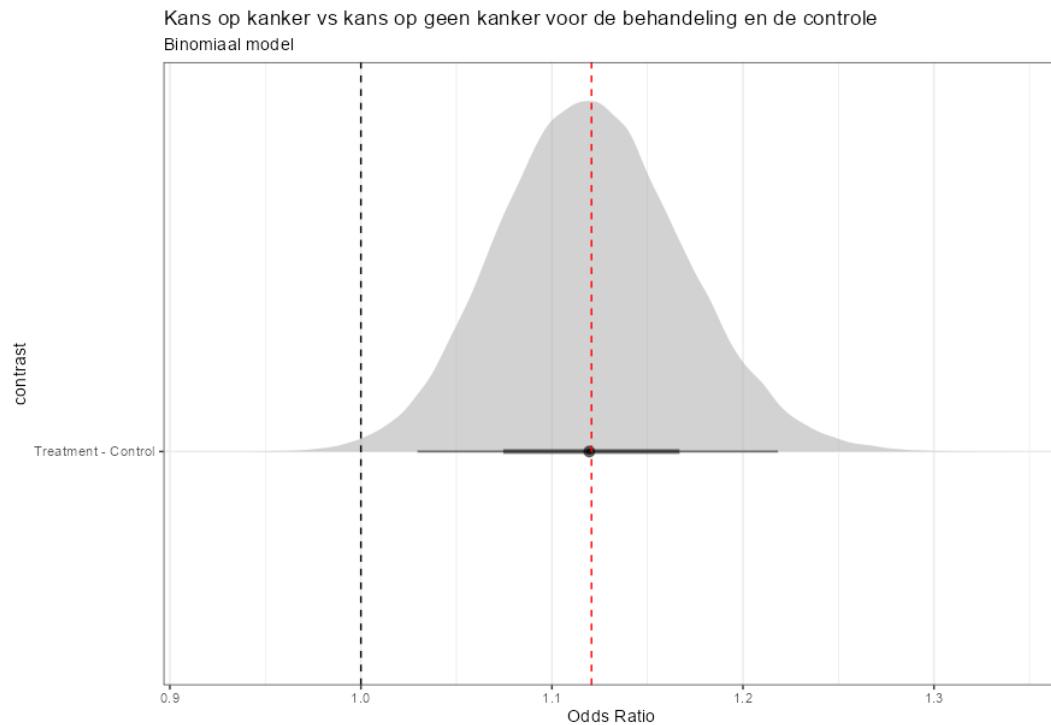
We zien die resultaten in **Figuur 163** en **Figuur 164** beneden. Beiden zijn afgebeeld in frequentieverdelingen en tonen de kans op kanker voor de controle en de behandelgroep. In **Figuur 163** zien we dat de kans groter is voor de behandelgroep dan voor de controlegroep, maar voor een accurate berekening is **Figuur 164** de juiste grafiek. Deze geeft de odds-ratio weer wat de ratio is tussen de kans op kanker versus de kans op geen kanker voor de behandelgroep én de controlegroep<sup>129</sup>. In dit geval is de odds-ratio ongeveer 1.12 wat heel klein is. De kansen om kanker te krijgen in de controlegroep en de behandelgroep is 0.073

<sup>129</sup> Als de kans op kanker in de behandelgroep 3/10 en in de controlegroep 2/10 dan is de odds ratio  $(3/10) / (7/10)$  gedeeld door  $(2/10) / (8/10) = 0.428 / 0.25 = 1.712$

en 0.0816, respectievelijk. Dat maakt de odds-ratio van 1.12. In **Tabel 40** zien we dat de hypothese dat de coëfficiënt voor behandeling groter is dan nul het beste past bij de data.



**Figuur 163.** Kans op kanker voor de controlegroep en de behandel (Treatment) groep.



**Figuur 164.** De odds-ratio voor het verschil tussen de controlegroep en de behandelgroep.

Effect	Hypothese 1	Hypothese 2	Likelihood Ratio
Behandeling	0	0.11	N/A
Behandeling	< 0	0.11	0
Behandeling	> 0	0.11	230.97

**Tabel 40.** De LR voor de hypothese dat de coëfficiënt voor de behandeling nul is, kleiner dan nul of groter dan nul.

De grote LR voor de hypothese dat de coëfficiënt voor behandeling groter is dan nul maakt dat het zinvol is om meer variabelen mee te nemen. We kunnen de data dus nog verder opsplitsen en de data zoals gebruikt zie je hieronder.

Studie <chr>	Group <chr>	Soort <chr>	Geslacht <chr>	Cases <dbl>	N <int>	N_mean <dbl>	ratio_sum <dbl>	ratio_mean <dbl>
1 Atkinson_a	Control	CD-1	Man	52	400	50	0.13	1.04
2 Atkinson_a	Control	CD-1	Vrouw	34	350	50	0.0971	0.68
3 Atkinson_a	Treatment	CD-1	Man	151	1200	50	0.126	3.02
4 Atkinson_a	Treatment	CD-1	Vrouw	84	1047	49.9	0.0802	1.68
5 Atkinson_b	Control	Sprague-	Man	42	800	50	0.0525	0.84
6 Atkinson_b	Control	Sprague-	Vrouw	16	294	49	0.0544	0.327
7 Atkinson_b	Treatment	Sprague-	Man	73	2123	33.2	0.0344	2.20
8 Atkinson_b	Treatment	Sprague-	Vrouw	26	804	33.5	0.0323	0.776
9 Brammer	Control	Wistar	Man	41	424	53	0.0967	0.774
10 Brammer	Control	Wistar	Vrouw	90	306	51	0.294	1.76
11 Brammer	Treatment	Wistar	Man	135	1254	52.2	0.108	2.58
12 Brammer	Treatment	Wistar	Vrouw	262	936	52	0.280	5.04
13 Enemoto	Control	Sprague-	Man	34	799	49.9	0.0426	0.681
14 Enemoto	Control	Sprague-	Vrouw	8	300	50	0.0267	0.16
15 Enemoto	Treatment	Sprague-	Man	78	2388	49.8	0.0327	1.57
16 Enemoto	Treatment	Sprague-	Vrouw	36	894	49.7	0.0403	0.725
17 Knezevich and	Control	CD-1	Man	22	438	48.7	0.0502	0.452
18 Knezevich and	Control	CD-1	Vrouw	32	430	47.8	0.0744	0.670
19 Knezevich and	Treatment	CD-1	Man	90	1344	49.8	0.0670	1.81
20 Knezevich and	Treatment	CD-1	Vrouw	95	1312	48.6	0.0724	1.96
21 Kumar	Control	Swiss	Man	10	200	50	0.05	0.2
22 Kumar	Control	Swiss	Vrouw	19	100	50	0.19	0.38
23 Kumar	Treatment	Swiss	Man	56	456	38	0.123	1.47
24 Kumar	Treatment	Swiss	Vrouw	69	300	50	0.23	1.38
25 Lankas	Control	Sprague-	Man	19	780	48.8	0.0244	0.390
26 Lankas	Control	Sprague-	Vrouw	24	291	48.5	0.0825	0.495
27 Lankas	Treatment	Sprague-	Man	80	2372	49.4	0.0337	1.62
28 Lankas	Treatment	Sprague-	Vrouw	84	885	49.2	0.0949	1.71
29 Stout and Ruecker	Control	Sprague-	Man	22	792	49.5	0.0278	0.444
30 Stout and Ruecker	Control	Sprague-	Vrouw	6	300	50	0.02	0.12
31 Stout and Ruecker	Treatment	Sprague-	Man	150	2356	49.1	0.0637	3.06
32 Stout and Ruecker	Treatment	Sprague-	Vrouw	46	900	50	0.0511	0.92
33 Sugimoto	Control	CD-1	Man	20	400	50	0.05	0.4
34 Sugimoto	Control	CD-1	Vrouw	26	400	50	0.065	0.52
35 Sugimoto	Treatment	CD-1	Man	112	1200	50	0.0933	2.24
36 Sugimoto	Treatment	CD-1	Vrouw	98	1200	50	0.0817	1.96
37 Suresh	Control	Wistar	Man	109	397	49.6	0.275	2.20
38 Suresh	Control	Wistar	Vrouw	26	267	44.5	0.0974	0.584
39 Suresh	Treatment	Wistar	Man	301	1005	41.9	0.300	7.19
40 Suresh	Treatment	Wistar	Vrouw	84	651	36.2	0.129	2.32
41 Takahashi	Control	CD-1	Man	0	150	50	0	0

42 Takahashi	Control	CD-1	Vrouw	3	50	50	0.06	0.06
43 Takahashi	Treatment	CD-1	Man	10	450	50	0.0222	0.2
44 Takahashi	Treatment	CD-1	Vrouw	11	150	50	0.0733	0.22
45 Wood_a	Control	CD-1	Man	28	408	51	0.0686	0.549
46 Wood_a	Control	CD-1	Vrouw	31	408	51	0.0760	0.608
47 Wood_a	Treatment	CD-1	Man	94	1224	51	0.0768	1.84
48 Wood_a	Treatment	CD-1	Vrouw	66	1224	51	0.0539	1.29
49 Wood_b	Control	Wistar	Man	40	406	50.8	0.0985	0.788
50 Wood_b	Control	Wistar	Vrouw	52	306	51	0.170	1.02
51 Wood_b	Treatment	Wistar	Man	104	1224	51	0.0850	2.04
52 Wood_b	Treatment	Wistar	Vrouw	168	918	51	0.183	3.29

Uiteindelijk maak ik 8 verschillende modellen waarvan de resultaten zichtbaar zijn in **Tabel 41**:

1. Een model zonder verklarende factoren.
2. Een model met *Behandeling* als verklarende factor.
3. Een model met *Geslacht* als verklarende factor.
4. Een model met *Soort* als verklarende factor.
5. Een model met *Behandeling* en *Geslacht* als verklarende factoren.
6. Een model met *Behandeling*, *Geslacht* en *Soort* als verklarende factoren.
7. Een model met *Behandeling*, *Geslacht*, de interactie tussen *Behandeling* en *Geslacht*, en *Soort* als verklarende factoren.
8. Een model met *Behandeling*, *Soort*, de interactie tussen *Behandeling* en *Soort* en *Geslacht* als verklarende factoren.

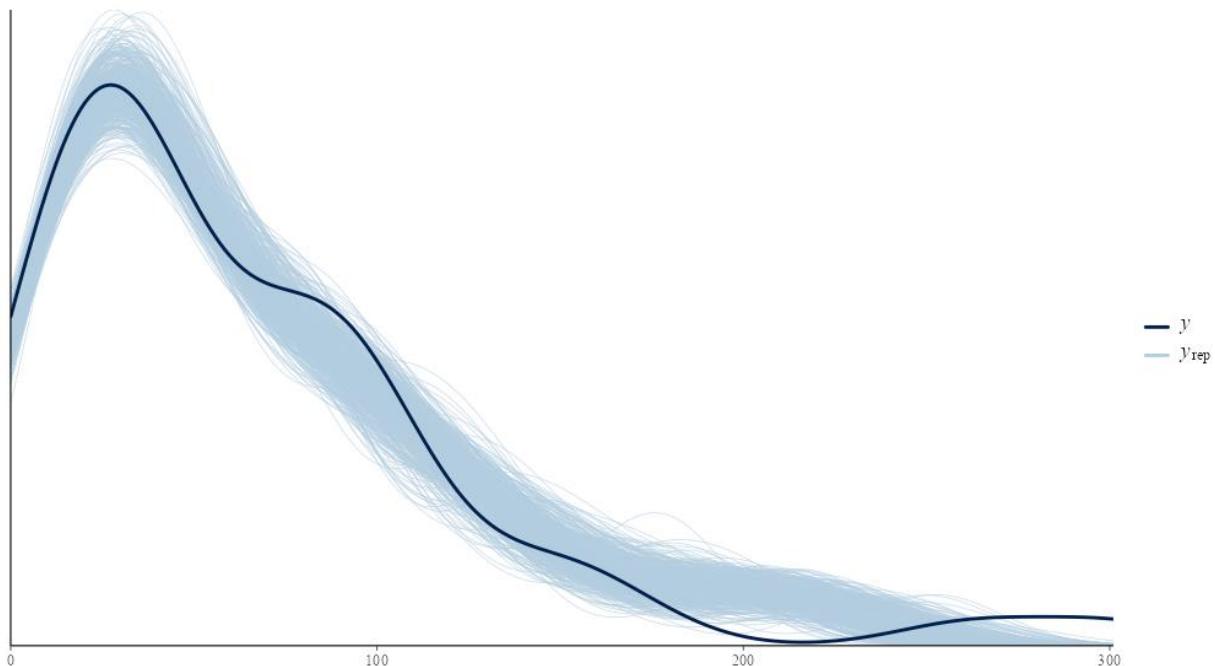
**Tabel 41** is niet heel makkelijk te interpreteren want geen enkel model staat direct boven alle modellen uit. Zeker als we naar de laatste drie modellen kijken is het maar lastig kiezen. Toch heb ik gekozen voor model 8 omdat he door de bocht genomen de hoogste LR-waarden heeft. Maar het is niet gek om te denken dat iemand ook voor model 6 of model 7 zou kunnen kiezen.

		Noemer							
		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Deler	Model 1	1	3.57	69,400	2390	236,000	7.05e+08	2.50e+08	5.74e+08
	Model 2	0.280	1	19,500	821.93	66,200	1.98e+08	7.11e+07	1.63e+07
	Model 3	<0.0001	<0.0001	1	0.042	3.40	10,100	3,660	8,390
	Model 4	<0.0001	0.001	23.69	1	80.59	240,000	86,100	197,000
	Model 5	<0.0001	<0.0001	0.294	0.012	1	2,980	1,070	2,450

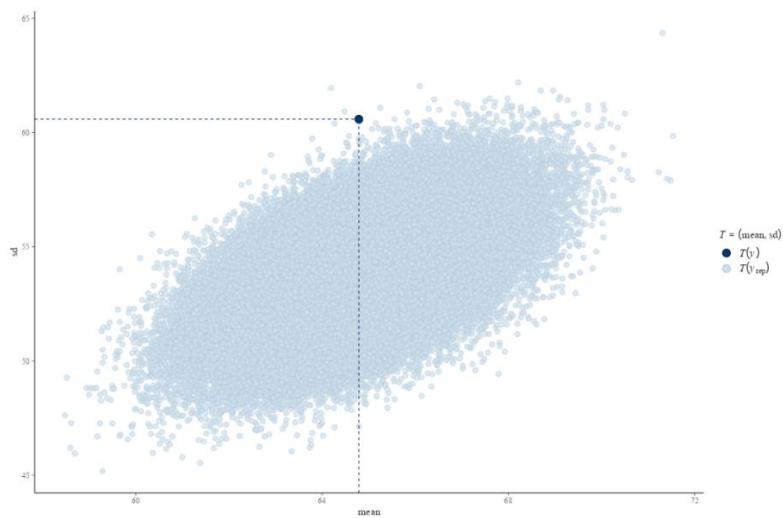
	<i>Model 6</i>	<0.0001	<0.0001	<0.0001	<0.0001	0.00034	1	0.360	0.826
	<i>Model 7</i>	<0.0001	<0.0001	0.00002	<0.0001	0.0009	2.78	1	2.29
	<i>Model 8</i>	<0.0001	<0.0001	0.00012	<0.0001	0.0004	1.21	0.436	1

**Tabel 41.** De LR-waarden voor de 8 modellen zoals hierboven beschreven.

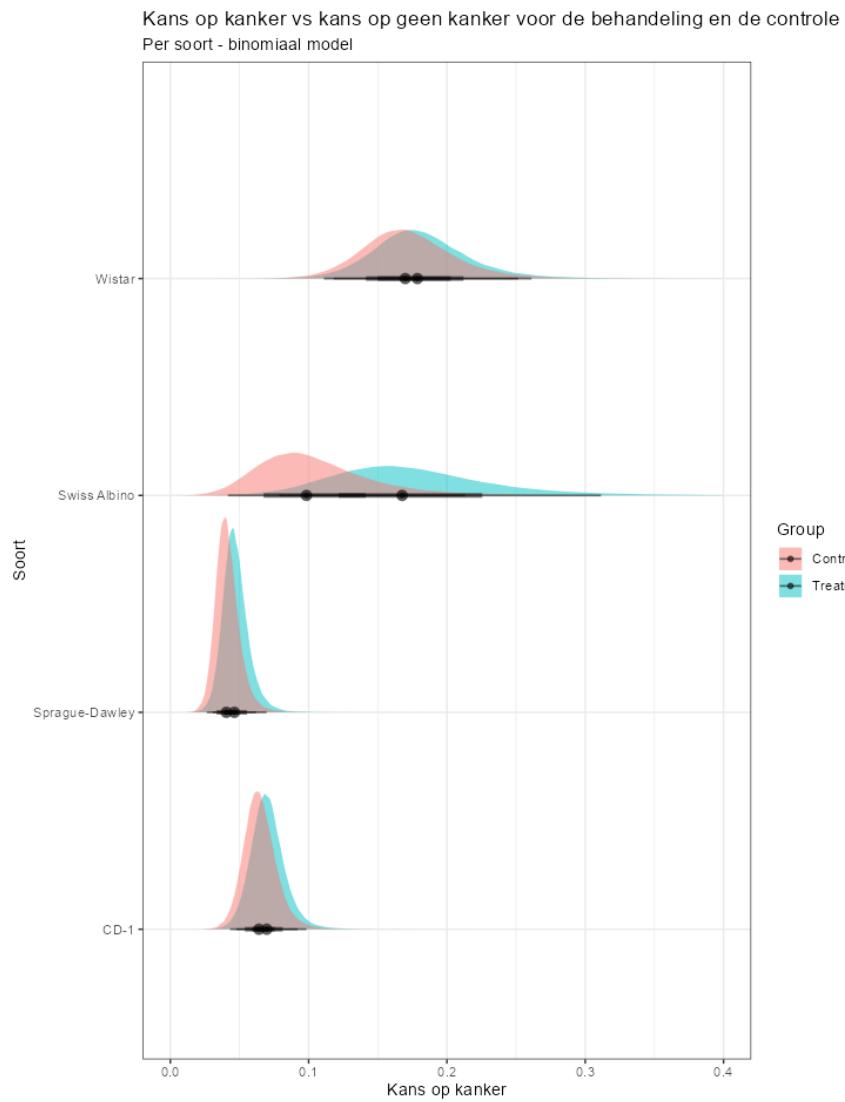
Onderaan staan de figuren (**Figuur 165** en **Figuur 166**) zoals we die al eerder hebben gezien, maar nu voor model 8. Deze figuren laten een ander beeld zien dan **Figuur 161** en **Figuur 162**. We zien nu dat de posterior op sommige plekken niet overeenkomt met de geobserveerde waarden. Het geobserveerde gemiddelde en de standaard deviatie zien we niet terug in de posterior trekkingen uit model 8. Dat is iets om in het achterhoofd te houden. Toch wil ik doorgaan met de uitkomsten en die zijn zichtbaar in **Figuur 167** en in **Figuur 168**. Specifieke focus ligt in deze figuren op het verschil tussen de controlegroep en de behandelgroep voor elke soort. Dat heb ik gedaan omdat model 8 een interactie meeneemt tussen groep en soort dier. Dan is het ook verstandig om eventuele odds-ratios per groep uit te rekenen<sup>130</sup>.

**Figuur 165.** De verdeling vanuit de posterior (dunne lijnen) en de geobserveerde verdeling. Vooral aan het eind zitten er wat deviaties.

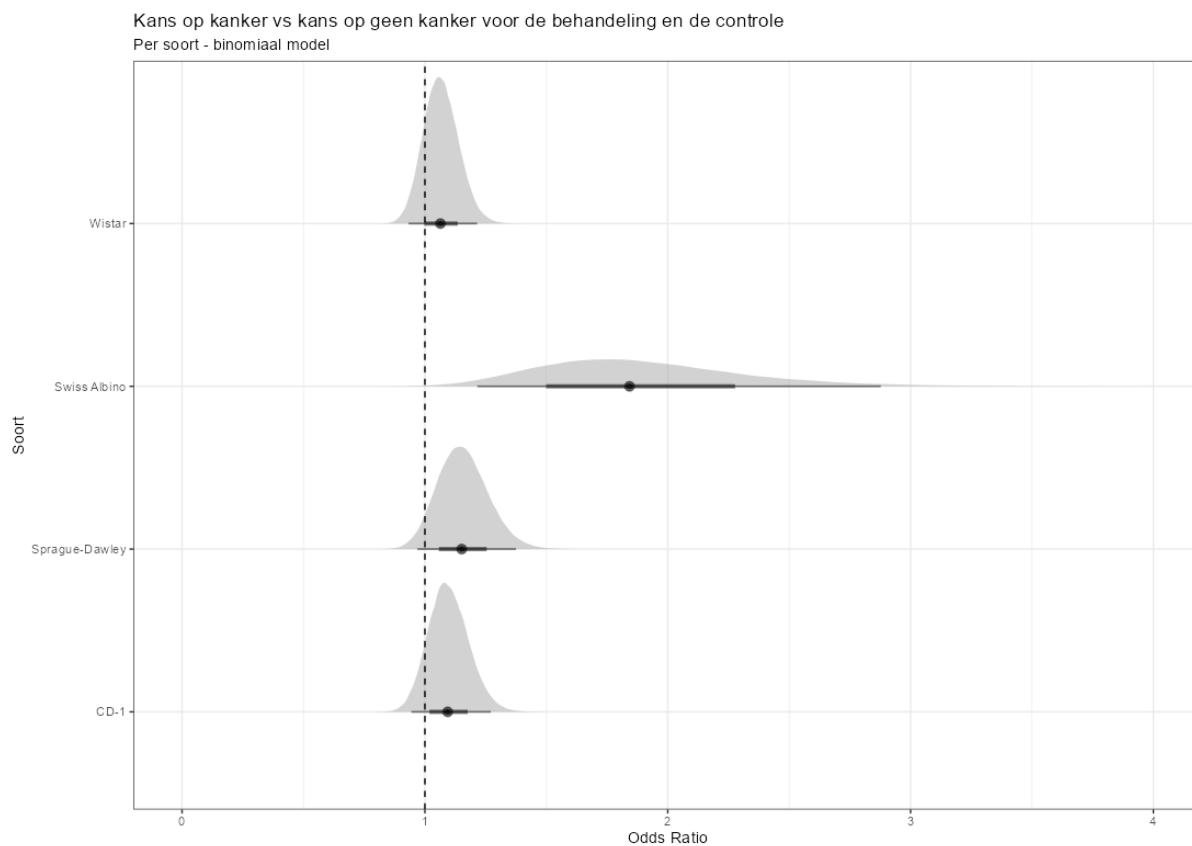
<sup>130</sup> Uiteindelijk ben ik alleen geïnteresseerd in het effect van de groep op de kans op kanker, maar met een interactie toegevoegd kan ik een marginaal effect voor behandeling niet meer zinvol uitrekenen.



**Figuur 166.** De posterior trekkingen van het gemiddelde en de standaarddeviatie ten opzichte van de geobserveerde getallen.



**Figuur 167.** De posterior verdelingen voor de controlegroep en de behandelgroep per soort.



**Figuur 168.** De odds-ratios voor behandeling voor elke soort.

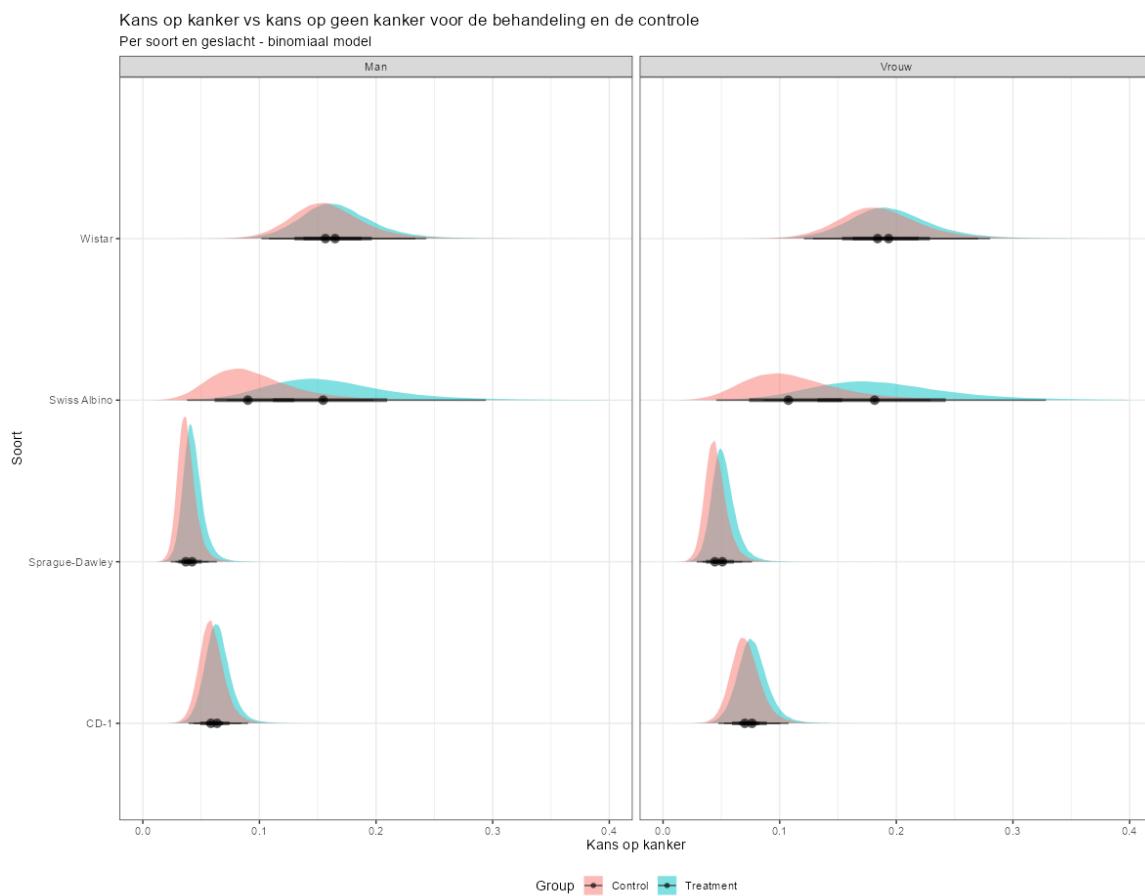
We zien dat de odds-ratios voor elke soort klein is, behalve voor de Swiss Albino ratten. Het is daarom zinvol om daar verdere toetsing op los te laten en uit te rekenen wat de LR is voor de hypothese dat het effect van behandeling in die groep nul is, kleiner dan nul of groter dan nul. De resultaten zijn zichtbaar in **Tabel 42**. Het is ook hier duidelijk dat de coëfficiënt groter is 0 wat betekent dat een ‘behandeling’<sup>131</sup> krijgen de kans op kanker verhoogt.

Soort	Effect	Hypothese 1	Hypothese 2	Likelihood Ratio
Swiss-Albino	Behandeling	0	0.52	N/A
Swiss-Albino	Behandeling	< 0	0.52	0.91
Swiss-Albino	Behandeling	> 0	0.52	97.34

**Tabel 42.** De LR voor de hypothese dat de coëfficiënt voor de behandeling nul is, kleiner dan nul of groter dan nul in Swiss-Albino ratten.

<sup>131</sup> Behandeling klinkt hier wat dubbel, maar helpt voor even in de interpretatie van het binaire karakter van de data.

In **Figuur 169** zien we de kans op kanker voor de behandeling en controlegroep per soort en per geslacht. Wat opvalt is wederom het grote verschil in de Swiss-Albino ratten. In **Tabel 43** staan de kansen uitgewerkt: zowel marginaal over geslacht heen als per geslacht. Wat opvalt is dat de kans op kanker in de behandelgroep een hele kleine toename laat zien ten opzichte van de controlegroep (**Tabel 43**).

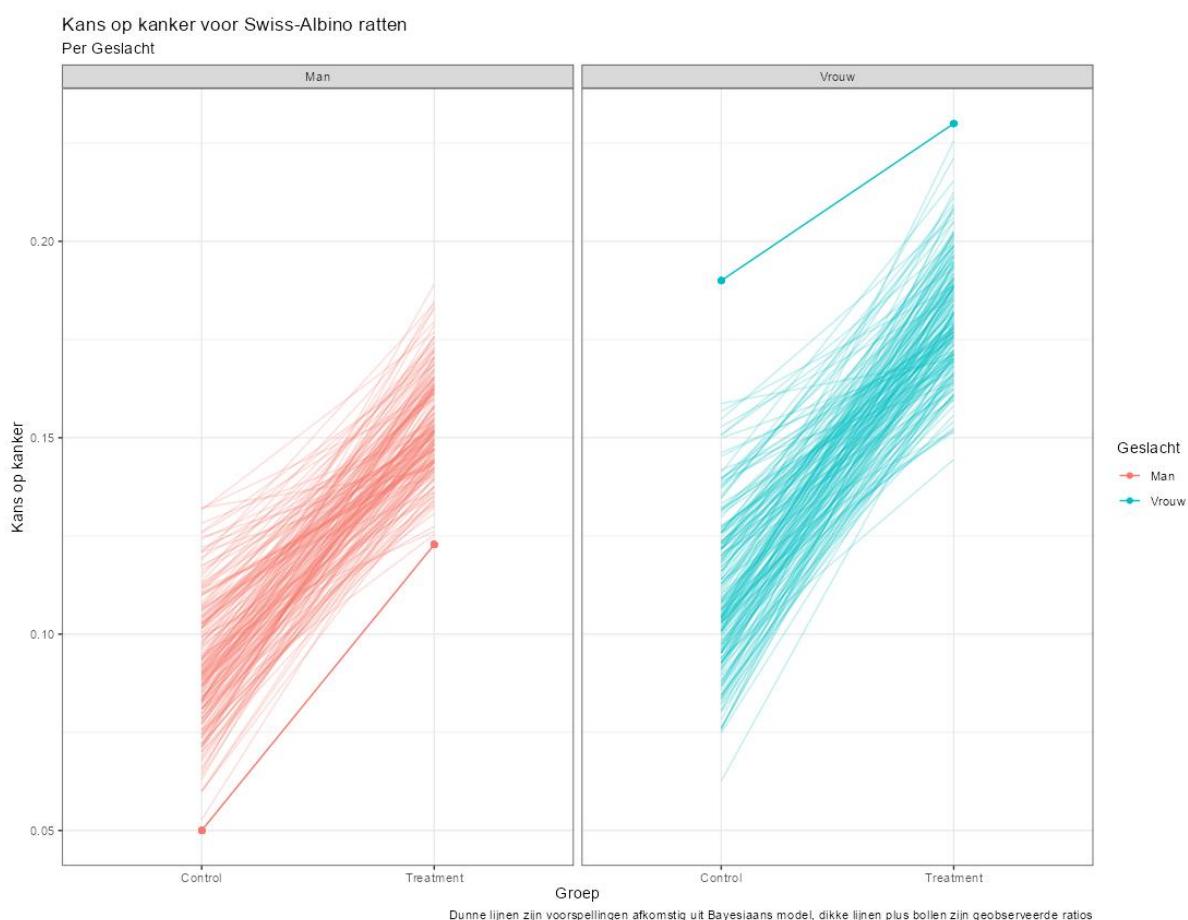


**Figuur 169.** De kans op kanker, berekend vanuit het model, per soort en per geslacht.

Soort	Groep	Geslacht	P(kanker) model	$\Delta$	P(kanker) data	$\Delta$
Swiss-Albino	Controle	Beiden	0.090	0.064	0.097	0.068
Swiss-Albino	Behandeling	Beiden	0.154		0.165	
Swiss-Albino	Controle	Mannen	0.090	0.064	0.05	0.073
Swiss-Albino	Behandeling	Mannen	0.154		0.123	
Swiss-Albino	Controle	Vrouwen	0.101	0.08	0.19	0.04
Swiss-Albino	Behandeling	Vrouwen	0.181		0.23	

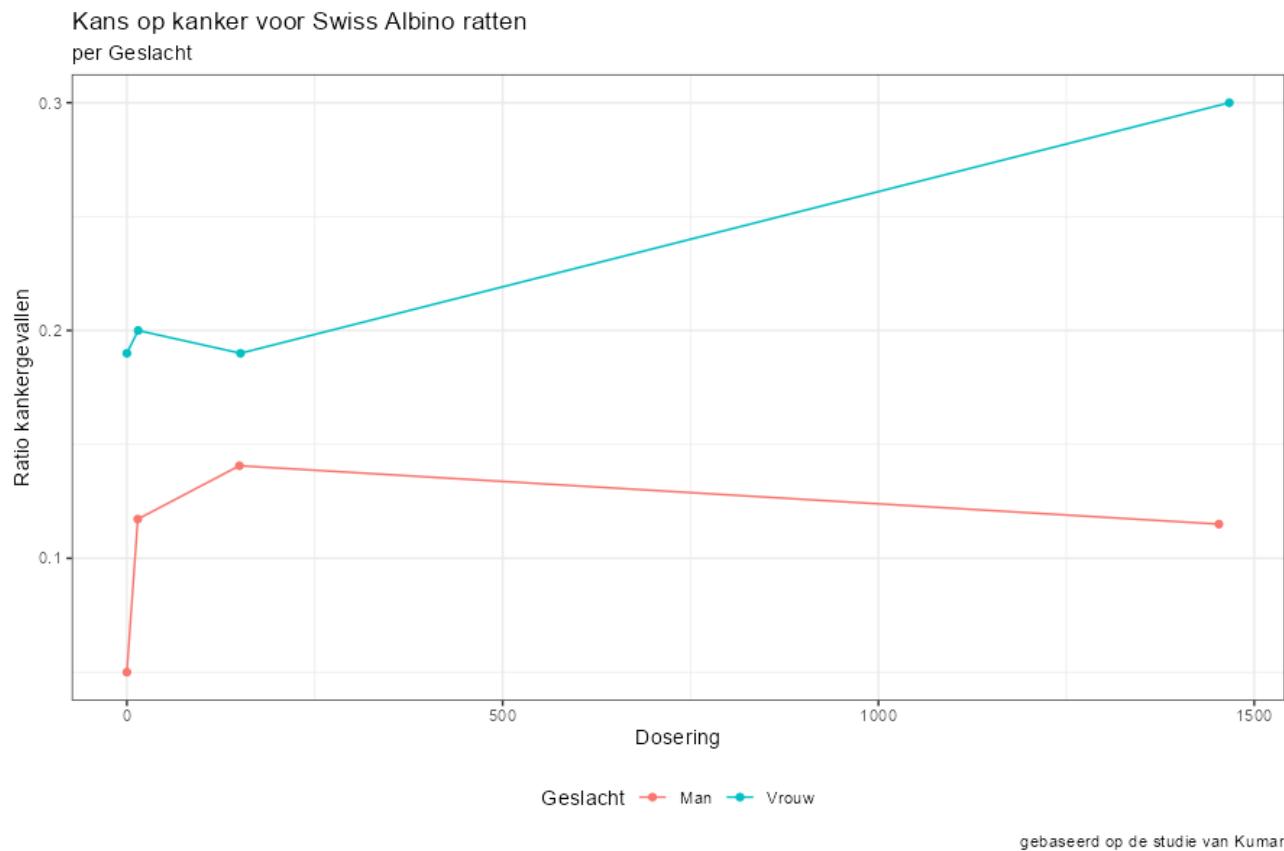
**Tabel 43.** De kans op kanker voor Swiss-Albino ratten per groep en per geslacht.

In **Figuur 170** zien we de kans op kanker zoals geobserveerd voor Swiss-Albino ratten en zoals gemodelleerd. Wat duidelijk te zien is, is dat het model de relatie aardig modelleert, maar de absolute getallen kloppen niet helemaal. Dat komt omdat het Bayesiaanse model ook een Mixed Model is. Mixed Models hebben de neiging om bij grote variatie meer naar ‘het midden te migreren’ en zo uitschieters te dempen. Dit noemt men ook wel shrinkage. Als we teruggaan naar de bron van de geobserveerde data valt dit wellicht te verklaren, want er is maar één studie die Swiss Albino ratten heeft geïncludeerd en dat is de studie van Kumar. Het is daarom zinvol om die studie in zijn geheel te tonen.



**Figuur 170.** De kans op kanker voor mannelijke en vrouwelijke Swiss-Albino ratten in de controlegroep en de behandelgroep. Het model wijkt af met een haast vaste constante wat een gevolg is van het type model.

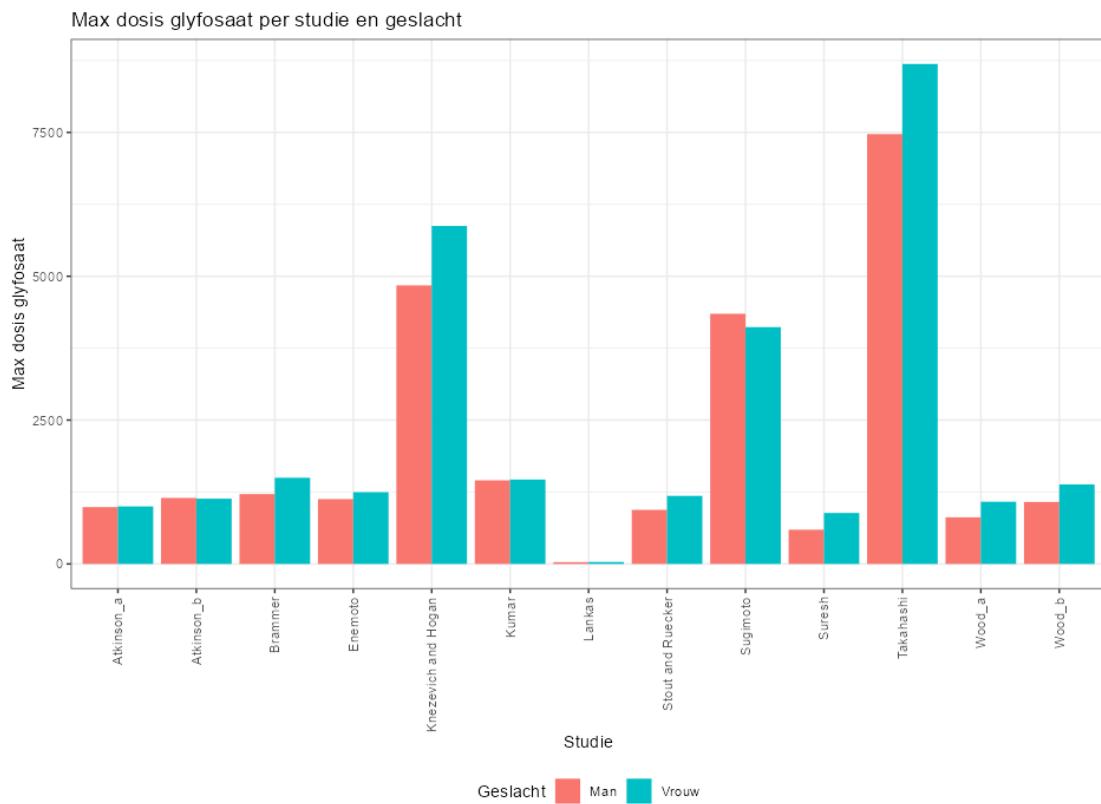
De studie staat afgebeeld in **Figuur 171** en toont voor vrouwen een stijging van ongeveer 12% tussen de controlegroep en de hoogste dosering. Voor mannen zien we eenzelfde stijging. De relatie is echter niet geheel duidelijk in de vorm wat mogelijk verklaart waarom het model wel de richting laat zien maar worstelt met de juiste getallen.



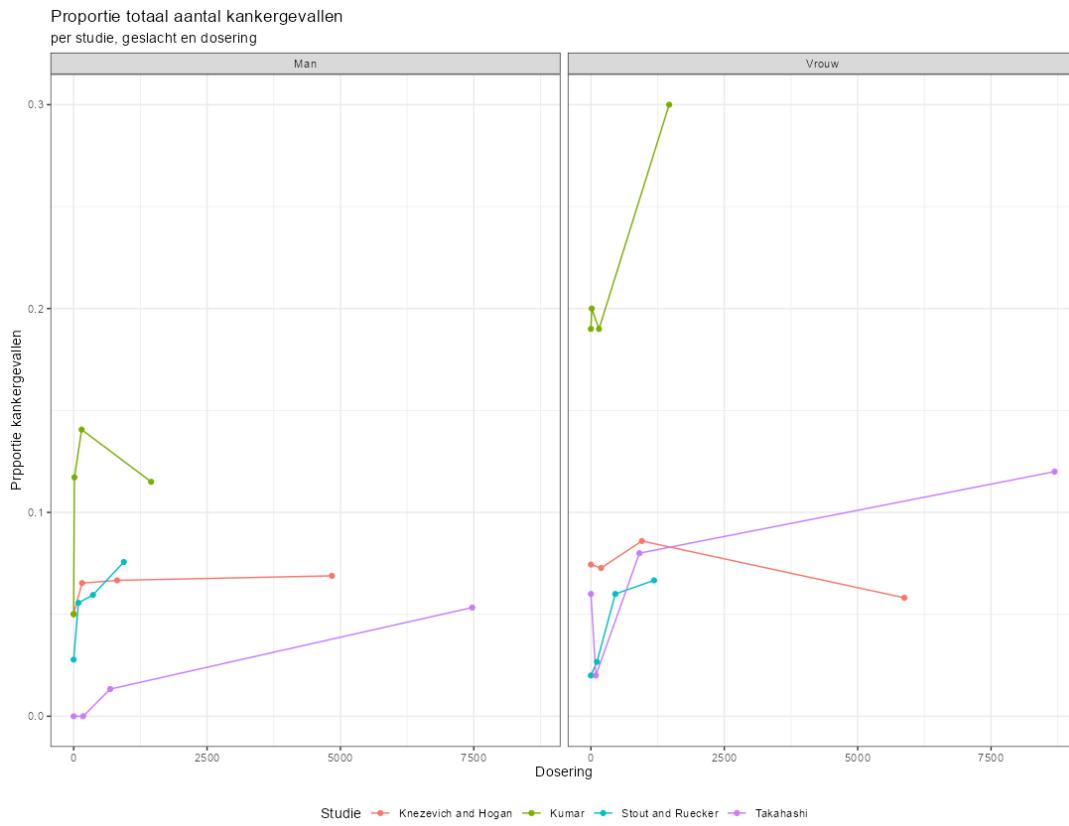
**Figuur 171.** Kans op kanker voor Swiss Albino ratten per geslacht – gebaseerd op de studie van Kumar die de enige studie is met Swiss Albino ratten.

Wellicht is er iets bijzonders aan deze studie omdat de Swiss-Albino ratten als enige een positief effect van behandeling laat zien. Een kijkje in de dosering en hoe de maximale dosering in de studie van Kumar zich verhoudt tot de maximale dosering in de andere studies toont niets bijzonders. Drie andere studies laten een grotere maximale dosering zien. Die drie studies staan afgebeeld in **Figuur 172**. Wat ik nu kan doen is de proef op de som nemen en de proportie kankergevallen voor die drie studies laten zien én de studie van Kumar. Om te laten zien hoe lastig het is om met proporties te werken toon ik dezelfde grafiek drie keer: de eerste keer zonder onzekerheid, de tweede keer met onzekerheid door middel van verticale strepen per dosering én de derde keer door dit gebied helemaal te arceren. Dat is, respectievelijk, **Figuur 173**, **Figuur 174** en **Figuur 175**.

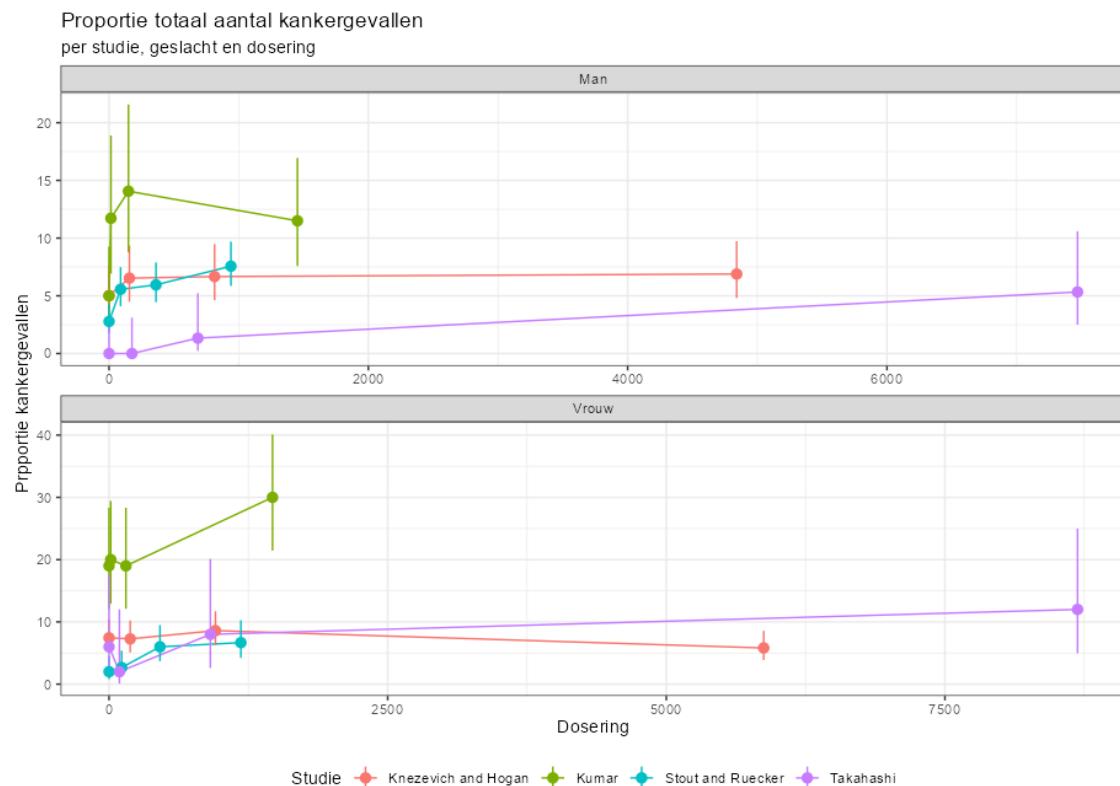
Deze figuren laten zien dat het verschil tussen de hoogste dosis en de controlegroep marginaal is die verdwijnt wanneer we context meenemen: de onzekerheid is gewoonweg te groot. De studie van Kumar, en dan met name in de vrouwelijke Swiss Albino ratten, lijkt een echt effect te tonen.



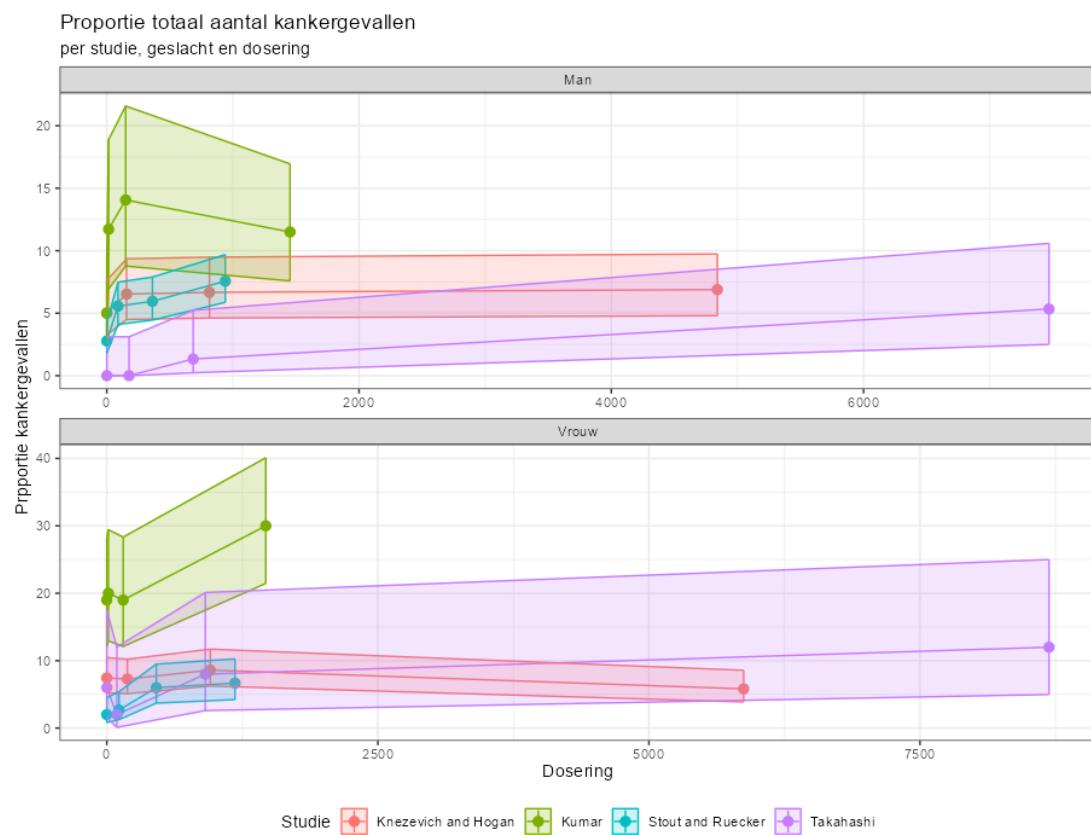
**Figuur 172.** Max dosis glyfosaat per studie en geslacht.



**Figuur 173.** Kans op kanker voor vier studies zonder onzekerheidsmarges.



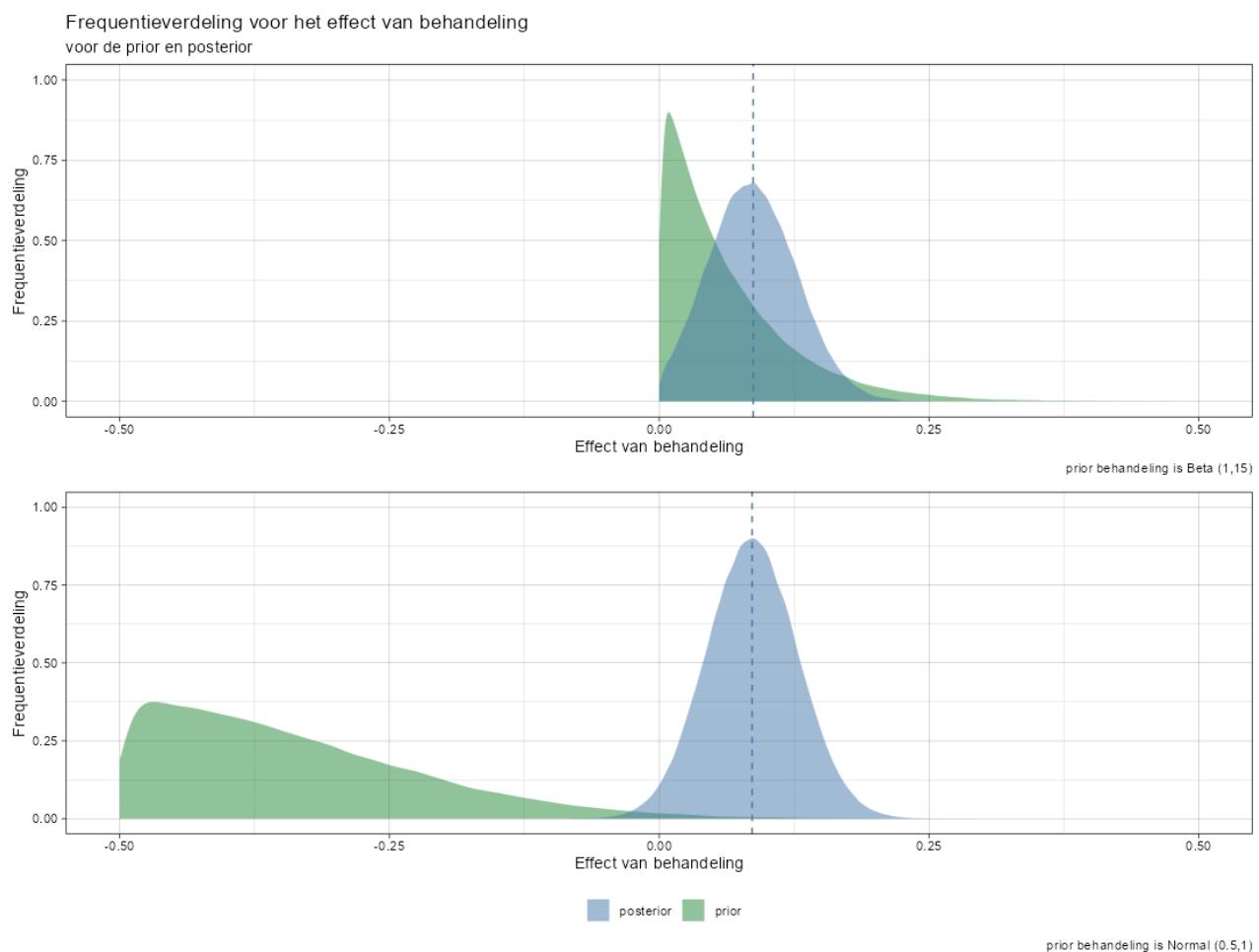
**Figuur 174.** Kans op kanker voor vier studies met onzekerheidsmarges.



**Figuur 175.** Kans op kanker voor vier studies met onzekerheidsmarges én onzekerheidsbanden.

## Het effect van de prior

Mij rest nog één laatste ding voordat ik dit hoofdstuk en daarmee ook dit rapport ga afsluiten en dat is het tonen van de eventuele invloed van de prior. Nu heb ik uiteindelijk voor model 7 gekozen en uit dat model rolden schattingen voor alle hyperparameters. De prior heb ik toen weggelaten, maar ik wil toch graag de invloed ervan laten zien. Dat kan ik kort doen, door voor model 1 te kiezen en dat is een model waarin ik alleen de behandeling meeneem als verklarende variabele. Het resultaat staat afgebeeld in **Figuur 176**.



**Figuur 176.** De invloed van twee priors op de posterior. De Beta prior gaat er van uit dat behandeling de kans op kanker verhoogt. De normaalverdeling is een prior met de aanname dat glyfosaat de kans op kanker verkleint.

Wat deze afbeelding laat zien is dat de kans op kanker, zoals geschat vanuit de data, nagenoeg niet beïnvloed wordt door de prior. Dat zien we omdat de posterior hetzelfde blijft. Eigenlijk is dat ook niet gek want als we alle tabellen uit *Supplementary Material 2* meenemen en dan ook nog eens alle rijen optellen, dan hebben we enorme aantallen voor

zowel het aantal keren dat kanker is geobserveerd en ook het totaal aantal observaties. Dat komt omdat de tabel per tumor rekent en daarmee een enorme inflatie aan getallen toont. Omdat we niet weten welk dieren welke kanker hadden, of dat er wellicht meerdere dieren waren met meerdere kankersoorten kunnen we daar maar lastig voor corrigeren. Uiteindelijk hebben we wel, door de bocht genomen, ongeveer 5200 observaties<sup>132</sup>. Dat ik dus een prior toevoeg die uitgaat van een negatief effect van de behandeling (de Beta prior) of van een positief effect (de Normaal prior) maakt dan ook niets meer uit<sup>133</sup>. We zullen onze bevindingen moeten zoeken in de data zoals we die hebben ontvangen vanuit Portier. Op mijn mail met verzoek om de exacte codes te delen is nooit reactie gekomen.

## Wat kunnen we concluderen?

Door middel van de Bayesiaanse statistiek zijn we uit de wereld van de statistische significant gestapt en hebben onze intrede gedaan in de wereld van bewijs. Met de Likelihood Ratio (LR) als belangrijkste metriek hebben we gekeken of we een dose-response relatie konden bouwen op basis van het model dat het meest wordt ondersteund door de data. Het lijkt erop dat een dose-response relatie in de vorm van een regressie geen relatie laat zien tussen dosering en glyfosaat, maar een binomiaal model toont uiteindelijk toch een relatie op basis van een model dat ook soort en geslacht meeneemt. De relatie is voornamelijk zichtbaar in Swiss-Albino ratten, maar daar hebben we maar één studie van. Door de bocht genomen is de kans op kanker, in die populatie, na behandeling voor glyfosaat 8% hoger bij de vrouwen en 6% hoger bij de mannen (op basis van het model). Deze bevinding staat los van de gekozen prior. Samengevat betekent dit dat we een traditionele dose-response analyse los moeten laten.

---

<sup>132</sup> 13 studies \* gemiddeld genomen vier doseringen \* twee geslachten \* 50 dieren = 5200 observaties.

<sup>133</sup> Een eventuele opdeling van de data zou de prior sterker kunnen maken, maar dat is eigenlijk alleen geldig als ik per tumorsoort zou kijken.

## Conclusie en aanbevelingen

Wat ik in feite getracht heb te doen, is het op verschillende manieren proberen te herhalen van de bevindingen van Portier door gebruik te maken van de data zoals gerapporteerd door Portier. Dit rapport is dus bovenal een herhaling van onderzoek dat is uitgevoerd en wat wordt aangehaald als hét bewijs dat glyfosaat kankerverwekkend is.

Kort door de bocht genomen lukt het mij niet om zijn bevindingen exact te repliceren. Dat betekent dus niet dat ik geen statistisch significante verschillen kan vinden als ik dezelfde methoden gebruik als Portier. Die zie ik ook, maar wanneer ik de methodiek van Portier hanteer treden er wel problemen op die niet worden geadresseerd.

In het algemeen rapporteert Portier per geslacht per studie welke kancersoort er wel of niet is opgetreden per dosering. Dit maakt dat de studie van Portier meer dan 200 statistische toetsingen kent. Het herhaaldelijk statistisch testen van eenzelfde dataset is echter een katalysator voor het vinden van vals positieven. Dit komt door de aannames die haast inherent zijn aan de frequentistische statistiek. Wanneer ik hiervoor corrigeer verdwijnen alle statistisch significante effecten.

Wat verder mist is dat de studies verschillen in welke soorten kanker wordt gerapporteerd. Het lijkt er sterk op dat elke analyse berust op het vinden van een soort kanker in welke dosering dan ook waarnaar voor elke dosering een analyse wordt gedaan. Wanneer een kancersoort in zijn geheel uitblijft wordt dat soms wel gerapporteerd, maar men is hier niet consequent in. De analyse van Portier wordt dus bovenal gedaan op het niveau van de tumor. Niet alleen ondervinden we dan het probleem van de vals positieven, maar we werken ook met proporties die als hoger worden gerapporteerd dan ze daadwerkelijk zijn. Dit komt omdat bij het samenvoegen van de studies, wat later ook door Portier wordt gedaan, het uitblijven van kanker niet wordt meegenomen in de berekening van de kans op kanker.

In het algemeen is het uitermate lastig gebleken om een zogenaamde dose-respons analyses uit te werken. Traditionele non-lineaire analyses mislukken en met behulp van meer lineaire technieken zie ik een hoop onzekerheid. De relatie tussen glyfosaat en kanker verschilt vaak en veel. Verder is er een substantiële kans op kanker voor de nul-dosering (de controlegroep). Dit alles maakt dat het zoeken naar een model dat recht doet aan de geobserveerde data, én aan de modelassumpties, buitengewoon lastig is.

Ik heb echt moeten zoeken om die relatie te vinden. Door de bocht genomen lukt het mij niet om met behulp van de frequentistische statistiek een relatie aan te tonen tussen dosering en kanker. De bevindingen zijn vaak niet statistisch significant wanneer ik tweezijdig toets. Een uitstap naar een eenzijdige toets voegt daar weinig aan toe én maakt dat we moeten aannemen dat het schatten van de relatie tussen dosering en kans op kanker een harde grens heeft in het schatten van de relatie. Ik voeg dan een assumptie die zich maar moeilijk laat verdedigen. Een eenzijdige toets heeft namelijk niet zo veel te maken met de richting van de relatie, maar eerder met het afkappen van de onzekerheid. In een dossier als deze, waarin de onzekerheid groot is, kan dat geen juiste methodiek zijn voor het bepalen van een relatie.

Een overstep van frequentistische statistiek naar Bayesiaanse statistiek laat zien dat modellen die dosering meenemen als verklarende variabele niet per se beter passen bij de data. Pas als we de dosering opdelen in een controlegroep en een behandelgroep lukt het mij om een relatie te tonen tussen glyfosaat en de kans op kanker. Althans, het model dat het effect van glyfosaat meeneemt wordt meer door de data ondersteund.

Uiteindelijk lijkt het erop dat alleen in de Swiss Albino ratten, op basis van één studie en dan met name bij vrouwen, de relatie tussen glyfosaat en de kans op kanker duidelijk is: een toename van 8% vanuit het model. De toename op kanker, vanuit de data, bedraagt dan 4%. Beide getallen kennen een aanzienlijke onzekerheidsmarge. Daarmee kunnen ze niet doorslaggevend zijn voor het gehele dossier.

We kunnen denk ik met dit rapport concluderen dat onderzoek naar glyfosaat beter moet en beter kan, maar daarvoor moet de data ook op het niveau van het dier worden gemeten waarbij ook wordt gekeken naar de factor tijd. Dat ontbreekt nu. Verder hebben we het hier over dierproeven en niet over menselijke studies.

Deze bevindingen zijn een stuk milder dan de uitspraken vanuit de BNNVARA/ZEMBLA reportage waarin zo sterk werd opgeroepen om de tweezijdige toets te vervangen door de eenzijdige toets. Het uitblijven van die vervanging zou wijzen op het moedwillig negeren van bewijs dat glyfosaat kankerverwekkend is. Nu ben ik geen toxicoloog en heb dus voornamelijk gekeken naar de statistiek achter de studie van Portier die zo vaak werd aangehaald in deze reportage en vervolgartikelen. Over de biologische *mode of action* laat ik mij in dit rapport niet uit. Mij ging het er boven alles om, om de relatie

tussen data, statistiek en gemaakte uitspraken beter te duiden. Dit rapport is dus maar één enkel onderdeel in een groter dossier.

Ik hoop van harte dat dit rapport laat zien dat statistiek bedrijven op data meer is dan wisselen van toets om een aanname gewicht te geven. Het is bovenal keuzes maken gebaseerd op de data zoals verkregen en de inzichten die deze data geven. Een essentieel onderdeel daarin is, en blijft, het visualiseren van de data. Dit rapport telt bijna 176 afbeeldingen. In het werk van Portier vinden we meer dan 200 statistische toetsen maar nul afbeeldingen. Ik denk dat het toevoegen van één enkele afbeelding, namelijk het tonen van de relatie tussen dosering en kanker per studie én geslacht, voldoende was geweest om te duiden waarom het uitvoeren van zoveel statistische toetsen problematisch zou zijn. Het vervangen van een tweezijdige toets door een eenzijdige toets zou hier niets aan veranderen: ik zou alleen maar voor meer vals positieven moeten corrigeren.

Dit rapport zegt **niet** dat er geen relatie is tussen glyfosaat en kanker, maar laat wel zien dat een analyse voor glyfosaat berust op veel aannames. Het moeten maken van aannames is normaal in het modelleren van data, maar brengt ook kwetsbaarheden met zich mee. Ik heb geprobeerd die kwetsbaarheden te tonen. In de studie van Portier zijn deze haast onzichtbaar.

Het lijkt er sterk op dat meer werk nodig is voor het tonen van een verband tussen glyfosaat en kanker, mits we ervan uit mogen gaan dat de data zoals meegenomen afkomstig zijn uit betrouwbare bronnen. Ik kan dat niet goed beoordelen, maar ben ervan uit gegaan dat als Portier deze studies selecteert ik dat ook kan doen.

Het meerwerk hoeft niet direct te leiden tot meer studies, maar kan ook een her-analyse zijn van de data. Het probleem met zowel de analyse van Portier als die van mij is dat wij beiden kijken naar de incidentie van kanker aan het einde van het onderzoek. Veel relevanter was het geweest als we hadden geweten (of althans IK had geweten) welk dier wanneer welke kanker zou krijgen. Op die manier zou ik in staat zijn om te zien of er dieren zijn die meerdere typen kanker krijgen en wanneer. Nu is het zo dat een dier dat niet sterft ook niet kan worden onderzocht op kanker (althans vaak niet) wat een zogenaamde survival analyse bemoeilijkt. Toch zou informatie op niveau van het dier én op niveau van tijd de kwaliteit van de analyse vergroten.

Mocht er uiteindelijk toch worden gekozen voor meerwerk dan is het belangrijk om voor een setting te kiezen die veel lijkt op de meest gebruikt setting. In afwezigheid van een

duidelijk signaal is het zinvol om een studie op te zetten waarin nieuwe informatie direct gestapeld kan worden op oude informatie. Deze informatie zou men dan in een Bayesiaanse analyse kunnen verwerken.

Samengevat lijkt er nog veel onwetendheid te zijn over statistiek en modelleren. De BNNVARA/ZEMBLA reportage met haar haast bombastische uitspraken over het kiezen van ‘de verkeerde toets’ sterkt helaas het idee dat statistiek een veld is waarin het draait om goed of fout. Niets is minder waar. Het draait meestal om kansen, maar bovenal draait het om het verwerken van informatie zodat deze kunnen worden gebruikt voor een gesprek.

## Figuren

<b>Figuur 1.</b> Een theoretische normaalverdeling met gemiddelde 0 en standaard deviatie 1. Het kenmerk van de normaalverdeling is zijn ‘bel’-vorm.....	19
<b>Figuur 2.</b> De theoretische verdeling van massa in een normaalverdeling. Ongeveer 68% van de gegevens vallen binnen één standaard deviatie van het gemiddelde. Bij twee standaarddeviaties is dat ongeveer 95%. .....	21
<b>Figuur 3.</b> De normaalverdeling met een gemiddelde van 184 cm en een standaard deviatie van 7 cm. Dit is waarschijnlijk hoe de verdeling van lengte bij de Nederlandse mannen verdeeld is. ....	22
<b>Figuur 4.</b> De normaalverdeling met een gemiddelde van 170 cm en een standaard deviatie van 6,3 cm. Dit is waarschijnlijk hoe de verdeling van lengte bij de Nederlandse vrouwen verdeeld is. ....	23
<b>Figuur 5.</b> Theoretische kansverdeling van de normaalverdeling met een gemiddelde 0 en standaard deviatie 1. Onderaan staan de cumulatieve percentages. ....	24
<b>Figuur 6.</b> De plek waar een lengte van 186 cm past in een verdeling van lengtes van Nederlandse mannen. ....	24
<b>Figuur 7.</b> De cumulatieve kansverdeling bij een gemiddelde van 184 cm met 7 cm standaarddeviatie. ....	25
<b>Figuur 8.</b> De plek waar een lengte van 200 cm past in een normaalverdeling van lengtes van Nederlandse mannen. ....	26
<b>Figuur 9.</b> De plek waar een lengte van 200 cm past in een normaalverdeling van lengtes van Nederlandse .....	27
<b>Figuur 10.</b> De plek waar een lengte van 150 cm past in een normaalverdeling van lengtes van Nederlandse mannen. ....	27
<b>Figuur 11.</b> De plek waar 186 cm past in een normaalverdeling van lengte van Nederlandse mannen met een gemiddelde van 170 cm en een standaard deviatie van 5 cm.....	28
<b>Figuur 12.</b> De verdeling van lengte waarbij de gegevens van mannen en vrouwen zijn gecombineerd. De gegevens zijn gesimuleerd. ....	32
<b>Figuur 13.</b> De verdeling van lengte waarbij de gegevens van mannen en vrouwen zijn gecombineerd. De gegevens zijn gesimuleerd. De rode stippellijn geeft aan waar de 200 cm valt in die verdeling. ...	33
<b>Figuur 14.</b> De verdeling van lengte waarbij de gegevens van mannen en vrouwen zijn gecombineerd. De gegevens zijn gesimuleerd. Het turquoise gedeelte laat zien wat het vlak is wat 200cm of meer is. ....	34
<b>Figuur 15.</b> De verdeling van Nederlandse mannen en vrouwen in eenzelfde grafiek. Zichtbaar zijn de verschillende pieken en de overlap. Beiden zijn van betekenis. ....	35
<b>Figuur 16.</b> De plek van 175 cm in de verdeling van Nederlandse mannen en van vrouwen. ....	36
<b>Figuur 17.</b> De lengte van 175 in de verdeling van mannen en die van vrouwen. Het rode gedeelte is het percentage mannen of vrouwen die kleiner of gelijk zijn aan 175 cm. Het turquoise vlak is het percentage dat groter of gelijk is. Op basis van deze massa kun je niet bepalen of iemand een man of vrouw is. Je kunt hoogstens het percentage uitrekenen waarin we iemand met deze lengte zien gegeven de verdeling van lengtes. ....	37
<b>Figuur 18.</b> De verdeling van verschillen tussen de groep Nederlandse mannen en de groep Nederlandse vrouwen. ....	39
<b>Figuur 19.</b> De verdeling van verschillen. Ongeveer 27% van de waarden zijn negatief. ....	39

<b>Figuur 20.</b> De lengte 200 cm in de frequentieverdeling van Nederlandse mannen. De waarde 200 cm komt .....	42
<b>Figuur 21.</b> De 2,5% grenswaarden links en rechts van het gemiddelde afgevinkt met een rode stippellijn. De zwarte stippellijn is het gemiddelde en de blauwe stippellijn laat zien waar de 200 cm valt.....	43
<b>Figuur 22.</b> De plek van 184 cm in een frequentieverdeling van 200 cm bij een standaarddeviatie van 7 cm.....	47
<b>Figuur 23.</b> De verdeling van lengtes bij mannen en vrouwen afgebeeld in een boxplot. Linksboven zien we de p-waarde die hoort de bij independent samples t-test. ....	49
<b>Figuur 24.</b> De verdeling van lengtes voor mannen en vrouwen, met daarbij informatie over het gemiddelde (m), de standaard deviatie (s), het verschil (diff) en het gestandaardiseerde verschil (smd). .....	50
<b>Figuur 25.</b> De verdeling van gemiddeldes voor mannen en vrouwen. Het valt duidelijk te zien dat er tussen die gemiddeldes geen observaties zijn. Er is daadwerkelijk een kloof tussen twee torens. ....	52
<b>Figuur 26.</b> Gesimuleerde verdeling van gemiddeldes. De blauwe lijn is één willekeurig gekozen gemiddelde. De zwarte lijn laat het gemiddelde van alles gemiddeldes zien (184 cm) en de rode lijnen tonen het 95% betrouwbaarheidsinterval. ....	53
<b>Figuur 27.</b> De verdeling van gesimuleerde verschillen met in het midden de zwarte lijn die het gemiddelde verschil toont (13.4 cm). De rode lijnen zijn de 95% betrouwbaarheidsintervallen. We zien geen enkele keer de waarde 0. ....	54
<b>Figuur 28.</b> De verdeling van verschillen wanneer de spreiding van beide groepen een factor 20 is van de originele spreiding. We zien nu voor het eerst de blauwe lijn van ‘geen verschil’ maar deze valt buiten de grenswaarde zodat we alsnog de nulhypothese van ‘geen verschil’ verwerpen. ....	56
<b>Figuur 29.</b> De verdeling van lengte bij mannen en vrouwen zoals bezien wanneer we 30 trekkingen doen uit een voor ons bekende gemiddelde én spreiding. Het geobserveerde gemiddelde en spreiding wijken af van de bekende waarden.....	57
<b>Figuur 30.</b> Relatie tussen standaard deviatie en aantal observaties per steekproefgrootte.....	58
<b>Figuur 31.</b> Relatie tussen de schatting van het gemiddelde en de steekproefgrootte. De blauwe lijn is het theoretisch gemiddelde (184 cm). De zwarte en rode lijn zijn schattingen gebaseerd op 1 steekproef per steekproefgrootte en het gemiddelde van 1000 steekproeven, respectievelijk. ....	59
<b>Figuur 32.</b> Een weergave van het G*Power programma waarmee we kunnen laten zien wat de kracht van een studie is om een effect te vinden, gegeven het gemiddelde van twee groepen, hun standaard deviatie, de groeps grootte en de gekozen $\alpha$ en $\beta$ waarden. ....	62
<b>Figuur 33.</b> De theoretische verdeling van twee groepen onder een vooraf bepaald gemiddelde, spreiding en groeps grootte. ....	62
<b>Figuur 34.</b> Het verschil tussen twee groepen op basis van de effectgrootte. De plot rechtsonder lijkt het meest op de data zoals daadwerkelijk verzameld. Dit komt omdat ik ben begonnen met de verdeling links (rood) en daar een effectgrootte van twee heb opgeteld. Dit is ook de effectgrootte die we zelf hebben berekend. ....	64
<b>Figuur 35.</b> Relatie tussen aantal observaties per groep, de effectgrootte tussen groepen en de kracht van een studie bij een $\alpha$ waarde van 0.05 en een $\beta$ waarde van 0.2. ....	65
<b>Figuur 36.</b> Relatie tussen effectgrootte (Cohen’s d) en het aantal observaties per groep. De stippellijnen laten zien wat de minimale grootte per groep moet zijn onder een $\alpha$ van 0.05 en een $\beta$ van 0.2.....	66

<b>Figuur 37.</b> De relatie tussen de $\alpha$ waarde en de power ( $1 - \beta$ ) van een studie bij verschillende aantallen observaties per groep. Omdat de effectgrootte zo groot is (2) maakt de keuze voor de $\alpha$ waarde eigenlijk niet meer uit.....	67
<b>Figuur 38.</b> De relatie tussen de $\alpha$ waarde, het aantal observaties per groep en de power van de studie.....	68
<b>Figuur 39.</b> De relatie tussen de $\alpha$ en $\beta$ waarde van een studie en de kracht om een effect te vinden. ....	69
<b>Figuur 40.</b> Dekkingsgraad bij een $\alpha$ van 5%. De simulatie laat zien dat we een dekkingsgraad van 96% hebben (4/100 simulaties is niet gedekt). .....	71
<b>Figuur 41.</b> De dekkingsgraad van een simulatie van munt opgooien. Hoewel de theoretische verdeling 50/50 is duurt het een tijd voordat de simulatie ook in de buurt komt. Dit laat zien dat 50/50 een model is en niet per se de werkelijkheid representeert Uiteindelijk, zo laat de theorie ook zien, zal een munt bij oneindig veel keer munt opgooien de beoogde 50/50 zien. Tenzij de munt niet zuiver is. ....	72
<b>Figuur 42.</b> De dekkingsgraad van het betrouwbaarheidsinterval bij 80, 90, 95 en 99%. Wat opvalt is dat deze dekkingsgraad ook afhankelijk is van de steekproefgrootte. Voor elke steekproefgrootte hebben we 10 herhaalde steekproeven genomen.....	73
<b>Figuur 43.</b> Voorbeeld van gluren. We zien hier het aantal observaties dat nodig is om de geschatte waarde $\alpha$ te halen.....	75
<b>Figuur 44.</b> Grenswaarde (blauwe stippellijn) als functie van tweezijdig en eenzijdig toetsen. ....	78
<b>Figuur 45.</b> De alternatieve hypothese is bij tweemaal eenzijdig toetsen er een van ‘geen verschil’ waarbij ‘geen verschil’ betekent dat er een verschil is wat binnen de klinische grenzen valt. ....	81
<b>Figuur 46.</b> Verschillen vormen van testen op basis van het betrouwbaarheidsinterval van een verschil. De grenswaarde is nu niet alleen de grens van ‘geen verschil’, maar ook de grens die gezet wordt op -0.5 en 0.5. Het wordt er niet makkelijker op zo. ( <a href="https://en.wikipedia.org/wiki/Equivalence_test#/media/File:Equivalence_Test.png">https://en.wikipedia.org/wiki/Equivalence_test#/media/File:Equivalence_Test.png</a> ).....	81
<b>Figuur 47.</b> Verschil tussen mannen en vrouwen in lengte waarbij de klinische grens op -1 én 1 is gezet. Het resultaat in de onderste grafiek laat zien dat het 99% betrouwbaarheidsinterval zowel de 1 als de 0 overschrijdt. Daarmee is het verschil tussen mannen en vrouwen niet statistisch significant én niet statistisch equivalent. We kunnen dus niet zomaar de nulhypothese van ‘geen verschil’ aannemen ook al is het verschil statistisch significant. ....	82
<b>Figuur 48.</b> Verdeling man en vrouwen over alle jaren heen. ....	86
<b>Figuur 49.</b> Verdeling per jaar over de geslachten heen.....	86
<b>Figuur 50.</b> Verdeling man en vrouw per jaar.....	87
<b>Figuur 51.</b> Verdeling jaren per geslacht. ....	87
<b>Figuur 52.</b> De verdeling van lengte (als boxplot) per jaar per geslacht op twee verschillende manieren met de mogelijke testen ertussen. Het zijn een hoop vergelijkingen die we kunnen maken.....	88
<b>Figuur 53.</b> p-waardes als functie van de vergelijking en test. ....	89
<b>Figuur 54.</b> Gemiddelde waarde en 95% betrouwbaarheidsinterval. ....	90
<b>Figuur 55.</b> Voorbeeld van de data van één enkele studie zoals gerapporteerd door Portier. OP basis van deze tabel, en de andere 12 tabellen, zal ik de analyses trachten na te bootsen. ....	92
<b>Figuur 56.</b> Proportie kanker per dosering, en per dosering én geslacht. ....	93
<b>Figuur 57.</b> Proportie kanker zoals verkregen vanuit <b>Figuur 55</b> . ....	94
<b>Figuur 58.</b> Het aantal keer dat één kancersoort voorkomt in 12 studies. Bijvoorbeeld: Mammary Gland Adenomas is in 3 studies gezien. Dit kan verschillen van de 3 studies waarin Mammary Gland Adenocarcinomas is gezien. ....	98

<b>Figuur 59.</b> Proportie kanker per dosering én proportie kanker per dosering en geslacht. Duidelijk te zien dat het gros van de studies data heeft verzameld tot ongeveer 2000 mg/kg/dag. De lineaire schaal is waarschijnlijk niet de beste schaal om mee te meten.....	99
<b>Figuur 60.</b> Dose-response curve waarbij de dosering wordt afgezet tegenover het aantal kankergevallen. De blauwe lijn is de lijn afkomstig van een wiskundig model met als data de rode bolletjes: dit zijn alle observaties. ....	100
<b>Figuur 61.</b> Dose response per geslacht – exercitie is gelijk aan <b>Figuur 60</b> .....	100
<b>Figuur 62.</b> Aantal dieren met kanker en hoeveelheid dieren per dosis (bovenste figuur);proportie kankergevallen per dosis (middelste figuur); én absolute verschil in proporties tussen de nuldosering en de andere doseringen (onderste figuur). ....	101
<b>Figuur 63.</b> Absolute verschil in proporties tussen de nuldosering en de andere doseringen per geslacht. ....	102
<b>Figuur 64.</b> Absolute verschil in proporties tussen de nuldosering en de andere doseringen per geslacht én diersoort. ....	103
<b>Figuur 65.</b> Absolute verschil in proporties tussen de nuldosering en de andere doseringen per tumorsoort.....	104
<b>Figuur 66.</b> Absolute verschil in proporties tussen de nul-dosering en de andere doseringen per tumorsoort en per geslacht. ....	105
<b>Figuur 67.</b> Totaal aantal kankergevallen per dosering per studie. ....	106
<b>Figuur 68.</b> Totaal aantal kankergevallen per studie, dosering en geslacht. ....	107
<b>Figuur 69.</b> Totaal aantal kankergevallen in verhouding tot groepsgrootte per studie, dosering en geslacht. ....	108
<b>Figuur 70.</b> Boxplot die de verdeling van p-waardes laat zien per methode én geslacht. ....	111
<b>Figuur 71.</b> Boxplot die de verdeling van p-waardes laat zien per methode, geslacht én studie. ....	112
<b>Figuur 72.</b> Boxplot die per methode het aantal significante resultaten laat zien voor $p \leq 0.05$ . Daarmee is het een visuele representatie van <b>Tabel 18</b> . De eenzijdige CA test zoals door mij uitgevoerd laat het grootste aantal significante resultaten zien. Daarna komen de resultaten van Portier: het aantal ligt tussen de eenzijdige en tweezijdige CA test. ....	112
<b>Figuur 73.</b> Boxplot die per methode en per studie het aantal significante resultaten laat zien voor $p \leq 0.05$ .....	113
<b>Figuur 74.</b> Heatmap die het aantal statistisch significante bevindingen laat zien per tumorsoort én per methode. De heatmap is zo gemaakt dat een rij alleen wordt aangemaakt als er ook maar één statistisch significant resultaat is per tumorsoort. Er is uiteraard naar meer tumorsoorten gekeken. Wat opvalt is dat de gerapporteerde Portier resultaten bijna altijd de meest significante verschillen laat zien per tumorsoort.....	113
<b>Figuur 75.</b> Dose-response relatie voor vier studies met data voor Alveolar-Bronchiolar Adenomas. Een vijfde studie ontbreekt. Zouden we deze studie toevoegen dan zou de grafiek er niet anders uitzien. Daar zit dus niet het verschil. ....	120
<b>Figuur 76.</b> Verdeling van de gerapporteerde p-waarden. De stippellijn links is de grens van statistisch significantie op 0.05.....	129
<b>Figuur 77.</b> Per studie het aantal toetsen dat onder een gespecifieerde grenswaarde viel. Dit zijn resultaten zoals door Portier gerapporteerd. ....	129
<b>Figuur 78.</b> Per studie en geslacht het aantal toetsen dat onder een gespecifieerde grenswaarde viel. Dit zijn resultaten zoals door Portier gerapporteerd. ....	130

<b>Figuur 79.</b> Verdeling van het aantal gerapporteerde p-waarden met (adjusted) of zonder (reported) correctie .....	131
<b>Figuur 80.</b> Per studie en geslacht het aantal toetsen dat onder een gespecifieerde grenswaarde viel met of zonder correctie methode. Dit zijn resultaten zoals door Portier gerapporteerd. ....	132
<b>Figuur 81.</b> Per studie en grenswaarde het aantal toetsen dat onder een gespecifieerde grenswaarde viel met of zonder correctie methode. Dit zijn resultaten zoals door Portier gerapporteerd. ....	133
<b>Figuur 82.</b> Dose-response relatie op basis van alle data. Gekeken is naar het aantal gerapporteerde tumorgevallen, los van het aantal geïncludeerde dieren. Zoals te zien valt is het haast onmogelijk om een dose-response curve te maken. ....	136
<b>Figuur 83.</b> Dose-response curve met op de x-as de dosering op de log schaal. Elke lijn is een studie. Op de Y-as is de ratio van het aantal kankergevallen per het aantal dieren per dosering. ....	137
<b>Figuur 84.</b> Dose-response curve met op de x-as de dosering op de log schaal. Elke lijn is een studie. Op de Y-as is de ratio van het aantal kankergevallen per het aantal dieren zoals geïncludeerd per dosering. Een verder onderverdeling is gemaakt per geslacht (kolom) en de duur van de studie (rijen). ....	138
<b>Figuur 85.</b> Ratio kankergevallen als functie van de dosering per studie (log schaal). Nu opgedeeld per geslacht, soort en duur van de studie. Opvallend is het aantal rechte lijnen. Ook kunnen we een aantal flinke dose-response relaties zien .....	139
<b>Figuur 86.</b> Ratio kankergevallen als functie van de dosering per studie. Elke lijn is een studie. Deze keer is de nul-dosering weggelaten. ....	140
<b>Figuur 87.</b> Ratio kankergevallen als functie van de dosering per studie. Elke lijn is een studie. Deze keer is de nul-dosering weggelaten. Opsplitsing per soort, duur van de studie en geslacht.....	140
<b>Figuur 88.</b> Resultaten van een LMM. Geslacht en soort lijken statistisch significant te zijn. ....	142
<b>Figuur 89.</b> Verdeling van restwaarden afkomstig van een LMM model. Hier wordt grotendeels aan voldaan, maar zien we op het einde tot een waaier-effect. Mogelijkerwijs is het opgeteld aantal kankergevallen niet de beste y-variabele om dosering aan te koppelen.....	143
<b>Figuur 90.</b> Voorspellingen afkomstig van een LMM model met als y-variabele het opgeteld aantal kankergevallen (sum Cases). De lijn laat de voorspelde waarde zien als functie van de dosering (log schaal), geslacht en de duur van de studie. De lijnen laten de gemiddelde voorspelling zien, de band eromheen is de onzekerheid. ....	144
<b>Figuur 91.</b> Voorspellingen afkomstig van een LMM model met als y-variabele het opgeteld aantal kankergevallen (sum Cases). De lijn laat de voorspelde waarde zien als functie van de dosering (log schaal), geslacht en soort. ....	145
<b>Figuur 92.</b> Resultaat van een LMM, gelijk aan <b>Figuur 88</b> , .....	146
<b>Figuur 93.</b> Restwaarden van een LMM met natural splines. De waaier op het einde is te groot, wat maakt dat .....	146
<b>Figuur 94.</b> Voorspellingen vanuit het model per geslacht, soort en duur. De relatie is duidelijk lineair gemodelleerd, maar de nulhypothese (coëfficiënt=0) kan niet worden vervangen. Daarvoor zijn de onzekerheidsbanden te groot. ....	147
<b>Figuur 95.</b> Een LMM met natural splines die gek gedrag laat zien. Dit is geen behulpzaam model...148	
<b>Figuur 96.</b> Resultaat van een LMM, gelijk aan <b>Figuur 88</b> , maar dan zonder de nul-dosering. ....	149
<b>Figuur 97.</b> Voorspellingen vanuit het model per geslacht, soort en duur. De nul-dosering is verwijderd. ....	149
<b>Figuur 98.</b> Resultaat van een LMM, gelijk aan <b>Figuur 92</b> , .....	150
<b>Figuur 99.</b> Non-lineaire relatie op basis van natural splines vanuit een model zonder .....	150

<b>Figuur 100.</b> Knipsel uit de studie van Portier (pagina 4). ....	152
<b>Figuur 101.</b> De verdeling van mogelijke uitkomsten bij het 50 keer opgooien van een munt afkomstig uit een binomiaal verdeling met kans 50%. ....	154
<b>Figuur 102.</b> Formules die horen bij een binomiaal verdeling. De uitkomsten zijn gebaseerd op 50 observaties en een verwachte kans van 50%. ....	155
<b>Figuur 103.</b> Proportie kankergevallen bij nul-dosering waarbij de deler het totaal aantal mogelijke observaties is. ....	156
<b>Figuur 104.</b> Proportie kankergevallen bij nul-dosering waarbij de deler het gemiddelde aantal mogelijke observaties is. ....	156
<b>Figuur 105.</b> Resultaat van een GLMM met een binomiale verdeling. ....	157
<b>Figuur 106.</b> Resultaat van een GLMM met ....	159
<b>Figuur 107.</b> Resultaat van een GLMM met een binomiale verdeling ....	160
<b>Figuur 108.</b> De voorspellingen uit het GLMM model zoals weergegeven in <b>Figuur 106</b> . ....	161
<b>Figuur 109.</b> De beoordeling van het GLMM model met binomiale verdeling. Goed te zien is de afstand tussen wat het model voorspeld aan proportie kankergevallen en wat daadwerkelijk is geobserveerd. ....	162
<b>Figuur 110.</b> De beoordeling van het GLMM model met binomiale verdeling en met splines. Dit model laat een vele sterkere correlatie zien tussen geobserveerde en voorspelde proporties kankergevallen, maar neigt sterk naar overfitting. ....	163
<b>Figuur 111.</b> De voorspelde waardes afkomstig van het GLMM model met binomiale verdeling en met splines. ....	164
<b>Figuur 112.</b> De beoordeling van het GLMM model met binomiale verdeling. De blauwe lijn toont het gemiddelde, of marginale, model. De rode lijn toont de daadwerkelijke GLMM met alle conditionele effecten. ....	165
<b>Figuur 113.</b> Voorspelde proporties aan kanker voor een GLMM model met binomiale verdeling en een interactie tussen dosering en geslacht. ....	165
<b>Figuur 114.</b> De correlatie tussen geobserveerde en voorspelde proporties kankergevallen uit een GLMM model met interactie tussen dosering en geslacht. ....	166
<b>Figuur 115.</b> Samenvatting van een GLMM model met binomiale ....	167
<b>Figuur 116.</b> Tabel 3 uit de Portier studie waarin voor CD-1 muizen studies worden gecombineerd per tumorsoort en geslacht. ....	168
<b>Figuur 117.</b> Resultaten van een incorrect GLMM model. ....	169
<b>Figuur 118.</b> De studies zoals in de data weergegeven voor mannelijke CD-1 muizen en Kidney Adenomas. Dit is de data die past bij de allereerste rij uit <b>Figuur 116</b> . ....	169
<b>Figuur 119.</b> Resultaten van een model waarin ik data heb toegevoegd. ....	170
<b>Figuur 120.</b> Tabel 4 uit de Portier studie. ....	173
<b>Figuur 121.</b> Relatie tussen dosering en de proportie Skin Basal Cell Tumors voor mannelijke Sprague-Dawley ratten. ....	174
<b>Figuur 122.</b> Tabel 5 uit de Portier studie. ....	175
<b>Figuur 123.</b> Verdeling van het gerapporteerde aantal kankergevallen. Deze verdeling lijkt sterk op de traditionele verdeling die we zien bij Poisson verdelingen. ....	176
<b>Figuur 124.</b> De frequentieverdeling bij een gemiddelde van vier vanuit een Poisson verdeling. In een Poisson verdeling staat de variantie gelijk aan het gemiddelde, dus vier. ....	177
<b>Figuur 125.</b> Bevindingen van een Poisson model met als y-variabele het opgeteld aantal kankergevallen. ....	178

<b>Figuur 126.</b> Het voorspelde aantal kankergevallen afgezet tegen het geobserveerde aantal kankermodellen op basis van een Poisson model.....	179
<b>Figuur 127.</b> Beoordeling van het Poisson model. De blauwe lijn toont het gemiddelde, of marginale, model. De rode lijn toont de daadwerkelijke GLMM met alle conditionele effecten. ....	179
<b>Figuur 128.</b> Voorspelde waarden uit het Poisson model met de interactie tussen geslacht en dosering.....	180
<b>Figuur 129.</b> Voorbeeld van een verdeling op basis van de negatieve binomiale verdeling. ....	182
<b>Figuur 130.</b> Formules die bij de negatieve binomiale verdeling horen plus de uitkomsten bij $r=20$ en $p=0.8$ .....	183
<b>Figuur 131.</b> Relatie tussen restwaarden van het model en het aantal kankergevallen.....	184
<b>Figuur 132.</b> Relatie tussen restwaarden van het model en de dosering (op de log schaal). ....	185
<b>Figuur 133.</b> Simulatie op basis van negatief binomiaal model om te bezien of het model meer nullen heeft dan we voorspellen én of het model meer variantie heeft dan we meenemen. ....	186
<b>Figuur 134.</b> Schattingen voor elke parameter in een specifiek Poisson Hurdle model. ....	190
<b>Figuur 135.</b> Schattingen voor elke parameter in een specifiek Poisson Hurdle model. ....	192
<b>Figuur 136.</b> Voorbeeld van een normaalverdeling met gemiddelde 1.03 en standaard deviatie 0.3 die zou kunnen gelden als prior voor de parameter Dosering.....	196
<b>Figuur 137.</b> Frequentieverdeling van de prior, likelihood en posterior van de Dosering parameter. ....	197
<b>Figuur 138.</b> Frequentieverdeling van de prior, likelihood en posterior van de Dosering parameter. ....	199
<b>Figuur 139.</b> Posterior check van het model. De donkere lijn is de geobserveerde waarde – de dunnen lijnen zijn trekken uit de verdelingen afkomstig van het model. ....	202
<b>Figuur 140.</b> Posterior check van het model waarbij zichtbaar wordt waar de afwijking tussen observatie en modelvoorspelling het grootst is. ....	202
<b>Figuur 141.</b> Posterior check van het model waarbij zichtbaar wordt waar de afwijking tussen observatie en modelvoorspelling het grootst is. ....	203
<b>Figuur 142.</b> De posterior verdeling van elke parameter in het model (links). Rechts zien we of de verdelingen, zoals deze tot stand zijn gekomen, op een manier is die zekerheid biedt aan de betrouwbaarheid van die verdelingen. ....	204
<b>Figuur 143.</b> Voorspellings (lijnen) en geobserveerde waarden in de relatie tussen dosering en het aantal kankergevallen. ....	204
<b>Figuur 144.</b> Voorspellings (lijnen) en geobserveerde waarden in de relatie tussen dosering en het aantal kankergevallen. In dit voorbeeld neem ik alleen de marginale (gemiddelde) parameter waarden mee. Ik neem in dit voorbeeld niet de correctie voor studie mee. ....	205
<b>Figuur 145.</b> Voorspellings en geobserveerde waarden waarbij de correctie voor studie wel wordt meegenomen. Duidelijk zichtbaar, in vergelijking met <b>Figuur 144</b> , is de toename in spreiding vanuit de schattingen.....	206
<b>Figuur 146.</b> Voorspellings voor mannen en vrouwen voor de relatie tussen dosering en aantal kankergevallen. ....	207
<b>Figuur 147.</b> Relatie tussen dosering en aantal kankergevallen, per studie en per geslacht waarbij de gekleurde lijnen de voorspellings zijn per geslacht en de zwarte lijnen en bolletjes de geobserveerde waarden zijn. Het gaat hier om totale kankergevallen per dosering per geslacht en per studie.....	207
<b>Figuur 148.</b> De posterior verdeling van de parameter Dosering voor mannen en vrouwen. ....	208
<b>Figuur 149.</b> Verdeling van het aantal kanker gevallen (boven, en verdeling van het totaal aantal kanker gevallen per studie, geslacht en dosering (onder).....	212

<b>Figuur 150.</b> De verdeling van de posterior voor een leeg hurdle model (lichte lijnen), en de verdeling van de geobserveerde tumorgevallen (donkere lijn).....	213
<b>Figuur 151.</b> De verdeling van de posterior voor een leeg hurdle model (lichte bollen), en de verdeling van de geobserveerde tumorgevallen (donkere bol). .....	213
<b>Figuur 152.</b> Grafiek die relatie laat zien tussen dosering en aantal tumorgevallen, per geslacht en studie, op de originele schaal. De geobserveerde data (zwarte lijnen) lijken meer zeker dan de modelvoorspellingen aangeven. .....	216
<b>Figuur 153.</b> Verdeling van het aantal kankergevallen bij de nul-dosering.....	218
<b>Figuur 154.</b> Drie verdelingen op basis van de geobserveerde data. Van boven naar beneden: de normaalverdeling, de gamma verdeling en de gamma-poisson verdeling. De gamma-poisson heet ook wel de negatieve binomiale verdeling en die hebben we al veelvuldig voorbij zien komen.....	219
<b>Figuur 155.</b> Verdeling van het aantal kankergevallen per geslacht en soort voor de nul-verdeling (controlegroep). .....	219
<b>Figuur 156.</b> Verdeling van aantal kankergevallen over alle doseringen heen. .....	220
<b>Figuur 157.</b> Verdeling van het aantal kankergevallen per nul-dosering en dosering waarbij dosering is samengevoegd tot één groep.....	221
<b>Figuur 158.</b> Verdeling van het aantal kankergevallen bij wel of geen dosering waarbij de y-waarde (hier de x-as) de ratio is van het totaal aantal kankergevallen gedeeld door het totaal aantal observaties. ....	221
<b>Figuur 159.</b> Verdeling van het aantal kankergevallen per geslacht en soort over de glyfosaatverdelingen heen. ....	222
<b>Figuur 160.</b> Betrouwbaarheid van elke hyperparameter in model 2.....	224
<b>Figuur 161.</b> Posterior verdeling ten opzichte van geobserveerde verdeling (donkere lijn). .....	224
<b>Figuur 162.</b> Eenzelfde figuur als <b>Figuur 161</b> maar dan platgeslagen in twee variabelen (gemiddelde en standaard deviatie). .....	225
<b>Figuur 163.</b> Kans op kanker voor de controlegroep en de behandel (Treatment) groep.....	226
<b>Figuur 164.</b> De odds-ratio voor het verschil tussen de controlegroep en de behandelgroep. ....	226
<b>Figuur 165.</b> De verdeling vanuit de posterior (dunne lijnen) en de geobserveerde verdeling. Vooral aan het eind zitten er wat deviaties. ....	229
<b>Figuur 166.</b> De posterior trekkingen van het gemiddelde en de standaarddeviatie ten opzichte van de geobserveerde getallen. ....	230
<b>Figuur 167.</b> De posterior verdelingen voor de controlegroep en de behandelgroep per soort. ....	230
<b>Figuur 168.</b> De odds-ratios voor behandeling voor elke soort.....	231
<b>Figuur 169.</b> De kans op kanker, berekend vanuit het model, per soort en per geslacht.....	232
<b>Figuur 170.</b> De kans op kanker voor mannelijke en vrouwelijke Swiss-Albino ratten in de controlegroep en de behandelgroep. Het model wijkt af met een haast vaste constante wat een gevolg is van het type model. ....	233
<b>Figuur 171.</b> Kans op kanker voor Swiss Albino ratten per geslacht – gebaseerd op de studie van Kumar die de enige studie is met Swiss Albino ratten.....	234
<b>Figuur 172.</b> Max dosis glyfosaat per studie en geslacht.....	235
<b>Figuur 173.</b> Kans op kanker voor vier studies zonder onzekerheidsmarges. ....	235
<b>Figuur 174.</b> Kans op kanker voor vier studies met onzekerheidsmarges.....	236
<b>Figuur 175.</b> Kans op kanker voor vier studies met onzekerheidsmarges én onzekerheidsbanden. ...	236

**Figuur 176.** De invloed van twee priors op de posterior. De Beta prior gaat er van uit dat behandeling de kans op kanker verhoogt. De normaalverdeling is een prior met de aanname dat glyfosaat de kans op kanker verkleint. .... 237

## Tabellen

<b>Tabel 1.</b> De gegevens zoals verkregen uit de berichtgeving van het Radboud UMC, maar dan in tabelvorm verwerkt. ....	20
<b>Tabel 2.</b> De vier waardes die ik invoer om te bepalen of de nulhypothese (200 cm) past bij een observatie van 184 (7) cm. ....	45
<b>Tabel 3.</b> De uitkomst van de one-sample t-toets. ....	46
<b>Tabel 4.</b> Resultaten van een independent samples t-test. ....	48
<b>Tabel 5.</b> De waarden zoals verkregen uit het Radbouw UMC afgezet tegen de waarden verkregen uit simulaties. Één enkele steekproef bevat 2000 simulaties: 1000 gesimuleerde observaties voor mannen en 1000 gesimuleerde observaties voor vrouwen. ....	52
<b>Tabel 6.</b> Het 95% betrouwbaarheidsinterval bij een gelijk gemiddelde, maar een steeds groter wordende spreiding. Naarmate de spreiding van één of beide groepen groter wordt zal ook het betrouwbaarheidsinterval toenemen. Hoe groter de spreiding rondom een gemiddelde hoe minder betrouwbaar is dat gemiddelde. ....	55
<b>Tabel 7.</b> Relatie tussen het verwerven of behouden van de nulhypothese in relatie tot $\alpha$ en $\beta$ . ....	61
<b>Tabel 8.</b> Betrouwbaarheidsinterval als functie van de spreiding van twee groepen en de $\alpha$ waarden. ....	70
<b>Tabel 9.</b> Gesimuleerde dekkingsgraad als functie van het aantal observaties in één steekproef en aantal herhaalde steekproeven. ....	74
<b>Tabel 10.</b> De uitkomsten met dezelfde data maar met drie verschillende alternatieve hypotheses: wel een verschil, mannen kleiner dan vrouwen en mannen groter dan vrouwen. Het betrouwbaarheidsinterval is altijd oneindig ( $\infty$ ) aan de andere kant waarvoor getoetst wordt. ....	79
<b>Tabel 11.</b> Betrouwbaarheidsinterval bij eenzijdig toetsen wanneer de $\alpha$ waarde handmatig wordt aangepast naar $0.05 / 2$ . De p-waarde blijft hetzelfde, maar het betrouwbaarheidsinterval schikt zich naar de tweezijdige toets. ....	79
<b>Tabel 12.</b> Betrouwbaarheidsinterval als een functie van de $\alpha$ waarde en de richting van de toets....	80
<b>Tabel 13.</b> De gemiddelde (en de standaard deviatie) lengte van mannen en vrouwen voor 1930, 1960, 1980 en 2001. ....	84
<b>Tabel 14.</b> Kans op een vals-positieve als functie van de $\alpha$ waarde en het aantal testen $\tau$ op dezelfde data.....	85
<b>Tabel 15.</b> p-waardes als functie van de vergelijking en de test. ....	89
<b>Tabel 16.</b> De p-waarde zoals gerapporteerd in de Portier studie voor Knezovich & Hogan, mannelijke muizen én Kidney Adenomas (original pathology) staat bovenaan. Ook rapporteren we de p-waarde van vijf verschillende testen zoals uitgevoerd in het statistiek programma R. Geen van de uitkomsten is gelijk aan de gerapporteerde p -waarde uit het artikel. ....	95

<b>Tabel 17.</b> De p-waarde zoals gerapporteerd in de Portier studie voor Knezevich & Hogan, vrouwelijke muizen én Spleen Composite Lymphosarcoma Ook rapporteren we de p-waarde van vijf verschillende testen zoals uitgevoerd in het statistiek programma R. Geen van de uitkomsten evenaart de gerapporteerde p -waarde uit het artikel.....	96
<b>Tabel 18.</b> Aantal significante resultaten, op basis van een andere grenswaarde. Elke methode heeft 183 toetsen uitgevoerd per grenswaarde. Onderaan staat het aantal statistisch significante resultaten zoals gerapporteerd door Portier. NA: not applicable oftewel de toets kan niet worden uitgevoerd. ....	111
<b>Tabel 19.</b> Aantal studies met mannelijke CD-1 muizen per tumorsoort. Wat deze tabel laat zien is dat in vijf studie waarin mannelijke CD-1 muizen werden onderzocht er in totaal 9 unieke tumorsoorten werden gezien, maar niet elke studie heeft bericht over elke tumorsoort. Dit zou wel moeten. ....	118
<b>Tabel 20.</b> Aantal tumoren per studie met mannelijke CD-1 muizen.....	119
<b>Tabel 21.</b> Aantal studies met mannelijke CD-1 muizen met data over Alveolar-Bronchiolar Adenomas. De bovenste rij is wat we kunnen opmaken uit de studie van Portier. De onderste rij als we de studie zouden toevoegen met geen incidentie. Dit verandert de ratios. ....	120
<b>Tabel 22.</b> Aantal studies met mannelijke CD-1 muizen die data hebben gerapporteerd over Kidney Adenomas (original pathology). De bovenste rij is wat we kunnen opmaken uit de studie van Portier. De onderste rij als we de studie zouden toevoegen die niks rapporteert over deze tumor. ....	121
<b>Tabel 23.</b> Aantal rijen met informatie per Geslacht / Ras / Soort.....	121
<b>Tabel 24.</b> Het aantal unieke tumoren per combinatie geslacht/ras/soort én het aantal studies waarin we deze combinatie zien. Het nut van deze getallen is dat we kunnen achterhalen hoe vaak een combinatie alle mogelijke tumorsoorten rapporteert. Bijvoorbeeld: in de vijf studies met mannelijke CD-1 muizen verwachten we dat elke studie getallen rapporteert over elk van de negen individuele tumoren. ....	122
<b>Tabel 25.</b> Aantal unieke tumoren per combinatie én in elke studie. De eerste rij laat zien dat Atkinsons 1993a mannelijke CD-1 muizen includeert. In die combinatie zijn negen unieke tumoren gerapporteerd over vijf studies heen. In de studie van Atkinson 1993a vinden we maar acht tumoren. Er ontbreekt er dus één. De studie van Stout & Ruecker rapporteert in de mannelijke Sprague-Dawley ratten 16 tumoren terwijl in die combinatie 21 unieke tumoren zijn gerapporteerd. Er ontbreken er dus zes.....	123
<b>Tabel 26.</b> De resultaten van Stout & Ruecker zoals gerapporteerd in de Supplementary Material 2 file van Pointer. Zichtbaar zijn de 5 significante vergelijkingen (dikgedrukt): drie bij de mannen en twee bij de vrouwen. ....	124
<b>Tabel 27.</b> Per tumorsoort de discrepantie tussen daadwerkelijk gerapporteerd en wat men had moeten rapporteren. Dit is alleen voor de Sprague-Dawley ratten.....	125
<b>Tabel 28.</b> De p-waarde zoals gerapporteerd in Pointer voor Skin Epithelioma (Keratoacanthomas) in Atkinson 1993b. Daaronder staat de p-waarde met de CA-test als we de resterende drie studies met nul incidentie toevoegen. De door mij berekende p-waarde verandert dan. Hierbij moet wel gesteld worden dat ik de gerapporteerde p-waarde nooit heb weten te repliceren.....	126
<b>Tabel 29.</b> Aantal significante resultaten per grenswaarde en geslacht. ....	130
<b>Tabel 30.</b> Aantal statistische testen wel of niet gecorrigeerd per grenswaarde.....	131
<b>Tabel 31.</b> De p-waarden voor mannelijke en vrouwelijke CD-1 muizen zoals gerapporteerd in Portier en zoals gevonden door het door mij gehanteerde GLMM model.....	172
<b>Tabel 32.</b> De p-waarden voor mannelijke en vrouwelijke Sprague-Dawley ratten zoals gerapporteerd in Portier en zoals gevonden door het door mij gehanteerde GLMM model.....	173

<b>Tabel 33.</b> De gevonden p-waarden door Portier en door het GLMM model wat ik heb gebruikt. Het gaat hier om mannelijke Wistar ratten. ....	175
<b>Tabel 34.</b> Classificatie van de Likelihood Ratio (LR) in termen van bewijskracht. ....	209
<b>Tabel 35.</b> Vergelijking tussen vijf modellen op basis van de Likelihood Ratio (LR). Model 3 is het model wat het best wordt verklaard gegeven de data en de priors. ....	210
<b>Tabel 36.</b> Bewijs voor de hypothese dat de coëfficiënt voor Dosering exact nul is, kleiner dan nul of groter dan nul. De grootste bewijskracht ligt voor de hypothese dat de coëfficiënt exact 0 is, gegeven de priors en de data.....	210
<b>Tabel 37.</b> De Likelihood Ratio (LR) matrix voor 7 modellen. Voor Model 6 is er het meeste bewijs, hoewel Model 6 en Model 3 haast evenveel bewijs krijgen toegekend. Dat betekent dat op basis van de data het maar lastig blijft wat de rol van de dosering is. ....	214
<b>Tabel 38.</b> Bewijs voor de hypothese dat de coëfficiënt voor Dosering exact nul is, kleiner dan nul of groter dan nul. De grootste bewijskracht ligt voor de hypothese dat de coëfficiënt groter dan 0 is, maar het bewijs is zwak.....	215
<b>Tabel 39.</b> De LR voor model 1 en model 2. ....	223
<b>Tabel 40.</b> De LR voor de hypothese dat de coëfficiënt voor de behandeling nul is, kleiner dan nul of groter dan nul. ....	227
<b>Tabel 41.</b> De LR-waarden voor de 8 modellen zoals hierboven beschreven. ....	229
<b>Tabel 42.</b> De LR voor de hypothese dat de coëfficiënt voor de behandeling nul is, kleiner dan nul of groter dan nul in Swiss-Albino ratten.....	231
<b>Tabel 43.</b> De kans op kanker voor Swiss-Albino ratten per groep en per geslacht.....	232