COMMISSIE
COGEM
GENETISCHE
MODIFICATIE

BEZOEKADRES:
A. VAN LEEUWENHOEKLAAN 9
3721 MA BILTHOVEN

POSTADRES:
POSTBUS 578
3720 AN BILTHOVEN

TEL.: 030 274 2777
FAX: 030 274 4476
INFO@COGEM.NET
WWW.COGEM.NET

Aan de staatssecretaris van
Infrastructuur en Milieu
Mevrouw S.A.M. Dijksma
Postbus 20901
2500 EX Den Haag

DATUM 09 december 2015
KENMERK CGM/151209-01
ONDERWERP Signalerende brief 'Statistiek bij de risicobeoordeling van gg-gewassen'

Geachte mevrouw Dijksma,

Statistiek is een essentieel hulpmiddel bij de analyse en interpretatie van onderzoeksresultaten. Statistische analyses hebben echter onbedoeld ook de status gekregen van keurmerk voor de correctheid van de conclusies die onderzoekers aan hun onderzoek verbinden. Daardoor wordt bij omstreden onderwerpen als genetische modificatie de manier waarop de gegevens statistisch zijn geanalyseerd, onderwerp van discussie. Bij deze discussie spelen verschillende elementen een rol, zoals of de gebruikte analysemethode de juiste is, of het aantal geteste organismen voldoende is om een statistische analyse te kunnen uitvoeren, of een statistisch verband van belang is en of op basis van de resultaten van de statistische analyse geconcludeerd kan worden dat er sprake is van een risico.

Omdat ook de COGEM bij het beoordelen van de risico's van genetisch gemodificeerde (gg-) gewassen gebruik maakt van de uitkomsten van statistische analyses, wil zij met deze brief toelichten hoe statistiek bij de risicobeoordeling van gg-gewassen wordt gebruikt en wat de beperkingen zijn van het gebruik van statistiek. Verduidelijking van de rol van statistiek bij het beoordelen van risico's is van belang voor een transparante en voor iedereen inzichtelijke risico-beoordeling.

Belang van statistiek voor biologisch onderzoek

Wanneer er onderzoek wordt uitgevoerd met levende organismen, zoals dieren of planten, of met materiaal van biologische oorsprong, moet er rekening worden gehouden met de natuurlijke variatie tussen organismen in een populatie. Omdat experimenten noodzakelijkerwijs met beperkte aantallen organismen worden uitgevoerd, worden de onderzoeksresultaten door de aanwezige natuurlijke variatie beïnvloed. Verschillen in levensduur, gewicht, gevoeligheid voor ziekten e.d. zijn bronnen van natuurlijke variatie. Ook kan er door verschillende

proefomstandigheden, bijvoorbeeld de plek van proefdieren binnen een dierverblijf of de locatie van een ‘plot’ bij een veldproef, variatie ontstaan.

Natuurlijke variatie zorgt voor ‘ruis’ en bemoeilijkt het ontdekken van daadwerkelijke verschillen tussen behandelingen of groepen. Bij het uitvoeren van experimenten wordt geprobeerd om de natuurlijke variatie zoveel mogelijk te verkleinen. Dit gebeurt ondermeer door organismen te gebruiken van dezelfde oorsprong en leeftijd, door onderzoek onder gelijke omstandigheden uit te voeren en door het onderzoek een aantal keren te herhalen. Daarnaast wordt door randomisatie, waarbij organismen willekeurig over verschillende behandelingen of locaties worden verdeeld, de natuurlijke variatie tussen behandelingen of groepen genivelleerd. Ondanks deze maatregelen zal in de onderzoeksresultaten altijd enige natuurlijke variatie tot uiting komen.

Vanwege deze natuurlijke variatie is statistiek onmisbaar bij onderzoek met biologisch materiaal. Met behulp van statistiek kan geanalyseerd worden of een verschil uitstijgt boven de natuurlijke variatie. Om hier uitspraken over te kunnen doen, is het wel van belang dat een experiment met een voldoende groot aantal experimentele eenheden (individuen, plots) wordt uitgevoerd en dat het meerdere keren wordt herhaald.

Er zijn vele verschillende statistische methoden die gebruikt kunnen worden om gegevens te analyseren. Op dit moment worden gegevens vrijwel altijd m.b.v. klassieke statistiek geanalyseerd, maar er is toenemende belangstelling voor de Bayesiaanse statistiek waarbij in tegenstelling tot de klassieke statistiek ook voorkennis (‘prior informatie’) bij de analyse wordt betrokken. De keuze voor een bepaalde statistische methode en de manier waarop de statistische analyse is uitgevoerd, zijn terugkerende elementen in de discussie die wordt gevoerd na publicatie van alarmerende resultaten bij studies met gg-gewassen.[1] Een ander terugkerend element in de discussie is de interpretatie van de uitkomsten van statistische analyses (met name de zogenaamde p-waarden). Wanneer de resultaten van de statistische analyses verkeerd worden geïnterpreteerd, kan dit tot verkeerde conclusies leiden.[2,3] Daarnaast is ook de relevantie van statistisch significante verschillen een onderwerp van discussie. Statistisch significante verschillen zijn niet altijd relevant voor de risicobeoordeling. Zo kan het statistisch significante verschil een eigenschap betreffen die niet van invloed is op een mogelijk risico. In dat geval zal het statistisch significante verschil de uitkomst van de risicobeoordeling niet veranderen.

Hoe wordt statistiek gebruikt bij het beoordelen van risico’s van gg-gewassen?

Bij de risicobeoordeling van gg-gewassen wordt gebruik gemaakt van statistiek om te analyseren of er verschillen zijn tussen het gg-gewas en de controle (de isogene of ouderlijn). Omdat de aanwezigheid van een statistisch significant verschil niet altijd betekent dat er ook sprake is van een risico, wordt bij significante verschillen tussen het gg-gewas en de controle nagegaan of het waargenomen verschil op een risico zou kunnen wijzen.

Het uitvoeren van een statistische analyse en de interpretatie van de resultaten

De statistische analyse begint met het formuleren van een uitgangshypothese (de zogenaamde 'nulhypothese') en een alternatieve hypothese. Ten behoeve van de risicobeoordeling van gg-gewassen worden normaal gesproken twee verschillende statistische toetsen uitgevoerd. Er wordt getoetst of het gg-gewas verschilt van de controle ('difference testing') en er wordt getoetst of het gg-gewas overeenkomt met de controle en verschillende conventionele rassen ('equivalence testing').[4] Deze twee statistische toetsen zijn complementair. Door de gegevens zowel met 'difference testing' als 'equivalence testing' te analyseren, wordt inzichtelijk hoe het gg-gewas zich tot de controle en tot conventionele rassen verhoudt.

Na het formuleren van de nulhypothese en de alternatieve hypothese worden de resultaten van de uitgevoerde experimenten, statistisch geanalyseerd. Er wordt getoetst wat de kans op het gevonden verschil is wanneer de nulhypothese waar is. In het geval van 'difference testing' betekent dit dat wordt berekend wat de kans op het gevonden verschil is wanneer het gg-gewas en de controle niet van elkaar verschillen. Deze zogenaamde overschrijdingskans wordt weergegeven met een p-waarde, die tussen de 0 en de 1 ligt. Een p-waarde van 0,05 betekent dat er 5% kans is dat het waargenomen verschil op toeval berust.

Meestal wordt de nulhypothese verworpen wanneer de p-waarde lager is dan 0,05. Bij een p-waarde hoger dan 0,05 wordt bij 'difference testing' (nulhypothese: geen verschil) gezegd dat er geen statistisch significant verschil is, terwijl bij een p-waarde lager dan 0,05 wordt gesteld dat er sprake is van een statistisch significant verschil. Bij 'equivalence testing' (nulhypothese: niet gelijkwaardig) wordt bij een p-waarde lager dan 0,05 juist gesteld dat er sprake is van een statistisch significante overeenkomst.

De grens van 0,05 voor statistische significantie is een compromis tussen het aantal herhalingen dat bij een onderzoek nodig is en de kans op fout-positieve resultaten. Om lagere p-waarden en dus een hogere betrouwbaarheid van het verkregen resultaat te verkrijgen, is een groter aantal herhalingen nodig. Het aantal herhalingen dat nodig is om de p-waarde steeds verder te verlagen stijgt echter snel. Wanneer een hogere p-waarde wordt gebruikt om de nulhypothese te verwerpen zijn minder herhalingen noodzakelijk om een verschil aan te tonen en zullen er bij eenzelfde aantal herhalingen meer statistisch significante verschillen worden gevonden, maar zal ook de kans op fout-positieve resultaten stijgen. Hoewel sommigen ervoor pleiten om een lagere p-waarde grens (bijvoorbeeld 0,005) te gebruiken zodat de kans op fout-positieve resultaten afneemt,[5] wijzen anderen erop dat dit bij een gelijkblijvend aantal herhalingen betekent dat de kans op fout-negatieve resultaten toeneemt.[6] Een p-waarde van 0,05 wordt op dit moment beschouwd als het beste compromis tussen het aantal benodigde herhalingen en de kans op fout-positieve en fout-negatieve resultaten.

De drempel van 0,05 voor statistische significantie is volledig ingeburgerd en wordt gebruikt als grens tussen belangrijke en niet-belangrijke resultaten. In de loop van de tijd is het belang dat

aan de p-waarde wordt gehecht gestegen. Een p-waarde van minder dan 0,05 wordt beschouwd als bewijs dat er echt een verschil bestaat tussen de onderzochte groepen.

Aandachtspunten bij de interpretatie van resultaten van statistische analyses

Er is echter een aantal aspecten waar bij de interpretatie van statistische analyses rekening mee gehouden moet worden. Overigens zijn deze aspecten niet alleen relevant bij het beoordelen van risico's van gg-gewassen, maar ook bij andere situaties waar de resultaten van statistische analyses gebruikt worden.

1) Toevallige statistische significantie

Eén aspect is dat een verschil dat aan de voorwaarden voor statistisch significantie voldoet er in werkelijkheid niet hoeft te zijn. Bij een p-waarde lager dan 0,05 wordt er in het geval van 'difference testing' gesteld dat er een statistisch significant verschil is gevonden. De p-waarde geeft echter de kans dat de gevonden resultaten worden verkregen wanneer de nulhypothese waar is. In het geval van 'difference testing' betekent een p-waarde van 0,04 dat er 4% kans is dat de waargenomen verschillen op toeval berusten en er in werkelijkheid geen verschil bestaat.

Dit aspect is vooral een aandachtspunt wanneer er binnen één onderzoek veel verschillende gegevens (steekproeven, waarnemingen) die onderling gerelateerd zijn, statistisch geanalyseerd worden. De kans dat bij veel verschillende metingen of bepalingen in een experiment één van de metingen onterecht als significant verschillend wordt aangemerkt, neemt namelijk sterk toe.
Bij een p-waarde van 0,04 is de kans dat een meting significant is omdat de groepen echt van elkaar verschillen 0,96. Bij 10 verschillende statistisch significante metingen is de kans dat er bij alle 10 de statistisch significante metingen ook daadwerkelijk een verschil bestaat tussen de groepen $0{,}96^{10}=0{,}66$ en dus 66%. Er is dan 34% kans dat één of meerdere van de metingen onterecht als significant verschillend wordt aangemerkt.

2) Significantie geeft geen informatie over de grootte van het verschil

Een tweede aspect is dat bij het bepalen van statistische significantie de grootte van het verschil geen rol speelt. Enerzijds kunnen zeer kleine verschillen statistisch significant zijn wanneer er weinig natuurlijke variatie is of wanneer ze consistent in vele herhalingen worden teruggevonden. Anderzijds zijn bij weinig herhalingen of veel natuurlijke variatie zelfs grote verschillen niet statistisch significant.
Wanneer er bijvoorbeeld met duizenden proefpersonen wordt onderzocht of er een verschil bestaat tussen de lengte van mensen die in 1970 of 1990 geboren zijn, dan kan een verschil van 0,5 centimeter al statistisch significant zijn. Wanneer ditzelfde onderzoek met slechts tien proefpersonen wordt uitgevoerd, dan zal zelfs een verschil van tien centimeter waarschijnlijk niet als statistisch significant uit het onderzoek naar voren komen.

Om ervoor te zorgen dat het onderscheidend vermogen ('statistical power') van een onderzoek groot genoeg is en een onderzoek in staat is om een eventueel verschil statistisch significant te

detecteren, worden er eisen gesteld aan het minimale aantal herhalingen.[7,8] Het onderscheidend vermogen van een onderzoek wordt naast het aantal herhalingen ook bepaald door het minimale effect dat het onderzoek moet kunnen aantonen.
Om het belang van een statistisch significant verschil op waarde te kunnen schatten, is het van belang dat naast informatie over statistische significantie ook bekend is hoe groot het aangetoonde verschil is (de effect-grootte).

3) Statistisch significant betekent niet automatisch een milieurisico
Een derde aspect is dat een statistisch significant verschil kan wijzen op een milieurisico, maar ook van geen betekenis kan zijn voor de risicobeoordeling. Hiervoor zijn een aantal redenen:

a. Het statistisch significante verschil is geen werkelijk verschil, maar wordt onterecht als een verschil gerapporteerd als gevolg van de foutmarge die bij de statistische analyse wordt geaccepteerd (zie 1).
b. Het statistisch significante verband heeft geen oorzakelijke achtergrond. Bij statistische analyses kan er een verband worden gevonden dat er in werkelijkheid niet is. Zo is er bijvoorbeeld een statistisch significant verband tussen de gemiddelde consumptie van kaas en het aantal mensen dat in hun lakens verstrikt raakt.[9] Er is echter geen oorzakelijk verband tussen deze gegevens.
c. Het statistisch significante verband wordt veroorzaakt door een derde factor die beide andere factoren beïnvloedt. Bij een correlatie tussen inkomen en het voorkomen van kanker zou leeftijd zo'n derde factor kunnen zijn. Oudere mensen verdienen meestal meer, maar krijgen ook vaker kanker.[10]
d. Er is een statistisch significant verschil, maar dit verschil is slechts zeer klein (zie 2). Een zeer klein verschil zal vaak geen gevolgen hebben voor de uitkomst van de risicobeoordeling. Dit is echter niet altijd het geval. Ook bij een zeer klein statistisch significant verschil moet daarom (net als bij een groot statistisch significant verschil) altijd door deskundigen geanalyseerd worden of het op een risico zou kunnen wijzen.

4) Data bewerking kan resultaat statistische analyse beïnvloeden
Een vierde aspect dat een punt van aandacht is bij de interpretatie van statistische analyses, is een fenomeen dat bekend staat als 'p-hacking'. Door de gegevens op een bepaalde manier te bewerken, bijvoorbeeld door bepaalde gegevens wel of juist niet in de analyse mee te nemen, door bepaalde gegevens wel of niet samen te voegen of door het aantal herhalingen tijdens het onderzoek aan te passen, kan op een gewenste p-waarde worden aangestuurd.[11] Het is daarom van belang dat bij bestudering van de uitgevoerde statistische analyses ook gekeken wordt naar de bewerkingen die op de gegevens zijn uitgevoerd en naar de oorspronkelijke gegevens zelf.

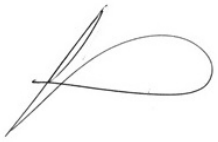
Hoe gaat de COGEM om met resultaten van statistische analyses?
Statistiek is een essentieel hulpmiddel bij de risicobeoordeling. Er zijn vele verschillende statistische methoden die elk onder bepaalde omstandigheden geschikt zijn voor de analyse van een bepaald type onderzoek. Het is belangrijk dat een methode wordt gekozen die geschikt is

voor de gebruikte proefopzet, het type gegevens, etc. De COGEM zal daarom altijd nagaan of de gebruikte statistische methode inderdaad geschikt is voor de proefopzet, het type gegevens en of voldaan wordt aan de voorwaarden voor het gebruik van de gekozen statistische methode. Zij zal hierbij onder andere gebruik maken van het onderzoek dat zij eerder heeft laten uitvoeren naar statistische methoden om gegevens van veldproeven te analyseren.[12]

In de wetenschappelijke literatuur wordt gediscussieerd over de waarde van statistische analyses. Vanwege de problemen die met het gebruik van p-waarden zijn geassocieerd (verkeerde interpretatie, geen informatie over de grootte van het effect, 'p-hacking', etc.) [2], wordt door sommige wetenschappers gesteld dat p-waarden, maar ook andere resultaten van statistische analyses met vergelijkbare problemen (betrouwbaarheidsintervallen, t-waarden, f-waarden etc.) niet meer gerapporteerd zouden moeten worden.[13]
De COGEM is echter van mening dat deze statistische analyses ondanks hun tekortkomingen nog steeds waardevolle informatie bieden. De COGEM vindt dat naast de resultaten van statistische analyses ook informatie moet worden verstrekt over de effect-grootte waarop getoetst is, het waargenomen gemiddelde, de standaardafwijking en het aantal herhalingen. Deze informatie is van belang voor de interpretatie van eventuele statistisch significante verschillen. Wanneer er statistisch significante verschillen of verbanden worden gevonden, zal de COGEM vanwege de eerder genoemde redenen verder onderzoeken of dit verschil op een milieurisico zou kunnen wijzen.

Omdat de gebruikte statistische analyse een terugkerend element is in de discussies die worden gevoerd na publicatie van alarmerende resultaten bij studies naar effecten van ggo's, heeft de COGEM met deze brief de aandacht willen vestigen op het belang én de beperkingen van statistiek. Ook heeft zij willen verduidelijken hoe zij bij de door haar uitgevoerde risicobeoordeling gebruik maakt van de resultaten van statistische analyses. De COGEM signaleert dat de resultaten van statistische analyses voor de risicobeoordeling geen eindpunt zijn, maar juist een startpunt voor verdere analyse en oordeelsvorming.

Hoogachtend,



Prof. dr. ing. Sybe Schaap
Voorzitter COGEM

c.c. Drs. H.P. de Wijs, Hoofd Bureau ggo
Mr. J.K.B.H. Kwisthout, Ministerie van IenM

Referenties

1. COGEM (2013). Waar rook is, is vuur? Omgaan met de uitkomsten van alarmerende studies over de veiligheid van ggo's. COGEM signalering CGM/131031-01
2. Nuzzo R (2014). Statistical errors. Nature. 506: 150-152
3. Woolston C (2015). Psychology journal bans P values. Nature. 519: 9
4. EFSA GMO Panel (2010). Scientific Opinion. Guidance on the environmental risk assessment of genetically modified plants. EFSA Journal 8(11): 1879
5. Johnson VE (2013). Revised standards for statistical evidence. Proc Natl Acad Sci U S A. 110(48): 19313-19317
6. Pericchi L *et al.* (2014). Adaptive revised standards for statistical evidence. Proc Natl Acad Sci U S A. 111(19): E1935
7. OECD (2009). Guidelines for the testing of chemicals. Test no. 453: Combined chronic toxicity\carcinogenicity studies. www.oecd-ilibrary.org/content/book/9789264071223-en (bezocht: 1 december 2015)
8. EFSA (2010). Scientific opinion on statistical considerations for the safety evaluation of GMOs. EFSA Journal 8(10): 1250
9. Vigen T (2015). Spurious correlations. www.tylervigen.com/spurious-correlations (bezocht: 24 november 2015)
10. MacKinnon DP *et al.* (2000). Equivalence of the mediation, confounding and suppression effect. Prev Sci. 1(4): 173
11. Simmons JP *et al.* (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 22(11):1359-1366
12. Semenov AV (2012). The use of statistical tools in field testing for effects of genetically-modified (GM) plants on non-target organisms (NTO) COGEM onderzoeksrapport CGM 2012-06
13. Trafimov D & Marks M (2015). Editorial, Basic Appl Soc Psych. 37(1): 1-2