S 15

J. Kuhlmeijk.

W
IA

Rapport over onderzoek van kakkerlakken.

1. Het probleem:

Het volgende probleem werd door Dr. Drion, T.N.O., Bewerking Waarnemingsuitkomsten, aan het MC voorgelegd.

N kakkerlakken worden ingespoten met een bepaalde, voor alle gelijke, hoeveelheid gif (g gram) en gewogen. Noem het gewicht van de kakkerlakken $\underline{h}$ ($\underline{h}$ is een stochastische grootheid, die op de collectie waaruit de N kakkerlakken genomen zijn een bepaalde verdeling bezit). Als maat voor de toegediende gifhoeveelheid wordt, op biologische gronden, genomen: $\underline{x} = \frac{g}{\underline{h}}$ (noem dit de gifmaat). Een gedeelte van de kakkerlakken sterft, een gedeelte blijft in leven. Ondersteld wordt, dat iedere kakkerlak een bepaalde gifmaat juist kan verdragen en bij toediening van een grotere hoeveelheid sterft. Deze "kritieke gifmaat" geven we aan met $\underline{y}$, een op de kakkerlakken-collectie stochastische grootheid. Gemeten worden: de N toegediende gifmaten $x_1 \ldots\ldots x_N$ en bij ieder daarvan wordt genoteerd: 1 (voor overlevend) of 0 (voor overleden). Gevraagd wordt, wat met zo weinig mogelijk onderstellingen uit een dergelijke proef geconcludeerd kan worden omtrent de verdeling van $\underline{y}$, i.h.b. wordt gezocht naar een vertrouwens-interval voor de 90% en 99%-quantilen.

2. Onderstellingen.

Het probleem is in wezen een probleem van de "probit-analysis". Meestal wordt dan de onderstelling gemaakt, dat $\underline{y}$ logarithmisch normaal verdeeld is. Daar de methode er dan toe leidt, dat tot in de staart van deze verdeling geëxtrapoleerd moet worden, is het niet duidelijk in hoeverre de uitkomst betrouwbaar is. Daarom is het volgens Dr.Drion wenselijk deze onderstelling niet te gebruiken. Wij zullen dus trachten iets te bereiken met uitsluitend de onderstelling, dat $\underline{y}$ een continue verdeling bezit.

3. Een mogelijke methode.

Indien wij beschikken over een steekproef van N waarnemingen van $\underline{y}$, zouden wij zonder moeite betrouwbaarheidsintervallen voor de quantilen kunnen berekenen. Dit is echter niet het geval. Wij beschikken slechts over de waarden $x_1 \ldots\ldots x_n$ en over de kennis, dat

$y_i \geq x_i$ is, als x_i het kenmerk 1 draagt en
$y_i \leq x_i$ is, " x_i " " 0 " .

Zij nu $\underline{X}$ een willekeurig getal en zij:

n_1 het aantal waarden x_i met $x_i \geq \underline{X}$ een kenmerk 1 en
n_0 " " " x_i " $x_i \leq \underline{X}$ " " 0 .

Dan zijn er minstens n_1 y-waarden rechts van $\mathfrak{X}$ en minstens n_0 links.

Noemen wij de verdelingsfct. van $\underline{y}$: $F_{\underline{y}}$ ($F_{\underline{y}}(y)$ is dan de kans, dat een willekeurig gekozen kakkerlak bij gifmaat $\leq y$ in leven zal blijven) en noemen we $F_{\underline{y}}(\mathfrak{X}) = p$, dan is de kans op het vinden van minstens n_1 y-waarden rechts van $\mathfrak{X}$ gelijk aan

$$\varphi = \sum_{n_1}^{N} \binom{N}{\nu} p^{N-\nu} q^{\nu}$$

en de kans op het vinden van minstens n_0 y-waarden links van $\mathfrak{X}$ is

$$\psi = \sum_{n_0}^{N} \binom{N}{\nu} p^{\nu} q^{N-\nu}$$

Zij nu: $\underline{B}_x = \underset{\varphi \leq \frac{\varepsilon}{2}}{\mathrm{Min}} \{1 - F_{\underline{y}}(\mathfrak{X})\}$ en $\underline{A}_x = \underset{\psi \leq \frac{\varepsilon}{2}}{\mathrm{Max}} F_{\underline{y}}(\mathfrak{X})$

(waarin dus $\underline{A}_x$ en $\underline{B}_x$ stochastisch zijn, daar de getallen n_1 en n_0 dit zijn) dan is:

$$P\left[\underline{A}_x \leqq F_{\underline{y}}(\mathfrak{X})\right] = 1 - \frac{\varepsilon}{2} \quad \text{en} \quad P\left[\underline{B}_x \geqq F_{\underline{y}}(\mathfrak{X})\right] = 1 - \frac{\varepsilon}{2} \,,$$

dus $P\left[\underline{A}_x \leqq F_{\underline{y}}(\mathfrak{X}) \leqq \underline{B}_x\right] = 1 - \varepsilon$, zoals gemakkelijk na te rekenen is.

We vinden derhalve een betrouwbaarheidsstrook voor $F_{\underline{y}}$ en mogen zoals bij betrouwbaarheidsgrenzen steeds het geval is, bij een gegeven $F_{\underline{y}}$ daaruit het betrouwbaarheidsinterval voor de bijbehorende $\underline{y}$-waarde aflezen.

A_x (de ondergrens) maakt een sprong naar boven indien een punt x_i met kenmerk 0 (van links naar rechts gaande) gepasseerd wordt; B_x (de bovengrens) als een punt x_i met kenmerk 1 gepasseerd wordt. Overigens zijn zij constant. $A_x = 0$ voor $x = -$ tot de eerste x_i met kenmerk 0 en $B_x = 1$ vanaf de laatste x_i met kenmerk 1 tot $x = +\infty$.

Deze grenzen A_x en B_x kunnen wellicht nog verscherpt worden, daar men ook wel iets kan zeggen over het aantal y-waarden tussen twee getallen $\mathfrak{X}$ en $\mathfrak{Y}$ in. Dit is hier niet nader uitgewerkt, daar dit rapport slechts ter oriëntering dient; bovendien is de hier aangeduide methode niet vrij van bezwaren, zoals in het volgende punt besproken wordt.

4. Bezwaren tegen de methode.

Deze methode zal in het algemeen zeer onscherp zijn, als gevolg van het feit, dat vóor de y_i slechts boven- en ondergrenzen bekend zijn. Voor hoge quantilen (b.v. 99%) zal men slechts als N zeer groot is een bovengrens kunnen vinden; een ondergrens vindt men al bij kleinere N.

Een bezwaar van meer principiële aard is, dat, indien de proef wordt ingericht als onder punt 1 beschreven, de lichte kakkerlakken alle een grote en de zware een kleine gifmaat krijgen toegediend, zodat de gifmaat niet toevallig verdeeld is met

betrekking tot het gewicht. Op die wijze sluipt impliciet de onderstelling in, dat de verdeling van y onafhankelijk is van die van het gewicht, dus, bij constante g, van die van x. Immers de berekening in het vorige punt is uitgevoerd met een zelfde $F_{y}(x)$ voor lichte en zware kakkerlakken.Het is echter de vraag of deze onderstelling gerechtvaardigd is. Men kan deze onderstelling uitschakelen door b. de kakkerlakken van tevoren te wegen en in een aantal gewichtsklassen van niet te grote lengte in te delen en dan de kakkerlakken in te spuiten met een gifhoeveelheid, die ~~omgekeerd~~ evenredig is met het gemiddelde van hun gewichtsklasse. Dan wordt de gifmaat op toevallige wijze over de dieren verspreid en bovendien bereikt men, dat de waarden $x_1, \ldots\ldots, x_N$ niet ver uit elkaar liggen, hetgeen het resultaat aanzienlijk verscherpt. (Het opeendringen van de x-waarden door uitsluitend kakkerlakken van één gewichtsklasse in te spuiten, is slechts zinvol indien de reeds genoemde onderstelling van onafhankelijkheid van y en gewicht (h) wordt ingevoerd, maar dit kan dus vermeden worden).
Indien men het experiment op deze wijze uitvoert kan men bovendien zorgen, dat alle x-waarden komen te liggen in de buurt van de plaats waar men het gezochte quantilepunt vermoedt. In dat geval is een behoorlijke kans op het vinden van een niet te groot betrouwbaarheidsinterval niet onaanneemlijk. Men zou bovendien gemakkelijk een methode kunnen ontwikkelen, waarbij men doorgaat met de proefneming (in groepjes van 10 b.v.) tot men een voor het doel voldoend klein interval gevonden heeft. Daarbij doet men dan goed ieder nieuw groepje proefdieren uit verschillende gewichtsklassen te kiezen, zodat ieder nieuw groepje als een "random sample" uit de totale kakkerlakkencollectie kan worden beschouwd.
Voor de voorlopige schatting van het gezochte quantile-punt kan men bovenstaand procédé gebruiken, indien al een experiment is uitgevoerd. Daarbij moet men dan de onafhankelijkheid van h en y maar aannemen, hetgeen voor een voorlopige schatting geen groot bezwaar is.

5. Een andere methode.

In het bovenstaande is een methode aangegeven om iets te weten te komen over de verdeling van de y, de kritieke gifmaat, die juist nodig is om een kakkerlak te doden. In deze verdeling is de gewichtsverdeling van de kakkerlakken verwerkt, d.w.z. indien men deze verandert (b.v. door selectie) is men niet zeker, of de F_{y} niet eveneens verandert. Men kan het probleem echter ook anders stellen, door de vraagstelling als volgt te formuleren:

Hoeveel gif moet een willekeurig gekozen kakkerlak toegediend worden, opdat de kans, dat hij sterft gelijk aan een gegeven p is. Dit komt ook meer overeen met de bij het onderzoek van insecticiden vaak toegepaste methode om als maat voor de toegediende gifhoeveelheid de concentratie (of de logarithme daarvan) te nemen, vgl. b.v. D.J.Finney, Probit Analysis (1947) Cambr.Univ.Press.

Men vraagt dan dus naar de p-quantile van $\underline{g}^*$, (en niet van de kritieke gif<u>maat</u>), waarbij $\underline{g}^*$ de <u>kritieke gifhoeveelheid</u> voorstelt, die, evenals $\underline{y}$, stochastisch over de kakkerlakken verdeeld is.

~~Deze~~ Van g kan men zonder moeite een groot aantal malen dezelfde waarde geven en men kan dan, daar men nu met het alternatief levend of dood te maken heeft, bij de gekozen g-waarde betrouwbaarheidsgrenzen voor de $F_{\underline{g}^*}(g)$ bepalen. Deze $F_{\underline{g}^*}(g)$ stelt de genoemde kans voor.

Indien men nu een betrouwbaarheidsinterval van bepaalde lengte wenst, waarbij de lengte zo gekozen wordt, dat men tevreden is met de gekozen waarde van g als $F_{\underline{g}^*}(g)$ tussen p en $p - \varepsilon$ ligt, kan men de methode der sequential analysis van Bernard en Wald toepassen. De kakkerlakken komen dan overeen met een partij goederen, die volgens een bepaald criterium (dat overeenkomt met al of niet de gifhoeveelheid g overleven) wordt gekeurd. Een dode kakkerlak komt overeen met een goedgekeurd product en een overlevende met een afgekeurd. Wordt de partij nu volgens Wald afgekeurd, dan is de voor g genomen hoeveelheid te klein, wordt hij goedgekeurd, dan kan hij misschien kleiner genomen worden. Deze methode heeft het grote voordeel geen verdere onderstellingen dan de continuïteit van de verdeling van $\underline{g}^*$ nodig te hebben en exact en willekeurig scherp te zijn, terwijl er een relatief klein aantal proeven voor nodig is. Bovendien heeft men een bijzonder fraaie beheersing over de foutenkansen. (Zou men deze methode echter willen toepassen op de $\underline{y}$ in plaats van de $\underline{g}^*$, dan zou men iedere kakkerlak moeten inspuiten met een hoeveelheid ~~omgekeerd~~ evenredig met zijn gewicht, hetgeen waarschijnlijk bezwaarlijk is).

6. <u>Opmerkingen.</u>

Deze laatste methode is niet nader uitgewerkt, aangezien het ons niet bekend is, of zij zinvol is. Bovendien is zij door Wald in zijn boek "Sequential Analysis" uitvoerig beschreven. Vermoedelijk is het de bedoeling een bepaald verdelgingsmiddel voor kakkerlakken te toetsen of verschillende met elkaar te vergelijken. De vraag of $\underline{y}$ of $\underline{g}^*$ de geschiktere variabele is zal er nu van afhangen of de hoeveelheid gif, die een kakkerlak bij deze verdelgingspoging opneemt, afhankelijk is van zijn gewicht, of niet. Is dit niet het geval, dan is de in punt 5 aangegeven methode de aangewezene, anders die van punt 3 met de in punt 4 aangegeven modificatie, tenminste, als over deze samenhang tussen gewicht en gifopname iets bekend is. Deze vragen zouden echter van tevoren met de onderzoeker besproken moeten worden.

Het is waarschijnlijk mogelijk de in punt 5 genoemde methode met de in 3 en 4 beschrevene te combineren. De sequential test kan waarschijnlijk zeer voordelig zijn (vooral wat het aantal proeven betreft) om een voorlopige schatting van het gezochte quantile-punt van de y te verkrijgen. Voor deze schatting n.l. is het niet nodig, dat alle toegediende gifmaten precies gelijk zijn. Bovendien zou men eventueel kunnen trachten de bovengrens van een hoog quantile, die met de eerste methode niet gemakkelijk te bepalen is, met de laatstgenoemde te verscherpen, door een speciale sequential test uit te voeren, waarbij men de gewichtsklassen wat extra smal neemt. Het is, zoals aan het eind van §3 reeds is opgemerkt, zeer wel mogelijk dat de in punt 3 en 4 beschreven onscherpe methode verbeterd kan worden zonder dat veel extra onderstellingen nodig zijn.
