Wiskundige statistiek in het landbouwkundig onderzoek

S. H. JUSTESEN,
Centrum voor Landbouwwiskunde, Wageningen

Overdruk uit het Landbouwkundig Tijdschrift
72ste jaargang no. 1, januari 1960

Wiskundige statistiek in het landbouwkundig onderzoek

S. H. JUSTESEN,
Centrum voor Landbouwwiskunde, Wageningen

In dit artikel laat schrijver de voornaamste statistische technieken de revue passeren. Een overzicht wordt gegeven van de mogelijkheden die het Centrum voor Landbouwwiskunde en de Afd. Bewerking Waarnemingsuitkomsten bieden. Het is de bedoeling dat de onderzoeker een indruk krijgt van de methoden die tot zijn beschikking staan en van de rekenhulpmiddelen waarvan hij gebruik kan maken.

Bij het landbouwkundig onderzoek vinden statistische methoden hoe langer hoe meer toepassing. Bij het vorderen van het onderzoek wordt het steeds meer noodzakelijk „op de kleintjes te letten", d.w.z. de kleine effecten mogen niet verwaarloosd worden. Men gebruikt statistische methoden om deze kleine effecten te kunnen onderscheiden van onvermijdelijke doch niet ter zake doende storende invloeden. Tevens streeft men voortdurend naar exacte formuleringen van landbouwkundige problemen waardoor een wiskundig statistische behandeling ervan mogelijk wordt.

Om hulp te bieden bij de wiskundige behandeling van landbouwkundige problemen werd in 1957 te Wageningen het Centrum voor Landbouwwiskunde opgericht. Het Centrum nam de taak op zich om voorlichting te geven over de toepassing van die methoden en over de uit te voeren berekeningen aan alle rijksinstellingen voor landbouwkundig onderzoek en soortgelijke instellingen ressorterende onder de Nationale Raad voor Landbouwkundig Onderzoek T.N.O. Het Centrum is er niet op berekend op grotere schaal rekenwerk voor anderen uit te voeren. Bij problemen die omvangrijk rekenwerk vereisen wordt verwezen naar de Afdeling Bewerking Waarnemingsuitkomsten van de Centrale Organisatie T.N.O. te Den Haag. Deze afdeling beschikt, naast een staf die desgewenst eveneens adviezen verstrekt, over een uitgebreide outillage voor het uitvoeren van berekeningen die de capaciteit van de onderzoeker te boven gaan. Een nieuwe aanwinst van die afdeling is de elektronische rekenautomaat ZEBRA waarmee zeer uitvoerige berekeningen snel kunnen geschieden. In het laatste gedeelte van dit artikel is een lijst opgenomen van de programma's die thans voor de ZEBRA beschikbaar zijn, welke lijst door het hoofd van de afdeling, de heer Th. J. D. Erlee is samengesteld.

OVERZICHT VAN STATISTISCHE TECHNIEKEN

1 *Variantie-analyse*

De methode der variantie-analyse is misschien wel de meest toegepaste sta-

tistische techniek bij het onderzoek. Zij vindt toepassing bij veldproeven, potproeven, proeven in kassen, proefnemingen in de bosbouw, enz. In het kort gezegd bestaat de methode uit het schatten van de bijdrage van verschillende oorzaken van variatie aan de totale variantie van een reeks waarnemingen. Bij een juiste wijze van proefnemen is het mogelijk te toetsen of deze verschillende bijdragen duidelijk groter zijn dan die welke geleverd wordt door de niet controleerbare oorzaken die men ook wel „toevallige" storingen noemt. De variantie wordt ontbonden in componenten die elk een bepaald effect representeren, zodat het mogelijk wordt de betekenis van de verschillende effecten tegen elkaar af te wegen.

In eenvoudige gevallen kan men gebruik maken van zgn. orthogonale proefschema's waarvan de statistische bewerking gemakkelijk en weinig tijdrovend zijn. Vaak is de vraagstelling echter zo uitgebreid dat gecompliceerdere schema's (zgn. niet-orthogonale schema's) meer doeltreffend zijn. Sommige hiervan – bijv. de rasterschema's van hogere orde – vereisen reeds zoveel rekenwerk dat beschikt moet worden over snelle elektrische rekenmachines. Het rekenwerk neemt echter onevenredig toe bij de zgn. „wilde" schema's waarbij het niet meer mogelijk is eenvoudige algebraïsche uitdrukkingen te geven voor de zuivere schattingen van de effecten. Iteratieve rekenmethoden bieden hier de oplossing, terwijl bij de reeds genoemde Afdeling Bewerking Waarnemingsuitkomsten T.N.O. met behulp van de ZEBRA de benodigde rekentijd tot een minimum wordt teruggebracht. Kennis van de mogelijkheden en keuze van een goed proefschema is voor de onderzoeker van groot belang, zodat het vaak aan te bevelen is, reeds voor het inzetten van de proef statistisch advies in te winnen.

2 *Regressie-berekening*

Vaak doet zich het geval voor dat men de waarde van een stochastische grootheid wenst te schatten uit de waarden die een aantal andere, bepalende variabelen aannemen. Men kan bijv. trachten de benodigde arbeidstijd voor het vellingswerk van een boom te schatten uit gegevens betreffende de afmetingen van de boom (dikte, hoogte enz.), de hoedanigheid van de boom (aard van de betakking), aard van het bos (ondergroei) enz.

Soms wil men de winst per ha van een landbouwbedrijf „verklaren" met behulp van grootheden als bedrijfsgrootte, veebezetting, kapitaalsinvestering enz. De vergelijking die deze betrekking representeert heet *regressie*vergelijking.

Het probleem is gewoonlijk om uit een groot aantal mogelijke bepalende variabelen een klein aantal uit te zoeken dat gezamenlijk een goede schatting van de gezochte grootheid levert. Hierbij spelen de partiële correlatiecoëfficiënten een rol en het is daarom nodig allereerst de correlatie-coëfficiënten van alle variabelen, de zgn. correlatie-matrix, te berekenen. Het uitzoeken van het „beste" [1] stel van k variabelen uit de n beschikbare, is een moeizame bezigheid omdat dit alleen is te vinden door alle mogelijke combinaties van k uit n te proberen. Is het aantal beschikbare variabelen bijv. 10 en wil men hieruit de beste 6 kie-

zen dan is het aantal mogelijkheden reeds 210, zodat het duidelijk is dat het rekenwerk zonder rekenautomaten onuitvoerbaar is. Men bedenke hierbij echter dat het rekenwerk zelfs met elektronische apparatuur bij een groot aantal variabelen fantastische afmetingen aanneemt; voor 20 variabelen zou het aantal combinaties van 6 reeds bijna 40 000 bedragen. Andere methoden van aanpak zijn in dat geval dus noodzakelijk.

3 *Bepalen van optimale voorwaarden*

Een bij technische procedé's veel voorkomend vraagstuk is het bepalen van de optimale voorwaarden voor een zeker produktieproces.

In de laatste tien jaar is een methode voor het bepalen van optimale voorwaarden tot in details uitgewerkt. Met behulp van een reeks na elkaar uitgevoerde proeven, waarbij men telkens op grond van de resultaten van voorafgaande proeven de bepalende factoren „instelt" kan men snel het vraagstuk van de optimale voorwaarden oplossen.

In het landbouwkundig onderzoek doen zich soortgelijke vraagstukken voor, doch de methode heeft nog vrijwel geen toepassing gevonden, gedeeltelijk omdat zij eenvoudig niet bekend is geworden. Wellicht ook omdat proeven in de landbouw veelal lang duren en dat daarom sequent proefnemen minder aantrekkelijk is. Toch doen zich zeker gevallen voor waarin de methode op zijn plaats zou zijn; men zou bijv. bij bemestingsproeven die enige jaren achtereen herhaald worden de mestgiften kunnen variëren op grond van de verkregen uitkomsten, met het doel na enige jaren een optimum combinatie op te sporen.

De benodigde berekeningen zijn dezelfde als die welke in de vorige paragraaf werden genoemd, nl. het bepalen van een aantal regressiecoëfficiënten, hetgeen neerkomt op het oplossen van een stelsel van lineaire vergelijkingen, de zgn. normaalvergelijkingen. Evenals bij de variantieanalyse hangt de hoeveelheid rekenwerk af van het proefschema, d.w.z. dat door geschikte keuze van de waarden van de verschillende factoren sterk op het rekenwerk bespaard kan worden.

4 *Lineaire programmering*

Bij het opstellen van een bedrijfsplan, moet een keuze worden gedaan tussen een aantal verschillende manieren, waarop de beschikbare produktiemiddelen kunnen worden aangewend. De verschillende produktieplannen leiden elk tot een andere samenstelling van de produktie die in het algemeen telkens een andere waarde (of winst) vertegenwoordigt. Het probleem is nu om, zonder voor een der noodzakelijke produktiemiddelen het beschikbare potentieel te overschrijden, dat produktieplan te kiezen dat de maximale winst oplevert. Als verondersteld mag worden dat het verbruik van een produktiemiddel evenredig is met de hoeveelheid hiermee geproduceerd goed en dat ook de winst een lineaire functie is van de hoeveelheden van de geproduceerde goederen, dan noemt men

---

[1] Onder „beste" wordt verstaan die welke schattingen levert die de waargenomen waarden het dichtst benaderen.

dit vraagstuk een lineair-programmeringsprobleem. Er is hierbij dus sprake van het maximaliseren van een lineaire functie onder de beperkende voorwaarden dat aan de beschikbare produktiemiddelen geen onvervulbare eisen worden gesteld. Een bijzonderheid is dat de geproduceerde hoeveelheid van elk goed een niet-negatieve waarde moet zijn, zodat alleen die oplossingen bruikbaar zijn die voor alle produkten in het plan een positieve waarde of de waarde nul bevatten.
Er bestaat een rekenschema dat, voor het geval er een oplossing bestaat, ook zeker tot de oplossing voert. Ook dit schema is voor de elektronische rekenautomaat geprogrammeerd.
Opgemerkt wordt dat de lineaire programmering niet een statistisch onderwerp is en dat het strikt genomen evenmin behoort tot wat men gewoonlijk onder landbouwkundig onderzoek verstaat. Het behoort echter zeker tot de onderwerpen die van belang zijn voor de landbouwvoorlichting, zoals ook uit het volgende voorbeeld blijkt. Bij het samenstellen van veevoeders worden een aantal eisen gesteld aan gehalte aan verschillende voedingsstoffen als eiwit, vet, koolhydraten, vitaminen, zouten enz., voorts aan de consistentie en smakelijkheid van het mengvoer. Het aantal beschikbare grondstoffen, dat in verschillende verhoudingen gemengd aan de eisen voldoet, is zeer groot. Gevraagd wordt het voordeligste mengsel aan te wijzen.

5 *Steekproef-onderzoek*

Bij het onderzoek van gemeenschappen is het vaak niet nodig en ook niet doeltreffend om alle individuen van de gemeenschap t.a.v. de eigenschappen waarin men geïnteresseerd is te onderzoeken. Veelal kan volstaan worden met schattingen, mits men tevens ingelicht is over de nauwkeurigheid van die schattingen. Een steekproef uit het geheel, de zgn. populatie of het universum, kan leiden tot goede schattingen. De wijze waarop de steekproef wordt genomen bepaalt in de eerste plaats of de schatting zuiver is, d.w.z. of het gemiddelde van een onbeperkt groot aantal van dergelijke steekproeven overeenkomt met de waarde van het universum; verder wordt door de gevolgde steekproeftechniek bepaald hoe nauwkeurig de schatting is, d.i. de kans op verschillen tussen steekproefwaarde en universumwaarde van gegeven grootte.
De vier voornaamste steekproefmethoden zijn:

a De toevallige steekproef. Hierbij wordt een steekproef van vastgestelde grootte op zodanige wijze gekozen dat ieder individu uit het universum een even grote kans heeft om deel van de steekproef uit te maken.

b De gerichte steekproef. Het universum wordt t.a.v. een gemakkelijk vast te stellen eigenschap, die bovendien gecorreleerd is met de te onderzoeken eigenschap, in klassen ingedeeld. De steekproef wordt nu zo over de klassen verdeeld, dat iedere klasse evenredig met zijn gewicht in het universum in de steekproef is vertegenwoordigd.
Gerichte steekproeven zullen in het algemeen aanzienlijk nauwkeuriger zijn dan toevallige steekproeven.

c Ratio-schattingen. Hier wordt eveneens gebruik gemaakt van de correlatie van eigenschappen. Er wordt dan bij benadering een vaste verhouding tussen de gezochte en de gemeten eigenschap verondersteld.

d Regressie-schattingen. Evenals bij de ratio-schatting berust de regressie-schatting op de correlatie tussen de gezochte eigenschap en een andere die gemakkelijk te be-

palen is. In plaats van een vaste verhouding wordt nu een lineaire regressie als grondslag van de schatting genomen.

De vier methoden zijn gerangschikt in volgorde van toenemende nauwkeurigheid doch tevens neemt de bewerkelijkheid zowel voor het verzamelen van de gegevens als voor de hoeveelheid rekenwerk toe. Bij enigszins grote steekproeven zijn mechanische rekenhulpmiddelen onontbeerlijk. Het gebruik van ponskaarten, die verwerkt kunnen worden door sorteermachines en rekenmachines kunnen daarbij veel hulp bieden. Vooral indien het onderzoek niet één maar verscheidene eigenschappen betreft is het gebruik van ponskaarten aan te bevelen. Steekproefonderzoek wordt veel toegepast bij economische en sociologische vraagstukken.

6 *Verdelingsvrije methoden*

Bij de meeste statistische schattingen en toetsingsmethoden worden er bepaalde veronderstellingen gemaakt over de kansverdeling van de beschouwde grootheid. Het kan zijn, dat dergelijke veronderstellingen weinig grond hebben, zodat het gewenst is de uitspraken – bijv. het aanwijzen van een betrouwbaarheidsinterval – onafhankelijk te maken van de kansverdeling. In de laatste 20 jaar zijn een groot aantal toetsingsmethoden ontwikkeld, waarvoor dit geldt. Enkele van de meest bekende zijn, de tekentoets, de toets van twee steekproeven van Wilcoxon, de symmetrietoets van Wilcoxon, de rangcorrelatie-coëfficiënten van Spearman en van Kendall, en de aanpassingstoets van Kolmogoroff. Behalve hun minder beperkte geldigheid hebben deze methoden het voordeel, dat zij slechts weinig rekenwerk vereisen.

DE PROGRAMMA'S VOOR DE ZEBRA

Hieronder volgt de specificatie van de rekenprogramma's[2], die voor het uitvoeren van statistisch rekenwerk met de elektronische rekenautomaat Zebra bij de A.B.W.–T.N.O. thans reeds beschikbaar zijn.

Een aantal dezer specificaties valt uiteen in twee delen: a. de maximum capaciteit van het programma, en b. de ervaring met het programma voor de benodigde tijd, bij het uitvoeren van een gegeven rekenwerk met de machine.

Bedacht moet worden, dat de in de specificaties genoemde tijden alleen een indruk geven van de tijden gedurende welke de Zebra zelf in beslag wordt genomen; de totale tijd, nodig voor het verwerken van een opdracht, bevat uiteraard meestal nog andere tijden: die voor het ponsen van de gegevens, voor het manipuleren met geponste banden, voor controlewerkzaamheden, voor het in speciale vorm uitschrijven van resultaten enz.

1 Het berekenen van correlatie-matrices

Programma 1 Maximum aantal variabelen 34; aantal waarnemingen per variabele onbeperkt. Bij 18 variabelen bedraagt de rekentijd ongeveer 1 minuut per waarneming. Zouden er dus 200 stellen van 18 waarnemingen zijn, dan zouden na 200 minuten alle 153 correlatie-coëfficiënten zijn berekend.

Programma 2 Maximum aantal variabelen 69; het produkt van het aantal variabelen

[2] Met uitzondering van het programma voor het opstellen van de seculair-vergelijking, dat nog niet beproefd is.

en het aantal waarnemingen per variabele mag niet groter zijn dan 6800. De waarnemingsgetallen moeten alle positief zijn en mogen het getal 1000 niet overschrijden. In het algemeen zal door het bijtellen van een constant getal en/of door het afkappen van een laatste cijfer zonder schade voor de nauwkeurigheid aan deze beide voorwaarden kunnen worden voldaan.
Bij 28 variabelen bedraagt de rekentijd ongeveer 1 minuut per waarneming. Dit programma is dus sneller dan het voorafgaande.

Programma 3 Maximum aantal variabelen 20. Bij dit programma kunnen behalve de correlatie-coëfficiënten ook de multipele regressievergelijkingen en de multipele correlatie-coëfficiënten met varianties berekend worden.

Bij elk dezer drie werkwijzen kunnen desgewenst de gecorrigeerde kwadraat- en produktsommen als tussenresultaat verkregen worden. Dit eist uiteraard enige additionele draaitijd van de machine.

2 Inverteren van de matrix

Orde van de matrix maximaal 38. Voor het inverteren van een matrix van de orde 18 is de rekentijd een half uur.

3 Bepaling van eigenwaarden en eigenvectoren van een matrix:

a *Door middel van een iteratiemethode.* Orde van de matrix maximaal 70. De rekentijden per iteratiestap zijn ongeveer als volgt:

orde 50 rekentijd ca. 2 min.
orde 40 rekentijd ca. 1 min.
orde 20 rekentijd ca. 15 sec.

De tijd voor de zgn. reductie bedraagt 1/5 tot 1/10 van de totale rekentijd.

b *Door het opstellen van de secularvergelijking.* Daar dit programma nog in bewerking is, zijn nog geen details bekend.

4 Simplex-methode voor lineaire programmering

Voor de capaciteit van het programma gelden de volgende voorbeelden:

760	grootheden	bij	5	voorwaarden
225	„	„	20	„
50	„	„	50	„
3	„	„	70	„

De rekentijd per iteratiestap is als volgt:

bij	30	grootheden	en	30	voorwaarden:	9 min.
„	20	„	„	20	„	: 3 „
„	10	„	„	10	„	: 1 „

5 Inorthogonale variantieanalyse met 2 criteria en herhalingen

Capaciteit: maximum aantal rijen en maximum aantal kolommen elk 200: maximum aantal waarnemingen 700 totaal.

Berekend worden:

1 Algemeen gemiddelde
2 Afzonderlijke kolomgemiddelden (gecorrigeerd voor rij-invloed)
3 Afzonderlijke rijgemiddelden (gecorrigeerd voor kolom-invloed)
4 Som van deviaties van afzonderlijke waarnemingen t.o.v. beste schatting (ter controle)
5 Restvariantie
6 Variantie „tussen rijgemiddelden"
7 Variantie „tussen kolomgemiddelden"
8 Afzonderlijke deviaties.