HET SAMENVATTEN VAN RASSENPROEVEN EN HET TOEPASSEN VAN VRUCHTBAARHEIDSCORRECTIES MET NIET-ORTHOGONALE METHODEN

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR IN DE LANDBOUWKUNDE OP GEZAG VAN DE RECTOR-MAGNIFICUS DR E. BROUWER, HOOGLERAAR IN DE PHYSIOLOGIE DER DIEREN, TE VERDEDIGEN TEGEN DE BEDENKINGEN VAN EEN COMMISSIE UIT DE SENAAT VAN DE LANDBOUWHOGESCHOOL TE WAGENINGEN, OP WOENSDAG, 15 JUNI 1949 TE 16 UUR

DOOR

G. HAMMING

TECHNISCHE HOOGE SCHOOL
BIBLIOTHEEK
DELFT

STAATSDRUKKERIJ — 's-GRAVENHAGE 1949

In de tijd dat ik voor de eerste maal verbonden was aan het Instituut voor Rassenonderzoek van Landbouwgewassen, heb ik me bezig mogen houden met een paar problemen uit het brede veld van de wiskundige verwerking van waarnemingsuitkomsten.

Het eenvoudigste probleem was dat van de vruchtbaarheidscorrecties. Met name bij de voedergewassen was het vaak nodig grote proefvelden aan te leggen met over de honderd rassen. Doordat zulke proefvelden vaak anderhalf tot twee ha groot werden, was niet altijd te vermijden, dat er hinderlijke onregelmatigheden in de grond aanwezig waren, die meestal vooraf op het oog niet waren te onderkennen. In de proefvelden waren dientengevolge vruchtbaarheidsverschillen van 30% niet zeldzaam. Aan de eis tot correctie was dus niet te ontkomen.

Nu wordt in de laatste jaren de proefveldverwerking geheel gedomineerd door de methoden van de school van R. A. FISHER. Deze methoden waren voor de bedoelde proefvelden evenwel niet bruikbaar, omdat bij de aanleg geen rekening gehouden was met de eisen van de FISHER-schema's.

Krachtens zijn grondslagen is de FISHER-methode voor het onderhavige probleem ook niet erg geschikt, omdat FISHER zich baseert op orthogonaliteit, terwijl het probleem van het vruchtbaarheidsverloop binnen de blokken in wezen inorthogonaal is.

Juist wegens die inorthogonaliteit is meer steun te verwachten van de grafische methoden, waaraan in Nederland vooral de naam van W. C. VISSER is verbonden. Van deze methode is dan ook een ruim gebruik gemaakt. Maar ook hierbij is er een bezwaar. Omdat de grafische methode meer afhankelijk is van de persoonlijke kijk op de resultaten, is het moeilijk een objectief oordeel over de fouten te krijgen.

In hoofdstuk IV is getracht aan dit bezwaar tegemoet te komen.

Een moeilijker probleem was dat van het samenvatten van de uitkomsten van rassenproeven, wanneer het aantal rassen wisselt. Dan kan namelijk niet „geFISHERd" worden. In hoofdstuk III is gezocht naar een weg door deze moeilijkheid.

Het bleek inderdaad mogelijk langs een uitvoerbare weg te komen tot samenvattende rasgemiddelden, die gelijk zijn aan de gemiddelden, die b.v. worden verkregen volgens de methode van hfdst. XII van Dr M. J. VAN UVEN: Mathematical treatment of the results of agricultural and other experiments. Deze laatste methode is in het onderhavige geval meestal niet te hanteren wegens het grote aantal onbekenden.

Zonder meer zouden deze gemiddelden dus een zeker vertrouwen mogen genieten. Maar helaas is er een complicatie, nl. de afhankelijkheid der rassen. De gebieden, waarop de conclusies van het proefveldonderzoek worden toegepast, zijn niet homogeen; er is een variatie in de milieufactoren. Als regel heeft ieder ras een eigen manier om op deze milieufactoren te reageren, maar soms zijn er twee of meer rassen, die in hun reacties sterk op elkaar gelijken. Bij deze rassen is er een zekere afhankelijkheid, die aanleiding geeft tot correlatieverschijnselen en die tevens oorzaak is, dat de berekende rasgemiddelden misleidend kunnen zijn.

Aan het probleem dat hierdoor ontstaat is een groot deel van hfdst. IV gewijd, maar dit neemt niet weg dat de formules die in dit verband gegeven zijn nog maar een eerste stap in de richting van een oplossing zijn. Aan de bijbehorende foutentheorieën is nog niet begonnen.

Over de eerste twee hoofdstukken kan ik kort zijn. Er worden allerlei begrippen in uitgewerkt, die in de laatste twee nodig zijn; er wordt soms ook extra lang stilgestaan bij onderwerpen, waarvan in de practijk bleek, dat er bij velen onduidelijkheid over heerste. Met name is in hfdst. II stilgestaan bij de vooronderstellingen van de vaak gebruikte aanname, dat de opbrengst van een bepaald ras op een bepaald proefveld mag worden gezien als de som van een bijdrage van een rasinvloed en een bijdrage van een proefveldinvloed.

Bizondere dank ben ik wel verschuldigd aan Prof. Dr M. J. VAN UVEN. Niet alleen vanwege de grondigheid en critische geest waarmee hij de wiskunde steeds doceerde, maar bovenal omdat hij op dezelfde manier ook dit werk heeft willen toetsen, waardoor een groot aantal onvolmaaktheden konden verdwijnen.

Evenzo ben ik grote dank verschuldigd aan Prof. Ir W. J. DEWEZ en Prof. Dr Ir J. C. DORST die beiden het manuscript van hun opmerkingen hebben voorzien, wat aan de algemene op-

bouw ten goede is gekomen, en verder een aantal onduidelijkheden heeft weggenomen.

Gaarne zal ik dan deze inleiding besluiten met mijn dank uit te spreken aan de Directeur van het I.V.R.O., Ir J. K. GROENEWOLT, die mij volledig tijd heeft gegeven voor het voltooien van deze studie, ook wanneer empirisch werk daarvoor soms moest wijken. Met groot genoegen denk ik verder aan de prettige sfeer terug, die ik op het I.V.R.O. van alle kanten heb mogen ontmoeten.

OPMERKING

Bij het schrijven van deze studie is aangenomen, dat de lezer bekend is met

Dr M. J. VAN UVEN: Mathematical Treatment of the Results of Agricultural and other Experiments
(in de tekst vaak genoemd: het leerboek van VAN UVEN)

en met de hoofdzaken van differentiaal- en integraalrekening, de FISHER-analyse en de correlatierekening.

INHOUD

Blz.

I – ENKELE ALGEMENE BEGRIPPEN

A – IDEALE FORMULES EN NOODFORMULES

101 – HET VINDEN VAN EEN FORMULE DIE DE GEGEVENS VERKLAART 1
102 – HET VINDEN VAN EEN FORMULE, DIE EEN BETROUWBARE VOORSPELLING TOELAAT 5
103 – WISKUNDIGE FORMULERING VAN DE VEREISTE METHODE 7
104 – HET BEGRIP „INVLOED" 8
105 – HET VEREISTE AANTAL VRIJHEIDSGRADEN 10
106 – HET UITSCHAKELEN VAN ONBEKENDEN. 11

B – FOUTEN EN GEWICHTEN BIJ NOODFORMULES

111 – HET METEN VAN DE STRIJDIGHEID 13
112 – VOORBEELD VAN HET GEBRUIKEN VAN EEN NOODFORMULE. 15
113 – DE ASYMPTOTISCHE WAARDE 17
114 – DE FOUTEN BIJ HET GEBRUIKEN VAN EEN NOODFORMULE 19
115 – HET TOEKENNEN VAN GEWICHTEN 24
116 – DE VARIANS-ANALYSE 27
117 – HET GEBRUIK VAN DE VARIANS 29
118 – HET KIEZEN VAN EEN MONSTER EN EEN FORMULE . 34

C – CORRELATIEREKENING EN LIJNVEREFFENING

121 – INLEIDING 36
122 – DE BETEKENIS VAN DE KANSVERDELING VOOR DE CORRELATIEREKENING. 39
123 – DE MAATEENHEDEN VAN x EN y 42
124 – DE REGRESSIELIJNEN 45
125 – VERGELIJKING VAN DE METHODEN VAN CORRELATIEREKENING EN LIJNVEREFFENING 49
126 – HET PSEUDO-CORRELATIEKARAKTER VAN DE LIJNVEREFFENING 51
127 – HET VERSCHIL TUSSEN CORRELATIE EN LIJNVEREFFENING 53
128 – DE KEUZE VAN DE JUISTE VEREFFENINGSLIJN . . . 55

Blz.

II – ENIGE BIZONDERHEDEN VAN HET RASSENONDERZOEK

201 – De zin van het proefveld 61
202 – Het condenseren van de invloeden en het normaliseren van de gegevens 62
203 – Het transformeren van bepaalgegevens in rekengegevens 66
204 – Contrôle op de juistheid van de transformatiefunctie 71
205 – Het meten van de milieugunstigheid 72
206 – De noodzaak het milieu te „wegen" 78

III – HET SAMENVATTEN VAN RASSENPROEVEN

301 – De interactie tussen ras en proefveld 81
302 – De correlatie van de schijnbare fouten van onafhankelijke rassen. 89
303 – Afhankelijkheid van rassen. 95
304 – Het samenvatten van proeven met wisselend aantal onafhankelijke rassen. 97
305 – Contrôle op de gevolgde methode. 105
306 – Het toekennen van gewichten 106
307 – Het schatten van de fouten van de berekende rasverschillen 111
308 – Contrôle op de empirische methode 117
309 – Het bepalen van de correlatie tussen de rassen 120
310 – Analyse van de correlatieverschijnselen 125
311 – Het samenvatten van proeven met wisselend aantal gedeeltelijk afhankelijke rassen . . . 130
312 – Nader onderzoek van de ontwikkelde methoden 141
313 – Onderzoek van de interpretatiefout 148

IV – HET BEREKENEN VAN VRUCHTBAARHEIDSCORRECTIES OP AFZONDERLIJKE PROEFVELDEN

401 – De aard van de vruchtbaarheidscorrectie . . . 152
402 – Nadere bepaling van de gewenste methode . . 156
403 – Het berekenen van de vruchtbaarheidslijn . . 162
404 – De toelaatbaarheid van een algemene vruchtbaarheidscorrectie 166

Blz.
405 – Over de voorwaarden voor correctie in één blok 171
406 – Samenvatting van de voorwaarden voor toelaatbaarheid van een vruchtbaarheidscorrectie . . 175
407 – Voorbeeld van het toetsen van een vruchtbaarheidslijn 180
408 – Precisering van de vruchtbaarheidslijn 184
409 – Practisch voorschrift voor de vruchtbaarheidscorrectie 186
410 – Het verbruikt aantal vrijheidsgraden 189
411 – De bepaling van de toevallige fout 192
412 – Vruchtbaarheidscorrecties bij proeven in vierkantsverband 195

I – ENKELE ALGEMENE BEGRIPPEN

A – IDEALE FORMULES EN NOODFORMULES

101 – Het vinden van een formule, die de gegevens verklaart

Iemand, die aan de landbouwers voorlichting zal geven aangaande de rassenkeuze, dient daarbij een zo nauwkeurig mogelijke voorspelling te kunnen doen over de resultaten, die de verschillende rassen het volgend jaar zullen geven.

Hiertoe is in de eerste plaats nodig dat hij zijn ervaring inventariseert. Hij gaat na welke opbrengsten of andere eigenschappen van de diverse rassen hem bekend zijn, en met welke bizonderheden ze eventueel in verband kunnen worden gebracht.

Ieder getal dat bekend is aangaande het object van onderzoek (dus b.v. aangaande een opbrengst) willen wij als *gegeven* aanduiden. De bizonderheden die eventueel kunnen dienen om het ontstaan van het gegeven te verklaren willen wij *omstandigheden* noemen.

Het is nu de taak van de voorlichter een algemene wet te vinden, die het ontstaan van het gegeven verklaart aan de hand van de omstandigheden en de mogelijkheid biedt, aan de hand van toekomstige omstandigheden de toekomstige gegevens zo nauwkeurig mogelijk te voorspellen. Deze wet willen wij aanduiden als de *conclusie*. Vaak zal de conclusie geformuleerd worden in de vorm van een formule.

Van deze formule zal de voorlichter gebruik maken om een antwoord te geven op de *vragen* van de practijk.

Wij willen aan de hand van een voorbeeld nagaan, waarop gelet moet worden bij het ontwerpen van een formule.

Veronderstel dat er een are tarwe is verbouwd in de Wageningse eng in het jaar 1942. Veronderstel verder dat als opbrengst 30 kg is verkregen. Het getal 30 is dus het *gegeven*.

Dit gegeven moet dus dienen om antwoord te geven op allerlei practische *vragen*, b.v.

a. Hoe *vruchtbaar* is de Wageningse eng?

b. Hoe *groeizaam* was het weer in 1942?

c. Welk *opbrengstvermogen* heeft tarwe.
d. Welke *betekenis* had het speciale ras dat verbouwd werd.
e. enz.

Dit zijn een aantal vragen, waarop men graag antwoord wil ontvangen. Voor het beantwoorden van al die vragen is men aangewezen op het éne gegeven. Met behulp daarvan zou bijv. geconcludeerd moeten kunnen worden:

f. De Wageningse eng kan 30 kg/are opbrengen.
g. Het jaar 1942 gaf oogsten van 30 kg/are.
h. Tarwe brengt 30 kg/are op.
i. Het verbouwde ras kan gemakkelijk een opbrengst van 30 kg/are bereiken.

Of men zou de vragen moeten combineren en b.v. als volgt redeneren:

j_1	De Wageningse eng is niet erg vruchtbaar, de vruchtbaarheid is	27 kg/are
j_2	Het gewas tarwe is niet geschikt voor de eng, negatieve opbrengststijging	–4 kg/are
j_3	Het jaar 1942 heeft nogal meegewerkt, de groeizaamheid was	5 kg/are
j_4	Het gekozen ras was extra goed, productieoverschot	2 kg/are
	som	30 kg/are

Wanneer zo één of meer vragen in onderlinge samenhang worden bestudeerd, willen wij spreken van een *vraagstuk*. Onder de letters *f, g, h, i* en *j* is dus telkens één vraagstuk opgelost.

Het is niet mogelijk een vraagstuk kwantitatief op te lossen zonder in de conclusie één of meer getallen te noemen. *Een getal dat in een conclusie noodzakelijk is, willen wij een u i t k o m s t noemen.*

In de conclusies *f—i* staat dus telkens 1 uitkomst; in conclusie j_1—j_4 staan er 4. Weliswaar zijn in conclusie j_1—j_4 5 getallen genoemd, maar er zijn slechts 4 *noodzakelijk*; de som behoeft niet vermeld te worden, daar die zonodig uit de andere vier kan worden afgeleid. *De uitkomsten moeten dus onderling onafhankelijk zijn.*

Wij willen nu nagaan of al deze conclusies toelaatbaar zijn.

In de conclusies *f—h* wordt het gegeven telkens als het gevolg van één omstandigheid beschouwd, terwijl de andere omstandigheden worden genegeerd. Dit negeren gebeurt nog op ongelijke

wijze. Formule *f* spreekt van „kunnen": De Wag. eng *kan* 30 kg/are opbrengen. Het verband tussen de Wag. eng en de 30 kg/are is vrij los. Er blijft zo ruimte voor andere invloeden.

De conclusies *g* en *h* daarentegen suggereren algemeen geldigheid. Tussen het jaar of het gewas enerzijds en de opbrengst anderzijds wordt een strak verband gelegd dat andere invloeden uitsluit. Niettegenstaande dit verschil wordt toch in alle drie conclusies één omstandigheid als *oorzaak* van het gegeven aangeduid. Volgens *f* is het de grond, die de 30 kg opbrengt (zij het misschien via een bepaald gewas), volgens *h* daarentegen is het gewas de gever van de opbrengst (zij het dan ook met behulp van de grond).

Wiskundig is er tegen geen van de drie conclusies bezwaar in te brengen. Het is niet een wiskundig probleem, of de opbrengst uiteindelijk te danken is aan de grond of aan het gewas. Maar er is wel bezwaar tegen, dat alle drie conclusies tegelijk worden aanvaard. Als gever van de ene opbrengst van 30 kg/are kunnen niet drie verschillende „omstandigheden" worden aangewezen, die ieder voor zich het verschijnsel geheel zullen verklaren. Wil men de verschillende invloeden erkennen, dan dienen ze gezamelijk in één conclusie vermeld te worden.

Het zou dan b.v. mogelijk zijn een conclusie in de gedaante van j_1—j_4 op te stellen. In deze conclusie worden 4 invloeden erkend, en aan ieder wordt een bepaalde werking toegeschreven. Bij nadere beschouwing blijkt echter dat de uitkomsten j_1—j_4 niet uit het gegeven geconcludeerd zijn. Er is reeds eerder waargenomen dat de Wageningse eng niet vruchtbaar is, dat tarwe daar niet het aangewezen gewas is enz. Zou de opbrengst van 30 kg/are zonder nadere kennis over de vier invloeden zijn verdeeld, dan had dit slechts als willekeur kunnen worden bestempeld.

Om te weten wat men met zijn gegevens mag doen, zal men daarom moeten letten op het wiskundig begrip „*vrijheidsgraad*". *Het aantal vrijheidsgraden geeft aan, hoeveel gegevens men onafhankelijk van elkaar heeft verkregen.* In het bovengenoemde voorbeeld is het aantal vrijheidsgraden dus 1, aangezien er slechts 1 gegeven is. Het is nu een wiskundige wet, *dat men niet meer uitkomsten mag verkrijgen dan men vrijheidsgraden heeft.* In bovengenoemd voorbeeld was er één vrijheidsgraad, daarom mag er ook slechts één getal in de conclusie worden genoemd.

Wil men verschillende invloeden erkennen, dan moet men ze

voorstellen als samenwerkende, om het ene gegeven voort te brengen, b.v.:

k. tarwe bracht in 1942 in de Wageningse eng 30 kg/are op;

l. de Wageningse eng bracht in 1942 per are 30 kg tarwe op.

In de conclusies *k* en *l* wordt meer rekenschap gegeven van de begeleidende verschijnselen dan in *g—i*. Toch hebben ze *wiskundig* geen enkel voordeel. Conclusie *f* verklaart het verschijnsel van 30 kg/are evengoed als conclusie *l*; *h* doet het evengoed als *k*. Van uit wiskundig oogpunt bezien mag er over de grond en het weer evengoed gezwegen worden als over de stand van de sterren en de schijngestalten der maan, die ook wel met de opbrengst in verband zijn gebracht.

Conclusie *h* is wiskundig evengoed als conclusie *k*. Maar men moet opletten, er niet meer in te lezen dan er gezegd wordt. In conclusie *k* mag niet gelezen worden, dat ongeveer hetzelfde resultaat verkregen zou zijn in 1943, of op een andere grond. In conclusie *h* mag men niet impliciet als uitkomst lezen, dat de grond *dus* geen invloed heeft. „Geen invloed" wil wiskundig zeggen een invloed 0; dit zou een tweede getal zijn.

Hier ligt het bezwaar van conclusie *i*. Daar wordt gezegd dat het verbouwde ras gemakkelijk 30 kg/are op kan brengen. Door het woord „gemakkelijk" worden de niet-ras-omstandigheden als tamelijk ongunstig gestempeld. Deze ongunstigheid kan niet uit de 30 kg/are zijn gebleken, tenzij er meerdere gegevens beschikbaar waren, waarmee het getal 30 kon worden vergeleken. Dit is echter niet in overeenstemming met onze veronderstelling.

Deze discussie kan als volgt worden samengevat.

Bij het formuleren van een conclusie zal men scherp moeten onderscheiden, wat uit de gegevens is afgeleid, en wat op andere gronden wordt beweerd.

De aard van de conclusie is vanuit wiskundig oogpunt onbelangrijk; slechts wordt vereist, dat ze zodanige uitkomsten aangeeft, dat die gezamelijk kwantitatief rekenschap geven van alle gegevens.

De conclusie mag niet meer uitkomsten bevatten, dan er vrijheidsgraden beschikbaar zijn.

Bij deze laatste regel kan nog worden opgemerkt dat in de practijk meestal getracht wordt uitkomsten te *vergelijken*. In dat geval is er een vergelijkingspunt nodig, dat ook een uitkomst is en dus 1 vrijheidsgraad verbruikt. Op grond hiervan kan men vaak lezen dat het aantal beschikbare vrijheidsgraden 1 lager is

dan het aantal gegevens. De vrijheidsgraad voor het vergelijkingspunt is er dan alvast afgetrokken.

102 – HET VINDEN VAN EEN FORMULE, DIE EEN BETROUWBARE VOORSPELLING TOELAAT

De eisen die aan het eind van 101 verzameld zijn, mogen vrij streng lijken, toch valt er gemakkelijk aan te voldoen. Immers, de aard van de conclusie wordt vanuit wiskundig oogpunt onbelangrijk genoemd. Wanneer men de vragen als volgt zou formuleren: ,,Wat heeft de eerste waarneming opgeleverd, wat de tweede, enz.'' dan kon men al deze vragen beantwoorden, door de gegevens rechtstreeks als uitkomst in te vullen.

Door zo te handelen zou scherp naar voren worden gebracht, wat de gegevens waren; plaats voor vooroordeel zou er niet zijn. Verder zou van alle gegevens kwantitatief rekenschap zijn gegeven, en het aantal uitkomsten zou niet boven het aantal vrijheidsgraden uitkomen.

Toch zou een conclusie van deze structuur niet bevredigen. Er moeten ook eisen aan gesteld worden, die rekenschap geven van het doel, waartoe de conclusie wordt opgesteld. In 101 is reeds uiteengezet dat het doel is, een formule te vinden, die voorspellingen toelaat. Dit voegt aan de eisen van 101 een aantal nieuwe toe.

In de eerste plaats eist dit, dat de v r a g e n van het vraagstuk opnieuw gesteld moeten kunnen worden.

In de tweede plaats moet aan de hand van de uitkomsten voorspeld kunnen worden, welk gegeven zal worden verkregen door een nieuwe waarneming binnen het kader van het vraagstuk.

Tenslotte moet het gegeven dat verkregen wordt, wanneer de waarneming werkelijk wordt verricht, gelijk zijn aan het voorspelde gegeven.

Aan de hand van deze drie eisen willen wij de conclusie *f—k* opnieuw beoordelen. De conclusie *i* en *j* waren reeds verworpen.

Conclusie *f* luidde: De Wageningse eng kan 30 kg/are opbrengen. Dit is een antwoord op de vraag: Hoe vruchtbaar is de Wageningse eng? Deze *vraag* voldoet inderdaad aan de gestelde eis. Niet alleen in 1942 kon die vraag gesteld worden, maar hij kan steeds worden herhaald. Aan de eerste voorwaarde is dus voldaan.

De tweede voorwaarde eist, dat nu ook het resultaat van een nieuwe waarneming moet kunnen worden voorspeld. Veronderstel dat in 1950 de waarneming zal worden herhaald. Welk gegeven

zal worden verkregen? Dat is niet bekend. Het *kan* 30 kg/are zijn. Maar met dat woord ,,kunnen'' wordt tevens uitgedrukt dat het ook anders kan. Er is geen zekerheid. Om deze reden moet conclusie *f* worden verworpen.

Conclusie *g* luidt: Het jaar 1942 gaf oogsten van 30 kg/are. Dit beantwoordt de vraag, hoe groeizaam het weer in 1942 was. Tegen deze *vraag* moet men reeds bezwaren hebben. Deze vraag kan niet opnieuw gesteld worden. Het jaar 1942 is voorbij. Wanneer het jaar in rekening moet worden gebracht, kan er b.v. gezegd worden: een nat jaar brengt zoveel op, of een droge zomer zoveel. Natte jaren en droge zomers kunnen zich herhalen, maar 1942 niet. Ook conclusie *g* zal dus moeten worden verworpen.

In conclusie *h* staat tenslotte: Tarwe brengt 30 kg/are op. Deze conclusie beantwoordt de vraag, welk opbrengstvermogen tarwe heeft. Ook deze vraag is aanvaardbaar. Dit opbrengstvermogen kan opnieuw worden onderzocht.

Wanneer wij nagaan of aan de hand van de uitkomst het resultaat van een nieuwe waarneming kan worden voorspeld, dan is het succes bevredigend. ,,Tarwe brengt 30 kg/are op. Het gegeven dat verkregen zal worden zal dus luiden: 30 kg/are.

Maar nu wordt in de derde plaats nog geëist, dat deze voorspelling juist zal blijken, wanneer die waarneming inderdaad wordt gedaan. Het is algemeen bekend, dat er meer kans is dat de nieuwe waarneming een afwijkend gegeven verschaft. Daarom lijkt ook de laatste conclusie verwerpelijk.

Doch nu dient men te bedenken, wat reeds in 101 is geëist. Er moet duidelijk onderscheiden worden, wat uit de gegevens wordt geconcludeerd, en wat ,,vooroordeel'' is. De zekerheid dat een nieuw gegeven wellicht niet met de conclusie zal overeenstemmen is niet uit dit ene gegeven verkregen maar uit andere, die formeel niet meegeteld hebben bij het verwerken van de gegevens.

Die oudere gegevens, die bij de onderzoeker tot ,,vooroordeel'' geworden zijn, worden licht meegeteld, omdat men graag wil controleren of aan de laatste eis werd voldaan. Voor een bevredigende conclusie is het dus nodig, dat er gecontroleerd wordt, of een nieuw gegeven zich inderdaad verdraagt met de uitkomsten, die met behulp van de oude gegevens zijn berekend.

Dit maakt dat er uiteindelijk meer gegevens nodig zijn, dan aan het einde van 101 werd verondersteld. Hierdoor wordt de laatste eis van 101 verzwaard. Die moet luiden: *De conclusie moet*

m i n d e r uitkomsten bevatten, dan er vrijheidsgraden beschikbaar zijn; of omgekeerd: *Er moeten m e e r vrijheidsgraden zijn, dan er uitkomsten in de conclusie vermeld worden.*

103 – WISKUNDIGE FORMULERING VAN DE VEREISTE METHODE

Door de eisen die in 102 zijn geformuleerd is de oplossing, die aan het begin van die paragraaf gegeven werd, wel helemaal onhoudbaar geworden. Alleen al de vragen: ,,wat heeft de eerste waarneming opgeleverd?'', ,,wat de tweede?'' enz. zijn verwerpelijk, omdat ze niet opnieuw gesteld kunnen worden. Een nieuwe waarneming zal nooit weer de eerste kunnen zijn.

Maar ook wordt aan de eis, dat het aantal uitkomsten kleiner moet zijn dan het aantal vrijheidsgraden, niet voldaan, daar een conclusie als bovengenoemd die aantallen automatisch gelijk maakt.

Wanneer die laatste eis maar niet geformuleerd was, zou het probleem nog tamelijk eenvoudig zijn. Er zijn talloze manieren om van n gegevens n uitkomsten af te leiden. Men kan b.v. n eerste graadsvergelijkingen opstellen, waarbij telkens een ander gegeven in het rechterlid wordt geplaatst en een andere combinatie van de onbekenden met telkens wisselende coëfficienten in het linkerlid. Er is niet veel geluk voor nodig om zo n onafhankelijke niet-strijdige vergelijkingen te krijgen. Met behulp van een dergelijk willekeurig stel vergelijkingen kan men steeds een aantal uitkomsten verkrijgen, die aan de voorwaarden van 101 voldoen.

Nu komt 102 met de eis dat de vragen van het vraagstuk opnieuw gesteld moeten kunnen worden. Dit hoeft niet zo streng te worden opgevat, dat ieder van de n vergelijkingen weer toepasselijk moet kunnen zijn op een toekomstige situatie; maar het houdt wel in, dat met behulp van de n onbekenden een nieuwe vergelijking moet kunnen worden opgebouwd, die op een toekomstige situatie kan slaan.

De eis dat met behulp van deze vergelijking het resultaat van een nieuwe waarneming moet kunnen worden voorspeld is nogal eenvoudig. Substitutie van de berekende uitkomsten in de nieuwe vergelijking geeft vanzelf een waarde in het rechterlid. Deze waarde moet als gegeven door de nieuwe waarneming worden geleverd. Uit de wijze van ontstaan van deze nieuwe vergelijking volgt dat zij afhankelijk zal zijn van de n andere.

Tenslotte is geëist dat deze waarneming ook inderdaad verricht wordt en met de voorspelling klopt. *De eis is dus dat over een aantal onderling afhankelijke vergelijkingen wordt beschikt, voor men over de formule (conclusie) tevreden mag zijn.* Door deze eis moet de willekeur bij het opstellen van vergelijkingen worden bestreden.

Feitelijk staat men bij het trekken van een conclusie voor tweeërlei taak. In de eerste plaats moet men zorgen voor de juiste bouw van de formule en daarmee voor de juiste formulering van het vraagstuk. In de tweede plaats moet men zorgen voor een juiste beantwoording van de gestelde vragen.

Door de juiste bouw van de formule wordt de conclusie kwalitatief goed, door de juiste beantwoording van de gestelde vragen kwantitatief.

Voor het *kwantitatief* beantwoorden van een vraagstuk met n vragen zijn slechts n gegevens nodig, doch voor het aanvaarden van de formule in *kwalitatief* opzicht is het nodig de afhankelijkheid van een overtollig aantal vergelijkingen te bewijzen.

Strict genomen zou bewezen moeten worden, dat alle mogelijke overtollige vergelijkingen van de eerste n afhankelijk waren. Immers, zo lang het denkbaar is een vergelijking vast te stellen die strijdig is met de voorgaanden is de juistheid van de conclusie onzeker. Absolute zekerheid is natuurlijk nooit bereikbaar.

104 – Het begrip „invloed"

Alvorens dieper in te gaan op de mogelijkheid een aantal afhankelijke vergelijkingen te verkrijgen, zullen hun bestanddelen nader moeten worden bestudeerd.

Het rechterlid is erg eenvoudig: het vermeldt de waarde van een bepaald gegeven.

Met het linkerlid is het ingewikkelder. In 103 is de mogelijkheid genoemd hiervoor een eerste-graadsvorm te nemen, waarin telkens een aantal van de onbekenden voorkomen. Maar het bestaan van deze mogelijkheid is alleen zeker, wanneer er geen overtollige gegevens zijn die strijdigheid kunnen veroorzaken. De gedachte een eerste-graadsvorm te nemen is willekeurig. Door deze willekeur kan de conclusie fout zijn in kwalitatief opzicht. Aan het eind van 103 is besproken dat deze willekeur juist moet worden verhinderd door de boventallige gegevens. Over de bouw van het linkerlid is dus niets bekend.

Om de gedachten te bepalen volgt hier zo maar een willekeurige vergelijking die voor een bepaald vraagstuk zou kunnen gelden

$$(104.1) \qquad a + \beta b + \lg(\gamma + c) + \frac{d^{bg \sin \delta}}{\lg\left(\varepsilon + \frac{\zeta}{\eta} + \vartheta e\right)} = y.$$

In deze formule kan voor y telkens een waargenomen gegeven gesubstitueerd worden. In het linkerlid moet drieerlei worden onderscheiden:

1e de onbekenden a, b, c, d en e, waarvan de waarden moeten worden berekend. Het zijn deze waarden die in 101 als ,,uitkomsten'' zijn gedefinieerd. Deze uitkomsten zijn voor alle vergelijkingen gelijk;

2e de constanten β, γ, δ, ε, ζ, η en ϑ, die voor een bepaalde vergelijking a-priori vast staan, en voor een andere a-priori weer anders zijn. Twee vergelijkingen met precies dezelfde waarden voor de constanten zouden strijdig zijn bij ongelijke y en identiek bij gelijke y. Strijdigheid is ongeoorloofd, en identiteit van twee vergelijkingen brengt de oplossing van het probleem niet dichterbij. Vandaar de bewering, dat de waarden van de constanten voor de ene vergelijking a-priori anders moeten zijn dan voor de andere. Deze constanten zijn in 101 als ,,omstandigheden'' aangeduid;

3e de bouw, die de samenwerking van al deze grootheden regelt, en de formule kwalitatief goed maakt.

Tot dusver hebben wij al onze aandacht gericht op de uitkomsten, die wij met behulp van de gegevens willen verkrijgen. Hierdoor hebben wij neiging de uitkomsten te zien als functie van de gegevens, en de gegevens als functie van de uitkomsten.

Rekentechnisch is dit juist. Maar wanneer even nuchter bovenstaande formule (104.1) wordt bezien, blijkt het overigens helemaal niet redelijk. De eenmaal verkregen uitkomsten zijn de parameters, de constanten, die voor alle vergelijkingen gelden. Dat de gegevens telkens anders mogen zijn, wordt gemotiveerd door het wisselvallig karakter van de ,,constanten'' β, γ, δ, ε, ζ, η en ϑ. Feitelijk ligt de verhouding zo: Voordat het vraagstuk is opgelost, zijn a, b, c, d en e onbekende constanten en α, β, γ, δ, ε, ζ, η en ϑ bekende variabelen.

Het is wenselijk hier even nadrukkelijk bij stil te staan. Doordat

de min of meer toevallig opgetreden waarden van de veranderlijken bekend zijn, nemen die gedurende de berekening de plaats van constanten in; doordat de constanten onbekend zijn moet hun waarde gezocht worden, hierdoor krijgen zij als onbekende de plaats van veranderlijken. Dit mag ons niet verhinderen in te zien, dat formule (104.1) y geeft als een functie van de veranderlijken β, γ, δ, ε, ζ, η en ϑ.

Deze veranderlijken willen wij als i n v l o e d e n *aanduiden. De bepaalde getalwaarden, die voor één vergelijking (dienende ter verklaring van één gegeven) aan de verschillende invloeden worden toegekend, hebben wij reeds aangeduid met de naam* o m s t a n d i g h e d e n *van het gegeven.* Een verandering van een invloed (m.a.w. het optreden van een andere omstandigheid) veroorzaakt dus een verandering van het gegeven.

Dit geldt niet voor de constanten. Immers een constante kan niet veranderen. Wanneer dit ogenschijnlijk toch gebeurde, zou dit betekenen, dat er tot dusver een invloed in verborgen was gebleven, doordat die invloed (toevallig) constant was. Het zou betekenen dat het probleem niet juist was opgelost. De bouw van de formule zou verkeerd gekozen zijn, en de oplossing kwalitatief onjuist.

105 – Het vereiste aantal vrijheidsgraden

Aan het eind van 102 is het vereiste aantal vrijheidsgraden in verband gebracht met het benodigd aantal uitkomsten. Er moesten meer vrijheidsgraden dan uitkomsten zijn. Nu moet er op gelet worden of er ook verband bestaat tussen het vereiste aantal vrijheidsgraden en het aantal invloeden. Daartoe volgen hieronder twee formules.

De formule

$$y = ax^5 + bx^4 + cx^3 + dx^2 + ex + f$$

geeft y als functie van één invloed x; doch er zijn zes uitkomsten: a, b, c, d, e en f. Er zijn zes gegevens nodig om de uitkomsten te bepalen, en verder nog een aantal om de formule te controleren op haar juistheid. Toch is hier maar één invloed.

Daarentegen geeft de formule

$$y = uvxz$$

y weer als functie van vier invloeden. Er zijn evenwel geen gegevens nodig om uitkomsten te bepalen daar er geen onbekende para-

meters zijn. Het eerste gegeven kan reeds dienen om de bouw van de formule te *controleren.*

We zien dat het aantal uitkomsten en het aantal invloeden onafhankelijk van elkaar zijn. Het minimum aantal benodigde gegevens wordt uitsluitend bepaald door de uitkomsten.

In de practijk hoeft een formule gewoonlijk niet alleen gecontroleerd te worden, doch hij moet eerst nog worden ontworpen. Voor dat doel zal het aantal gegevens wel een veelvoud moeten zijn van het aantal invloeden. Dit zal niet nader worden nagegaan.

Er is namelijk een belangrijker complicatie aan het probleem. Om afhankelijkheid van de formules te verkrijgen moeten de getalwaarden van alle omstandigheden nauwkeurig bekend zijn, evenals de getalwaarde van het gegeven. Dit is in de landbouwwetenschap nooit het geval.

Iedere waarneming voegt daardoor een reeks speciale vragen toe aan de algemene vragen van het probleem: In welke mate is de ene omstandigheid verkeerd becijferd, en in welke mate de andere?, tenslotte in welke mate het gegeven? De beantwoording van deze vragen brengt een aantal *correctieconstanten* in de formules, *die telkens maar eenmaal voorkomen.* De waarden ervan moeten als uitkomst uit de berekening worden gevonden. Deze correctieconstanten moeten niet alleen rekenschap geven van vergissingen en onnauwkeurigheden, maar ook van de toevalsonbepaaldheid die in sommige opzichten een fundamentele eigenschap van de natuur is.

Omdat ieder gegeven minstens één correctieconstante eist, zal het beschikbare aantal vrijheidsgraden steeds beneden het vereiste aantal uitkomsten blijven. Door deze omstandigheid is het ideaal van een aantal afhankelijke vergelijkingen onbereikbaar. Integendeel, de vergelijkingen zullen onbepaald blijven; daar er meer onbekenden dan gegevens zijn.

106 – Het uitschakelen van onbekenden

Waar het niet mogelijk is, het aantal gegevens zo hoog op te voeren, dat het aantal onbekenden overtroffen wordt, zal het aantal onbekenden moeten worden verlaagd. Nagegaan moet dus worden welke onbekenden moeten vervallen.

Reeds in 102 is geëist van de vragen van het probleem, dat ze opnieuw gesteld moeten kunnen worden. De berekende constanten moeten in meer dan een vergelijking voorkomen en zo

bijdragen tot het verklaren van meer dan een gegeven. Aan deze voorwaarde wordt niet voldaan door de correctieconstanten.

De correctieconstanten komen daarom wel in de eerste plaats in aanmerking om weggelaten te worden. De formule die ontstaat door uit de ,,juiste'' formule, die tot afhankelijkheid van de vergelijkingen zou moeten leiden, de correctieconstanten weg te laten willen wij als *ideale formule* aanduiden. Wanneer in de ideale formule aan de constanten de ,,ware'' waarden zijn toegekend zouden wij van *ware formule* willen spreken.

Het is duidelijk dat de bouw van de ideale formule niet verandert door het toenemen van het aantal gegevens. Daarom moet het mogelijk geacht worden, voldoende gegevens te verzamelen om de ideale formule te berekenen.

Toch wordt in de landbouwwetenschap vaak niet met de ideale formule gewerkt. En wel omdat die formule niet bekend is. Het landbouwkundig onderzoek is nog in volle gang. De invloeden die de plantengroei bepalen zijn nog niet alle bekend; van de wisselwerking tussen die invloeden is nog een groot deel duister. In die situatie is het niet mogelijk de ideale formule op te bouwen.

Toch moet de een of andere formule gebruikt worden. Zo een formule, die in bouw van de ideale afwijkt, willen wij aanduiden als *noodformule.*

Een noodformule kan tweeërlei doel hebben. In de eerste plaats stelt men zulke noodformules op om al tastende tot de ideale te komen. Dit is het doel van het wetenschappelijk onderzoek.

Voor de landbouwvoorlichter is een ander doel belangrijker. Het is de taak van het rassenonderzoek: te voorspellen welke gegevens het volgend jaar verzameld zullen kunnen worden. De practijk moet vooraf weten welke ervaringen ze met een bepaald ras zal hebben. Deze ervaringen moet de voorlichter voorspellen, zonder dat hij de ideale formule weet.

Hij heeft dus een noodformule nodig, die de vereiste voorspellingen zo nauwkeurig mogelijk toelaat. Hoewel deze formule onjuist is, zal ze *moeten* worden gebruikt in de practijk; immers de landbouw gaat door.

Wij komen dus tot het volgend overzicht van de besproken formules:

(106.1) De *juiste formule* geeft steeds afhankelijkheid maar bestaat niet.

De *ware formule* is kwalitatief en kwantitatief goed.
De *ideale formule* is kwalitatief goed.
De *noodformule* is kwalitatief onjuist.

In wiskundige symbolen kunnen deze formules als volgt worden omschreven:

juiste formule: $y_k = f\,((x_k - t_{x_k}), (z_k - t_{z_k}) \ldots . a_o, b_o, c_o \ldots) + t_{y_k}$
ware formule: $y = f\,(x, z, \ldots . a_o, b_o, c_o, \ldots .)$
ideale formule: $y = f\,(x, z, \ldots . a, b, c, \ldots .)$
noodformule: $y = \varphi\,(x, \ldots .)$

In deze formules worden door t_{x_k}, t_{y_k}, t_{z_k} .... aangeduid de correctieconstanten die volgens het eind van 105 bij het omstandighedencomplex k nodig zijn. Door $a_o, b_o, c_o \ldots .$ worden ware waarden van de parameters aangeduid door a, b, c parameters met onbekende waarden. Verder wordt door f aangeduid de formule met de juiste bouw, door φ de formule met de onjuiste bouw.

Het blijkt dat bij de juiste formule een identiteit ontstaat wanneer bij het omstandighedencomplex k het verkregen gegeven y_k vergeleken wordt met de getalswaarde van het rechterlid, indien daarin naast de omstandigheden $x_k, z_k \ldots .$ en de ware parameterwaarden $a_o, b_o, c_o, \ldots .$ ook de ware correctie constant en $t_{x_k}, t_{y_k}, t_{z_k} \ldots .$ zijn opgenomen.

De ware formule geeft nooit identiteit, maar is voor alle omstandigheden de beste benadering van de juiste.

De ideale formule is gelijk aan de ware, behalve dat de waarden van de parameters niet bekend zijn, deze formule is ideaal als *middel* om de ware te vinden.

De noodformule is verkeerd gebouwd. Bij gevolg komen de parameters a, b, c er misschien niet eens in voor, terwijl ook een deel van de omstandigheden, b.v. z, onvermeld kunnen zijn.

Dit laatste type formule komt in de practijk zeer veel voor en zal dus nog onze speciale aandacht moeten vragen.

B – FOUTEN EN GEWICHTEN BIJ NOODFORMULES

111 – HET METEN VAN DE STRIJDIGHEID

Wanneer er niet met de *juiste* formule gewerkt wordt, doch met de *ideale* formule of een *nood*formule, zal er strijdigheid optreden zodra er meer gegevens zijn dan er uitkomsten gevraagd worden.

Het berekenen van uitkomsten uit strijdige vergelijkingen wordt

vereffening genoemd. Met behulp van de uitkomsten kan nu *uit de vergelijkingen* worden berekend welke gegevens bij een bepaald omstandighedencomplex verwacht mogen worden. Deze „berekende gegevens" willen we als „*verwachtingen*" of „*verwachte gegevens*" aanduiden.

Deze verwachtingen zullen niet gelijk zijn aan de bijbehorende *waargenomen gegevens*. Er zal een *afwijking* overblijven tussen beide. Deze afwijkingen stellen ons in staat de *strijdigheid* in een getal vast te leggen. Een veel gebruikte maat voor de strijdigheid is de *varians*.

Wanneer b.v. vier rassen in drie blokken met elkaar worden vergeleken is het met de FISHER-methode mogelijk de volgende variansen te berekenen.

1e De varians van het gehele proefveld
2e „ „ „ de rasverschillen
3e „ „ „ de blokverschillen
4e „ „ „ de interactie tussen rassen en blokken.

De eerste varians meet de strijdigheid, die ontstaat wanneer men alle opbrengsten identiek stelt. De tweede meet de strijdigheid die ontstaat wanneer men de rassen identiek verklaart. De derde doet hetzelfde wanneer men de blokken voor identiek verklaart. De vierde tenslotte meet de strijdigheid die ontstaat wanneer men verklaart dat alle rassen *gelijk* op de eventuele blokverschillen reageren.

Wanneer de mening dat de rassen of de blokken identiek zijn juist is, werkt men dus met een *ideale* formule. *In dat geval* is de varians een maat voor de *toevallige* fouten. De varians gedraagt zich dan vaak in overeenstemming met bekende foutenwetten.

Wanneer de aanname evenwel onjuist is werkt men met een *noodformule*. In dat geval meet de varians naast de *toevallige* ook de *systematische* fouten. Zonder meer mag dan niet aangenomen worden dat de *varians* de *toevalswetten* volgt.

In verband hiermee willen wij alleen die varians, die *toevallige* fouten meet, aanduiden door σ^2. De varians die is opgebouwd uit *toevallige* en *systematische* fouten duiden wij aan als ψ^2.

Aan de hand van een voorbeeld willen wij nu nagaan welke moeilijkheden zich bij een *varians* van een *noodformule* kunnen voordoen.

112 – VOORBEELD VAN HET GEBRUIKEN VAN EEN NOODFORMULE,

Veronderstel dat onder de volgende omstandigheden x de bijbehorende gegevens y zijn waargenomen. Zowel de gegevens als de omstandigheden zijn *foutloos* in cijfers uitgedrukt.

$x =$	1089	1156	1296	1369	1521	1681	1764	1936
$y =$	142.56	146.88	155.52	159.84	168.48	177.12	181.44	190.08

Deze cijfers worden vereffend op de lijn $y = mx + q$ (met de formules voor x zeker, y onzeker) onder *aanname* dat dat de *ideale* formule is. De varians duiden we dus aan door σ^2. Hieronder volgt de berekening

	x	y	u	v	uu	uv	vv
	1089	142.56	–387.5	–22.68	150 156.25	8 788.50	514.3824
	1156	146.88	–320.5	–18.36	102 720.25	5 884.38	337.0896
	1296	155.52	–180.5	–9.72	32 580.25	1 754.46	94.4784
	1369	159.84	–107.5	–5.40	11 556.25	580.50	29.1600
	1521	168.48	44.5	3.24	1 980.25	144.18	10.4976
	1681	177.12	204.5	11.88	41 820.25	2 429.46	141.1344
	1764	181.44	287.5	16.20	82 656.25	4 657.50	262.4400
	1936	190.08	459.5	24.84	211 140.25	11 413.98	617.0256
som	11812	1321.92			634 610.00	35 652.96	2006.2080
gemid.	1476.5	165.24					

$$\overline{m} = \frac{[uv]}{[uu]} = 0.05618$$

$$\overline{q} = \overline{y} - \overline{m}\overline{x} = 82.36$$

$$\sigma^2 = \frac{[vv] - \frac{[uv]^2}{[uu]}}{n-2} = 0.5321$$

$$\sigma_{\overline{m}}^2 = \frac{\sigma^2}{[uu]} = 0.000\,000\,8384$$

$$\sigma_{\overline{m}} = 0.0009$$

$$\sigma_{\overline{q}}^2 = \left(\frac{1}{n} + \frac{\overline{x}^2}{[uu]}\right)\sigma^2 = 3.56\,\sigma^2 = 1.8943$$

$$\sigma_{\overline{q}} = 1.38$$

We vinden dus als resultaat dat

$$\sigma^2 = 0.5321$$

$$\overline{m} = 0.05618 \pm 0.0009$$

$$\overline{q} = 82.36 \pm 1.38$$

Zowel de $\overline{m}$ als de $\overline{q}$ zijn dus ,,zeer betrouwbaar'' afwijkend van 0.

Omdat de lijn iets krom is willen wij voor nadere informatie ook eens proberen de lijn te vereffenen op $y = a \lg x + b$. De berekening blijft dezelfde als de voorgaande, slechts wordt de x door $\lg x$ vervangen.

De berekening geeft het volgend resultaat:

$$\sigma^2 = 0.4804$$

$$\sigma_{\bar{a}}^2 = 8.63$$

$$\sigma_{\bar{b}}^2 = 86.4720$$

$$\overline{a} = 190 \pm 3$$

$$\overline{b} = -435.35 \pm 9.30$$

Ook de uitkomsten voor a en b maken in het licht van hun middelbare fout ,,zeer betrouwbaar'' de indruk niet gelijk aan 0 te zijn.

Niet alleen zijn de uitkomsten van beide formules alle ,,zeer betrouwbaar'' ongelijk 0, maar ze zijn ook lijnrecht met elkaar in tegenspraak. Wanneer door de q met practische zekerheid wordt uitgedrukt dat de lijn de positieve Y-as snijdt, terwijl de andere formule beweert dat de lijn de negatieve Y-as asymptotisch nadert en reeds voor de waarde $x = 1$ ($\lg x = 0$) zeer zeker een sterk negatieve waarde heeft ($b = -435.35$), klopt er iets niet.

Eerst wordt bewezen dat de constante term positief is ($q = 82.26$), dan wordt bewezen dat hij negatief is ($b = -435.35$). Tweemaal is bijgevolg ,,zeer betrouwbaar'' aangetoond dat er een constante term is. In werkelijkheid liggen alle punten foutloos op de lijn $y = 4.32\sqrt{x}$. Deze lijn gaat door de oorsprong en heeft geen constante term!

113 – De asymptotische waarde

Het is duidelijk dat de berekende middelbare fout van de uitkomsten hier niet inlicht over de afwijking tussen de berekende en de *ware* waarde van de constanten. De constanten van een noodformule hebben geen ware waarde, want de noodformule is onjuist.

Wel kan er gesproken worden van hun *asymptotische waarde*. Wanneer er een oneindig aantal waarnemingen verwerkt werden, die toevallig gekozen waren, zouden zeer bepaalde uitkomsten berekend worden. Deze waarden worden asymptotische waarden genoemd. Wij willen die waarden nu berekenen voor bovenstaand voorbeeld, waarbij wij als noodformule $y = mx + q$ kiezen. Het resultaat van de berekening willen wij als *asymptotische verwachting* aanduiden.

Bij de waarneming w, verricht onder de omstandigheid x_w hoort als ware waarde (y_o) van het gegeven

$$y_{o_w} = 4.32\sqrt{x_w}$$

De waarde volgens de noodformule is

$$y_w = mx_w + q$$

De afwijking is dus

$$4.32\sqrt{x_w} - mx_w - q. \tag{113.1}$$

Volgens de methode der kleinste kwadraten moeten wij de m en de q nu zodanig kiezen, dat de som van de kwadraten van deze afwijkingen zo klein mogelijk wordt. In het onderhavige geval is het aantal afwijkingen oneindig en de kwadraatsom dus ook. Daarom moeten wij werken met het gemiddelde.

Wanneer wij voor de overzichtelijkheid 4.32 vervangen door c, is het kwadraat van (113.1):

$$m^2x_w^2 - 2cm\, x_w^{\frac{3}{2}} + (c^2 + 2mq)\, x_w - 2cq\, x_w^{\frac{1}{2}} + q^2$$

Met deze formule kunnen wij niet verder werken, wanneer we niet eerst ons universum nader begrenzen. Wij stellen daartoe dat de x steeds tussen 1000 en 2000 ligt (wat in ons voorbeeld ook het geval was) en dat alle waarden van x in dat gebied evenveel kans hebben voor te komen. Nu kunnen wij de relatieve frequentie van de waarde x_w gelijk stellen aan $\frac{1}{1000}\, dx_w$, immers $\int_{1000}^{2000} \frac{1}{1000}\, dx_w = 1$.

Het gemiddelde van de kwadraten zal nu gelijk zijn aan

$$\frac{1}{1000}\int_{x=1000}^{2000}(m^2x^2 - 2cmx^{\frac{3}{2}} + (c^2 + 2mq)\,x - 2cq\,x^{\frac{1}{2}} + q^2)\,dx$$

Dit is

$$\frac{1}{1000}\left[\tfrac{1}{3}m^2x^3 - \tfrac{4}{5}cm\,x^{\frac{5}{2}} + (\tfrac{1}{2}c^2 + mq)\,x^2 - \tfrac{4}{3}cq\,x^{\frac{3}{2}} + q^2x\right]_{1000}^{2000}.$$

Wanneer wij stellen

$$x_a = 2000$$
$$x_b = 1000$$

wordt deze vorm

$$(113.2)\quad \frac{1}{1000}\Big\{\tfrac{1}{3}m^2\,(x_a^3 - x_b^3) - \tfrac{4}{5}cm\left(x_a^{\frac{5}{2}} - x_b^{\frac{5}{2}}\right) +$$
$$+ (\tfrac{1}{2}c^2 + mq)\,(x_a^2 - x_b^2) - \tfrac{4}{3}cq\left(x_a^{\frac{3}{2}} - x_b^{\frac{3}{2}}\right) + q^2\,(x_a - x_b).$$

Om na te gaan voor welke waarde van m en q deze vorm zo klein mogelijk wordt, moeten wij hem partieel naar m en q differentiëren. Door die differentiaalquotienten gelijk aan 0 te stellen zullen wij in dit geval het minimum vinden.

Wij vinden zo de vergelijkingen

$$\frac{1}{1000}\Big\{\tfrac{2}{3}m\,(x_a^3 - x_b^3) - \tfrac{4}{5}c\left(x_a^{\frac{5}{2}} - x_b^{\frac{5}{2}}\right) + q\,(x_a^2 - x_b^2) = 0$$

$$\frac{1}{1000}\Big\{\,m\,(x_a^2 - x_b^2) - \tfrac{4}{3}c\left(x_a^{\frac{3}{2}} - x_b^{\frac{3}{2}}\right) + 2q\,(x_a - x_b) = 0$$

Door $x_a = 2000$ en $x_b = 1000$ te substitueren wordt dit (wanneer wij tevens de eerste vergelijking door 1000 delen)

$$\frac{14000}{3}\,m - \tfrac{4}{5}\,(4\sqrt{2} - 1)\,c\sqrt{1000} + 3q = 0$$
$$3000\,m - \tfrac{4}{3}\,(2\sqrt{2} - 1)\,c\sqrt{1000} + 2q = 0.$$

Wij lossen hieruit op

$$m = \frac{3c\,(12 - 8\sqrt{2})}{5\sqrt{1000}}$$

$$q = \frac{c\,(128\sqrt{2} - 172)\,\sqrt{1000}}{15}.$$

Substitutie van $c = 4.32$ geeft

$$(113.3)\qquad \hat{m} = 0.05625273179$$
$$\hat{q} = 82.14233671$$

Dit zijn dus de asymptotische verwachtingen van de constanten van de onjuiste noodformule. Wanneer wij deze asymptotische verwachtingen in de noodformule invullen willen wij spreken van de *asymptotisch verwachte noodformule.*

114 – De fouten bij het gebruiken van een noodformule

Nu verschillende begrippen zijn besproken is het mogelijk na te gaan welke soorten fout er gemaakt kunnen worden, wanneer de juiste formule wordt vervangen door een noodformule. Dit vervangen willen wij in gedeelten doen (Zie voor de namen der formules 106.1).

a. In de eerste plaats vervangen wij de juiste formule door de *ware* formule. Hierbij worden dus de correctieconstanten weggelaten.

Doordat deze correctieconstanten bij iedere waarneming weer andere waarden hebben kan hun grootte niet stuk voor stuk worden vastgesteld. Daarom tracht men bij het nemen van proeven zoveel voorzorgen te nemen, dat de gegevens en de omstandigheden zo nauwkeurig mogelijk bekend zijn. Soms kan de orde van grootte van deze correctieconstanten vooraf geschat worden; dit kan aanleiding geven tot het toekennen van ,,gewichten'' (Zie voor de theorie hierover het leerboek van Van Uven, hfdst. III). Verder is er niets aan te doen.

Daarom hoopt men deze onbekende correctieconstanten als ,,toevallige'' fouten op te mogen vatten, die gezamelijk aanleiding geven tot een *toevallige afwijking* bij de waarneming. In formule uitgedrukt is deze afwijking t_o bij een waarneming w

$$t_{o_w} = y_w - y_{o_w}$$

waarbij y_w weergeeft de waarde van het gegeven, zoals het is waargenomen en y_{o_w} de ,,ware waarde'' er van.

Wanneer er geen andere fouten gemaakt zijn, dus wanneer met de ideale formule gewerkt is, is deze t_o de enige oorzaak van de varians. In 111 hebben wij reeds gezegd dat wij in dit geval de varians door σ^2 zullen aanduiden.

Deze varians wordt gewoonlijk genoemd de *toevalsvarians.* De wortel hiervan (σ) is volgens de bekende foutentheorieën de beste schatting van de asymptotische middelbare waarde van de ware toevallige fouten van de gegevens.

b. In de tweede plaats vervangen wij de ware formule door de asymptotisch verwachte noodformule. Dit komt er op neer dat het verband tussen omstandigheden en gegevens onjuist geïnterpreteerd wordt. Wij zouden hier dan ook van *interpretatiefout* willen spreken. De afwijking t_i tengevolge van de interpretatiefout kan als volgt in formule worden weergegeven.

$$t_{i_w} = \hat{y}_w - y_{o_w}$$

waarbij $\hat{y}_w$ de waarde voorstelt, die het gegeven y zou moeten hebben bij de waarneming w volgens de asymptotisch verwachte noodformule, terwijl de y_{o_w} op dezelfde manier als boven kan worden omschreven.

Van deze afwijkingen t_i mag niet worden aangenomen dat ze toevallig verdeeld zijn. Bij bepaalde omstandigheden behoort een bepaalde interpretatieafwijking. Dit is b.v. te zien wanneer men voor het voorbeeld uit 112 de afwijkingen berekent en die tegen de omstandigheden x afzet. Bij gebruikmaking van de noodformule $y = mx + q$ vindt men dat de afwijking positief is bij lage x, daarna negatief wordt om tenslotte weer positief te zijn.

Wanneer de gevonden afwijkingen een systematisch verband vertonen met een of meer invloeden spreekt men van systematische fout. Dergelijke afwijkingen zijn dus het gevolg van een interpretatiefout. De invloeden, die bij de systematische fout een rol spelen zijn in de noodformule op een onjuiste wijze opgenomen, of geheel weggelaten.

Een verwachting van het gemiddeld kwadraat van de interpretatieafwijkingen t_i laat zich voor ons voorbeeld met behulp van formule (113.2) berekenen. Wanneer daarin de asymptotische verwachtingen voor m en q (zie 113.3) worden gesubstitueerd, wordt de waarde 0.5144 gevonden.

Naast de symbolen ψ^2 en σ^2, die we in 111 bespraken, willen wij voor de varians tengevolge van de interpretatiefout nog het symbool φ^2 invoeren. De wortel φ kan dan worden aangeduid als middelbare interpretatiefout.

c. In de derde plaats moet de asymptotisch verwachte noodformule worden vervangen door de *berekende*. Ook in ons voorbeeld was er verschil tussen beide, zoals de volgende vergelijking leert

(114.1)	Berekende waarde volgens 112		Asymptotische waarde volgens 113 en 114
$\overline{m}$	0.05618	$\hat{m}$	0.05625
$\overline{q}$	82.36	$\hat{q}$	82.14
ψ^2	0.5321	$<\psi^2>(=\varphi^2)$	0.5144

De varians werd in 112 nog als σ^2 aangeduid. Daar wij aan de vereffeningsformules het karakter van noodformules herkenden hebben wij nu de notatie ψ^2 genomen.

In (114.1) blijkt dat er onderscheid is tussen de berekende varians ψ^2 en de asymptotisch verwachte interpretatievarians φ^2. Hierbij speelt de toevalsvarians geen rol, want er is uitgegaan van foutloze waarden, en er is bovendien voor gezorgd dat ook de afrondingsfouten niet te groot werden.

De oorzaak van het verschil ligt hier, dat volgens onze aanname alle omstandigheden tussen $x = 1000$ en $x = 2000$ evenveel kans hebben op te treden, terwijl van al die mogelijke waarden slechts 8 verwerkelijkt zijn in de cijfers van 112. Deze 8 waarden hebben nu de taak het gehele gebied tussen $x = 1000$ en $x = 2000$ te vertegenwoordigen. Ze zijn er als het ware een monster uit. Maar het monster is niet zo ideaal mogelijk getrokken. Dit blijkt als wij het gebied van x in 8 gelijke delen verdelen en daar naast vermelden welke x als vertegenwoordiger moet worden beschouwd (zie onderstaande tabel).

1000		1500	
	) 1089		) 1521
1125		1625	
	) 1156		) 1681
1250		1750	
	) 1296		) 1764
1375		1875	
	) 1369		) 1936
1500		2000	

Het blijkt dat de omstandigheid $x = 1369$ zelfs buiten zijn gebied ligt; verder liggen er maar een paar redelijk in het midden. De fout die hier gemaakt is zouden wij als *monsterfout* willen aanduiden.

Wanneer het monster volgens toeval getrokken wordt doet deze fout misschien aan de toevallige fout denken. Maar de monsterfout is opzettelijk te beïnvloeden, en daarom veel gevaarlijker. Wij willen dit aantonen aan de hand van vier voorbeelden (steeds

gebruiken we punten die voldoen aan $y = 4.32\sqrt{x}$). De varians duiden wij aan met ψ^2.

(114.2)

	$x =$	$y =$	$u =$	$v =$	
	1024	138.24	−476	−28.08	$[uu] = 829280$
	1089	142.56	−411	−23.76	$[uv] = 46647.38$
	1406	162.00	−94	−4.32	$[vv] = 2631.3984$
	1483	166.32	−17	0	
	1521	168.48	21	2.16	$\overline{m} = 0.05625 \pm 0.0012$
	1561	170.64	61	4.32	$\overline{q} = 81.95 \pm 1.88$
	1936	190.08	436	23.76	
	1980	192.24	480	25.92	$\psi^2 = 1.2436$
gem.	1500	166.32			

(114.3)

	$x =$	$y =$	$u =$	$v =$	
	1156	146.88	−338	−19.44	$[uu] = 539\,790$
	1225	151.20	−269	−15.12	$[uv] = 30\,270.24$
	1261	153.36	−233	−12.96	$[vv] = 1\,698.2784$
	1333	157.68	−161	−8.64	
	1641	174.96	147	8.64	$\overline{m} = 0.05608 \pm 0.0005$
	1723	179.28	229	12.96	$\overline{q} = 82.54 \pm 0.74$
	1764	181.44	270	15.12	
	1849	185.76	355	19.44	$\psi^2 = 0.1316$
gem.	1494	166.32			

(114.4)

	$x =$	$y =$	$u =$	$v =$	
	1024	138.24	−295	−18.09	$[uu] = 416\,542$
	1089	142.56	−230	−13.77	$[uv] = 24358.32$
	1156	146.88	−163	−9.45	$[vv] = 1427.0904$
	1260	153.36	− 59	−2.97	
	1332	157.68	13	1.35	$\overline{m} = 0.05848 \pm 0.0010$
	1406	162.00	87	5.67	$\overline{q} = 79.19 \pm 1.40$
	1521	168.48	202	12.15	
	1764	181.44	445	25.11	$\psi^2 = 0.4463$
gem.	1319	156.33			

(114.5)

	$x =$	$y =$	$u =$	$v =$	
	1226	151.20	-449	-25.11	$[uu] = 446\,954$
	1483	166.32	-192	-9.99	$[uv] = 24\,140.16$
	1561	170.64	-114	-5.67	$[vv] = 1\,305.7848$
	1641	174.96	-34	-1.35	
	1723	179.28	48	2.97	$\overline{m} = 0.05401 \pm 0.00086$
	1849	185.76	174	9.45	$\overline{q} = 85.84 \pm 1.45$
	1936	190.08	261	13.77	
	1981	192.24	306	15.93	$\psi^2 = 0.3276$
gem.	1675	176.31			

In deze vier voorbeelden hebben wij met de gebruikelijke formules de middelbare fout van de uitkomsten uit de ψ^2 berekend, zonder daarmee te zeggen dat zo iets geoorloofd is. Deze middelbare fouten zullen wij in het volgende tabelletje (114.6) aanduiden als $\psi_{\overline{m}}$ en $\psi_{\overline{q}}$. In die tabel willen wij onze uitkomsten als volgt samenvatten:

(114.6)

	Asymptotische verwachting	Voorb. 114.2	Voorb. 114.3	Voorb. 114.4	Voorb. 114.5	($_3$=114.4 $_4$=114.5)	Verschil tussen 114.4 en 114.5
$\hat{m}$	0.05625						
$\overline{m}$		0.05625	0.05608	0.05848	0.05401		
$\lvert\overline{m}-\hat{m}\rvert$		—	0.00017	0.00223	0.00224	$\lvert\overline{m}_3-\overline{m}_4\rvert$	0.00447
$\psi_{\overline{m}}$		0.0012	0.0005	0.0010	0.00086	$\overline{m}_3-\overline{m}_4$	0.0013
$\frac{\lvert\overline{m}-\hat{m}\rvert}{\psi_{\overline{m}}}$		—	0.34	2.23	2.6	$\frac{\lvert\overline{m}_3-\overline{m}_4\rvert}{\psi_{\overline{m}_3-\overline{m}_4}}$	3.39
$\hat{q}$	82.14						
$\overline{q}$		81.95	82.54	79.19	85.84		
$\lvert\overline{q}-\hat{q}\rvert$		0.19	0.40	2.95	3.70	$\lvert q_3-q_4\rvert$	6.65
$\psi_{\overline{q}}$		1.88	0.74	1.40	1.45	$\psi_{\overline{q}_3-\overline{q}_4}$	2.00
$\frac{\lvert q-\hat{q}\rvert}{\psi_{\overline{q}}}$		0.1	0.54	2.1	2.55	$\frac{\lvert\overline{q}_3-\overline{q}_4\rvert}{\psi_{\overline{q}_3-\overline{q}_4}}$	3.32
ϕ^2	0.5144						
ψ^2		1.2436	0.1316	0.4463	0.3276		

Het blijkt dat de uitkomsten van de eerste twee voorbeelden bizonder goed zijn; de varians wordt echter totaal verkeerd berekend. De varians van het eerste voorbeeld is ongeveer tienmaal die van het tweede. Wij willen deze speciale vorm van monsterfout aanduiden als *deviatiefout*.

In het derde en vierde voorbeeld zijn de afwijkingen veel beter geschat, doch hier deugen de uitkomsten niet. De verschillen tussen de uitkomsten van het derde en het vierde voorbeeld zijn zelfs „zeer betrouwbaar", hoewel ze voor hetzelfde gebied gelden en het uitgangsmateriaal „foutloos" was. Deze vorm van monsterfout willen wij aanduiden als *parameterfout*, omdat de uitkomsten, de berekende waarden van de parameters van de noodformule, afwijken van de asymptotische verwachting daarvan.

In 116 en 117 zullen deze fouten nader worden bestudeerd.

115 – Het toekennen van gewichten

Evenals bij de toevallige fouten kan men ook bij de monsterfouten trachten hun invloed te corrigeren door het toekennen van gewichten. Men zou als volgt kunnen redeneren: Het gebied dat bemonsterd moet worden ligt tussen $x = 1000$ en $x = 2000$. Hieruit zijn 8 waarnemingen getrokken, die ieder een deel van dit gebied vertegenwoordigen. Wegens de monsterfout vertegenwoordigt ieder niet een even groot deel. De grootte van het vertegenwoordigd gebied zou men als gewicht van de waarneming kunnen nemen.

In ons geval is de grootte van het gebied als volgt te schatten: De omstandigheden x worden gerangschikt volgens opklimmende grootte. Neem nu aan dat de grenzen van twee delen telkens midden tussen twee omstandigheden liggen. Ter illustratie zijn in tabel (115.1) de gewichten van voorbeeld (114.2) berekend.

Om kleine getalwaarden voor het gewicht te krijgen is in dit geval het vertegenwoordigd deel gedeeld door 20. Op deze manier zijn voor alle 4 voorbeelden gewichten berekend, deze zijn vermeld in tabel (115.2). Het gewicht is voorgesteld door g.

Met deze gewichten zijn de berekeningen overgedaan. In tabel (115.3) zijn de resultaten vermeld. De voorbeelden waarbij gewichten zijn gebruikt zijn aangeduid als (114.2*a*, 114.3*a*) enz. Ter vergelijking zijn de uitkomsten zonder gewicht er nog weer

(115.1)

x	Grenzen	Vertegen-woordigd deel	Gewicht
	1000		
1024	1056	56	3
1089	1247	191	10
1406	1444	97	5
1483	1502	58	3
1521	1541	39	2
1561	1748	207	10
1936	1958	210	10
1980	2000	42	2

(115.2)

Voorbeeld 114.2		Voorbeeld 114.3		Voorbeeld 114.4		Voorbeeld 114.5	
x	g	x	g	x	g	x	g
1024	3	1156	3	1024	1	1226	18
1089	10	1225	1	1089	1	1483	8
1406	5	1261	1	1156	2	1561	4
1483	3	1333	3	1260	2	1641	4
1521	2	1641	3	1332	2	1723	5
1561	10	1723	1	1406	2	1849	5
1936	10	1764	1	1521	4	1936	3
1980	2	1849	3	1764	8	1981	2

naast vermeld. De waarden van ψ^2 zijn herleid op het gemiddeld gewicht, dat aan de waarnemingen werd toegekend.

Het valt op dat de schattingen van de ψ^2 in de voorbeelden (114.2 en 114.3) niet veel zijn verbeterd; de uitkomsten ($\overline{m}$ en $\overline{q}$) van de voorbeelden (114.4 en 114.5) daarentegen wel. Het wil ons voorkomen dat dit zijn oorzaak vindt in de keuze van het monster. Als algemene conclusie zouden wij willen trekken, dat *het toekennen van gewichten stellig invloed heeft, maar niet afdoende helpt.*

Een andere conclusie is misschien nog belangrijker. Wanneer bij het gebruik van een noodformule gewichten worden toegekend op grond van de toevallige fouten, zullen deze gewichten invloed krijgen op de monsterfouten. Omdat de monsterfouten met de

(115.3)

	Asymptotische waarde	114.2	114.2*a*	114.3	114.3*a*	114.4	114.4*a*	114.5	114.5*a*
$\hat{m}$	0.05625								
$\bar{m}$		0.05625	0.05629	0.05608	0.05607	0.05848	0.05700	0.05401	0.05529
$\lvert\bar{m} - \hat{m}\rvert$		—	0.00004	0.00017	0.00018	0.00223	0.00075	0.00224	0.00096
$\psi_{\bar{m}}$		0.0012	0.0011	0.0005	0.0006	0.0010	0.00087	0.00086	0.00075
$\frac{\lvert\bar{m} - \hat{m}\rvert}{\psi_{\bar{m}}}$		—	0.04	0.34	0.3	2.23	0.86	2.6	1.28
$\hat{q}$	82.14								
$\bar{q}$		81.95	81.85	82.54	82.55	79.19	81.20	85.84	83.71
$\lvert\bar{q} - \hat{q}\rvert$		0.19	0.29	0.40	0.41	2.95	0.94	3.70	1.57
$\psi_{\bar{q}}$		1.88	1.68	0.74	0.89	1.40	1.30	1.45	1.15
$\frac{\lvert\bar{q} - \hat{q}\rvert}{\psi_{\bar{q}}}$		0.1	0.17	0.54	0.46	2.1	0.7	2.55	1.37
ϕ^2	0.5144								
ψ^2		1.2436	1.0246	0.1316	0.1929	0.4463	0.3756	0.3276	0.3049

systematische interpretatiefouten samenhangen kan dit zeer ongewenst zijn.

116 – De variansanalyse

Onder leiding van Prof. R. A. Fisher heeft zich een mathematische school ontwikkeld, die zich toelegt op de zgn. variansanalyse (analysis of variance).

De grondgedachte van bovengenoemde school is de volgende: Wanneer er verschillende fouten *onafhankelijk* van elkaar werken, die allen bijdragen tot de varians, dan kan de totale varians verkregen worden door de som te nemen van de variansen die iedere fout op zich zelf veroorzaakt.

Wanneer er geen monsterfout is mag in ons geval dus gesteld worden

$$\psi^2 = \sigma^2 + \varphi^2. \tag{116.1}$$

De toevallige fout is immers naar zijn aard onafhankelijk van iedere andere foutenbron. (Misschien niet altijd wat de absolute grootte betreft, maar zeker wat het teken aangaat).

Wanneer er wel een monsterfout is gemaakt zouden wij deze formule graag uitbreiden met een paar termen B en C, die rekenschap konden geven van de deviatie- en de parametervarians. Wij zouden dan krijgen

$$\psi^2 = \sigma^2 + \varphi^2 + B + C. \tag{116.2}$$

Dit mag echter niet zonder meer. De monsterfout is niet onafhankelijk van de interpretatiefout. Wanneer er geen interpretatiefout is, kan er geen monsterfout zijn. ndien onze gegevens inderdaad werden vereffend op de lijn $y = c\sqrt{x}$, zou het onbelangrijk zijn in welk gebied voor x onze waarnemingen lagen.

Aangenomen moet dus worden dat de B en de C afhankelijk zijn van de φ^2. In verband hiermee willen wij ze voorlopig aanduiden als *deviatie- en parameter t e r m.*

Naar zijn aard is de σ^2 van deze termen evenzeer onafhankelijk als van φ^2. Voor de eenvoudigheid willen wij daarom tijdelijk aannemen dat $\sigma^2 = 0$. Wanneer wij ons bovendien het speciale geval denken dat $C = 0$, wordt formule (116.2) gereduceerd tot

$$\psi^2 = \varphi^2 + B. \tag{116.3}$$

Aangezien de minimumwaarde van $\psi^2 = 0$ is, volgt hieruit

dat de minimumwaarde van B gelijk aan $-\varphi^2$ is. Ook hieruit blijkt duidelijk dat de B geen „varians" is.

Wanneer wij nu definiëren

(116.4) $$\frac{B}{\varphi^2} = A$$

dan blijkt dat

(116.5) $$A \geqslant -1$$

Een maximumwaarde van A laat zich niet vaststellen. Die is afhankelijk van de bouw van de ideale en de noodformule en van het onderzochte gebied.

In verband met (116.4) kan (116.3) worden geschreven als

(116.6) $$\psi^2 = (1 + A)\,\varphi^2$$

Wanneer wij weer $\sigma^2 = 0$ stellen krijgt (116.2) nu de gedaante

(116.7) $$\psi^2 = (1 + A)\,\varphi^2 + C$$

Ook deze vorm kan niet negatief worden dus

(116.8) $$C \geqslant -(1 + A)\,\varphi^2$$

Van C is niet alleen een minimumwaarde te vinden, maar ook een maximumwaarde. De C geeft rekenschap van de parameterfout, d.w.z. van het feit dat de berekende noodformule afwijkt van de asymptotisch verwachte. Wanneer de vereffening met de methode der kleinste kwadraten (kleinste varians) is uitgevoerd, houdt dit dus in dat de varians ten opzichte van de berekende noodformule (ψ^2) kleiner is dan de varians ten opzichte van ieder ander, en dus ook ten opzichte van de *asymptotisch verwachte noodformule*. Laatstgenoemde varians is in (116.6) gegeven.

Hieruit volgt dus

$$\psi^2 \leqslant (1 + A)\,\varphi.^2$$

of

(116.9) $$C \leqslant 0$$

Wanneer wij nu een grootheid Z definiëren als

(116.10) $$Z = -\frac{C}{(1 + A)\,\varphi^2}$$

volgt uit (116.8) en (116.9)

(116.11) $$0 \leqslant Z \leqslant 1.$$

In verband met (116.7) en (116.10) kan (116.2) nu worden geschreven als

$$\psi^2 = \sigma^2 + (1 - Z)(1 + A)\varphi^2 \tag{116.12}$$

117 – HET GEBRUIK VAN DE VARIANS

Het bestuderen van een vraagstuk kan tweeërlei doel hebben. In de eerste plaats is het mogelijk dat men een algemene wet wil omschrijven. In de tweede plaats kan het doel zijn een toekomstig gegeven te voorspellen. Naarmate het eerste doel beter bereikt is, willen wij de conclusie *nauwkeuriger* noemen, naarmate het tweede doel beter te bereiken valt, noemen wij de formule *bruikbaarder*.

Wanneer *met een ideale formule* is gewerkt, zijn de *nauwkeurigheid* en de *bruikbaarheid* alsvolgt uit de varians (σ^2) af te leiden.

De *nauwkeurigheid* wordt gemeten door de middelbare fout van alle berekende parameters (uitkomsten). Wij veronderstellen de methoden hiertoe als bekend.

De *bruikbaarheid* hangt af van de *middelbare fout* van de enkele waarneming (σ) en van de *nauwkeurigheid* der formule.

De σ geeft nl. slechts aan wat de middelbare afwijking is tussen het waargenomen, of waar te nemen, en het *ware* gegeven, d.w.z. het gegeven dat uit de ware formule voor de gegeven omstandigheden kan worden berekend.

Maar men kan de ware formule niet vinden. De *berekende* formule wijkt door een te kort aan nauwkeurigheid iets van die ware af. Met behulp van de middelbare fouten der verkregen uitkomsten kan nu worden nagegaan, wat de middelbare afwijking is tussen het *verwachte* gegeven, dat voor een bepaald omstandigheden-complex k uit de berekende formule kan worden afgeleid, en het *ware* gegeven dat voor diezelfde omstandigheden uit de *ware* formule zou zijn berekend. Deze middelbare afwijking willen wij aanduiden door U_k. Deze U_k heeft natuurlijk ook invloed op het slagen van een voorspelling. Hoe groter de U_k, des te slechter de voorspelling.

Nu is de toevallige fout, die bij een volgende waarneming gemaakt wordt, onafhankelijk van de middelbare fouten van de verkregen uitkomsten. De afwijking V_k, die het middelbaar verschil aangeeft tussen een voorspelling voor bepaalde omstandigheden en de eventuele waarneming, die onder deze omstandigheden zal volgen, kan dus worden berekend uit de formule

(117.1) $$V_k^2 = \sigma^2 + U_k^2$$

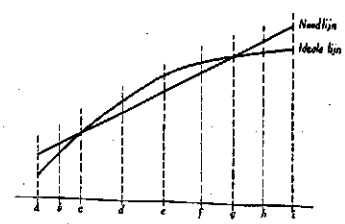
Naarmate de V_k kleiner is, is de *bruikbaarheid* van de berekende formule voor de betreffende omstandigheden groter.

Wanneer met een noodformule is gewerkt, heeft het geen zin over *nauwkeurigheid* te spreken. Er kan dan slechts sprake zijn van *bruikbaarheid*. Wij willen nu nagaan wat het verband is tussen de varians van (116.12) en de bruikbaarheid van de noodformule.

Hierbij moet er aan gedacht worden dat er over afwijkingen tussen ideale en noodformules geen gedetailleerde wetten zijn op te stellen. Een noodlijn blijft van de ideale afwijken omdat de onderlinge relatie *niet* nauwkeurig bekend is. Voor een zeer eenvoudig verband willen wij trachten een paar regels op te stellen in de verwachting, dat in ingewikkelde gevallen min of meer analoge regels zullen gelden.

Bij de uiteenzetting willen wij gebruik maken van figuur (117.2)

(117.2)



waarin een ideale lijn en een asymptotisch verwachte noodlijn zijn getekend. De snijpunten van de ideale en de noodlijn hebben als abscis $x = c$ en $x = g$. De uiterste grenzen van x zijn a en i; het gemiddelde is e. De waarden b, d, f en h liggen ongeveer midden tussen eerstgenoemde in.

Wij kunnen nu formule (116.12) het best bestuderen door ons weer voor te stellen dat er geen toevallige fouten zijn. Die zijn toch onafhankelijk van alle andere.

Eerst willen wij letten op de invloed van A. Daartoe nemen wij aan dat de asymptotisch verwachte noodlijn (toevallig) precies is berekend, dus dat $Z = 0$.

Geval (117.3) Voor het geval $A = -1$ liggen alle waargenomen punten op de snijpunten, waarvoor $x = c$ en $x = g$.

In dit geval is de vorm $(1 + A) = 0$, zodat $\psi^2 = \sigma^2$. Wanneer men hieruit zou concluderen, dat de berekende waarde ψ de middelbare afwijking tussen een *waar te nemen* en een (met de nood-

formule) *voorspeld* gegeven zou aangeven, zou men verwaarlozen, dat voor andere waarden van x dan c en g de interpretatiefout zeker mee moet tellen. Met name in de buurt van $x = a$ of e zou de afwijking veel te laag geschat zijn. In dit geval zou de bruikbaarheid van de formule dus te hoog worden aangeslagen.

Geval (117.4) Wij willen nu aannemen dat de waarnemingspunten zich van $x = c$ en $x = g$ naar weerskanten verspreiden. De helft van c gaat naar b, de andere helft naar d; die van g verdelen zich zo in de richting naar f en naar h. Het is mogelijk dat dit verschuiven zo plaats vindt, dat de zwaartepunten van beide groepen zich practisch niet verplaatsen, zodat de berekende lijn niet verandert. Dan zal de A toch groter worden. *De grootte van A hoeft de bruikbaarheid van de noodformule niet te beïnvloeden.*

Geval (117.5) Het vergroten van A gaat door tot de punten zich bevinden in $x = a$, $x = i$ en in de buurt van $x = e$. Dan heeft de A zijn maximum bereikt, maar is de asymptotisch verwachte noodlijn nog behouden gebleven, zodat nog geldt $Z = 0$. Formule (116.12) wordt dan $\psi^2 = \sigma^2 + (1 + A)\,\varphi^2$. Zou men nu de ψ gelijkstellen aan de boven aangeduide middelbare afwijking tussen waar te nemen en voorspeld gegeven, dan zou de afwijking te groot worden geschat. Deze afwijking zou voor de extreme gevallen $x = a$, e of i gelden, niet voor de tussenliggende omstandigheden. In dit geval zou de bruikbaarheid van de formule dus te laag worden aangeslagen.

Uit het voorgaande is duidelijk *dat aan het berekenen van de bruikbaarheid vooraf moet gaan de (misschien onbetrouwbare) aanname dat $A = 0$ is.*

Geval (117.6) Wij willen nu de Z nader bestuderen. Daartoe denken wij ons eerst dat de beide groepen punten in $x = c$ en $x = g$ bij elkaar blijven en gezamelijk verschuiven; b.v. van c naar b en van g naar f. I dit geval blijft de vorm $(1 + A)$ niet nul, want de A reageert op de afstand tot de asymptotisch verwachte lijn. De vorm $(1 - Z)\,(1 + A)\,\varphi^2$ blijft daarentegen wel nul, want alle punten zullen op de berekende noodlijn liggen. De Z moet dan dus 1 zijn.

Strikt genomen is de Z dus geen maat voor de afwijking tussen de berekende en de asymptotisch verwachte noodformule. Hij geeft aan hoever de waargenomen punten van de *berekende* lijn verwijderd zijn. Wanneer die punten met de berekende lijn samenvallen is de $Z = 1$. Aangezien in geval (117.3) de punten samen-

vallen met de asymptotisch verwachte lijn, die dan tegelijk de berekende is, zou men kunnen verdedigen, dat de Z dan ook 1 is en niet nul, zoals wij boven aannamen. Geval (117.3) is een limietgeval, die zich ongelijk laat belichten al naar de manier, waarop de limiet wordt bereikt. De waarde van Z is in dit speciaal geval onbepaald.

Geval (117.7) Nu moet gelet worden op de afwijkingen die er zullen zijn tussen de berekende en de asymptotisch verwachte noodformule. Laat hun varians zijn E^2. Het verband tussen E^2 enerzijds en de grootheden A en Z anderzijds moet nu nader worden bestudeerd. Wij laten daartoe de beide groepen punten zich zeer geleidelijk uit $x = c$ en $x = g$ verwijderen. De A zal dus langzaam groter worden. En even langzaam groeit de *kans* dat een berekende noodlijn beduidend van de asymptotisch verwachte afwijkt. Laten wij eens aannemen dat op een gegeven ogenblik de punten zich willekeurig over de waarden $x = b$, d, f en h hebben verdeeld.

Laten de punten b, d, f en h nauwkeurig zo gedefinieerd zijn, dat voor deze waarden van x de afstand tussen asymptotisch verwachte noodlijn en ware lijn φ bedraagt. Dan zal de waarde van A bij het bereiken van deze punten tot 0 zijn gestegen.

In dit geval *kunnen* alle punten zich geconcentreerd hebben in b en f, of d en f. In beide gevallen zal de $Z = 1$ zijn, doch de gemiddelde afwijking tussen de berekende en asymptotisch verwachte noodlijn zal geheel ongelijk zijn. *Het verband tussen de gezochte afwijking en de grootheden A en Z is tamelijk los.*

Toch is er wel enig verband: zolang de punten zich over $x = b$, d, f en h verdelen, kan de berekende noodlijn de ideale lijn niet bij $x = a$ en $x = e$ snijden. Wij kunnen dus zeggen dat de *kans* op een grote E^2 groter wordt naar mate A groter wordt.

Geval (117.8) Wanneer de punten zeer evenredig over $x = b$, d, f en h verdeeld zijn zal de asymptotisch verwachte noodlijn wel behouden zijn gebleven. De Z zal dus nul zijn. We willen nu de Z laten toenemen terwijl we de A constant houden. Daartoe laten we telkens 1 punt van h naar f gaan en b.v. van d naar b. Door deze verplaatsingen zal de noodlijn draaien en de Z groter worden. Dit gaat zo door tot alle punten in f en b zijn. Dan is de $Z = 1$. Hieruit blijkt dat *bij constante A de gezochte kwadraten* (E^2) *groter worden naarmate de Z groter wordt.* Aangezien de kans op een grote E^2 ook groter wordt naarmate de φ^2 groter is kunnen we zeggen dat

$$E^2 = P\,(A,\ Z,\ \varphi^2)$$

waarbij P een zodanige kansfunctie is dat van E^2 asymptotisch verwacht mag worden dat hij groter wordt met A, Z en φ^2.

Wij willen nu nagaan hoe groot het kwadraat (V^2) van de *afwijkingen tussen waar te nemen en voorspeld gegeven* wordt. In 116 zijn de volgende bestanddelen gevonden:

Het eerste bestanddeel is de toevallige varians σ^2.

Het tweede bestanddeel, hiervan onafhankelijk is de interpretatievarians φ^2.

Een derde bestanddeel zou de deviatieterm kunnen zijn. Dit is echter niet het geval. Zolang de asymptotisch verwachte noodformule goed berekend wordt uit een verkeerd gekozen monster zal deze verkeerde keuze van het monster geen kwaad doen. De grootheid A speelt in dit verband dus geen rol.

Als laatste bestanddeel komt de parameterterm naar voren. De parameterfout heeft de berekende varians ψ^2 *verlaagd*. Maar deze fout maakt tevens, dat de berekende lijn minder goed voor het gehele gebied geldt. De V^2 zal er dus door worden *verhoogd*, met het bedrag $P\,(A,\ Z,\ \varphi^2)$.

Als totale waarde van V^2 wordt dus verkregen

$$V^2 = \sigma^2 + \varphi^2 + P\,(A,\ Z,\ \varphi^2) \tag{117.9}$$

Vergelijking van deze formule met (116.12) toont duidelijk aan, dat een grote Z de V^2 verhoogt en dus de gevonden conclusie slechter maakt, terwijl tegelijkertijd de ψ^2 wordt verlaagd. Wanneer de ψ als „middelbare fout" wordt opgevat wordt de middelbare fout dus kleiner naarmate de conclusie slechter wordt. Wij kunnen het ook zo zeggen: *Wanneer er een parameterfout wordt gemaakt is er een sterke tendens dat de gevonden conclusie des te beter lijkt naarmate hij slechter is.*

Om de formules (116.12) en (117.9) aan elkaar gelijk te mogen stellen is het nodig dat zowel de Z als de A gelijk zijn aan 0. Pas wanneer dit het geval is kan uit de ψ^2 de V^2 worden voorspeld, zodat een uitspraak kan worden gedaan over de bruikbaarheid. *Om aan deze voorwaarden te voldoen is het nodig het gehele gebied zojuist mogelijk te bemonsteren.*

Hoewel dit ideaal gemakkelijk te stellen is, is er toch moeilijk aan te voldoen. Een noodformule wordt juist gebruikt omdat men de stof niet voldoende beheerst. Men weet vaak niet welke invloeden werkzaam zijn, hoe kan men ze dan juist in het monster opnemen?

Er zal altijd een afwijking zijn tussen *berekende* en *asymptotisch verwachte* noodformule, d.w.z. de Z en de A zullen nooit 0 zijn. Daarom zal men de grootte van A en Z moeten schatten. Dit is echter vaak onmogelijk. De interpretatiefout φ is een *systematische* fout, die men niet kent en waarvan men gewoonlijk dus ook geen foutenwetten kent. Pas wanneer men een *bewuste* reden heeft om aan te nemen dat de φ^2 de wetten van σ^2 volgt, kan men trachten te schatten in hoeverre de berekende noodlijn van de asymptotisch verwachte afwijkt. Voor ieder speciaal geval zal daarvoor de juiste weg moeten worden gevonden.

Dat uit de φ^2 geen betrouwbaarheid kan worden berekend is ook de reden waarom de ,,zeer betrouwbare'' conclusies van 112 zo zeer onjuist konden zijn. Het behoeft geen nader betoog dat de conclusies daar alsvolgt hadden moeten luiden:

Wanneer men aanneemt dat $y = mx + q$ de ideale formule is, neemt men met zeer grote waarschijnlijkheid impliciet aan dat m en q positief zijn.

Wanneer men daarentegen aanneemt dat $y = a \lg x + b$ de ideale formule is, neemt men daarmee met zeer grote waarschijnlijkheid impliciet aan, dat a positief en b negatief is.

Omdat het onmogelijk is beide formules als ideaal aan te nemen spreken de hoge (schijnbare) betrouwbaarheid van q en b elkaar niet tegen.

118 – Het kiezen van een monster en een formule

Uit het voorgaande laten zich enkele regels afleiden waaraan men zich in de practijk moet houden.

In de eerste plaats moet men zorgen dat het zwaartepunt van het monster samenvalt met het zwaartepunt van het onderzochte gebied.

Dit laat zich het gemakkelijkst aantonen door van geval (117.7) uit te gaan. Daar is verondersteld, dat de waarnemingspunten zich over $x = b$, d, f en h hadden verdeeld. Nu liet zich denken dat in werkelijkheid de punten slechts in twee van die vier plaatsen waren, b.v. $x = b$ en f, of d en f.

Wanneer de punten over d en f of b en h verdeeld zijn is de afwijking tussen berekende en asymptotisch verwachte noodlijn kleiner dan bij verdeling over b en f of d en h. Bij verdeling over b en d of f en h zou de afwijking nog veel groter zijn. In dezelfde mate als de afwijkingen groeiden, zou ook de verschuiving van het zwaartepunt groter worden.

Wanneer evenveel punten in b als in h liggen, valt x wel ongeveer

in *e*. Zijn de punten evenredig over *b* en *f* verdeeld, dan ligt het gemiddelde lager. Door een onevenredige verdeling *kan* het gemiddelde evenwel nog naar *e* verschoven worden. Wanneer de punten in *b* en *d* liggen *kan* dit niet meer. Door te zorgen dat het zwaartepunt van de omstandigheden van het monster goed ligt vermijdt men dus de zeer extreme fouten.

In de tweede plaats mag men niet extrapoleren. Wanneer wij in figuur (117.1) de lijnen verlengen tot buiten $x = a$ en $x = i$, zien wij dat de afwijking zeer snel toeneemt. De φ^2 is buiten het onderzochte gebied groter dan hier binnen. De ψ^2 is daar dus groter dan werd berekend. Wij kunnen deze eis ook zo formuleren: *Zorg dat het monster uit het gehele onderzochte gebied wordt genomen en niet uit een speciaal deel.*

In de derde plaats moet men zorgen dat de Z zo klein mogelijk is. Daarvoor is het nodig dat men meer *omstandigheden* onderzoekt dan men uitkomsten verwacht. De noodlijn $y = mx + q$ wordt door twee punten bepaald, en heeft daarom zeker twee snijpunten met de ideale lijn. Wanneer nu alle waarnemingen op twee willekeurig gekozen omstandigheden betrekking hebben zullen deze beide omstandigheden de plaats van de snijpunten bepalen. De Z zal dan dus ongeveer 1 zijn.

Aan het eind van 102 hebben wij reeds geëist, dat er meer vrijheidsgraden, en dus meer waarnemingen moesten zijn, dan er uitkomsten verwacht werden. Nu wordt hier dus de eis aan toegevoegd, dat al deze waarnemingen op verschillende omstandigheden betrekking moeten hebben. *Het is nodig dat de onderzochte omstandigheden zo gevariëerd mogelijk zijn.*

Door sterk gevarieerde omstandigheden te zoeken heeft men het gevaar dat de A positief wordt. Wij hebben reeds gezien dat de $A = 0$ moet zijn (zie conclusie aan het eind van geval (117.5). Bij het variëren van de omstandigheden moet men zich dus hoeden voor overdrijving.

Overigens heeft men de *zekerheid* dat de ψ^2 (116.12) door een grote A wordt verhoogd, en slechts de *kans* dat dit met V^2 (117.9) het geval is (zie conclusie na geval (117.4). Een te grote A heeft dus de tendens te veroorzaken, dat de V^2 te groot geschat wordt en de bruikbaarheid te klein, een te kleine A veroorzaakt het omgekeerde. Een te grote A zal daarom wellicht niet zo funest zijn als een te kleine. Wij zouden daarom willen eisen: *Tracht te zorgen dat het monster het onderzochte gebied goed weerspiegelt; maar*

bedenk dat het beter is, dat er wat teveel extreme gevallen worden opgenomen, dan te weinig. (Extreem wil zeggen: ver van de snijpunten).

Wanneer dus het monster aan al de bovenstaande eisen voldoet, moet getracht worden, de *bruikbaarste* formule te zoeken. Daartoe moet men de noodformule als een geheel beoordelen door middel van de varians ψ^2.

Wanneer alle voorzorgen zijn genomen, willen wij aannemen dat de ψ een redelijke schatting is van de middelbare afwijking tussen een *waar te nemen* en een *voorspeld* gegeven. Meer dan een redelijke schatting mag men niet verwachten daar men A en de Z niet in de hand heeft. Wij willen nu als principe voor het kiezen van een noodformule opstellen: *Kies die noodformule, die de kleinste varians geeft.*

Ook in het licht van formule (116.12) lijkt dit principe redelijk, daar *aller*kleinste varians betekent dat $\psi^2 = \sigma^2$. Indien de ideale formule zich onder de beproefde formules bevindt zou zich op grond hiervan laten verwachten dat deze automatisch de voorkeur geniet. Toch is dit niet helemaal waar. Formule (116.12) is een *asymptotische* formule. Dubbele producten tussen φ en σ zijn *asymptotisch* 0, maar kunnen in een speciaal geval wel negatief zijn. Wanneer men slechts één geval onderzoekt, zal er allicht een noodformule te vinden zijn die een negatief dubbel product geeft. Deze noodformule krijgt met bovenstaand criterium de voorkeur. Pas wanneer men een voldoend aantal proeven op gelijke wijze bewerkt mag men verwachten dat de ideale formule door de kleinste varians wordt aangewezen.

Wanneer wij dus nu eindelijk een keuze willen doen tussen de beide noodformules van 112, geven wij aan de formule $y = a \lg x + b$ de voorkeur, omdat die een varians laat zien van 0.4804 tegen 0.5321 voor de lijn $y = mx + q$. Deze voorkeur geldt natuurlijk uitsluitend het onderzochte gebied $1000 < x < 2000$. Voor $x = 0$ is het extrapoleren van $y = a \lg x + b$ veel funester dan van $y = mx + q$.

C – CORRELATIEREKENING EN LIJNVEREFFENING

121 – INLEIDING

Het komt vaak voor dat twee variabele grootheden x en y een onderlinge samenhang vertonen. Zo is er b.v. samenhang

tussen de lengte van een staaf en de temperatuur er van; er is samenhang tussen de lengte van de pink van een volwassen man en de lengte van zijn duim.

Het is gebruikelijk de mogelijke gevallen al naar de manier van samenhang in twee groepen te splitsen: het *functioneel* verband en het *stochastisch* verband. Ingeval van een *functioneel* verband wordt aangenomen dat bij een bepaalde x slechts 1 bepaalde y behoort; bij een *stochastisch* verband wordt aangenomen dat bij een bepaalde x verschillende waarden van y kunnen behoren, welke waarden ieder een bepaalde kans hebben op te treden.

Ingeval van een *functioneel* verband wordt gebruik gemaakt van de methode der *lijnvereffening*, in geval van een *stochastisch* verband van de *correlatierekening*. Dit lijkt een scherp verschil, doch is het niet, aangezien lijnvereffening en correlatierekening in hun methoden nauw verwant zijn. Het is dan ook niet mogelijk inzicht te verkrijgen in de ene methode zonder kennis te nemen van de andere.

Er is niet alleen een nauwe samenhang tussen de methoden, het verschil tussen een functioneel en een stochastisch verband is ook maar zeer betrekkelijk. Immers zelfs bij een staaf waarvan lengte en temperatuur onderling vergeleken worden moet men erkennen dat het niet waar is dat bij een bepaalde temperatuur een zeer bepaalde lengte behoort. De warmtebeweging van de ijzeratomen is een statistisch verschijnsel, waaraan kansverdelingen inhaerent zijn, en de variaties in lengte die het gevolg zijn van de warmtebeweging zijn minstens aan dezelfde kanswetten onderworpen. Ook bij deze staaf moet men dus spreken van een stochastisch verband, aangezien bij een bepaalde temperatuur niet een minitieus bepaalde lengte *behoort*. Wanneer men het verband nagaat tussen de waargenomen lengte en de waargenomen temperatuur komt hierbij nog een ander statistisch verschijnsel nl. de kansverdeling van de meetfouten. Deze kansverdeling kan men desnoods als onbehoorlijk verwerpen. Maar dit neemt niet weg dat er bij onze staaf naast het functioneel verband ook een essentieel stochastisch verband is. Zo moet er omgekeerd bij correlatierekening steeds naast het stochastisch verband sprake zijn van een functioneel element.

De besproken staaf gaf dus *bijna* een functioneel verband te zien. Een iets „stochastischer" verband zal men vinden bij het volgend onderzoek: Men gaat na wat de samenhang is tussen de

leeftijd van een kind en zijn armlengte. Wanneer men allerlei kinderen in het onderzoek betrekt zal blijken dat niet alle kinderen van 3 jaar dezelfde armlengte hebben. Hier is duidelijk een stochastisch verband. Toch is hier niet steeds plaats voor correlatierekening. Wanneer er b.v. kinderen van 0—20 jaar in het onderzoek worden betrokken, moeten de resultaten met de methode der lijnvereffening verwerkt worden. (De reden staat aan het eind van 122 aangegeven).

Dit geldt ook wanneer men niet de leeftijd als één van de variabelen neemt, maar voor kinderen van 0—20 jaar b.v. het verband tussen de lengte van arm en voet nagaat. Niettegenstaande het stochastisch verband moet ook hier lijnvereffening worden toegepast.

Wij willen nu trachten in de sfeer van de correlatierekening te komen door het functioneel element, nl. het groeien van het kind met al zijn ledematen uit te schakelen.

Wanneer men kinderen neemt van dezelfde leeftijd, en dan het verband tussen lengte van arm en voet bestudeert, zal er zeker van correlatierekening gebruik kunnen worden gemaakt. Maar .... wat is „dezelfde leeftijd"? Moet de leeftijd tot op de dag gelijk zijn, of is een speling van een maand toegelaten, een paar jaar misschien nog? Het is niet mogelijk te zeggen: hier houdt de lijnvereffening op en begint de correlatierekening!

Voor de scheiding is zeker niet beslissend in hoeverre het functioneel element verdwenen is. Immers wanneer men kinderen neemt die op de dag nauwkeurig 3 jaar zijn, dan mag men zeker gebruik maken van de correlatierekening; maar een functioneel verband is hier aanwezig. Zou men deze kinderen rangschikken naar toenemende lichaamslengte, dan zullen korte kinderen meestal korte armen *en* korte voeten hebben. Door het sorteren naar leeftijd is de invloed van het groeien niet geheel uitgeschakeld. Het ene kind groeit sneller dan het ander.

Wij kunnen zelfs zeggen dat zonder functioneel element geen correlatie mogelijk is. *Toevallige* fouten zijn immers niet gecorreleerd. In zekere zin is de correlatiecoëfficient een maat voor de verhouding waarin toevallig en functioneel element dooreen geweven zijn.

Onze conclusie moet dus zijn *dat zowel in de correlatierekening als bij de lijnvereffening gewerkt wordt met een universum waarin een functioneel element ligt, waaraan de leden van het universum stochastisch gebonden zijn.*

Blijft de vraag: Wat is het verschil tussen de twee methoden. Dit zal in de volgende paragrafen nagegaan worden.

122 – De betekenis van de kansverdeling voor de correlatierekening

Wanneer telkens aan een lid van een universum twee grootheden x en y gemeten kunnen worden, dan noemen wij x en y aan elkaar gecorreleerd wanneer x en y ieder aan een kansverdeling zijn onderworpen.

In het vervolg zal slechts rekening worden gehouden met het geval dat x en y beide normaal verdeeld zijn.

Het is nu mogelijk uit het universum die leden uit te zoeken die een bepaalde x gemeenschappelijk hebben. Indien men voor zo'n „ondergroep" de gemiddelde y bepaalt zal deze gemiddelde y een functie zijn van de gemeenschappelijke x.

Er laten zich nu drie gevallen onderscheiden.

1e Wanneer x groter wordt, wordt de gemiddelde y groter. Dit heet *positieve correlatie.*

2e Wanneer de x groter wordt, wordt de gemiddelde y kleiner. Dit heet *negatieve correlatie.*

3e Wanneer de x verandert, verandert de gemiddelde y niet. Dit kan worden aangeduid als *geen correlatie,* maar moet als grensgeval in de beschouwingen worden betrokken. Vandaar dat we deze mogelijkheid in de definitie van correlatie niet hebben uitgesloten.

Wij willen nu een paar eigenschappen beschouwen van de *normale positieve* correlatie. Daartoe veronderstellen wij als bekend, dat in een correlatiediagram plaatsen van gelijke kansdichtheid verbonden kunnen worden door een ellips. In figuur (122.1) is zulk een ellips in beeld gebracht.

Wanneer wij de waarde van de kansdichtheid van iedere ellips langs een derde as (Z-as) hadden afgezet, zouden wij een driedimensionale figuur hebben gekregen, die de kansdichtheid (z) weergaf als functie van x en y. Wij veronderstellen weer als bekend dat iedere vertikale doorsnede door deze figuur, evenwijdig aan de Y-as, weer een normale kansverdeling laat zien van de waarden van y bij de constant gehouden waarde van x. Al deze vertikale doorsneden hebben dezelfde (partiële) standaardafwijking.

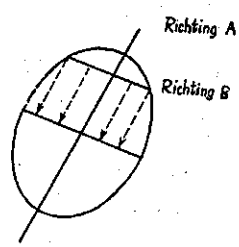
Zo kunnen wij ook van de vertikale doorsneden evenwijdig aan de X-as zeggen dat ze allen normale verdelingen te zien geven met dezelfde (partiële) standaardafwijkingen.

Wij kunnen zelfs nog verder gaan en zeggen dat de vorm van de figuur onafhankelijk is van de ligging van de X en Y coördinaten. Wij mogen het coördinatenstelsel laten draaien om de Z-as. Aangezien in iedere positie de doorsneden, evenwijdig aan de X- of Y-as, een normale kansverdeling laten zien, kunnen wij in het algemeen zeggen dat een vertikale doorsnede volgens *iedere* koorde van de kansellips een normale verdeling te zien geeft. Hier kan weer als bizonderheid aan worden toegevoegd, dat evenwijdige doorsneden steeds dezelfde standaardafwijking hebben.

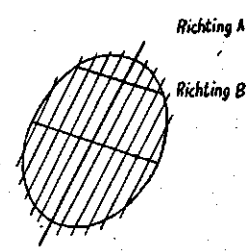
Wij willen nu twee toegevoegde richtingen beschouwen van de kansellips. Uit de analitische meetkunde is bekend dat een middellijn in de ene richting (A) alle koorden die in de toegevoegde richting (B) lopen middendoor deelt. De normale kansverdelingen die in de vertikale doorsneden volgens de koorden in de richting B worden gevonden, hebben dus allen hun top boven de middellijn in de richting A.

Wij kunnen nu alle doorsneden in de richting B zodanig naar de middellijn in die richting verschuiven, dat ieder punt evenwijdig aan de richting A wordt verplaatst (zie fig. 122.1). Wij krijgen dan boven deze middellijn een aantal normale verdelingen

(122.1)



(122.2)



gesuperponeerd, die allen hun top boven het middelpunt der ellips hebben en allen dezelfde standaardafwijking hebben. De kanskrommen zijn dus congruent wanneer men het kansoppervlak binnen iedere doorsnede = 1 stelt en op eenzelfde schaal standaardiseert.

Inplaats van de doorsneden te verschuiven kunnen wij ook het grondvlak in smalle, onderling even brede, strookjes indelen, zodanig dat de overeenkomstige delen der in bovenbedoelde zin congruente kanskrommen in de verschillende doorsneden in dezelfde strook komen te liggen (zie fig. 122.2). De grenzen der stroken lopen dan in de richting A.

Uit het voorgaande volgt nu, dat het aantal punten dat in iedere

strook verwacht mag worden, aan de normale verdeling moet beantwoorden.

Maar ook binnen iedere strook zijn de punten normaal verdeeld, omdat ook in doorsneden in de richting A normale verdeling optreedt.

Het voorgaande kunnen wij in het kort alsvolgt samenvatten. (122.3).

a. Het correlatiediagram kan op willekeurige wijze in smalle maar even brede evenwijdige stroken verdeeld worden. De richting waarin de stroken lopen zij A.

(Deze stroken moeten zo smal zijn dat ze als benadering van een koorde kunnen gelden).

b. De verdeling van de punten die binnen een strook vallen zal aan de normale verdeling beantwoorden.

c. De aantallen punten, die binnen de opeenvolgende stroken vallen zullen ook aan de normale verdeling beantwoorden.

d. De lijn, die de gemiddelden uit iedere strook verbindt zal lopen in de richting B, die toegevoegd is aan richting A.

Wij willen de richting A aanduiden als „richting van middelen", een term die ook door Ir W. C. VISSER wordt gebruikt. De lijn die de gemiddelden verbindt, duiden wij aan als de vereffeningslijn.

De vereffeningslijn moet rekenschap geven van het functioneel element, dat in het materiaal voorhanden is (zie slot van 121); de kansverdeling die we vinden volgens de „richting van middelen" moet rekenschap geven van het stochastisch verband dat er nu eenmaal is.

Uit het voorgaande zal duidelijk zijn dat de richting van middelen geheel willekeurig kan zijn, en dat dus ook iedere vereffeningslijn mogelijk is. De vraag is nu: welke vereffeningslijn is de meest juiste?

Om dit na te gaan kan men als criterium stellen, dat men het functioneel element zo sterk mogelijk wil schatten, en het stochastisch verband zo zwak mogelijk. Dit criterium is hierom redelijk omdat het (asymptotisch) niet mogelijk is een deel van het stochastisch verband bij het functioneel element te trekken. Daarentegen is het wel mogelijk, door verkeerde interpretatie een deel van het functioneel element als niet functioneel, en dus als stochastisch te zien.

Ons criterium komt dus hierop neer, dat men die richting van middelen zoekt, die de kleinste standaardafwijking *binnen* de strook geeft.

Bij het zoeken van die richting kan men gebruik maken van de wet dat de standaardafwijking in evenwijdige stroken gelijk is, zodat het voldoende is die stroken onderling te vergelijken, waarin het middelpunt van de ellips is gelegen.

Als maat voor de standaardafwijking van een bepaalde strook kan genomen worden de afstand van het middelpunt tot de omtrek van de ellips binnen de strook. (Dit vloeit voort uit het feit dat de ellips punten van gelijke kansdichtheid verbindt, dus vergelijkbare punten. De genoemde afstanden zijn dus evenredig met de standaardafwijkingen).

Het is duidelijk dat deze afstand het kortst is wanneer de strook loopt langs de korte as van de ellips. De toegevoegde richting van de korte as geeft dus de beste vereffeningslijn. Dit is de lange as.

Wij kunnen dus concluderen dat de lange as het geschiktst is om het functioneel element van het materiaal weer te geven.

In verband met deze keuze kunnen wij nu zeggen: *Het is typerend voor een normale correlatie dat er niet alleen een normale kansverdeling is in de richting loodrecht op de vereffeningslijn, maar ook in de lengterichting van deze lijn.*

Om de beide normale kansverdelingen gemakkelijk te kunnen onderscheiden willen wij het voetpunt van de loodlijn, die uit een waarnemingspunt op de vereffeningslijn wordt neergelaten aanduiden als de *totaalindruk* van dat waarnemingspunt. Wij hebben dus een normale verdeling van de totaalindrukken en een normale verdeling van de waarnemingspunten rond hun totaalindruk.

123 – De maateenheden van x en y

In de vorige paragraaf is uiteengezet dat de lange as het geschiktst is om het functioneel element van het materiaal weer te geven. Deze conclusie wordt enigszins verwonderlijk wanneer men bedenkt dat de lange as niet tegen projectie bestand is. Hiermee wordt dit bedoeld: Wanneer men uit de gegeven ellips een nieuwe figuur afleidt, door de afstand van ieder punt van de ellips tot de X-as b.v. te halveren, dan zal de nieuwe figuur weer een ellips zijn, maar de middellijn die op deze manier uit de lange as van de oude ellips wordt verkregen zal niet meer de lange as zijn. De nieuwe lange as ligt in dit voorbeeld horizontaler dan de gevonden middellijn.

Wij komen dus tot de conclusie dat de meest plausibele weergave van het functioneel element afhankelijk is van de maat-

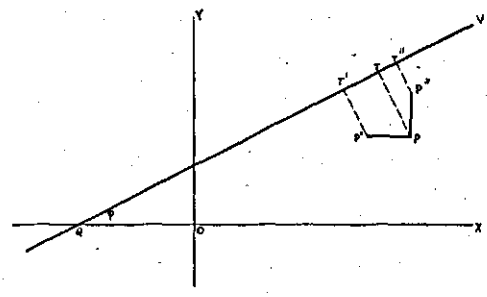
eenheden, waarin x en y zijn uitgedrukt; een conclusie, die niet zonder meer plausibel is.

De betekenis van deze conclusie kan aan de hand van een voorbeeld worden toegelicht. Veronderstel dat de correlatie is onderzocht tussen de lengten van vaders en hun volwassen oudste zonen. Veronderstel verder dat de lange as aangeeft dat bij een lengtetoename van de vaders van 1 cm ook een lengtetoename van de zoons van 1 cm behoort.

Wanneer nu de lengte van de vaders in cm wordt uitgedrukt en de lengte van de zoons in mm dan mag uit bovenstaande aanname niet afgeleid worden, dat bij een lengtetoename van de vaders van 1 cm een lengtetoename van de zoons van 10 mm behoort. De lengtetoename van de zoons, in mm uitgedrukt, zal groter zijn.

Om deze abnormaliteit te verklaren moeten wij nagaan welke invloed de x en de y uitoefenen op de *totaalindruk*, die aan het eind van 122 gedefinieerd is.

(123.1)



In figuur (123.1) is een waarnemingspunt P getekend en de vereffeningslijn QV, die een hoek φ met de X-as maakt. Het punt T geeft de totaalindruk van P weer.

Nagegaan moet dus worden in welke mate de totaalindruk T verandert, wanneer een der coördinaten x of y van P varieert. Uit de figuur valt gemakkelijk af te leiden dat een verandering in de x van P tot P' een verandering in de totaalindruk van T tot T' geeft die gelijk is aan

(123.2) $$\overline{TT'} = \overline{PP'} \cos \varphi;$$

evenzo laat zich vinden bij een verandering in de y van P tot P''

(123.3) $$TT'' = \overline{PP''} \sin \varphi.$$

Wanneer de veranderingen $\overline{PP'}$ en $\overline{PP''}$ aan elkaar gelijk zijn verhouden hun invloeden op de totaalindruk T zich dus als

$$\cos \varphi : \sin \varphi.$$

Naarmate de hoek φ kleiner wordt heeft dus de x meer invloed op de totaalindruk, en de y minder.

In de correlatierekening hangt de waarde van de hoek φ tennauwste samen met de standaardafwijking van x en y. Naarmate de standaardafwijking van x groter wordt in verhouding tot die van y wordt de hoek φ kleiner. De bovenstaande regel kan dus ook alsvolgt worden geformuleerd.

Naarmate de standaardafwijking van x groter wordt ten opzichte van die van y wordt de invloed van x op de totaalindruk groter en die van y kleiner

of, nog anders gezegd

Die coördinaat heeft de grootste invloed op de totaalindruk die de grootste standaardafwijking heeft.

Wat dit betekent willen wij toelichten aan de hand van de reeds besproken correlatie tussen de lengte van vaders en hun volwassen oudste zonen. Verondersteld werd dat de lange as aangaf dat bij een lengtetoename van de vaders van 1 cm ook een lengtetoename van de zoons van 1 cm behoorde. In dit geval is de hoek φ dus 45° en de invloed van vaders en zoons op de totaalindruk gelijk.

Wanneer daarentegen als maateenheid voor de lengte der vaders de cm wordt genomen, en voor die van de zoons de mm, dan wordt de φ veel groter dan 45°, indien de lengten der zoons langs de Y-as zijn afgezet. In dit geval hebben de zoons dus meer invloed op de totaalindruk dan de vaders, wat in een verschuiving der lange as tot uiting komt.

Tot dusver hebben wij een aantal wiskundige wetten beschreven; nu moet nagegaan worden hoe die wetten in de practijk moeten worden toegepast. Met name moet de vraag gesteld worden of het zinvol is dat de ene coördinaat meer invloed heeft op de totaalindruk dan de ander.

Het is zonder meer duidelijk dat de beide coördinaten dan niet meer gelijksoortige plaatsen innemen, maar principieel in aard verschillen. Dit nu is bij correlatierekening bijna (en waarschijnlijk helemaal) nooit de bedoeling. Men wil de beide coördinaten in de regel gelijkwaardig zien.

Laten wij dit nagaan aan de hand van de correlatie tussen de lengte van een volwassen vrouw en de lengte van haar haar.

Het is duidelijk dat hier sprake is van een stochastisch verband, van de vrouwen van een bepaalde lengte hebben sommigen kort haar, anderen hebben het lang of halflang. Er zal echter ook een functioneel element moeten zijn. Wanneer een korte vrouw hetzelfde kapsel kiest als een lange zullen haar haren naar evenredigheid korter zijn.

Wanneer wij als maateenheid voor beide lengten de cm kiezen zal de standaardafwijking van de lengte van het haar vermoedelijk verreweg het grootst zijn, door het verschil in kapsel dat een vrouw kan kiezen. In verband met bovenstaande regel is dus duidelijk dat de lengte van het haar verreweg de grootste invloed heeft op de totaalindruk. Is er enige grond om de lengte van het haar meer bepalend te achten voor de totaalindruk dan de lengte van de vrouw zelf? Zo'n reden lijkt moeilijk te vinden. Veeleer zou men kunnen overwegen de lengte van de vrouw de grootste invloed te geven. De lengte van het lichaam lijkt toch wèl zo'n karakteristieke eigenschap als de lengte van het haar! Maar ook deze gedachte heeft een twijfelachtige waarde. Het gaat niet om de totaalindruk van de *vrouw*, maar om de totaalindruk van het waarnemingspaar „lengte van het lichaam van een vrouw, lengte van haar haar". En dit waarnemingspaar dient niet om de vrouw te bestuderen, maar om de algemene samenhang tussen lichaamslengte en haarlengte na te gaan. In verband met die laatste samenhang is het begrip totaalindruk gedefinieerd, en in verband met deze samenhang zijn beide waarnemingen gelijkwaardig. Ze behoren dus ook dezelfde invloed te hebben op de totaalindruk.

Onze conclusie aan het eind van 122 moeten wij dus in zoverre aanvullen, dat wij zeggen:

De lange as is het geschiktst om het functioneel element van het materiaal weer te geven, *wanneer de maateenheden zodanig gekozen zijn dat deze as een hoek van* 45° *met de X-as maakt.* In dit geval zijn de standaardafwijkingen van x en y gelijk.

Het probleem is dus niet hoe de lange as te vinden, maar hoe de juiste verhouding tussen de maateenheden te vinden.

124 – De regressielijnen

Een veel voorkomend probleem bij de correlatierekening is het volgende. Stel dat van een bepaald lid dat tot het universum behoort slechts de x of de y is bepaald, wat is dan de waarschijnlijkste waarde van de andere coördinaat? Bij de correlatie van de

lengten van vaders en zoons is de vraag dus deze: wanneer men de lengte van een bepaalde zoon weet, hoe lang is zijn vader dan vermoedelijk.

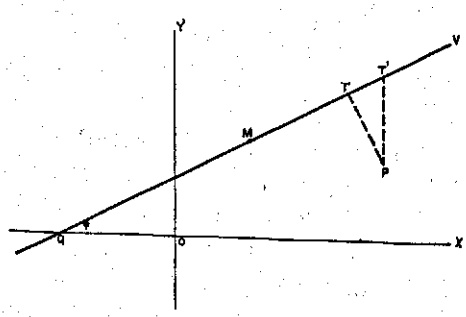
Om dit probleem op te lossen zou men de onbekende coördinaat kunnen berekenen door substitutie van de bekende in de formule van de lange as, waarvan wij vonden dat hij het functioneel verband het beste weergeeft. Zoals bekend mag worden verondersteld zou dit onjuist zijn. In dit geval moet gebruik worden gemaakt van de z.g.n. *regressielijnen*. Dit is duidelijk, wanneer men let op de regel van (122.3) sub *d*. Wanneer alleen de x bekend is weet men dat het waarnemingspunt ligt in een strook evenwijdig van de y as. De vereffeningslijn, die bij deze richting van middelen behoort loopt in een richting, toegevoegd aan de richting van de y as. Zo zal bij een bekende y als vereffeningslijn *die* genomen moeten worden, die loopt in de richting toegevoegd aan de x as. Het zijn deze lijnen die als *regressielijnen* bekend staan.

Hoewel de juistheid van het gebruik van de regressielijnen dus gemakkelijk is aan te tonen, lijkt het toch even verwonderlijk, dat de meest geschikte vereffeningslijn niet gebruikt mag worden.

Wij willen daarom nagaan waarom de lange as niet bruikbaar is om de ene coördinaat van een waarnemingspunt uit de ander af te leiden.

Wij noemen de coördinaten van het waarnemingspunt (zie fig. (124.1) (x_P, y_P); die van de totaalindruk (x_T, y_T)

(124.1)



Wanneer de vereffeningslijn QV vaststaat is de plaats van punt T volledig bepaald door de waarde van x_T. Het is gemakkelijk in te zien dat de stelling dat er een normale kansverdeling is in de lengterichting van de vereffeningslijn (zie eind 122) inhoudt dat de waarden van x_T normaal verdeeld zijn.

Wanneer naast de vereffeningslijn QV ook nog de totaalindruk

$T(x_T, y_T)$ bekend is wordt de ligging van het waarnemingspunt P volkomen bepaald door de waarde van x_P. Het is weer gemakkelijk in te zien dat de stelling dat er een normale kansverdeling is in de richting loodrecht op de vereffeningslijn (zie eind 122) inhoudt dat de waarden van x_P (bij vaststaande x_T) normaal verdeeld zijn. Het gemiddelde van deze waarden x_P is x_T, zoals uit de definitie van de vereffeningslijn blijkt.

Het probleem uit de bekende x_P de onbekende y_P af te leiden kan nu ook anders worden geformuleerd. Wij kunnen vragen: Welke totaalindruk hoort waarschijnlijk bij de bekende waarde van x_P? Immers wanneer naast de lijn QV de waarde van x_P en de totaalindruk (x_T, y_T) gegeven zijn is de y_P bepaald. Men moet dus trachten uit de gevonden waarde x_P de waarschijnlijkste waarde van x_T af te leiden.

Een eerste benadering van de oplossing vindt men, wanneer men met elkaar vergelijkt de mogelijkheden (bij vaststaande x_T)

(124.2) $$\begin{array}{l|l} a & x_P' = x_T - \alpha \\ b & x_P'' = x_T + \alpha. \end{array}$$

Omdat de x_P normaal verdeeld is om x_T is de kans op een afwijking $-\alpha$ steeds even groot als de kans op een afwijking $+\alpha$. Het lijkt daarom redelijk de x_P als schatting van x_T te nemen. Dit geeft dus de schatting dat de totaalindruk op de lange as moet liggen in T^I (zie fig. 124.1).

Een nauwkeuriger resultaat krijgt men wanneer men de volgende mogelijkheden vergelijkt

(124.3) $$\begin{array}{l|l} a & x_T' = x_P + \alpha \\ b & x_T'' = x_P - \alpha. \end{array}$$

Wanneer de x_P niet gelijk is aan de abscis van het middelpunt der kansellipsen x_M zal de kans op x_T' en x_T'' niet gelijk zijn. Die waarde van x_T, die het dichtst bij x_M ligt zal de meeste kans hebben, aangezien x_T normaal om x_M is verdeeld. Wanneer wij veronderstellen dat $x_P > x_M$ dan is de kans op x'_T kleiner dan de kans op x''_T; dit geldt voor alle mogelijke waarden van α.

In (124.2) hadden wij daarom moeten vergelijken de mogelijkheden

(124.4) $$\begin{array}{l|l} a & x_P = x_T' - \alpha \\ b & x_P = x_T'' + \alpha. \end{array}$$

Omdat in evenwijdige stroken de standaardafwijkingen gelijk

zijn (zie pag. 39, laatste alinea) is de kans dat er een afwijking $-a$ optreedt bij gegeven x_T' even groot als de kans dat er een afwijking $+a$ optreedt bij gegeven x_T''. *Maar de kans op het optreden van* x_T'' *is groter dan van* x_T' (*wanneer* $x_P > x_M$), *en daarom* is de kans op (124.4*b*) groter dan op (124.4*a*).

Wij vinden dus de volgende stelling:

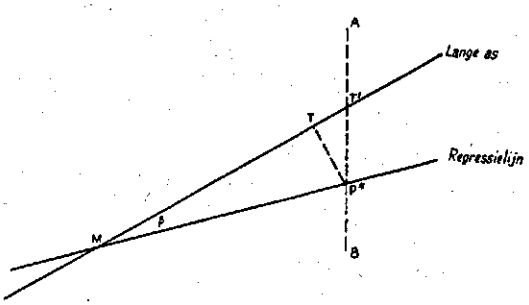
Van wege de kansverdeling in de richting van de lange as ligt een gevonden waarde x_P ($> x_M$) gemiddeld boven de bijbehorende x_T; een waarde x_P ($< x_M$) ligt gemiddeld beneden de bijbehorende x_T. De bijbehorende y_P ligt dus ook dichter bij de y_M dan door de lange as wordt aangegeven.

Onze conclusie is dus:

Het gebruik van de regressielijnen in de correlatierekening is nodig, om de invloed van de kansverdeling in de richting van de lange as tot gelding te brengen.

Voor de duidelijkheid is de onderlinge positie van de lange as en de regressielijn nog eens in beeld gebracht in fig. (124.5)

(124.5)



Verondersteld is dat uit de reeds beschikbare gegevens van een correlatiegeval berekend konden worden het middelpunt M, de lange as en de regressielijnen, waarvan een in tekening is gebracht. Nu is van een nieuw waarnemingspunt P de x_P vastgesteld. De vraag is, wat is de y_P?

Uit het gegeven x_P is bekend dat P moet liggen op de stippellijn AB. Indien men de y_P uit de lange as zou moeten berekenen zou men het waarnemingspunt T' aan moeten nemen. Omdat men de regressielijn moet gebruiken dient men de P in P^* aan te nemen. Hiermee neemt men dus tegelijk aan dat de totaalindruk van P in T valt. Er kan dus evengoed gezegd worden dat niet de plaats

van P is gekozen in P^*, maar dat de plaats van zijn totaalindruk is gekozen in T.

Waardoor wordt de keuze van T nu bepaald? Door de kansverdeling in de richting van de lange as, en de kansverdeling in de richting TP. Hoe korter de afstand MT, des te meer kans is er dat T als totaalindruk voorkomt; hoe korter de afstand PT, des te groter is de kans dat T de totaalindruk van P is. Omdat P op de lijn AB moet blijven kan TP alleen korter worden wanneer T naar T' nadert. Als T tussen M en T' ligt wordt MT daardoor juist langer en onwaarschijnlijker. De plaats van T is dus een compromis tussen het streven de afstand TM en de afstand TP zo klein mogelijk te houden. Wanneer de hoek tussen lange as en regressielijn β is, is gemakkelijk in te zien dat de waarde van de breuk $\frac{TM}{TP^*}$ steeds $\cot \beta$ is, dus constant.

Wij vinden dus als vaste regel van de regressielijn: *Wanneer er een verschil bestaat tussen een waargenomen waarde van x (x_P) en de gemiddelde waarde x_M, dan wordt dit verschil bij een bepaald correlatiegeval naar een vaste verhouding verdeeld over de invloed van de kansverdeling in de richting van de lange as, en de invloed van de kansverdeling loodrecht daarop.*

Niettegenstaande in alle stroken loodrecht op de lange as dezelfde standaardafwijking voorkomt, wordt dus groter afwijking aangenomen naarmate de strook verder van het middelpunt verwijderd is.

Tenslotte nog deze opmerking. De regressielijnen zijn ongevoelig voor de maateenheden van x en y, evenals de correlatiecoëfficient. Omdat dit de grootheden zijn waarmee in de correlatierekening meestal wordt gewerkt is het in de practijk gewoonlijk niet nodig op de maateenheden te letten. Wij hebben er in de vorige paragraaf uitvoerig bij stilgestaan in verband met de problemen van de lijnvereffening, die wij nu gaan bespreken.

125 – Vergelijking van de methoden van correlatierekening en lijnvereffening

Aan het eind van 122 werd uiteengezet, dat het typerend is voor de normale correlatie, dat er niet alleen een kansverdeling is in de richting loodrecht op de vereffeningslijn, maar ook in de lengterichting van die lijn.

Wij kunnen hier nu tegenoverstellen *dat het typerend is voor de*

lijnvereffening dat de kansverdeling in de lengterichting van de vereffeningslijn ontbreekt.

Niettegenstaande dit verschil worden bij de lijnvereffening dezelfde formules gebruikt als bij de correlatierekening, zoals uit het volgende zal blijken:

VAN UVEN bespreekt in hoofdstuk IX van zijn leerboek twee algemeen bekende mogelijkheden van lijnvereffening

1e één coördinaat (b.v. x) is vrij van fouten;

2e de fouten van x en y zijn even groot.

In het eerste geval wordt bij foutloze x als richting van middelen genomen de richting evenwijdig aan de Y-as. In het tweede geval de richting loodrecht op de vereffeningslijn. De formules die in het eerste geval gevonden worden zijn identiek met die van de regressielijnen uit de correlatierekening; de formules van het tweede geval geven bij correlatierekening de vergelijking van de lange as.

Wanneer men de toepassing van deze formules bij correlatierekening en lijnvereffening vergelijkt, vallen een paar trekken op.

Bij het gebruik maken van de lange as is er een verschil in het begrip *standaardafwijking* bij correlatie en lijnvereffening. Bij correlatie kan men onderscheiden tussen twee soorten standaardafwijking. In de eerste plaats de standaardafwijking van de ene variabele (b.v. x), die men vindt wanneer men alle gevonden waarden van x samenvat zonder te letten op de waarde van de andere variabele. In de tweede plaats de standaardafwijking van die waarden van x, die behoren bij waarnemingspunten, die de y gemeenschappelijk hebben. Deze laatste standaardafwijking wordt *partiële standaardafwijking* genoemd, de eerste zouden wij totale standaardafwijking kunnen noemen.

Bij correlaties geldt nu de wet dat de partiële standaardafwijking $\sqrt{1-\gamma^2}$ maal de totale standaardafwijking is voor iedere variabele. (Door γ duiden wij de correlatiecoëfficient aan). De verhouding tussen beide standaardafwijkingen is dus voor x en voor y gelijk.

Aangezien de x en y slechts dan gelijke totale standaardafwijkingen kunnen hebben wanneer de lange as onder een hoek van 45° met de X-as staat, moet deze voorwaarde dus ook vervuld zijn om gelijke *partiële* standaardafwijkingen te hebben.

Rekentechnisch komt met het begrip *partiële standaardafwijking* uit de correlatierekening het begrip middelbare fout van de enkele waarneming uit de lijnvereffening overeen. Hier is evenwel geen verband tussen middelbare waarnemingsfout en totale variatie

Bij gevolg kan bij lijnvereffening de vereffeningslijn wel een andere hoek dan 45° met de X-as maken.

Bovenstaand verschijnsel leidt tot eigenaardige consequenties. Op grond van het feit dat de fouten van x en y gelijk zijn, raadt VAN UVEN als richting van middelen de richting loodrecht op de vereffeningslijn aan. De invloed die de coördinaten x en y op de totaalindruk van het punt hebben, zal nu beantwoorden aan (123.2) en (123.3). Wanneer de daargenoemde hoek φ niet 45° is zal de ene coördinaat meer invloed hebben op de totaalindruk dan de ander, niettegenstaande beide coördinaten even goed zijn.

Ook de projectie geeft moeilijkheden. Wanneer de maateenheid langs de X-as gehalveerd wordt, worden de waarden van x numeriek tweemaal zo groot, de fout van x dus ook. Omdat de hoek φ hierdoor kleiner wordt zal de invloed van de x op de totaalindruk ook vergroten.

Wij vinden hier dus de regel terug die wij in 123 bij de correlatierekening vonden: *Wanneer door verandering van de maateenheden van de variabelen de middelbare fout van x groter wordt ten opzichte van die van y, wordt de invloed van x op de totaalindruk ook groter en die van y kleiner.* Het is duidelijk dat dit een onprettige eigenschap is.

Ook bij de regressielijnen moeten wij nog even stil staan. Bij de correlatierekening vonden wij dat het gebruik van de regressielijnen gemotiveerd werd door de kansverdeling in de lengterichting van de lange as. Daar deze kansverdeling bij lijnvereffening niet voorkomt, moet er nu een andere reden zijn voor het gebruik er van. VAN UVEN motiveert het gebruik van de regressielijnen door een voorkeur voor een zekere richting van middelen. In hoeverre dit juist is zullen wij in 128 nagaan.

126 – HET PSEUDO-CORRELATIEKARAKTER VAN DE LIJNVEREFFENING

In 125 hebben wij uiteengezet dat het typerend is voor lijnvereffening dat er wel een kansverdeling is in de richting van middelen, maar niet in de lengterichting van de vereffeningslijn. Wij kunnen dus een kansfiguur ontwerpen zoals in (128.1) is uitgebeeld.

De lijnen van gelijke kansdichtheid lopen evenwijdig aan de vereffeningslijn AB. Het valt gemakkelijk in te zien dat de richting van middelen onbelangrijk is. Wanneer in de richting loodrecht

op de vereffeningslijn een normale verdeling optreedt, treedt in iedere andere richting (behalve die van de vereffeningslijn) ook een normale verdeling op, met de top boven de vereffeningslijn AB.

Hier is dus een wezenlijk verschil met de correlatierekening, waar bij iedere richting van middelen een eigen middellijn behoort. Daar vindt men bij een richting van middelen evenwijdig aan de assen de regressielijnen; bij een richting van middelen loodrecht op de vereffeningslijn vindt men de lange as. Bij lijnvereffening moet men met alle methoden de lijn AB vinden. Wij kunnen dus zeggen dat bij de lijnvereffening de formules van de regressielijnen en de formule van de lange as „asymptotisch" steeds de lijn AB zullen aangeven.

Hier staat tegenover dat bij het uitwerken van een bepaald geval van lijnvereffening de drie oplossingen steeds ongelijk zijn. De lijn die berekend wordt met een vertikale richting van middelen zal steeds kleiner hoek maken met de positieve X-as dan de „lange as", de andere regressielijn maakt een nog grotere hoek.

Hoewel de drie oplossingen asymptotisch de lijn AB aanwijzen, vallen ze asymptotisch nooit samen.

De oorzaak van deze absurditeit ligt hierin dat het materiaal dat bij een lijnvereffening verwerkt wordt noodzakelijk meer op een geval van correlatie dan op een geval van lijnvereffening lijkt. Wij hebben gezien dat een lijnvereffening daar wordt toegepast waar in de lengterichting van de vereffeningslijn geen kansverdeling optreedt. Op grond hiervan zou men over de gehele lengte van de lijn dezelfde frequentie van de waarnemingen kunnen begeren. Om proeftechnische redenen zal men evenwel steeds een *frequentieverdeling* van de waarnemingen aantreffen, die met de normale kansverdeling althans dit gemeen heeft, dat buiten een bepaalde grens practisch geen waarnemingen meer voorkomen.

Dit alleen maakt reeds dat het waarnemingsmateriaal meer op een geval van correlatie dan van lijnvereffening gaat lijken. Het is dit *pseudo-correlatiekarakter* van het empirisch materiaal, dat een verschil tussen de verschillende oplossingen doet ontstaan. Dit verschil is dus steeds het gevolg van een fout; een fout in de frequentieverdeling van het empirisch materiaal. Een betrouwbaar verschil tussen de vereffeningslijnen die uit hetzelfde materiaal berekend worden, behoort dus onmogelijk te zijn. Formules die asymptotisch hetzelfde resultaat behoren te geven mogen onderling geen betrouwbare verschillen vertonen.

Bij de foutenberekening van de lijnvereffening wordt evenwel geen rekening gehouden met de fout in de frequentieverdeling, slechts op de afstanden tussen waarnemingspunt en vereffeningslijn wordt gelet. In de practijk is het dan ook *wel* mogelijk materiaal te verzamelen waarin de regressielijnen betrouwbaar ongelijk zijn. Wanneer bij een bepaald materiaal de beide regressielijnen berekend zijn en de lange as, dan zullen deze lijnen weinig veranderen, wanneer (bij gelijkblijvende frequentieverdeling in de richting van de vereffeningslijn en gelijkblijvende middelbare fout) het aantal waarnemingspunten sterk wordt opgevoerd. De middelbare fout van de berekende parameters der lijnen zal daardoor evenwel naar wens kunnen worden gedrukt. Met voldoende materiaal moet het dus mogelijk zijn betrouwbare verschillen aan te tonen tussen de verschillende lijnen, niettegenstaande alle lijnen asymptotisch de vereffeningslijn AB moeten opleveren.

127 – HET VERSCHIL TUSSEN CORRELATIE EN LIJNVEREFFENING

In het voorgaande hebben wij reeds enige pogingen gedaan het verschil tussen correlatie en lijnvereffening te benaderen. In 121 trachtten wij onderscheid te maken doordat wij het *stochastisch verband* wezenlijk achtten voor *correlatierekening* en het *functioneel verband* voor *lijnvereffening*. Bij nadere beschouwing bleek evenwel, dat zowel bij lijnvereffening als bij correlatierekening een stochastisch en een functioneel element voorkwamen.

In 125 dachten wij het verschil zo aan te kunnen geven dat er bij correlatierekening sprake zou zijn van een kansverdeling in de lengterichting van de lange as; deze kansverdeling zou bij lijnvereffening uitbreken. Maar in 126 vonden wij dat er a priori enige uitspraken konden worden gedaan over de frequentieverdeling van de totaalindruk van de waarnemingspunten bij lijnvereffening. En wat is een aprioristische uitspraak over een komende frequentieverdeling anders dan een kansverdeling?

Wij zullen nu trachten nog een scherpere omschrijving van het verschil te geven:

In gevallen waarin lijnvereffening moet worden toegepast dient men de onbekende coördinaat van een waarnemingspunt uit de bekende te berekenen met behulp van de vereffeningslijn, die het verband tussen beide coördinaten het beste weergeeft.

In gevallen waarin correlatie moet worden toegepast dient men de

onbekende coördinaat van een waarnemingspunt uit de bekende te berekenen met behulp van een regressielijn, die van de beste vereffeningslijn verschilt.

Deze beide stellingen zijn terstond duidelijk wanneer men zich 124 te binnen brengt. Om het practisch verschil te doorvoelen willen wij nog een paar voorbeelden bespreken.

Stel dat men uit een bepaald stadje van de volwassen mannelijke inwoners tussen 20 en 50 jaar het verband tussen lichaamslengte en armlengte heeft bepaald. Dan weet men dat de lange as van de correlatie-ellips het geschiktst is om de samenhang tussen beide lengten weer te geven.

Indien men nu een reus uit dat stadje ontmoet die men niet gemeten heeft, dan kan men de uitspraak doen dat hij „waarschijnlijk" *naar verhouding* korte armen heeft. Nauwkeuriger gezegd: De kans dat zijn armen *naar verhouding* kort zijn is groter dan de kans dat zijn armen naar verhouding lang zijn.

Waarom? Niet, omdat reuzen over het algemeen naar verhouding korte armen hebben. Het begrip *naar verhouding* wordt juist gebruikt in verband met de algemene indruk over reuzen, die door de lange as wordt weergegeven. Maar wel omdat er meer kans is dat men te maken hebben met een persoon, waarvan de lichaamslengte iets groter is dan met de totaalindruk overeenkomt, dan met iemand waarvan de lichaamslengte te klein is in verhouding tot de totaalindruk. De uitspraak berust op de kansverdeling van de totaalindruk. Wij kunnen er daarom de uitspraak naast zetten: Wanneer men van iemand uit dat stadje ontdekt dat hij erg lange armen heeft zal hij „waarschijnlijk" *naar verhouding* kort zijn.

Laten wij hier tegenover stellen een voorbeeld uit de lijnvereffening. Stel dat men de lengte van een staaf en zijn temperatuur vergelijkt. Hierbij kunnen wij aprioristische uitspraken doen over de frequentieverdeling. Een lagere temperatuur dan het absolute nulpunt wordt niet gevonden, een hogere dan het smeltpunt van de staaf ook niet. Wanneer wij de apparatuur van de onderzoeker kennen kan de uitspraak nog verder gaan. Wellicht kan niet gekoeld worden, zodat de ondergrens bij kamertemperatuur ligt. Wellicht wordt een thermometer gebruikt die niet verder gaat dan 200° C.

Laten wij aannemen dat op grond van de apparatuur vaststaat

dat de waarnemingen tussen de 20 en 200° C zullen vallen. Dan kan op grond van 126 worden vastgesteld dat men materiaal zal verzamelen met een pseudo-correlatiekarakter. Het zal mogelijk zijn een lange as en regressielijnen te berekenen. Wij willen aannemen dat op bepaalde gronden mag worden aangenomen dat de lange as de juiste vereffeningslijn is. Dan mag nu niet gebruik worden gemaakt van de regressielijnen om de ene coördinaat van een waarnemingspunt uit de ander te berekenen. Er kan nu niet gezegd worden: Wanneer er een temperatuur is gemeten van 198° C zal de staaf op dat moment „waarschijnlijk" *naar verhouding* kort zijn geweest. Ongetwijfeld zullen bij een bepaalde temperatuur verschillende lengten gevonden worden, en bij een bepaalde lengte verschillende temperaturen; maar er is geen kansverdeling van de totaalindruk. Wij kunnen niet zeggen: De gemeten temperatuur ligt boven het gemiddelde en dus is de kans dat de „ware" temperatuur 0.01° C lager lag groter dan de kans dat die temperatuur 0.01° C hoger lag. Zolang de temperaturen binnen het bereik van de apparatuur vallen zijn beide kansen gelijk bij normale verdeling der afwijkingen. De kans op een bepaalde totaalindruk neemt niet toe, naarmate die totaalindruk dichter bij het gemiddelde ligt, want binnen de mogelijkheden van de apparatuur zijn alle kansen gelijk.

128 – De keuze van de juiste vereffeningslijn

In 126 hebben wij gezien dat de formules uit de correlatierekening voor het vinden van de lange as en de regressielijnen bruikbaar zouden moeten zijn om de juiste vereffeningslijn bij lijnvereffening te vinden. Door de boven- en benedengrens, die empirisch niet te vermijden zijn, krijgt het materiaal evenwel een pseudo-correlatiekarakter waardoor de lijnen, die langs de verschillende wegen berekend zijn, niet samenvallen. Hier dringt zich de vraag op, welke formule nu in de practijk gebruikt moet worden.

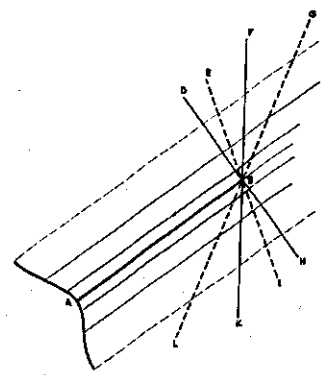
In 127 zagen wij dat het bij correlaties juist is, de ene coördinaat van een punt uit de ander af te leiden met behulp van een andere lijn dan de vereffeningslijn. Wij zagen tevens dat dit bij lijnvereffening niet mag. Daar moet steeds de vereffeningslijn worden gebruikt. We staan dus *uitsluitend* voor de vraag hoe de vereffeningslijn te kiezen.

Zoals wij reeds zagen vindt dit probleem zijn oorzaak in het

voorkomen van onvermijdelijke grenzen. Wij zullen daarom trachten een keuze tussen de mogelijke formules te maken door de grenzen te bestuderen.

Daartoe hebben wij in fig. (128.1) de kansverdeling weergegeven, die bij lijnvereffening verwacht mag worden. De figuur is aan de bovenkant op 4 manieren scherp begrensd. De grenslijn DH staat loodrecht op de vereffeningslijn; de grenslijn FK loopt even-

(128.1)



wijdig aan de Y-as, de lijnen EI en GL nemen nog andere posities in. Wij willen aannemen dat aan de onderkant grenslijnen lopen evenwijdig aan de genoemde. Hoe moet nu de juiste vereffeningslijn worden gekozen?

Het is duidelijk dat de richting van middelen evenwijdig aan de grenslijnen moet lopen. Dat is de enige richting die toelaat dat overal stroken worden gevormd over de gehele breedte van het waarnemingsmateriaal. Dit toch is noodzakelijk opdat het midden van iedere strook op AB valt. Wanneer DH de grenslijn aangeeft moet er dus gewerkt worden met de formule voor de lange as, geeft FK de grenslijn aan, dan is een van de regressielijnen nodig. Wanneer FK de grenslijn is en er zou toch gewerkt worden met de formule van de lange as, dan zou de lijn aan de bovenkant te hoog komen te liggen, is DH de grenslijn dan komt de ene regressielijn aan de bovenkant te laag.

Het blijkt tevens dat er nog meer mogelijkheden zijn. Loopt de grenslijn in de richting EI dan is de lange as te hoog en de regressielijn te laag; bij een grenslijn als GL is zelfs de regressielijn te hoog.

Is het nu mogelijk de vorm van de grenslijn a priori vast te

stellen? Soms wel. Laten wij weer blijven bij ons voorbeeld van de staaf. Het is denkbaar dat de gehele apparatuur toelaat te werken tussen 0 en 300° C, maar dat de thermometer ons slechts in staat stelt temperaturen waar te nemen van 20—200° C. Wanneer wij er naar streven het gehele gebied te onderzoeken zullen wij grenzen krijgen evenwijdig aan de staaflengte-as. Een temperatuur van 201° C wordt nooit waargenomen, bij de waargenomen temperatuur 200° C heeft de lengte nog geen enkele beperking van mogelijkheden.

De gevonden grenslijn is niet strict gebonden aan de toevallige fout van de waargenomen lengte en temperatuur. Het is zeer wel denkbaar dat de lengte zeer nauwkeurig wordt waargenomen en dat de thermometer zeer slecht is. Misschien is de temperatuur soms 198° C maar tracht de thermometer meer dan 200° C aan te wijzen, zodat de waarneming mislukt; misschien is een andere keer de temperatuur 201° maar blijft de thermometer bij 199° C staan, zodat deze temp. wordt genoteerd met een lengte die bij de temperatuur van 201° C hoort. De fout van de temperatuurwaarnemingen heeft hier geen invloed op de richting van de grenslijn. De grenslijn zal evenwijdig aan de as van de staaflengte verlopen, en de richting van middelen zal dus ook zodanig moeten zijn. In geval de lengtemetingen practisch foutloos zijn in verhouding tot de grote fouten van de temperatuur beveelt VAN UVEN evenwel een richting van middelen aan, evenwijdig aan de temperatuur-as. Wij zien dus dat de richting van middelen niet steeds uit de grootteverhouding van de fouten volgt.

Meestal is dit echter wel het geval. Wanneer er voldoende waarnemingen zijn, zodat het observeren van een *bepaald* „waar" punt op AB een groep toevallig verdeelde waarnemingen geeft, dan kunnen deze waarnemingen worden getypeerd door een kansellips met de middelbare fouten van x en y als lange en korte as. In de ellips van het meest extreme geobserveerde ware punt op AB als grenslijn kan nu worden beschouwd de middellijn, die toegevoegd is aan de richting van de vereffeningslijn. Wanneer x foutloos is ontaardt de ellips tot een lijn evenwijdig aan de Y-as en dus de genoemde middellijn ook. Wanneer x en y gelijke fouten hebben wordt de ellips een cirkel, en de bedoelde middellijn staat dan loodrecht op de vereffeningslijn.

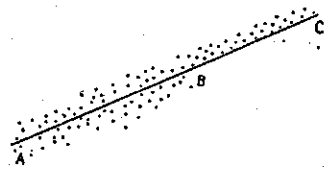
Men moet evenwel letten op de voorwaarde, dat er voldoende waarnemingen zijn, zodat om ieder punt van AB een normale

kansverdeling van waarnemingen mag worden verwacht. Meestal is het aantal waarnemingen zo gering dat aan deze voorwaarde niet wordt voldaan. Dan blijft de berekening asymptotisch natuurlijk wel goed, maar wellicht kan uit de bizonderheden van het waarnemingsmateriaal (een monster uit het universum van mogelijkheden) soms een conclusie worden getrokken over de vraag hoe de berekende lijn van de ware afwijkt.

Wij zullen daartoe naar een nieuw criterium moeten zoeken. Als zodanig zouden wij dit kunnen nemen: Wanneer de juiste vereffeningslijn *is* berekend zal binnen iedere strook, die in de richting van middelen uit de puntenbundel wordt genomen, de middelbare fout, die uit de punten boven de vereffeningslijn wordt berekend, gelijk moeten zijn aan de middelbare fout, die in dezelfde strook uit de punten beneden de vereffeningslijn wordt berekend. Een eventueel verschil tussen deze waarden moet *toevallig* zijn en dus onafhankelijk van de plaats van de strook.

Wanneer een strook in een willekeurige richting wordt beschouwd geldt dezelfde eis, doch dan moet bedacht worden dat aan de uiteinden onvolledige stroken kunnen ontstaan, waarvan b.v. de benedenhelft ontbreekt (b.v. DBF in fig. 128.1, wanneer FK de grenslijn is en DH de richting van middelen aangeeft). In zulke gevallen moet geëist worden dat het aanwezig deel harmonisch aansluit aan de verdere gegevens. Zo zal in fig. (128.2) de lijn

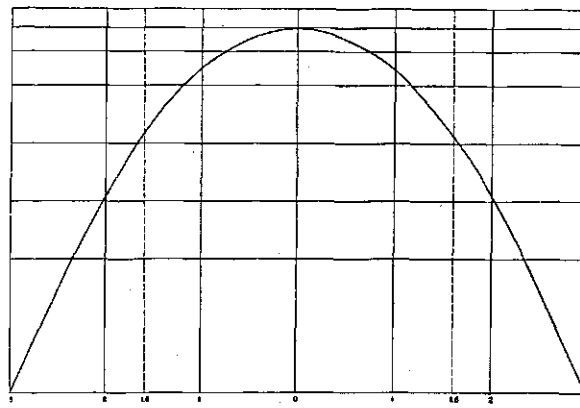
(128.2)



AC als juiste vereffeningslijn kunnen worden genomen, omdat de complete stroken aan de gestelde voorwaarde voldoen, en de niet complete stroken tussen B en C harmonisch bij het overige deel aansluiten.

Bij het gebruik van dit criterium gaan wij er dus toe over de juistheid van de lijn *achteraf* vast te stellen. Wanneer wij door de puntenbundel van (128.2) een lijn hadden berekend zou hij steiler gestaan hebben; dit geldt voor de lange as en de beide regressielijnen. Toch zijn we overtuigd dat de aangegeven lijn de beste is,

(128.3)



op grond van het nieuwe criterium. Dit criterium leidt dus uit de abnormaliteit van het monster bepaalde conclusies af over het universum.

Ons nieuwe criterium laat een volledig mathematische contrôle toe, maar is vooral gemakkelijk bij een *grafische* verwerking. Bij een grafische verwerking blijkt het nl. vaak gemakkelijk de niveaulijnen van kansdichtheid te bepalen, die op een afstand van 1 ½ à 2× de middelbare fout van de vereffeningslijn zijn verwijderd. Deze lijnen laten zich meestal veel gemakkelijker vinden dan de vereffeningslijn zelf.

Om dit aan te tonen hebben wij de grafiek van de normale verdeling op logarithmisch papier afgezet (zie fig. 128.3). Visueel ziet men nl. niet in de eerste plaats de frequenties op de verschillende plaatsen, maar de frequentie*verhoudingen*. Men ziet op de grafiek dat er visueel binnen de aangegeven grenzen weinig verschil is, en dat de dichtheid van punten buiten de grenzen snel afneemt. Dit snel afnemen blijkt ook wel hieruit dat buiten de niveaulijnen op een afstand van 1.6 σ nog 11% van de punten valt, buiten de lijnen van 1.8 σ nog 7% en buiten de lijnen van 2 σ nog slechts 4.5%.

Wanneer men aan weerszijden van de puntenbundel niveaulijnen heeft getrokken, moet men als vereffeningslijn nemen de lijn die er midden tussen doorloopt. Door deze keuze bereikt men dat in iedere strook, die dwars over de puntenbundel wordt gelegd, de middelbare fout aan weerszijden van de lijn gelijk geschat wordt. Daar is immers de afstand tussen niveaulijn en vereffeningslijn een maat voor. Tevens bereikt men dat onregelmatigheden in de verdeling der punten (b.v. einden van de empirische bundel)

geen invloed op de verkregen uitkomsten behoeven te hebben.

Het belangrijkste voordeel van een grafische verwerking is evenwel niet, dat bepaalde toevallige fouten daarmee kunnen worden uitgeschakeld. Vaak werkt een numerieke bewerking nauwkeuriger, en is dan kwantitatief beter.

Het nut van een grafische bewerking ligt vooral op kwalitatief gebied. Reeds bij het eenvoudig probleem van rechtlijnige vereffening is het moeilijk de juiste formule te vinden om bovenvermelde redenen. Wanneer een verkeerde formule wordt genomen heeft men in wezen te maken met een *noodformule,* met alle nadelen van dien. Bij kromlijnige vereffening is de kans op een juiste keuze nog veel kleiner, tenzij de keuze a priori theoretisch stevig gefundeerd is. Daar zal men dus nog eerder tot grafische verwerking moeten overgaan.

Wij willen eindigen met de volgende regel: *Een grafische verwerking is kwantitatief vaak slechter dan een numerieke, maar kwalitatief vaak beter, omdat er minder kans is dat er met een noodformule gewerkt wordt. Abnormale verdeling van het waarnemingsmateriaal kan soms reden zijn ook bij ideale formules een grafische verwerking te prefereren.*

II – ENIGE BIZONDERHEDEN VAN HET RASSENONDERZOEK

201 – De zin van het proefveld

Een landbouwgewas reageert in zijn opbrengst en andere eigenschappen op allerlei invloeden. Vele daarvan zijn met name bekend, en misschien zijn er ook nog vele onbekende. In ieder geval is de reactie van het gewas op al deze gevarieerde invloeden niet steeds bekend. Zelfs is de reactie van het ene ras in de regel anders dan die van het andere ras.

In deze situatie stelt het rassenonderzoek zich tot taak aan te geven welke rassen aan de landbouwpractijk moeten worden aanbevolen. Daartoe moet de weg gevonden worden door al deze onbekende en bekende invloeden met hun al of niet kwalitatief en kwantitatief bepaalde werkingen.

Een methode van rassenonderzoek, waarbij al deze moeilijkheden ten volle tot gelding komen, is de rassenenquête. Wanneer aan allerlei mensen gevraagd wordt naar hun practijkervaring met een bepaald ras, is aan de hand van de antwoorden een gemiddelde ervaring voor het onderzochte gebied vast te stellen. Onderzoekt men zo het gemiddelde van verschillende rassen, dan kunnen deze rassen worden vergeleken. Bij deze methode komt de onzekerheid, die nog bestaat aangaande de invloeden, ten volle tot gelding in de onzekerheid, waaraan de vergelijking der rassen onderworpen blijft.

Gelukkig is aan de moeilijkheden, zoals die werden uiteengezet, gedeeltelijk te ontkomen. Het moge waar zijn, dat de rassen ongelijk op de omstandigheden reageren, het is evenzeer waar, dat er in de reacties der verschillende rassen toch ook een grote mate van overeenstemming is. Wanneer van een perceel wordt gezegd dat het vruchtbaar is, dan wordt daarmee uitgedrukt, dat de omstandigheden voor alle rassen gunstig zijn; is een stuk land te nat of te zuur dan schaadt dit alle rassen in meer of mindere mate.

Van dit feit kan het rassenonderzoek gebruik maken doordat het niet vraagt naar de opbrengsten der verschillende rassen, maar naar de onderlinge opbrengst*verschillen*. Het is zeer wel

mogelijk dat van twee rassen niet op 10% nauwkeurig te zeggen valt, wat ze op een bepaald perceel zullen opbrengen, maar dat wel kan worden verzekerd dat het ene ongeveer 5% meer op zal brengen dan het andere. De oorzaak hiervan is dus deze, dat de diverse rassen gewoonlijk *globaal* op de uitwendige omstandigheden gelijk reageren. Een ras dat onder de ene omstandigheid extra veel opbrengt zal het vaak onder vele andere omstandigheden ook doen. Dit feit is voldoende bekend, zodat het niet verder behoeft te worden besproken. Wij zullen er evenwel van uit moeten gaan om de zin van het proefveldwezen te begrijpen.

Omdat de rassen globaal op de uitwendige omstandigheden gelijk reageren, kunnen de uitwendige omstandigheden globaal worden uitgeschakeld. Dit gebeurt op het rassenproefveld, waar alle rassen naast elkaar worden gezet onder practisch dezelfde omstandigheden. Over het gehele land verspreid worden onder allerlei omstandigheden zulke proefvelden aangelegd. Wanneer een bepaald rasverschil op al die proefvelden wordt geconstateerd, mag wel worden geconcludeerd dat het verschil tussen de rassen onder allerlei omstandigheden blijft bestaan, en dat de rassenkeuze onder al die omstandigheden ook niet moet veranderen.

In de practijk zal blijken dat de rasverschillen onder al die omstandigheden niet helemaal constant zijn, er blijft een zekere schommeling over. De werking van de omstandigheden is slechts *globaal* uitgeschakeld.

De overblijvende schommelingen zullen meestal ook van al de invloeden afhangen, die op de opbrengsten en andere gegevens werken. *Door de proefvelden wordt het aantal belangrijke invloeden niet verminderd; slechts wordt hun werking grotendeels uitgeschakeld, zodat met veel minder kennis toch een tamelijk goede rassenkeuze mogelijk is.*

202 – Het condenseren van de invloeden en het normaliseren van de gegevens

In de vorige paragraaf hebben we over talloze invloeden gesproken, waarvan men veel te weinig weet. Het ligt voor de hand dat in de practijk met al die onbekende en vage invloeden niet te werken valt. Daarom moeten ze worden samengevat tot een paar hanteerbare begrippen.

Omdat op de rassenproefvelden, waarvan wij de wiskundige bewerking nu nader bestuderen, alle rassen geheel onder dezelfde

omstandigheden worden verbouwd (voor zover men die in de hand heeft), willen wij alle invloeden, die op het proefveld werkzaam zijn, samenvatten als de *invloed van het milieu.*

Door een dergelijke schematisering krijgt de te gebruiken formule uitgesproken het karakter van een noodformule. Er zal daarom op gelet moeten worden, dat er een *interpretatiefout* wordt gemaakt. Deze zal met de verborgen invloeden en de variabiliteit de oorzaak zijn van het optreden van *strijdigheid* en dus van een varians.

Bij het rassenonderzoek willen wij dus in eerste instantie de opbrengst of de andere gegevens zien als functie van een ras- en van een milieuinvloed; terwijl wij door de varians trachten te meten in hoever ons pogen geslaagd is.

In formule uitgedrukt is onze wens dus dat wij kunnen stellen:

$$O_{kp} = \Phi\,(R_k; M_p), \tag{202.1}$$

waarbij O_{kp} voorstelt het gegeven (b.v. opbrengst) dat met ras k in het milieu p is verkregen; R_k stelt voor de invloed van ras k en M_p de invloed van het milieu p.

Omdat in deze formule alle invloeden tot twee gecondenseerd zijn, willen wij deze formule een *condensatieformule* noemen en de interpretatiefout, die hierbij gemaakt wordt, een *condensatiefout.*

In formule (202.1) wordt het gegeven O dus beschreven als een functie Φ van twee veranderlijken R en M. Over de bouw van de functie Φ wordt evenwel niets gezegd: Daardoor wordt er aan de formule een derde element van variatie toegevoegd. Om algemene beschouwingen over formule (202.1) te kunnen ontwikkelen is het nodig, dat de bouw stabiel is, en liefst eenvoudig.

Wij willen *aannemen* dat deze stabiele en eenvoudige bouw bereikt kan worden door de gegevens O_{kp} te transformeren (b.v. de logarithme er van te nemen). Wij krijgen dan een grootheid Ω_{kp}, die een functie (Θ) is van O_{kp}, dus (zie 202.1).

$$\Omega_{kp} = \Theta\,(O_{kp}) = \Theta\,\{\Phi\,(R_k\,;\,M_p)\} = \Lambda\,(R_k\,;\,M_p). \tag{202.2}$$

De gegevens O, zoals ze zijn waargenomen, noemen wij *bepaalgegevens;* de getransformeerde waarden ervan (Ω), waarmee de berekening zal worden volvoerd, de *rekengegevens.*

Evenals de O is de Ω een functie van uitsluitend R en M. Omdat de bouw van Θ nog onbepaald is kunnen wij aan de functie Λ

aprioristische eisen stellen, die hun compensatie in de Θ zullen vinden. De eis die wij willen stellen is deze:

Wanneer men in een bepaald milieu p_1 het rasverschil bepaalt van de rekengegevens Ω van de rassen k_1, k_2, k_3, ... dan moet men dezelfde getallen vinden als wanneer men die rasverschillen bepaalt in het milieu p_2, p_3, ...

Dus b.v. (wanneer Ω_{21} betekent het rekengegeven van ras k_2 in milieu p_1)

(202.3) $$\Omega_{21} - \Omega_{11} = \Omega_{22} - \Omega_{12}$$

Het verschil $\Omega_{21} - \Omega_{11}$ kunnen wij opvatten als de aangroeiing van Ω_{kp} wanneer men in milieu p_1 van ras verandert. Deze aangroeiing willen wij aanduiden door

(202.4) $$\triangle\, \Omega_{(k_2-k_1)1} = \Omega_{21} - \Omega_{11}$$

Dat het gaat over een verschil tussen ras k_2 en ras k_1 duiden wij aan door $(k_2 - k_1)$; dat de p de waarde p_1 houdt duiden wij aan door de p te vervangen door 1.

De aangroeiing $\triangle\, \Omega_{(k_2-k_1)1}$ mogen wij beschouwen als gevolg van een aangroeiing $\triangle\, R_{(k_2-k_1)_1}$ van de gunstigheid van de raseigenschappen R, wanneer men van ras k_1 op ras k_2 overgaat. Formule (202.3) is het gemakkelijkst te hanteren, wanneer deze aangroeiing er ook in vermeld wordt. Wij lezen hem dus als volgt (zie ook 202.4)

(202.5) $$\frac{\triangle\, \Omega_{(k_2-k_1)1}}{\triangle\, R_{(k_2-k_1)}} = \frac{\triangle\, \Omega_{(k_2-k_1)2}}{\triangle\, R_{(k_2-k_1)}} = \ldots = \frac{\triangle\, \Omega_{(k_2-k_1)p}}{\triangle\, R_{(k_2-k_1)}}.$$

(Door $\triangle\, \Omega_{(k_2-k_1)p}$ willen wij aanduiden dat de k varieert van k_1 naar k_2, en dat de p wordt vastgehouden op een willekeurige waarde). Aangezien de overgang van k_1 op k_2 slechts een voorbeeld is van een overgang van het ene ras op een ander, doch onze formulering voor alle overgangen toepasselijk moet zijn willen wij de indicatie $(k_2 - k_1)$ weglaten, en in 't algemeen zeggen dat $\frac{\triangle\, \Omega}{\triangle\, R}$ onafhankelijk moet zijn van het milieu, deze waarde moet algemeen geldig zijn.

Hoewel de raseigenschappen niet continu verlopen, zullen wij toch aannemen, dat wij het differentiequotient $\frac{\triangle\, \Omega}{\triangle\, R}$ mogen vervangen door het differentiaalquotient $\frac{\partial \Omega}{\partial R}$.

De waarde van $\frac{\partial \Omega}{\partial R}$ laat zich het best uit (202.2) afleiden door de vergelijking $\Omega_{kp} = \Lambda (R_k; M_p)$ partieel naar R te differentiëren. Dit geeft

$$\frac{\partial \Omega}{\partial R} = \Lambda_R (R; M);$$

$\frac{\partial \Omega}{\partial \Psi}$ zou dus een functie kunnen zijn van R en M beide. Wij hebben in (202.5) geëist dat het geen functie van M is, dus mag het uitsluitend een functie van R zijn. Stellen we deze functie voor door Ψ', dan geeft dit

$$\frac{\partial \Omega}{\partial R} = \Psi' (R).$$

Integreren van deze functie geeft

$$\Omega = \Psi (R) + \Gamma.$$

Wij zien dus dat onze eis meebrengt dat de Ω een zodanige functie is, dat hij bestaat uit de som van twee delen $\Psi (R)$ en Γ. Het eerste deel is een zuivere functie van R; het tweede deel is onafhankelijk van R (het verdwijnt bij differentiëren naar R), doch kan wel afhankelijk van M zijn, wat volgens (202.2) inderdaad het geval is. Γ bevat dus een functie van M, maar kan ook nog een onafhankelijk constante bevatten. Wij geven hem weer als $\Gamma (M) + C$.

Samenvattend kunnen wij dus zeggen dat onze eis is dat

$$\Omega_{kp} = \Psi (R_k) + \Gamma (M_p) + C.$$

Nu moet nog beslist worden over de bouw van Ψ en Γ.

De functie Ψ geeft weer dat de rekengegevens Ω afhangen van rasinvloeden R. Deze invloeden zijn vaak niet in getallen uit te drukken. De invloed wordt b.v. uitgeoefend door de bouw van een chromosoom. Het enige wat zich gemakkelijk uit laat drukken is de *werking* van die invloeden op het gegeven. Hiervoor is de $\Psi (R_k)$ een geschikte maat. Zo is de $\Gamma (M_p)$ een maat voor de werking van het milieu p.

In het vervolg willen wij $\Psi (R_k)$ weergeven door R_k en $\Gamma (M_p)$ door M_p. Onder R_k moet dan worden verstaan: de *bijdrage* van het ras aan het rekengegeven, onder M_p de *bijdrage* van het milieu aan dat gegeven. We vinden dus als te gebruiken formule

(202.6) $$\Omega_{kp} = R_k + M_p + C.$$

Deze formule is niet nieuw. In de variansanalyse wordt er regelmatig mee gewerkt. Nieuw is slechts dat wij het niet van de *bepaalgegevens* willen beweren doch van de *rekengegevens*. Niet altijd zal formule (202.6) voor de *bepaalgegevens* juist zijn. In vele gevallen kunnen er dan rekengegevens worden gevonden, die beter aan (202.6) beantwoorden. Dit lukt echter lang niet altijd! Hoe men kan trachten de rekengegevens uit de bepaalgegevens af te leiden zullen wij in 203 bespreken.

Nu volstaan wij met er op te wijzen, dat door de normalisatiefunctie Θ alle cijfers op gelijke wijze verwerkt kunnen worden. Er blijft geen onderscheid tussen opbrengst-, kwaliteits-, waarderings- en andere cijfers. Om niet te abstract te zijn, zullen wij in het vervolg niet van „het gegeven" spreken, doch van de opbrengst. Daarbij wordt onder „opbrengst" steeds verstaan het „*reken*gegeven", tenzij uitdrukkelijk anders vermeld wordt.

De waarde R_k willen wij aanduiden als *rasvoortreffelijkheid*, de waarde M_p als *milieugunstigheid*.

203 – Het transformeren van bepaalgegevens in rekengegevens

In formule (202.2) hebben wij neergeschreven

$$\Omega_{kp} = \Theta\,(O_{kp}). \tag{202.2}$$

d.w.z. de rekenopbrengst Ω_{kp} is een functie van de bepaalopbrengst O_{kp}. Deze functie willen wij aanduiden als *transformatiefunctie*. Het is nu onze taak de transformatiefunctie te vinden. Hiertoe moeten wij aansluiting zoeken bij onze eis uit formule (202.6)

$$\Omega_{kp} = R_k + M_p + C. \tag{202.6}$$

Wij beginnen met formule (202.2) te differentiëren en vinden dan

$$\frac{d\Omega_{\kappa\pi}}{dO_{\kappa\pi}} = \Theta'\,(O_{kp}). \tag{203.1}$$

Wij hebben de k en de p in het linkerlid door $\varkappa$ en π vervangen om aan te geven dat k en p lopende indices zijn geworden. De letters k en p blijven wij gebruiken wanneer ze worden „vastgehouden". In het linkerlid van (203.1) is dus uitgedrukt dat de aangroeiing mag plaats vinden door een verandering van k of van p of van beide. In het rechterlid hebben wij k en p

gehandhaafd, omdat de grootteverhouding van de veranderingen in Ω en O een functie is van de niet tegelijkertijd veranderende waarde O_{kp}.

Hoewel dus k en p in het linkerlid tegelijk mogen veranderen, kan er geen bezwaar tegen zijn een index tijdelijk vast te houden en te schrijven:

$$\frac{d\Omega_{kp}}{dO_{kp}} = \Theta'(O_{kp}) \tag{203.2}$$

$$\frac{d\Omega_{k\pi}}{dO_{k\pi}} = \Theta'(O_{kp}).$$

In de formules (203.2) wordt slechts uitgedrukt dat een formule (203.1) die geldt bij alle combinaties van rassen en proefvelden, ook geldt wanneer men zich tot een proefveld p of een ras k beperkt.

Voor de verdere verwerking is het gemakkelijk het differentiaalquotient te vervangen door het differentiequotient. Wij gaan dus niet een aangroeiing van de raseigenschappen of de proefveldeigenschappen bestuderen, maar het verschil tussen twee rassen $(k_1 - k_2)$ of twee proefvelden $(p_1 - p_2)$. In verband hiermee kunnen wij de aanduiding van de opbrengst, waarop de aangroeiing betrekking heeft, ook niet meer algemeen houden. Wij moeten aanduiden, dat wij te maken hebben met de rassen k_1 en k_2, eventueel de proefvelden p_1 en p_2, en schrijven daarom (203.2) als volgt:

$$\frac{\triangle \Omega_{(k_1-k_2)p}}{\triangle O_{(k_1-k_2)p}} = \Theta'(O_{k_{1,2}\,p}) \tag{203.3}$$

$$\frac{\triangle \Omega_{k(p_1-p_2)}}{\triangle O_{k(p_1-p_2)}} = \Theta'(O_{kp_{1,2}}).$$

Op dezelfde manier kunnen wij op grond van (202.6) neerschrijven

$$\frac{\triangle \Omega_{(k_1-k_2)p}}{\triangle R_{(k_1-k_2)}} = 1 \tag{203.4}$$

$$\frac{\triangle \Omega_{k(p_1-p_2)}}{\triangle M_{(p_1-p_2)}} = 1.$$

In verband met (203.4) lezen wij (203.3) nu alsvolgt

(203.5)

$$\frac{\triangle \Omega_{(k_1-k_2)p}}{\triangle O_{(k_1-k_2)p}} = \frac{\triangle \Omega_{(k_1-k_2)p}}{\triangle R_{(k_1-k_2)}} \cdot \frac{\triangle R_{(k_1-k_2)}}{\triangle O_{(k_1-k_2)p}} = \frac{\triangle R_{(k_1-k_2)}}{\triangle O_{(k_1-k_2)p}} = \Theta' (O_{k_{1,2}\, p})$$

$$\frac{\triangle \Omega_{k(p_1-p_2)}}{\triangle O_{k(p_1-p_2)}} = \frac{\triangle \Omega_{k(p_1-p_2)}}{\triangle M_{(p_1-p_2)}} \cdot \frac{\triangle M_{(p_1-p_2)}}{\triangle O_{k(p_1-p_2)}} = \frac{\triangle M_{(p_1-p_2)}}{\triangle O_{k(p_1-p_2)}} = \Theta' (O_{kp_{1,2}}).$$

Uit (203.5) is gemakkelijk af te leiden

(203.6)

$$\frac{\triangle O_{(k_1-k_2)p}}{\triangle R_{(k_1-k_2)}} = \vartheta (O_{k_{1,2}\, p})$$

$$\frac{\triangle O_{k(p_1-p_2)}}{\triangle M_{(p_1-p_2)}} = \vartheta (O_{kp_{1,2}})$$

waarbij

(203.7)

$$\vartheta (O_{k_{1,2}\, p}) = \frac{1}{\Theta' (O_{k_{1,2}\, p})}$$

$$\vartheta (O_{kp_{1,2}}) = \frac{1}{\Theta' (O_{kp_{1,2}})}.$$

In (203.6) staan nu twee condities waaraan voldaan moet zijn, wanneer men de juiste transformatiefunctie heeft gekozen. Om deze condities goed te begrijpen willen wij er een van in woorden formuleren.

In de eerste vergelijking van (203.6) staat het volgende uitgedrukt: Wanneer men van een bepaald proefveld p de opbrengsten van twee rassen vergelijkt, is er een verschil mogelijk in de *bepaalopbrengsten*; er kan een $\triangle O_{(k_1-k_2)p}$ bestaan. Wanneer er met *rekenopbrengsten* wordt gewerkt, is het mogelijk tussen die zelfde rassen een verschil in *rasvoortreffelijkheid* vast te stellen; men vindt een $\triangle R_{(k_1-k_2)}$. De verschillen tussen de bepaalopbrengsten en die tussen de rasvoortreffelijkheden zullen onderling samenhangen. Deze samenhang wordt op een zekere manier (ϑ), verband houdende met de gekozen transformatiefunctie (zie 203.7), bepaald door het opbrengstniveau $O_{k_{1,2}\, p}$ waarop beide rassen met hun bepaalopbrengsten op dit proefveld staan.

Doordat in de besproken vergelijking is uitgedrukt dat men zich moet beperken tot proefveld p, is de mogelijkheid gegeven dat de manier van samenhang (ϑ) op ieder proefveld anders wordt gekozen. Indien enigszins mogelijk moet men trachten dit te ver-

mijden. Wij eisen daarom, *dat de gekozen transformatiefunctie geldt voor alle proefvelden en alle rassen.*

In verband hiermee kunnen wij in (203.6) de p vervangen door de lopende index π en om gelijksoortige reden de k door $\varkappa$.

Wij schrijven dus

$$\frac{\triangle O_{(k_1-k_2)\pi}}{\triangle R_{(k_1-k_2)}} = \vartheta\,(O_{k_{1,2}\pi}) \tag{203.8}$$

$$\frac{\triangle O_{\kappa(p_1-p_2)}}{\triangle M_{(p_1-p_2)}} = \vartheta\,(O_{\kappa p_{1,2}}).$$

Hiermee hebben wij een paar vergelijkingen verkregen, die als volgt de geschiktste transformatiefuncties kunnen verschaffen.

Van alle proefvelden die men in de loop der jaren heeft verkregen kan men voor twee rassen k_1 en k_2 berekenen het verschil tussen de bepaalopbrengsten op ieder proefveld afzonderlijk ($\triangle O_{(k_1-k_2)\,p}$) en de gemiddelde bepaalopbrengst ($O_{k_{1,2}\,.\,p}$). Bij het kiezen van het rassenpaar moet men zorgen dat het verschil tussen beide rassen niet te groot en niet te klein is. Niet te groot, want $\triangle O_{(k_1-k_2)\,p}$ is een benadering van $dO_{\kappa p}$ en zou dus heel klein moeten zijn; en toch niet te klein, want de toevallige fouten mogen de systematische verschillen niet overheersen.

Van de bovenste formule is nu nog slechts de $\triangle R_{(k_1-k_2)}$ onbekend. Aangezien deze grootheid onafhankelijk moet zijn van de proefvelden kunnen wij hem vervangen door het symbool C. In dit geval is de C dus afhankelijk van het verschil tussen k_1 en k_2. Straks zullen wij waarden van C aantreffen die van andere verschillen afhankelijk zijn. Dan zullen we de C van een index moeten voorzien.

Omdat in de besproken formule slechts π veranderlijk is kunnen wij $\triangle O_{(k_1-k_2)\pi}$ als zuivere functie van $O_{k_{1,2}\pi}$ zien. Wanneer er een voldoend aantal punten is, moet aan de hand van een grafiek een functie gevonden kunnen worden, die het verband genoegzaam weergeeft. Op deze wijze vindt men de functie ϑ, die volgens (203.7) in Θ' en daarna in de transformatiefunctie Θ is om te zetten.

Wanneer er niet voldoende punten zijn, zal men verschillende rassenparen en verschillende proefveldenparen moeten onderzoeken, en de gegevens samenvoegen. Dit is om een andere reden ook reeds wenselijk, aangezien anders misschien voor één rassen-

paar een transformatiefunctie wordt genomen, die bij de andere rassen minder goed past.

Wij staan dus voor de vraag, hoe de gegevens moeten worden samengevat. Aanwezig zijn een aantal lijnen, die aan de volgende formules beantwoorden.

(203.9)

$$\begin{aligned}
\triangle O_{(k_1-k_2)\pi} &= C_1 \,.\, \vartheta\,(O_{k_{1,2}\pi}) \\
\triangle O_{(k_3-k_4)\pi} &= C_2 \,.\, \vartheta\,(O_{k_{3,4}\pi}) \\
\triangle O_{\kappa\,(p_1-p_2)} &= C_3 \,.\, \vartheta\,(O_{\kappa p_{1,2}}) \\
\triangle O_{\kappa\,(p_3-p_4)} &= C_4 \,.\, \vartheta\,(O_{\kappa p_{3,4}}) \\
&\text{enz.}
\end{aligned}$$

waarbij C_1, C_2, C_3 en C_4 weergeven de waarden van $\triangle R_{(k_1-k_2)}$, $\triangle R_{(k_3-k_4)}$, $\triangle M_{(p_1-p_2)}$ en $\triangle M_{(p_3-p_4)}$.

Doordat iedere C weer anders is zullen de lijnen een ongelijke vorm hebben. Als de lijnen recht zijn zal de één steiler staan dan de ander, zijn de lijnen krom, dan zal de kromming ongelijk zijn. Het is duidelijk dat het in deze omstandigheden niet gemakkelijk is een „gemiddelde" lijn te krijgen door de afzonderlijke lijnen over en aan elkaar te passen. Dit bezwaar is te ondervangen door de waarden van C te schatten. De waarde van C_1 is $\triangle R_{(k_1-k_2)}$, dus het verschil in rasvoortreffelijkheid tussen ras k_1 en ras k_2. Bij benadering zal hier voor kunnen worden genomen het verschil tussen de gemiddelde bepaalopbrengsten van ras k_1 en ras k_2 op de onderzochte proefvelden. Wanneer achteraf aan de figuren blijkt dat een betere onderlinge aansluiting wordt verkregen door enkele waarden van C anders te kiezen, dan kan voor ieder speciaal geval worden nagegaan, of een dergelijke verandering ook wenselijk is.

Wij komen dus tot de volgende methode

(203.10)

1e Deel de rassen in, in groepen van twee, die niet te veel en niet te weinig in opbrengst verschillen. Doe hetzelfde met de proefvelden.

2e Bepaal voor iedere groep de waarde

$$\triangle O\,(= \triangle O_{(k_1-k_2)\,p}) = O_{k_1 p} - O_{k_2 p}$$

of

$$\triangle O\,(= \triangle O_{k\,(p_1-p_2)}) = O_{kp_1} - O_{kp_2}.$$

3e Bepaal voor iedere groep de waarde

$$O\,(=O_{k_{1,2}p}) = \frac{O_{k_1 p} + O_{k_2 p}}{2}$$

of

$$O\,(=O_{kp_{1,2}}) = \frac{O_{kp_1} + O_{kp_2}}{2}.$$

4e Schat voor iedere groep de waarde

$$C\,(=\triangle R_{(k_1-k_2)}) \simeq \overline{O}_{k_1\pi} - \overline{O}_{k_2\pi}$$

$$C\,(=\triangle M_{p_1-p_2}) \simeq \overline{O}_{\kappa p_1} - \overline{O}_{\kappa p_2}.$$

(Door $\overline{O}_{k_1\pi}$ wordt aangeduid dat $O_{k_1 p}$ over p is gemiddeld).

5e Bepaal voor iedere groep de waarde van $\frac{C}{\triangle O}$.

6e Zet $\frac{C}{\triangle O}$ af als functie van O. Volgens (203.8) en (203.7) geldt nu

$$\frac{C}{\triangle O} = \frac{1}{\vartheta\,(O)} = \Theta'\,(O).$$

7e Tracht een functie Θ' te vinden die bij de gevonden lijn goed aansluit, en zoek de transformatiefunctie Θ door te integreren.

Wanneer men zijn transformatiefunctie voor practische doeleinden zoekt en niet voor theoretische, is er geen essentiële eis over de bouw, doch slechts de practische voorwaarde dat de formule de lijn behoorlijk moet benaderen en aan het integreren niet teveel moeilijkheden in de weg moet leggen.

204 – Contrôle op de juistheid van de transformatiefunctie

Bij het opstellen van de methode voor het vinden van een transformatiefunctie hebben wij als ideaal gesteld, dat alle rassen en alle proefvelden dezelfde vergelijkingsfunctie zouden krijgen. Met het stellen van dit ideaal is evenwel helemaal niet gezegd dat het ook inderdaad mogelijk is. Het is om deze reden wenselijk dat men achteraf controleert of men misschien ook een bepaald ras of een bepaald proefveld geweld heeft aangedaan.

Deze contrôle is mogelijk wanneer men met de verkregen transformatiefunctie rekenopbrengsten heeft bepaald, en deze cijfers heeft samengevat volgens de methoden van 304, zodat voor

ieder ras de voortreffelijkheid R_k bekend is, en voor ieder proefveld de gunstigheid M_p.

Met behulp van (202.6) laat zich nu voor ras k op veld p de waarde Ω_{kp} uitrekenen. (De invloed van C wordt in hfdst. III ook nader besproken).

Voor ieder ras en ieder proefveld laat zich nu controleren of het verband tussen O en Ω aan de transformatiefunctie Θ voldoet. Eventueel kan men de algemene functie Θ voor ieder ras en ieder proefveld vervangen door een speciale functie Θ_k of Θ_p.

Met het openen van deze mogelijkheid willen wij niet zeggen dat het nu ook steeds wenselijk is. Hierover zal de practijk moeten beslissen. Er zijn evenwel gevallen waarin wij wel zonder meer kunnen zeggen dat het wenselijk is van deze mogelijkheid gebruik te maken. Dit is het geval met schattingscijfers.

Allerlei kwalitatieve eigenschappen van het gewas worden vaak in schattingscijfers vastgelegd, b.v. smaak, strostevigheid, wintervastheid enz. Nu is het vaak moeilijk de betekenis van de cijfers nauwkeurig vast te leggen, de een gebruikt cijfers tussen 7 en 9, de ander tussen 4 en 10. Wanneer al deze cijfers moeten worden samengevat is het nodig eerst de schalen bij elkaar aan te passen.

In dit geval zijn de waarden van M onbelangrijk, omdat de grootte hiervan vaak meer van de waarnemer dan van het proefveld afhankelijk is. Men kan dus een willekeurige waarde van M aannemen. In verband hiermee is het in dit geval niet nodig een algemene transformatiefunctie op te stellen volgens de methode van 203. Van de verschillende waarden R_k mag men voorlopige cijfers schatten door de *bepaal*gegevens eerst samen te vatten, of door cijfers te nemen uit vroegere proeven. Uit de aangenomen waarde van M en de voorlopige waarden van R_k laten zich voorlopige waarden Ω_{kp} berekenen.

Voor ieder proefveld kan nu grafisch het verband worden nagegaan tussen O_{kp} en Ω_{kp}. Uit de grafiek wordt voor iedere O_{kp} de bijbehorende Ω_{kp} afgelezen. Deze rekengegevens ondergaan nu een definitieve bewerking volgens hfdst. III.

205 – Het meten van de milieugunstigheid

In formule (202.6) hebben wij tot uitdrukking gebracht, dat wij de opbrengst van een bepaald ras in een bepaald milieu wilden zien als de som van een bijdrage van het ras en een bijdrage van het milieu. Dit ideaal hebben wij ons durven stellen omdat wij

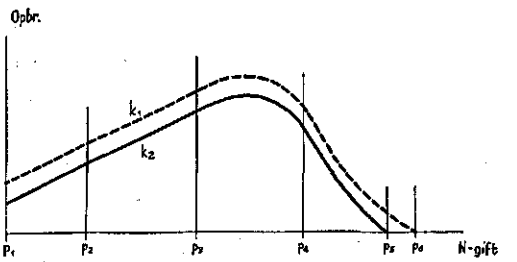
weten dat de verschillende rassen *globaal* gelijk op uitwendige omstandigheden reageren.

Wanneer die gelijkheid van reactie nu maar volkomen was, zou ieder ras in staat zijn de gunstigheid van de verschillende milieu's te meten door de variaties van zijn rekenopbrengst. Doch die gelijkheid van reacties gaat in de regel niet tot in details, doch blijft *globaal,* ook wanneer rekenopbrengsten worden gebruikt. Dit stelt ons voor speciale moeilijkheden, die wij nu moeten bestuderen.

Wij beginnen met ons een stikstofbemestingsproef te denken, waarin twee rassen, k_1 en k_2, onderzocht zijn bij verschillende stikstofhoeveelheden. Omdat verandering van stikstofhoeveelheid een verandering van milieu meebrengt, hebben wij als waarde van de milieu-index p de stikstofhoeveelheid genomen. In fig. (205.1) ziet men de bepaalopbrengsten O_{kp} afgezet als functie van p.

Wij zijn er van uitgegaan dat de stikstof eerst gunstig werkt, en dat later een overmaat aan stikstof schadelijk is. Om extreme gevallen te bestuderen hebben wij aangenomen dat de opbrengst tot 0 kan dalen, hoewel dat in de practijk niet erg waarschijnlijk is.

(205.1)



In dit voorbeeld ziet men dat het verschil van de bepaalopbrengsten van k_1 en k_2 bij een bepaalde stikstofgift overal gelijk is. De bepaalopbrengsten kunnen hier dus tevens als rekenopbrengsten worden gebruikt (zie 202.5).

Dit voorbeeld is wel uiterst eenvoudig. De gelijkheid van reactie op uitwendige omstandigheden is voor de rassen k_1 en k_2 volkomen. Hoe zullen wij nu de bijdrage van het ras van die van het milieu moeten scheiden?

Wij zullen uit moeten gaan van formule (202.6)

(202.6) $$O_{kp}\,(=\Omega_{kp}) = R_k + M_p + C.$$

De constante C kan naar wens over R_k en M_p worden verdeeld; ook kan hij geheel of gedeeltelijk onafhankelijk vermeld blijven. Dit geeft ons de vrijheid de nulpunten van waaruit R en M becijferd worden, willekeurig te kiezen. Wij willen nu het nulpunt van R onafhankelijk kiezen, en dan dat van M zo vaststellen dat C verdwijnt.

Voor wij het nulpunt van R kiezen moeten wij bedenken dat deze proef in zoverre abnormaal is, dat wij het milieu door de waarde van p (in dit geval de stikstofhoeveelheid) konden karakteriseren. Meestal zal p een willekeurige index zijn; dan moet het milieu door de M_p worden getypeerd.

Om ons dit nog weer duidelijk bewust te worden denken wij ons enige plaatsen in Nederland waar onze proef precies de meegedeelde resultaten zou opleveren. Inplaats van een bemestingsproef wordt op al deze „identieke" gronden een rassenproef aangelegd, dus een proef waarin de verschillende rassen bij één bepaalde N-bemesting worden vergeleken.

De toegediende N-bemesting hangt op iedere plaats van het inzicht van de boer af, omdat die in het algemeen de grond het best kent. Nu rekent de ene boer op een nat jaar en geeft weinig stikstof, de ander verwacht een droog jaar en geeft veel. Wanneer de proefveldresultaten moeten worden samengevat hebben wij dus te maken met verschillende milieu-invloeden. Het ene proefveld heeft de bemesting p_1 ontvangen, een volgende p_2 enz.

Nu zou men veronderstellen dat ook nu de meststofhoeveelheden p dienst zouden kunnen doen om het milieu te typeren, omdat door het verspreiden van de N-trappen over verschillende plaatsen met identieke omstandigheden niets veranderd is. Toch is dat niet het geval. Wanneer men de verschillende stikstofgiften naast elkaar op een proefveld geeft is men van de identiteit der gronden tamelijk zeker. Zijn diezelfde giften ruimtelijk verspreid, dan mag die identiteit in het algemeen niet worden verwacht. Een practische methode moet haar dus buiten beschouwing laten.

Wij zien immers dat ook reeds op identieke gronden wegens de invloed van de boer verschillende bemestingen zouden worden gegeven. Bovendien zijn in de practijk de gronden niet identiek. Aangezien de N-behoefte van verschillende gronden heel ongelijk is, zegt de gegeven N-bemesting in de practijk gewoonlijk niets over de gunstigheid van het milieu. Dit is de reden dat het milieu in het algemeen door M_p moet worden getypeerd.

Wij willen nu nagaan wat er gebeurt wanneer wij volgens de algemene principes proefvelden zouden samenvatten, die wel identiek waren, behalve dat ze verschillende N-bemestingen ontvingen. Dan zou in het geval van (205.1) de opbrengst van ras k_1 beter geschikt zijn voor het typeren van het milieu dan de opbrengst van ras k_2. Immers tussen p_5 en p_6 zijn er nog gunstigheidsverschillen, die niet door k_2 worden aangetoond, maar waar wel ras k_1 op reageert. In dit geval zouden wij daarom kunnen overwegen het nulpunt van de R zodanig te kiezen dat

(205.2) $$R_1 = 0.$$

Aangezien in ons geval bepaalopbrengst en rekenopbrengst gelijk zijn volgt hieruit:

(205.3) $$O_{1p} = \Omega_{1p} = M_p.$$

Wanneer wij nu het contante verschil tussen de opbrengsten van ras k_1 en ras $k_2 = A$ stellen, kunnen wij van O_{2p} zeggen dat

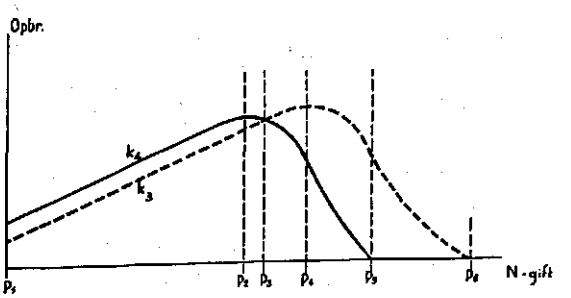
(205.4) $$O_{2p} = M_p - A$$

behalve wanneer $(M_p - A) < 0$; dan is $O_{2p} = 0$.

Een speciaal ras k_z dat tot taak heeft de milieugunstigheid M_p door middel van zijn opbrengst Ω vast te leggen, wordt een standaardras genoemd. In gevallen als fig. (205.1) weergeeft, is het productiefste ras dus vanuit dit oogpunt het geschiktst als standaardras.

Figuur (205.1) geeft echter een zeer speciaal geval weer, dat zelden voorkomt. Immers de rassen reageren in werkelijkheid slechts *globaal* gelijk op de uitwendige omstandigheden. Een grafiek als in fig. (205.5) is weergegeven, is in de practijk waarschijnlijker; wanneer wij tenminste weer aannemen, dat bij een hoge stikstofgift de opbrengst geheel kan uitblijven.

(205.5)



Beide rassen geven een stijgende bepaalopbrengst bij stijgende N-hoeveelheid, tot er een maximum bereikt wordt; dan daalt de opbrengst weer tot 0.

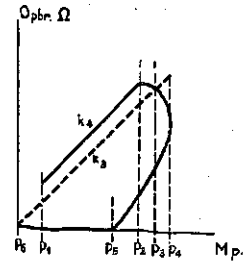
We zien dat in dit voorbeeld niet geldt, dat het verschil $(O_{4p} - O_{3p})$ constant is. Hier is het nodig de bepaalopbrengsten O in rekenopbrengsten Ω te transformeren. Omdat dit momenteel niet ons onderwerp is, willen wij aannemen dat in dit geval door een zodanige transformatie niets valt te bereiken, zodat ook nu de bepaalopbrengsten de meest geschikte rekenopbrengsten zijn.

Om het milieu te typeren zouden wij ook nu weer het productiefste ras (k_3) als standaardras kunnen nemen en stellen

(205.6) $$O_{3p} = \Omega_{3p} = M_p.$$

Aangezien de rasverschillen niet constant zijn kunnen wij de raseigenschappen het best onderzoeken door de opbrengsten van beide rassen tegen de milieugunstigheid M_p af te zetten en de opbrengstlijn van beide rassen te vergelijken (zie fig. 205.7)

(205.7)



De lijn voor ras k_3 is gemakkelijk te tekenen. Uit formule (205.6) volgt dat het een lijn is door de oorsprong onder een hoek van 45°, die zich uitstrekt van het minimum p_6 tot het maximum p_4. De lijn voor ras k_4 is moeilijker te construeren. Van p_1 tot p_2 (zie 205.5) ligt hij op een constante afstand boven die van k_3. Bij p_2 is het maximum bereikt dan buigt de k_4-lijn naar beneden, terwijl M_p nog stijgt. Bij p_3 snijden de k_3- en k_4-lijn elkaar, dan daalt de lijn regelmatig door tot bij p_4 het maximum van Ω_{3p} en dus van M_p is bereikt. Verder daalt de lijn van k_4 bij teruglopende M_p. Bij p_5 is de opbrengst van ras $k_4 = 0$, dan is de M_p nog vrij hoog. Verder volgt de k_4-lijn de M_p-as tot de oorsprong.

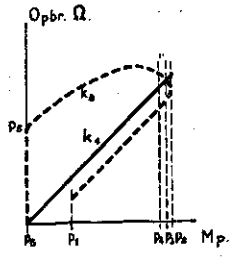
Wij zien dat ras k_3 een zeer regelmatige lijn oplevert, ras k_4 daarentegen een zeer onregelmatige. Dit ligt niet aan belangrijke rasverschillen; uit fig. (205.5) blijkt dat beide rassen ongeveer

gelijk reageren. De oorzaak is dat het ene ras tot standaard is verheven en het andere niet. Wij zouden ook ras k_4 als standaardras kunnen nemen en dus stellen

(205.8) $$O_{4p} = M_p$$

Dan zou blijken dat ras k_3 een onregelmatig verloop heeft, terwijl k_4 een rechte lijn geeft (zie fig. 205.9)

(205.9)



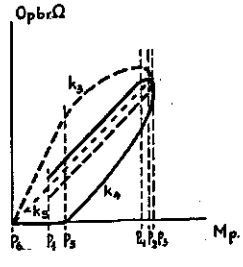
Het is natuurlijk onjuist dat een bepaald ras een extra prettige indruk maakt doordat het als standaardras is gekozen. In het algemeen moet men dus van het gebruik van standaardrassen afzien, wanneer niet alle onderzochte rassen kwalitatief gelijk reageren. Veel beter is het alle rassen gelijke kansen te geven en als „standaardras" het gemiddelde van de onderzochte rassen te nemen. Wij willen dit fictief ras als „*fictiefras*" aanduiden. Voor dit fictiefras k_s geldt dus

(205.10) $$\Omega_{sp} = \frac{[\Omega_{\kappa p}]}{m} = M_p.$$

Wanneer wij het aantal rassen door m voorstellen.

Hoeveel wij in gevallen als (205.5) vooruitgaan door het gebruik van een fictiefras moge blijken uit fig. (205.11), waar de opbrengsten van de rassen k_3 en k_4 zijn afgezet tegen de M_p zoals die bepaald wordt door het fictiefras.

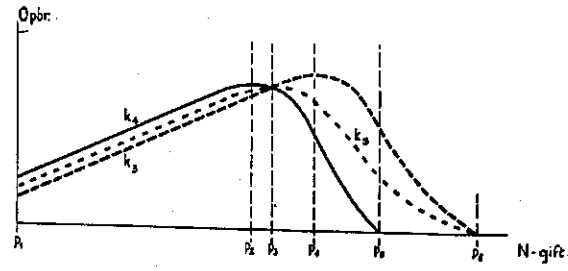
(205.11)



We zien dat beide rassen nu gelijksoortige lijnen vertonen. Ook valt het op dat in fig. (205.7) en (205.9) het niet-standaardras ongunstig voor de dag komt in vergelijking met (205.11) ten behoeve van het standaardras. Het werken met een fictiefras is dus veel eerlijker.

Om de eigenschappen van het fictiefras duidelijk in het licht te laten komen is de opbrengstlijn van het fictiefras (k_s) in fig. (205.5) ingetekend (zie fig. 205.12)

(205.12)



Men ziet hier dat het fictiefras in zijn reactie op stikstof het midden houdt tussen k_3 en k_4, dit is een gunstige eigenschap. Voorbij p_5 is de opbrengst van $k_4 = 0$; dan is $M_p = \frac{1}{2} O_{3p}$. Dit heeft tot gevolg dat het fictiefras evenlang op stikstof blijft reageren als k_3. De voordelen van het productieve standaardras (zie conclusie na (205.4)) worden dus ook bereikt met het fictiefras. In gevallen als weergegeven in (205.1) kan het fictiefras daarom evengoed dienst doen als het standaardras.

Wij kunnen daarom de algemene conclusie trekken, dat het wenselijk is steeds met een fictiefras te werken en nooit met een standaardras. De opbrengst van het fictiefras wordt daarbij verondersteld gelijk te zijn aan het gemiddelde van de rekenopbrengsten van alle onderzochte rassen.

206 – De noodzaak het milieu te „wegen"

In 201 hebben wij gezien dat het milieu uitermate gecompliceerd is. In 202 spraken wij de wens uit, een bepaald milieu toch door één getal M_p te typeren. In 205 voerden wij het „fictiefras" in, dat als opbrengst M_p zou moeten geven. Bij dit alles moeten wij dus de consequentie aanvaarden dat een bepaalde waarde van M_p verkregen kan worden bij allerlei verschillende omstandigheden. Men ziet dit reeds in fig. (205.12) waar p_2 en p_4 ongeveer dezelfde M_p ($= \Omega_s$) opleveren, zo ook p_1 en p_5. In deze verschillende

omstandigheden reageert een afzonderlijke ras niet steeds precies gelijk, zodat er verdubbeling van lijnen optreedt, wanneer de opbrengst Ω_{kp} van een bepaald ras tegen de milieugunstigheid M_p wordt afgezet (fig. 205.11).

Nu is hier nog maar sprake van één invloed (stikstof). Bij talloze invloeden wordt het aantal milieu's met gelijke gunstigheid veel groter. Voor een bepaald ras zal men dan geen verdubbeling van de lijn vinden maar een gehele lijnenbundel. Wanneer de omstandigheden continu veranderen zal men zelfs geen aparte lijnen kunnen onderscheiden. Ieder ras geeft een bepaald gebied waarbinnen waarnemingen mogen worden verwacht. Zolang men wil volstaan met het berekenen van globale rasverschillen verlangt men door dit gebied een lijn met de formule

$$\Omega_{kp} = M_p + R_k.$$

Door deze formule wordt weergegeven hoeveel ras k meer opbrengt dan het fictiefras.

Wanneer wij deze lijnen in fig. (205.11) voor beide rassen zouden intekenen, zou onze conclusie zijn dat de k_4 beslist slechter is dan de k_3. Wij zouden immers zowel voor k_3 als k_4 midden tussen beide takken door willen gaan. Wanneer wij de rassenkeuze aan de hand van fig. (205.5) moesten bepalen, zouden wij dezelfde conclusie trekken.

Bij een landbouwkundige beoordeling moet er evenwel rekening mee worden gehouden, dat de bemesting niet groter mag zijn dan de optimale, dus p_4. Beneden p_4 ligt de opbrengst van k_4 duidelijk gemiddeld hoger dan die van k_3. Zo zijn wij ook niet billijk: te lage stikstofgiften moeten ook vermeden worden. De toestand tussen p_1 en p_2 heeft daarom ook maar weinig waarde. Tussen p_2 en p_4 wint p_3 het weer. Aan de hand van fig. (205.5) is op deze wijze gemakkelijk de noodzaak en mogelijkheid van het „wegen" van milieu's te bespreken. Wanneer men moet werken met figuren als (205.11) is de noodzaak natuurlijk even groot, maar de mogelijkheid kleiner.

Men staat hier voor de practische toepassing van de theorieën die wij in hoofdstuk I ontwikkelden. Formule (202.6) is uitgesproken een noodformule. Wanneer men er mee werkt moet men niet alleen letten op de vrij onschuldige toevallige fouten, maar juist bizonder bedacht zijn op de systematische fouten, waaronder de monsterfouten zo'n grote plaats innemen.

Daarom moet men bedenken dat men het materiaal niet min of meer toevallig mag verzamelen, doch dat men zeer zorgvuldig een typisch monster uit het onderzochte gebied moet trekken naar tijd, plaats en omstandigheden. Het is onmogelijk aan deze eis te voldoen! Om een redelijk klimaatmonster te trekken heeft men reeds vele jaren nodig. Het beoordelen van een nieuw ras mag niet zoveel jaren duren. Het zal nodig zijn door het toedienen van gewichten deze monsterfouten te corrigeren. In 114 bleek, dat het moeilijk zal vallen een goed gewichtenstelsel te verzinnen. Het zal geen toelichting meer behoeven wanneer wij er nog even op wijzen dat de „toevallige fout" van een proefveld vaak geen maatstaf voor het gewicht zal mogen zijn, en het aantal herhalingen nog veel minder.

Uit het voorgaande moge duidelijk zijn, dat er vaak met monsterfouten gewerkt *moet* worden. Door het toedienen van gewichten is hier enig herstel mogelijk. Zodra de gegevens het toelaten zal evenwel een theoretische verdieping van de kennis moeten worden nagestreefd. Aan het eind van hfdst. III zal hierop iets dieper worden ingegaan.

III – HET SAMENVATTEN VAN RASSENPROEVEN

301 – De interactie tussen ras en proefveld

Bij het samenvatten van proefvelden komen een aantal problemen naar voren, die zich het gemakkelijkst laten toelichten, wanneer wij uitgaan van de schema's die gebruikt worden in de varians-analyse volgens de school van R. A. Fisher. Om deze reden willen wij beginnen met enige aspecten van een Fisher-analyse te bestuderen.

Wij gaan weer uit van formule (202.6)

$$\Omega_{kp} = R_k + M_p + C. \tag{202.6}$$

In 205 merkten wij reeds op, dat de C ons in staat stelt de nulpunten van R en M naar believen te kiezen. Daar gaven wij er de voorkeur aan de C met de M_p te laten samensmelten. In dit hoofdstuk kiezen wij liever de nulpunten zodanig dat

$$[R_\kappa] = 0$$
$$[M_\pi] = 0.$$

Evenals in (203.1) hebben wij de k door een $\varkappa$ vervangen, omdat hij een lopende index is, zo is de p vervangen door een π. Om aan te geven dat R en M vanuit hun gemiddelde worden geteld willen wij ze vervangen door resp. ϱ en μ, zodat wij lezen

$$[\varrho_\kappa] \equiv 0 \tag{301.1}$$
$$[\mu_\pi] \equiv 0.$$

Substitutie van ϱ en μ in (202.6) geeft

$$\Omega_{kp} = \varrho_k + \mu_p + C. \tag{301.2}$$

Wanneer wij het aantal rassen door m aangeven en het aantal proefvelden door n, geeft sommatie van (301.2) over k en p

$$[[\Omega_{\kappa\pi}]] = n\,[\varrho_\kappa] + m\,[\mu_\pi] + mnC.$$

of (zie 301.1)

$$C = \frac{[[\Omega_{\kappa\pi}]]}{mn} = \overline{\overline{\Omega}}_{\kappa\pi} = \overline{\overline{\Omega}}. \tag{301.3}$$

Substitutie van C in (301.2) geeft

$$\Omega_{kp} = \overline{\overline{\Omega}} + \varrho_k + \mu_p. \tag{301.4}$$

De waarden van ϱ en μ laten zich (in verband met 301.1) nu berekenen volgens de formules

$$\mu_p = \frac{[\Omega_{\kappa p}]}{m} - \overline{\overline{\Omega}} \tag{301.5}$$

$$\varrho_k = \frac{[\Omega_{k\pi}]}{n} - \overline{\overline{\Omega}}.$$

In de formules (301.4) en (301.5) zijn onze wensen aangaande het rekenschema vastgelegd. Via (202.6) hangen ze immers samen met 202, waar wij onze idealen formuleerden.

Maar misschien wordt aan deze wensen niet geheel voldaan. Hoewel wij wensen dat een rekenopbrengst alleen afhangt van totaalgemiddelde, rasvoortreffelijkheid en milieugunstigheid (zie 202), kan het zijn dat in werkelijkheid ook invloeden van het toeval aanwezig zijn. Het kan zelfs zijn dat ons streven, de juiste transformatiefunctie (zie 203) te vinden mislukt is, zodat er systematische fouten zijn. Hoe de fouten ook mogen zijn, wij moeten rekening houden met hun aanwezigheid.

Dit heeft b.v. tot gevolg dat de waarden μ_p en ϱ_k die wij volgens (301.5) berekenen niet gelijk zijn aan de waarden die wij wensen.

Wij zullen ze daarom aanduiden door μ_p' en ϱ_k', zodat (301.5) wordt:

$$\mu_p' = \frac{[\Omega_{\kappa p}]}{m} - \overline{\overline{\Omega}} \tag{301.6}$$

$$\varrho_k' = \frac{[\Omega_{k\pi}]}{n} - \overline{\overline{\Omega}}.$$

Bovendien zullen wij in formule (301.4) niet alleen de ϱ_k en μ_p door ϱ_k' en μ_p' moeten vervangen, maar bovendien nog door een μ_{kp} moeten aanduiden dat ook na die verandering nog fouten aanwezig kunnen zijn. Formule (301.4) wordt dus

$$\Omega_{kp} - \overline{\overline{\Omega}} = \varrho_k' + \mu_p' + u_{kp}. \tag{301.7}$$

In formule (301.7) hebben wij nu als belangrijke elementen overgehouden ϱ_k', μ_p' en $\overline{\overline{\Omega}}$, die allemaal afhankelijk zijn van toevallige fouten. Dit is ongewenst, vandaar dat wij liever werken met de asymptotische waarden ϱ_k, μ_p en $<\overline{\overline{\Omega}}>$.

Deze asymptotische waarden zouden worden verkregen, wanneer wij met onze m rassen *oneindig* veel proeven zouden kunnen nemen onder vermijding van monsterfouten (zie voor dit begrip 114). In dit geval zouden de *toevallige* fouten hun invloed verliezen.

Daarentegen zouden eventuele systematische fouten blijven bestaan. Zo b.v. fouten, die ontstaan wanneer er geen goede transformatiefunctie is te vinden; kortom steeds wanneer er met een noodformule gewerkt wordt. Wanneer wij nu kortheidshalve van „asymptotisch" spreken, bedoelen wij dus dat begrip dat in 113 „asymptotisch verwacht" is genoemd.

Wanneer wij deze asymptotische waarden in (301.7) invullen krijgen wij

$$\Omega_{kp} - <\overline{\overline{\Omega}}> = \hat{\varrho}_k + \hat{\mu}_p + t_{kp}. \tag{301.8}$$

De grootheden t willen wij aanduiden als „asymptotische fout".

Sommeren van (301.8) over k en p geeft

$$[[\Omega_{\kappa\pi}]] - mn <\overline{\overline{\Omega}}> = n\,[\hat{\varrho}_\kappa] + m\,[\hat{\mu}_\pi] + [[t_{\kappa\pi}]]. \tag{301.9}$$

De vorm $[\hat{\varrho}_\kappa]$ moet gelijk aan 0 zijn omdat „asymptotisch" slechts m rassen onderzocht worden. Dit blijkt uit de definitie van het „fictiefras" als gemiddelde van alle *onderzochte* rassen. De vorm $[\hat{\mu}_\pi]$ zal slechts dan 0 zijn, wanneer er geen „parameterfout" (zie 114) is gemaakt. Wij zullen dus met een mogelijke parameterfout rekening moeten houden.

Uit (301.9) laat zich nu gemakkelijk berekenen

$$\overline{\overline{\Omega}} - <\overline{\overline{\Omega}}> = \frac{[\hat{\mu}_\pi]}{n} + \frac{[[t_{\kappa\pi}]]}{mn}. \tag{301.10}$$

Trekken wij (301.10) van (301.8) af dan vinden wij

$$\Omega_{kp} - \overline{\overline{\Omega}} = \hat{\varrho}_k + \hat{\mu}_p - \frac{[\hat{\mu}_\pi]}{n} + t_{kp} - \frac{[[t_{\kappa\pi}]]}{mn}. \tag{301.11}$$

Middelen van deze formule over k geeft (zie ook 301.6)

$$\frac{[\Omega_{\kappa p}]}{m} - \overline{\overline{\Omega}} = \mu_p' = \hat{\mu}_p - \frac{[\hat{\mu}_\pi]}{n} + \frac{[t_{\kappa p}]}{m} - \frac{[[t_{\kappa\pi}]]}{mn}. \tag{301.12}$$

Zo is ook

$$\frac{[\Omega_{k\pi}]}{n} - \overline{\overline{\Omega}} = \varrho_k' = \hat{\varrho}_k + \frac{[t_{k\pi}]}{n} - \frac{[[t_{\kappa\pi}]]}{mn}. \tag{301.13}$$

Substitueren wij uit (301.12) en (301.13) μ_p' en ϱ_k' in (301.7) dan vinden wij

$$\Omega_{kp} - \overline{\overline{\Omega}} = \hat{\varrho}_k + \hat{\mu}_p - \frac{[\hat{\mu}_\pi]}{n} + \frac{[t_{k\pi}]}{n} + \frac{[t_{\kappa p}]}{m} - 2\,\frac{[[t_{\kappa\pi}]]}{mn} + u_{kp}. \tag{301.14}$$

Wanneer wij dit vergelijken met (301.11) vinden wij

$$t_{kp} - \frac{[[t_{\kappa\pi}]]}{mn} = \frac{[t_{k\pi}]}{n} + \frac{[t_{\kappa p}]}{m} - 2\,\frac{[[t_{\kappa\pi}]]}{mn} + u_{kp},$$

of

$$u_{kp} = t_{kp} - \frac{[t_{k\pi}]}{n} - \frac{[t_{\kappa p}]}{m} + \frac{[[t_{\kappa\pi}]]}{mn}. \tag{301.15}$$

Het blijkt dus dat het geen rechtstreekse schadelijke invloed heeft op de schijnbare fout wanneer $[\hat{u}_\pi] \neq 0$ is.

Bij verschillende uiteenzettingen die zullen volgen is het prettig vermeld te zien over hoeveel getallen gesommeerd wordt. Het gebruikelijke symbool daarvoor is $\sum\limits_{k=1}^{m} t_{kp}$. Hier is uitgedrukt dat over k wordt gesommeerd en dat k m waarden aan kan nemen. Deze schrijfwijze vinden wij evenwel minder gemakkelijk. We sluiten liever aan bij het sommatieteken [ ]. De vorm $[t_{\kappa p}]$ is de som van m getallen $t_{\kappa p}$, waarbij $\varkappa$ loopt van 1 tot m. Dit getal m willen wij plaatsen voor de vorm $[t_{\kappa p}]$, en vooraf laten gaan door een $\uparrow$. Door $\uparrow m\,[t_{\kappa p}]$ drukken wij dus uit dat $[t_{\kappa p}]$ de som is van m termen. Zo betekent $\uparrow\uparrow mn\,[[t_{\kappa\pi}]]$ dat $[[t_{\kappa\pi}]]$ de dubbele som is van mn termen.

Vaak zal $\varkappa$ alle waarden van 1 tot m mogen aannemen met uitzondering van een bepaalde, b.v. l. Onder $\uparrow(m-1)\,[t_{\lfloor l \rfloor p}]$ wordt nu verstaan de som van $(m-1)$ termen $t_{\kappa p}$, waarbij $\varkappa$ alle waarden van 1 tot m mag aannemen behalve l.

In deze symbolen uitgedrukt wordt formule (301.15)

$$\begin{aligned}
u_{kp} &= t_{kp} - \frac{\uparrow n\,[t_{k\pi}]}{n} - \frac{\uparrow m\,[t_{\kappa p}]}{m} + \frac{\uparrow\uparrow mn\,[[t_{\kappa\pi}]]}{mn} = \\
&= t_{kp} - \frac{\uparrow(n-1)\,[t_{k\lfloor p \rfloor}]}{n} - \frac{t_{kp}}{n} - \frac{\uparrow(n-1)\,[t_{\lfloor k \rfloor p}]}{m} - \frac{t_{kp}}{m} + \\
&+ \frac{\uparrow\uparrow(m-1)(n-1)\,[[t_{\lfloor k \rfloor \lfloor p \rfloor}]]}{mn} + \frac{\uparrow(m-1)\,[t_{\lfloor k \rfloor p}]}{mn} + \\
&+ \frac{\uparrow(n-1)\,[t_{k\lfloor p \rfloor}]}{mn} + \frac{t_{kp}}{mn} = \left(1 - \frac{1}{n} - \frac{1}{m} + \frac{1}{mn}\right) t_{kp} - \\
&- \left(\frac{1}{n} - \frac{1}{mn}\right)\uparrow(n-1)\,[t_{k\lfloor p \rfloor}] - \left(\frac{1}{m} - \frac{1}{mn}\right)\uparrow(m-1)\,[t_{\lfloor k \rfloor p}] + \\
&+ \frac{\uparrow\uparrow(m-1)(n-1)\,[[t_{\lfloor k \rfloor \lfloor p \rfloor}]]}{mn},
\end{aligned}$$

of

$$(301.16)\quad u_{kp} = \frac{(m-1)(n-1)}{mn} t_{kp} - \uparrow(n-1)\left[\frac{(m-1)\, t_{k\lfloor p\rfloor}}{mn}\right] -$$

$$- \uparrow(m-1)\left[\frac{(n-1)\, t_{\lfloor k\rfloor p}}{mn}\right] + \uparrow\uparrow(m-1)(n-1)\left[\left[\frac{t_{\lfloor k\rfloor\lfloor p\rfloor}}{mn}\right]\right].$$

Men ziet duidelijk dat de schijnbare fout u_{kp} afhankelijk is van alle asymptotische fouten t_{kp}. Wanneer er twee rassen op twee proefvelden worden onderzocht ($m = 2$, $n = 2$) krijgt formule (301.16) de vorm

$$u_{kp} = \tfrac{1}{4} t_{kp} - \tfrac{1}{4} t_{k\lfloor p\rfloor} - \tfrac{1}{4} t_{\lfloor k\rfloor p} + \tfrac{1}{4} t_{\lfloor k\rfloor\lfloor p\rfloor};$$

dan hebben alle asymptotische fouten evenveel invloed op u_{kp}. Worden m en n beide groter dan gaat t_{kp} snel domineren.

Uit de schijnbare fouten u pleegt de varians van de interactie tussen rassen en proefvelden berekend te worden. Daartoe worden eerst alle waarden u gekwadrateerd. Voor u_{kp} geeft dit

$$u_{kp}^2 = \frac{(m-1)^2(n-1)^2}{m^2n^2} t_{kp}^2 + \uparrow(n-1)\left[\frac{(m-1)^2\, t_{k\lfloor p\rfloor}^2}{m^2n^2}\right] +$$

$$+ \uparrow(m-1)\left[\frac{(n-1)^2\, t_{\lfloor k\rfloor p}^2}{m^2n^2}\right] + \uparrow\uparrow(m-1)(n-1)\left[\left[\frac{t_{\lfloor k\rfloor\lfloor p\rfloor}^2}{m^2n^2}\right]\right] -$$

$$- 2\uparrow(n-1)\left[\frac{(m-1)^2(n-1)}{m^2n^2} t_{kp}\, t_{k\lfloor p\rfloor}\right] + \ldots \text{ enz.}$$

Het is niet nodig alle termen verder op te schrijven. Voor de overzichtelijkheid willen wij alle coëfficienten buiten de sommatietekens brengen, zodat we krijgen

$$(301.17)\quad u_{kp}^2 = \frac{(m-1)^2(n-1)^2}{m^2n^2} t_{kp}^2 + \frac{(m-1)^2}{m^2n^2}\uparrow(n-1)\,[t_{k\lfloor p\rfloor}^2] +$$

$$+ \frac{(n-1)^2}{m^2n^2}\uparrow(m-1)\,[t_{\lfloor k\rfloor p}^2] + \frac{1}{m^2n^2}\uparrow\uparrow(m-1)(n-1)\,[[t_{\lfloor k\rfloor\lfloor p\rfloor}^2]] -$$

$$- \frac{2(m-1)^2(n-1)}{m^2n^2}\uparrow(n-1)\,[t_{kp}\, t_{k\lfloor p\rfloor}] + \ldots \text{ enz.}$$

Men ziet dat er eerst enige termen zuivere kwadraten komen;

dan volgen een groot aantal dubbele producten. De kwadraatsommen zijn natuurlijk alle positief. Verder is het een bekende stelling dat de asymptotische waarde van de som van de dubbele producten 0 mag worden gesteld, wanneer de beide factoren onafhankelijk van elkaar zijn. Laten wij b.v. de dubbele som $[[t_{k\lfloor p\rfloor}\, t_{\lfloor k\rfloor p}]]$ nemen. Deze waarde wordt verkregen door iedere waarde $t_{k\lfloor p\rfloor}$ te vermenigvuldigen met iedere waarde $t_{\lfloor k\rfloor p}$. Wij kunnen daarom zeggen

$$[[t_{k\lfloor p\rfloor}\, t_{\lfloor k\rfloor p}]] = [t_{k\lfloor p\rfloor}\, [t_{\lfloor k\rfloor p}]] = [t_{\lfloor k\rfloor p}]\, [t_{k\lfloor p\rfloor}].$$

De som van de producten is gelijk aan het product van de sommen. Wanneer alle waarden t onafhankelijk van elkaar zijn, zijn beide sommen asymptotisch gelijk aan 0 en het product dus ook.

In ons geval mogen wij er echter niet op rekenen dat de asymptotische fouten t onderling onafhankelijk zijn. Immers de waarden t zijn opgebouwd uit interpretatiefout en toevallige fout gezamenlijk. Voor zover de t uit toevallige fouten is voortgekomen is er onafhankelijkheid, maar voorzover er een interpretatiefout is gemaakt moet dit nog nader worden bewezen.

Wij willen daarom de waarden t splitsen in de bestanddelen i en τ, waarbij i rekenschap geeft van de interpretatiefout en τ van de toevallige fout. Wij stellen dus

$$t_{kp} = i_{kp} + \tau_{kp}. \tag{301.18}$$

Wanneer wij met behulp van deze formule t_{kp} in (301.17) substitueren krijgen wij naast de vormen met uitsluitend i of τ natuurlijk ook nog mengtermen $[i\tau]$. Van deze sommen kan in ieder geval worden gezegd dat ze asymptotisch gelijk aan 0 zijn, omdat de onafhankelijke toevallige fouten τ er in een oneven graad in voorkomen. Wij kunnen daarom (301.17) uiteen laten vallen in twee gelijke delen, waarbij in het ene deel de t vervangen is door i, in het tweede deel door τ.

Wij zullen eerst het deel met de *toevallige* fouten bestuderen. Hierbij vallen asymptotisch alle dubbele producten weg. Wanneer alle waarnemingen a-priori even betrouwbaar zijn, kunnen wij verder alle kwadraten τ^2 zien als een schatting van de middelbare *toevalsvarians* σ^2. Door de kwadraten t^2 in (301.17) te vervangen door σ^2 vinden wij

(301.19) Toevallig deel van $<u_{kp}^2> =$

$$= \frac{(m-1)^2(n-1)^2}{m^2n^2}\sigma^2 + \frac{(m-1)^2}{m^2n^2}\uparrow(n-1)[\sigma^2] +$$

$$+ \frac{(n-1)^2}{m^2n^2}\uparrow(m-1)[\sigma^2] + \frac{1}{m^2n^2}\uparrow\uparrow(m-1)(n-1)[[\sigma^2]] =$$

$$= \frac{(m-1)^2(n-1)^2 + (m-1)^2(n-1) + (n-1)^2(m-1) + (m-1)(n-1)}{m^2n^2}\sigma^2 =$$

$$= \frac{(m-1)(n-1)}{mn}\sigma^2.$$

Om de invloed van de *interpretatiefout* na te gaan, kunnen wij beter niet uitgaan van (301.17), maar van (301.15). Wanneer wij daarin de t door i vervangen krijgen wij voor het systematisch deel van u_{kp}

(301.20) Systematisch deel van $u_{kp} =$

$$= i_{kp} - \frac{[i_{k\pi}]}{n} - \frac{[i_{\kappa p}]}{m} + \frac{[[i_{\kappa\pi}]]}{mn}.$$

Om de grootheden i beter te leren kennen willen wij een ogenblik aannemen dat wij kunnen werken met opbrengsten Ω_{kp}^* die vrij zijn van *toevallige* fouten. Formule (301.8) zou dan luiden

(301.21) $$\Omega_{kp}^* - <\overline{\overline{\Omega}}> = \hat{\varrho}_k + \hat{\mu}_p + i_{kp}.$$

Middelen van deze formule over k geeft

(301.22) $$\frac{[\Omega_{\kappa p}^*]}{m} - <\overline{\overline{\Omega}}> = \frac{[\hat{\varrho}_\kappa]}{m} + \hat{\mu}_p + \frac{[i_{\kappa p}]}{m}.$$

Nu is het linkerlid de definitie van $\hat{\mu}_p$, terwijl $[\varrho_\kappa] = 0$ is. Hieruit volgt dat

(301.23) $$[i_{\kappa p}] = 0$$
$$[[i_{\kappa\pi}]] = 0.$$

Substitutie van de waarden uit (301.23) in (301.20) geeft

(301.24) Systematisch deel van $u_{kp} =$

$$= i_{kp} - \frac{[i_{k\pi}]}{n} = \frac{n-1}{n} i_{kp} - \frac{\uparrow(n-1)[i_{k\lfloor p\rfloor}]}{n}.$$

Voor wij formule (301.24) kwadrateren dienen wij wel na te gaan of alle voorkomende waarden $i_{k\pi}$ onderling onafhankelijk zijn. Dat deze onafhankelijkheid niet vanzelf spreekt volgt duidelijk uit (301.23) waar aangetoond wordt dat de waarden $i_{\kappa p}$ onderling

wel afhankelijk zijn. De som van een eindig aantal onafhankelijke waarden kan immers niet gegarandeerd 0 zijn.

De afhankelijkheid van de waarden $i_{\kappa p}$ vindt zijn oorzaak in de definitie van het fictiefras, die een asymptotisch getal $\hat{\mu}_p$ laat afhangen van een *eindig* aantal rassen (zie 301.22). Het asymptotisch getal $\hat{\varrho}_k$ hangt in tegenstelling daarmee niet af van een eindig aantal proefvelden. Het aantal proefvelden kan misschien naar plaats beperkt zijn, maar niet naar tijd; ieder jaar kan nieuwe verrassingen geven. Ook wanneer er geen toevallige fouten waren, zouden er oneindig veel waarnemingen nodig zijn om $\hat{\varrho}_k$ te berekenen. In verband hiermee zijn de waarden $i_{k\pi}$ inderdaad onderling onafhankelijk.

Bij het kwadrateren zullen de sommen der dubbele producten van verschillende waarden $i_{k\pi}$ daarom asymptotisch 0 worden, zodat alleen op de zuivere kwadraten behoeft te worden gelet. Uit (301.24) laat zich nu berekenen voor de asymptotische waarde van het systematisch deel van u_{kp}^2

(301.25) Systematisch deel van $<u_{kp}^2> =$

$$= \frac{(n-1)^2}{n^2} i_{kp}^2 + \uparrow(n-1)\left[\frac{i_{k\lfloor p\rfloor}^2}{n^2}\right].$$

De waarden $i_{k\pi}$ zijn systematische grootheden zonder toevallige fout, die zelf hun asymptotische waarde zijn. Met een bepaald ras op een bepaald proefveld in een bepaald jaar hoort een zeer bepaalde interpretatiefout gevonden te worden. Deze interpretatiefout is evenwel ook maar van zeer plaatselijk belang.

Bij globale onderzoekingen interesseert ons vaker de middelbare waarde der interpretatiefouten. Wanneer wij de asymptotische waarde daarvan voorstellen door j, kunnen wij iedere concrete i_{kp}^2 zien als een schatting van j^2. Substitutie van deze j^2 in formule (301.25) geeft

(301.26) Systematisch deel van $<u_{kp}^2> =$

$$= \frac{(n-1)^2}{n^2} j^2 + \frac{n-1}{n^2} j^2 = \frac{n-1}{n} j^2.$$

De bedoeling i_{kp}^2 te zien als een benadering van een asymptotisch gemiddelde waarde zullen wij vaak aanduiden door het symbool $<i_{kp}^2>$.

Uit (301.26) en (301.19) laat zich nu als totale waarde van $<u_{kp}^2>$ berekenen

$$(301.27) \qquad <u_{kp}^2> = \frac{n-1}{n} j^2 + \frac{(m-1)(n-1)}{mn} \sigma^2.$$

Als formule voor de interactievarians is gebruikelijk

$$\frac{\uparrow\uparrow mn\,[[u_{\kappa\pi}^2]]}{(m-1)(n-1)}.$$

Wanneer wij deze grootheid weergeven door $\psi_{\rho\mu}^2$ kunnen we schrijven:

$$(301.28) \qquad \psi_{\rho\mu}^2 = \frac{\uparrow\uparrow mn\,[[u_{\kappa\pi}^2]]}{(m-1)(n-1)} = \frac{m}{m-1} j^2 + \sigma^2.$$

Men ziet hier een middel de grootte van de middelbare interpretatiefout j te schatten, wanneer de interactievarians en de toevalsvarians bekend zijn. De toevalsvarians kan vaak berekend worden uit het verschil tussen de blokken (herhalingen) binnen de proefvelden. Hierop zullen wij in hfdst. IV dieper ingaan.

302 – De correlatie van de schijnbare fouten van onafhankelijke rassen

Wanneer men de rassen als volkomen onafhankelijk beschouwt, zijn de interpretatiefouten onderling afhankelijk. Dit volgt direct uit formule (301.23) waar staat

$$(301.23) \qquad [i_{\kappa p}] = 0$$

Hieruit is af te leiden

$$\uparrow(m-1)\,[i_{\lfloor k \rfloor p}] = -i_{kp}$$

of

$$(302.1) \qquad \frac{\uparrow(m-1)\,[i_{\lfloor k \rfloor p}]}{m-1} = -\frac{i_{kp}}{m-1}.$$

In deze formule wordt een verband gezien tussen de rassen $\lfloor k \rfloor$ en ras k. Om dit verband nader te bestuderen is het gemakkelijk een typische vertegenwoordiger voor de rassen $\lfloor k \rfloor$ te kiezen. Wij willen dit ras aanduiden als ras l.

Wanneer men nu weet, dat de interpretatiefout van ras k op proefveld p gelijk aan i_{kp} is, volgt uit (302.1) dat de gemiddelde waarde van de interpretatiefout van een ras $l\,(\neq k)$ op dat proefveld p gelijk is aan $-\frac{i_{kp}}{m-1}$. Wij willen de fout i_{lp} ontleden in zijn gemiddelde $-\frac{i_{kp}}{m-1}$ en de afwijking van dat gemiddelde a_{lp}, zodat

(302.2) $$i_{lp} = -\frac{i_{kp}}{m-1} + a_{lp}.$$

In deze formule wordt dus een verband gelegd tussen de waarden i_{lp} en i_{kp}. Het is nu de vraag of wij het verband mogen zien als een geval van correlatie, of als een „geval van lijnvereffening". Om dit uit te maken zullen wij het criterium aanleggen van 122.

Volgens dat criterium zou er van correlatie gesproken mogen worden, wanneer i_{lp} en i_{kp} beide normaal verdeeld waren. Wij hebben aan het eind van 117 uiteengezet dat het moeilijk kan zijn hierover een uitspraak te doen, omdat de interpretatiefouten *systematische* fouten zijn. Wij hebben daar geëist dat er een *bewuste* reden moet zijn om aan te nemen dat de besproken interpretatiefouten de toevalswetten volgen. Welnu, zo'n reden is hier te noemen. In hfdst. II werd besproken dat de milieugunstigheid M_p van talloze invloeden afhangt, die allen door het proefveld globaal zijn uitgeschakeld (zie 201 slot).

Nu is het een bekende wiskundige wet, dat er een normale verdeling optreedt, wanneer er vele onderling onafhankelijke invloeden zijn, waarvan niet één domineert. Doordat door het proefveld de invloeden globaal zijn uitgeschakeld, zal er van domineren *vermoedelijk* geen sprake meer zijn.

Wij komen dus tot de conclusie dat de interpretatiefouten i_{lp} en i_{kp} waarschijnlijk normaal verdeeld zijn, zodat het juist is de samenhang tussen i_{kp} en i_{lp} als een geval van correlatie op te vatten.

Wij willen nu trachten de correlatiecoëfficient te berekenen. Daartoe gaan wij uit van (302.2)

Substitutie van de i uit (302.2) in (301.18) geeft

$$t_{kp} = i_{kp} + \tau_{kp}$$

$$t_{lp} = -\frac{i_{kp}}{m-1} + a_{lp} + \tau_{lp}.$$

De asymptotische waarde $<t_{kp}\, t_{lp}>$ zal dus zijn

(302.3) $$<t_{kp}\, t_{lp}> = <i_{kp}\, i_{lp}> = -\frac{<i_{kp}^2>}{m-1} + <i_{kp}\, a_{lp}>.$$

De producten met τ worden immers asymptotisch 0 wegens onafhankelijkheid van de factoren.

Om de waarde van $<i_{kp}\, a_{lp}>$ te kunnen berekenen willen wij eerst formule (302.2) sommeren over l. Dit geeft

$$\uparrow(m-1)\,[i_{\lfloor k \rfloor p}] = -\,i_{kp} + \uparrow(m-1)\,[a_{lp}].$$

Vergelijking met (302.1) toont dat

$$\uparrow(m-1)\,[a_{lp}] = 0.$$

Wanneer wij $<i_{kp}\,a_{lp}>$ over l middelen fungeert i_{kp} als een constante. Deze term wordt dan dus ook 0.

Waar wij aan het begin van deze paragraaf volkomen onafhankelijkheid der rassen hebben aangenomen, sluit dit uit, dat een *bepaald* ras l *systematisch* in een andere verhouding tot ras k staat, dan het gemiddelde van alle rassen l. Voor een bepaald ras l moet $<i_{kp}\,a_{lp}>$ dus evenzeer 0 worden, als voor het gemiddelde van allen. (302.3) wordt dus

$$(302.4) \qquad <t_{kp}\,t_{lp}> = <i_{kp}\,i_{lp}> = -\frac{<i_{kp}^2>}{m-1}.$$

Op dezelfde manier laat zich bewijzen

$$<t_{kp}\,t_{lp}> = -\frac{<i_{lp}^2>}{m-1}.$$

Hieruit moet dus volgen dat de asymptotische waarden $<i_{kp}^2>$ en $<i_{lp}^2>$ aan elkaar gelijk zijn. Op deze weg voortgaande kunnen wij zeggen *dat de asymptotische waarde van de interpretatievarians voor alle rassen gelijk is*, mits er geen afhankelijkheid tussen bepaalde rassen bestaat. Deze asymptotische waarde hebben wij in formule (301.27) aangeduid door j^2. Wanneer wij dit ook doen in (302.4) vinden wij

$$(302.5) \qquad <t_{kp}\,t_{lp}> = -\frac{j^2}{m-1}.$$

In de practijk is het niet mogelijk de waarde van de asymptotische fout t_{kp} te bepalen. Er moet gewerkt worden met de schijnbare fout u_{kp}. Wij zullen daarom na moeten gaan wat de asymptotische waarde is van het product $u_{kp}\,u_{lp}$.

Uit (301.16) is het toevallig deel van u_{kp} te vinden door de t te vervangen door τ; in (301.24) is het systematisch deel van u_{kp} gegeven, zodat wij weten

$$(302.6) \quad u_{kp} = \frac{n-1}{n} i_{kp} - \uparrow(n-1)\left[\frac{i_{k\lfloor p \rfloor}}{n}\right] + \frac{(m-1)(n-1)}{mn}\,\tau_{kp} -$$

$$- \uparrow(n-1)\left[\frac{(m-1)\,\tau_{k\lfloor p \rfloor}}{mn}\right] - \uparrow(m-1)\left[\frac{(n-1)\,\tau_{\lfloor k \rfloor p}}{mn}\right] +$$

$$+ \uparrow\uparrow(m-1)(n-1)\left[\left[\frac{\tau_{\lfloor k \rfloor \lfloor p \rfloor}}{mn}\right]\right].$$

Voor u_{lp} is een soortgelijke formule op te stellen.

Bij het berekenen van $<u_{kp}\, u_{lp}>$ mogen wij bedenken dat alle producten $i\tau$ asymptotisch 0 worden; producten van twee toevallige fouten τ met ongelijke indices worden eveneens 0. Evenwel moet bedacht worden dat $[k]$ soms l betekent en dat $[l]$ de betekenis k kan hebben.

Rekening houdend met het bovenstaande vinden wij als waarde van $<u_{kp}\, u_{lp}>$

(302.7)

$$
\begin{aligned}
<u_{kp}\, u_{lp}> = & \frac{(n-1)^2}{n^2} <i_{kp}\, i_{lp}> - \uparrow(n-1)\left[\frac{(n-1)<i_{k\lfloor p\rfloor}\, i_{lp}>}{n^2}\right] - \\
& - \uparrow(n-1)\left[\frac{(n-1)<i_{l\lfloor p\rfloor}\, i_{kp}>}{n^2}\right] + \\
& + <\uparrow(n-1)\left[\frac{i_{k\lfloor p\rfloor}}{n}\right]\uparrow(n-1)\left[\frac{i_{l\lfloor p\rfloor}}{n}\right]> + \\
& + <\frac{(m-1)(n-1)}{mn}\tau_{lp} \times -\frac{n-1}{mn}\tau_{lp}> + \\
& - \uparrow(n-1)\left[<\frac{(m-1)\,\tau_{l\lfloor p\rfloor}}{mn} \times \frac{\tau_{l\lfloor p\rfloor}}{mn}>\right] - \\
& - <\frac{n-1}{mn}\tau_{kp} \times \frac{(m-1)(n-1)}{mn}\tau_{kp}> + \\
& + \uparrow(n-1)\left[<\frac{\tau_{k\lfloor p\rfloor}}{mn} \times -\frac{m-1}{mn}\tau_{k\lfloor p\rfloor}>\right].
\end{aligned}
$$

De termen met i laten zich nog aanzienlijk vereenvoudigen. Volgens (302.4) is $<i_{kp}\, i_{lp}> = -\dfrac{<i_{kp}^2>}{m-1} = -\dfrac{j^2}{m-1}$.

De waarde van $<i_{k\lfloor p\rfloor}\, i_{lp}>$ en $<i_{l\lfloor p\rfloor}\, i_{kp}>$ is 0, aangezien beide interpretatiefouten aan een ander proefveld ontleend zijn. Bij de bespreking van (301.24) zagen wij dat zelfs de waarden i_{kp} en $i_{k\lfloor p\rfloor}$ onderling onafhankelijk zijn. Deze onafhankelijkheid blijft natuurlijk gehandhaafd wanneer wij ook nog van ras verwisselen.

Om de waarde van

$$<\uparrow(n-1)\left[\frac{i_{k\lfloor p\rfloor}}{n}\right]\uparrow(n-1)\left[\frac{i_{l\lfloor p\rfloor}}{n}\right]>$$

te berekenen zullen wij de vorm voluit moeten schrijven. Dan lezen wij

$$<\uparrow(n-1)\left[\frac{i_{k[p]}}{n}\right]\uparrow(n-1)\left[\frac{i_{lp}}{n}\right]> = \frac{1}{n^2}<(i_{k_1}+i_{k_2}+\ldots)(i_{l_1}+i_{l_2}+\ldots)> =$$

$$= \frac{1}{n^2}\uparrow(n-1)[<i_{kq}\, i_{lq}>] + \text{een aantal dubbele producten die 0 worden.}$$

In verband met (302.4) en (302.5) is de waarde van deze vorm gelijk te stellen aan $-\frac{n-1}{n^2(m-1)}j^2$.

Wanneer wij verder iedere waarde τ^2 zien als een schatting van σ^2 vinden wij als asymptotische waarde van (302.7).

$$(302.8) \quad <u_{kp}\, u_{lp}> = -\frac{(n-1)^2}{n^2(m-1)}j^2 - \frac{n-1}{n^2(m-1)}j^2 -$$

$$-2\sigma^2\left(\frac{(m-1)(n-1)^2}{m^2n^2} + \frac{(n-1)(m-1)}{m^2n^2}\right) =$$

$$= -\frac{n-1}{n(m-1)}j^2 - \frac{2(m-1)(n-1)}{m^2n}\sigma^2.$$

Sommatie van (302.8) over p geeft

$$\uparrow n\,[<u_{k\pi}\, u_{l\pi}>] = -\frac{n-1}{m-1}j^2 - \frac{2(m-1)(n-1)}{m^2}\sigma^2$$

of

$$(302.9) \qquad <\frac{\uparrow n\,[u_{k\pi}\, u_{l\pi}]}{n-1}> = -\left(\frac{j^2}{m-1} + \frac{2(m-1)}{m^2}\sigma^2\right).$$

Men ziet hier dat de vorm

$$\frac{\uparrow n\,[u_{k\pi}\, u_{l\pi}]}{n-1}$$

wat zijn systematisch deel betreft de waarde aanneemt die op grond van (302.5) verwacht kan worden. Verrassend is evenwel de grote invloed van de toevallige fout. De coëfficient van σ^2 is bij niet te kleine m ongeveer het dubbele van die van j^2.

Om een correlatiecoëfficient te berekenen moeten wij de j^2 en σ^2 uit (302.9) wegdelen. Dit kan gedaan worden met behulp van de waarde $\psi_{p\mu}{}^2$ uit form. (301.28).

De j^2 en σ^2 komen in (302.9) evenwel in een andere verhouding voor dan in (301.28). In verband hiermee kunnen wij de correlatiecoëfficient wellicht het best berekenen volgens de formule

$$(302.10) \qquad \frac{\frac{1}{n-1} \uparrow n \left[u_{k\pi}\, u_{l\pi}\right]}{\psi_{\rho\mu}^2 + \frac{m-2}{m}\sigma^2} \simeq -\frac{1}{m}$$

[het $\simeq$ teken is nodig omdat van asymptotische naar empirische waarden is overgegaan].

Mocht de waarde van de breuk uit formule (302.10) inderdaad ongeveer $-\frac{1}{m}$ zijn, dan wordt daardoor het vermoeden van onafhankelijkheid van de onderzochte rassen gesteund. Wordt er een sterk afwijkende waarde gevonden, dan moet er aan afhankelijkheid van rassen worden gedacht.

Voor de volledigheid is het misschien wenselijk nog op twee punten te wijzen.

In de eerste plaats is onze conclusie uit formule (302.4), dat de asymptotische waarde van alle interpretatievariansen gelijk is, in werkelijkheid geen conclusie, maar de vooronderstelling, die de overgang van (302.1) naar (302.2) redelijk maakt. Wanneer de interpretatiefouten $i_{[k]p}$ wezenlijk ongelijk zijn kunnen wij ze niet laten vertegenwoordigen door een bepaalde i_{lp}.

In de tweede plaats zijn de formules voor de regressielijnen nl. (zie 302.2)

$$i_{lp} = -\frac{i_{kp}}{m-1} + a_{lp}$$

$$i_{kp} = -\frac{i_{lp}}{m-1} + a_{kp}$$

reeds vastgesteld voor over normale correlatie gesproken was. In dit geval zouden ook van regressielijnen gebruik zijn gemaakt, wanneer er geen sprake was van correlatie. Dat hier afgeweken wordt van de voorschriften voor lijnvereffening vindt hierin zijn oorzaak, dat i_{lp} geen zuivere functie van i_{kp} is. Dat hier regressielijnen worden genomen hangt samen met hetgeen wij bespraken in 124, naar aanleiding van figuur (124.5). Wij kwamen daar tot de conclusie dat de regressielijn de gevonden afwijkingen naar evenredigheid verdeelt over de verschillende kansverdelingen. Volgens (302.2) wordt zo een grote i_{lp} niet alleen geweten aan een grote i_{kp}, maar ook aan een grote waarde van de fouten $i_{[k\,l]p}$. Deze verdeling hangt weer samen met de aprioristische gelijkwaardigheid van alle interpretatiefouten.

303 – AFHANKELIJKHEID VAN RASSEN

Aardappelrassen worden langs vegetatieve weg vermeerderd en zijn dus klonen. Alle individuen van een ras zijn erfelijk gelijk. Nu worden in deze klonen soms mutanten gevonden. Van deze planten zijn één of enkele eigenschappen toevallig veranderd. Worden deze veranderingen op de proefvelden onderzocht, dan wordt de mutant als een apart ras beschouwd, althans proefveldtechnisch. Het spreekt vanzelf dat zo'n mutant sterk verwant is aan de oorspronkelijke kloon. Ook door andere oorzaken kan er verwantschap ontstaan. Zweedse graanrassen zullen vaak collectief anders op wintervastheid reageren dan Franse rassen. Kortom, het is op allerlei wijze mogelijk dat twee rassen extra veel op elkaar gelijken.

Wij willen nu voor enkele formules van de vorige paragraaf nagaan, hoe ze worden, wanneer tegelijk ras k wordt onderzocht en een knopmutant k', die niet voor de onderzochte eigenschap is gemuteerd. In dit geval kunnen wij stellen dat de interpretatiefouten van k en k' gelijk zijn, zodat

(303.1) $$i_{kp} = i_{k'p}$$

Formule (302.1) wordt nu

$$\uparrow(m-2)\,[\lfloor k\,k'\rfloor p] = -(i_{kp} + i_{k'p}) = -2\,i_{kp}$$

of

$$\frac{\uparrow(m-2)\,[i_{\lfloor kk'\rfloor p}]}{m-2} = -\frac{2}{m-2}\,i_{kp},$$

waarbij $\lfloor k\,k'\rfloor$ betekent, dat $\varkappa$ alle waarden aanneemt behalve k en k'. Eén ras uit deze groep willen wij als ras l aanduiden. Het is nu gemakkelijk in te zien dat

(303.2) $$\langle i_{kp}\,i_{k'p}\rangle = \langle i_{kp}^2\rangle$$

$$\langle i_{kp}\,i_{lp}\rangle = -\frac{2\,\langle i_{kp}^2\rangle}{m-2}.$$

Vergelijkt men deze formule met (302.4) dan ziet men dat de correlatie tussen twee verwante rassen algebraïsch hoger wordt dan $-\frac{1}{m-1}$, terwijl deze hogere correlatie meebrengt, dat de correlatie met een willekeurig onafhankelijk ras algebraïsch lager wordt.

Wanneer wij van ras l uitgaan in onze beschouwingen kunnen wij (302.1) schrijven

$$\uparrow(m-1)\,[i_{\lfloor l \rfloor p}] = -\,i_{lp}$$

of

$$\frac{\uparrow(m-1)[i_{\lfloor l \rfloor p}]}{m-1} = -\,\frac{i_{lp}}{m-1}.$$

Wanneer wij nu voor i_{kp} en $i_{k'p}$ een gemiddelde waarde berekenen zal het in beide gevallen zijn

$$\bar{i}_{kp} = -\,\frac{i_{lp}}{m-1}$$

$$\bar{i}_{k'p} = -\,\frac{i_{lp}}{m-1}.$$

Dit is niet in strijd met (303.1). Formule (302.4) gaat dus voor ons geval in zoverre op dat geldt

(303.3) $$<i_{kp}\, i_{lp}> = -\,\frac{<i_{lp}^2>}{m-1}.$$

Vergelijking van (303.2) en (303.3) leert

$$\frac{<i_{lp}^2>}{m-1} = \frac{2\,<i_{kp}^2>}{m-2}$$

of

(303.4) $$\frac{<i_{lp}^2>}{<i_{kp}^2>} = \frac{2\,(m-1)}{(m-2)}.$$

Wij zien zo, dat de middelbare waarde van i_{kp} kleiner is dan die van i_{lp}. In woorden uitgedrukt: *Wanneer twee of meer rassen afhankelijk van elkaar zijn, wordt daardoor de absolute grootte van hun interpretatiefout ten opzichte van die van de onafhankelijke rassen gedrukt.*

Eigenlijk is dit niet verwonderlijk. Wanneer men b.v. 10 verschillende knopmutanten van één kloon onderzoekt, die voor de onderzochte eigenschap niet zijn veranderd en één heel ander ras, dan zal het fictiefras, zoals gedefinieerd in (205.10) bijna identiek zijn met de soms mutanten gevende kloon. De afwijkingen tussen de mutanten en het fictiefras zullen dan ook veel kleiner zijn, dan die tussen het fictiefras en een vreemde kloon. Wij zien dus dat de formules (301.26), (301.28), (302.8) en (302.9) die alle asymptotische waarden $<i_{\kappa\pi}^2>$ gelijkstellen aan j^2 niet juist zijn, wanneer er afhankelijkheid tussen sommige rassen bestaat.

Toch behouden ze hun nut. Formule (301.28) is belangrijk om te controleren of er naast de toevallige fouten nog systematische voorkomen. Wanneer uit formule (301.28) blijkt dat er geen systematische fouten voorkomen, zijn ze ook niet verwant. Wanneer er wel systematische fouten voorkomen moeten ze toch nader worden onderzocht. Formule (302.9) vindt zijn bestemming in (302.10). Deze formule dient er voor om te controleren of er afhankelijkheid tussen de rassen bestaat of niet.

Het lijkt ons onbegonnen werk formule (302.10) zodanig te wijzigen, dat de weg wordt geopend voor een empirische bepaling van de mate van correlatie. In 309 wordt een betere mogelijkheid besproken.

304 – Het samenvatten van proeven met wisselend aantal onafhankelijke rassen

Veronderstel dat men in het jaar a van een bepaald gewas 5 rassen had, aan te duiden als k_1, k_2, ... k_5. In het volgend jaar $(a + 1)$ is ras k_1 van de rassenlijst afgevoerd, dat ras werd dus niet langer onderzocht. Daarentegen boden de kwekers twee nieuwe rassen aan (k_6 en k_7). Ook het volgend jaar $(a + 2)$ traden er mutaties op. Enz. In tabel (304.1) is samengevat welke rassen in de verschillende jaren onderzocht zijn.

(304.1)

ras / jaar	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9
a	×	×	×	×	×				
$a + 1$		×	×	×	×	×	×		
$a + 2$			×	×	×	×	×		
$a + 3$			×	×	×		×		
$a + 4$				×	×		×	×	×
$a + 5$					×		×	×	×

Moet nu de voortreffelijkheid van al deze rassen worden vergeleken?; en zo ja, hoe? Zie hier een paar vragen waarvoor het practische proefveldwerk de onderzoeker steeds stelt.

Om met de eerste vraag te beginnen: Wanneer er een nieuwe ziekte optreedt, waarvoor k_1 ongevoelig is en de andere rassen meer of minder vatbaar, dan is het zeker nodig voor alle rassen

de onderlinge voortreffelijkheidsverschillen te bepalen. Pas dan kan beoordeeld worden hoeveel oogstderving men moet accepteren wanneer men ras k_1 verbouwt om het risico van de ziekte te vermijden.

De onderlinge voortreffelijkheidsverschillen zullen bepaald moeten worden aan de hand van de gegevens van bovenvermelde proeven. Men kan niet een proef van het volgend jaar afwachten, omdat de practijk dan reeds over een goede richtlijn voor de rassenkeuze moet beschikken.

Bij het verwerken van bovenstaande proeven valt geen gebruik te maken van de FISHER-methode. Het schema is niet orthogonaal en kan dat ook niet gemaakt worden. Het is theoretisch mogelijk afgevoerde rassen steeds te blijven onderzoeken om de orthogonaliteit te handhaven, doch nieuw gekweekte rassen laten zich niet voor hun ontstaan op de velden plaatsen. Orthogonaliteit is dus onbereikbaar. Daarom is het nodig een methode te hebben voor het verwerken van niet-orthogonale schema's van bovenstaand type. Wanneer die methode eenmaal ontwikkeld is, kunnen ook binnen het jaar zonder bezwaar een variërend aantal rassen worden onderzocht op de diverse proefvelden. In Nederland wordt hiertoe vaak de behoefte gevoeld.

Wij willen nu nagaan hoe deze proeven kunnen worden samengevat. Daartoe moeten wij eerst letten op onze definitie van fictiefras, zoals gegeven in (205.10). Daar wordt geëist dat het fictiefras het gemiddelde is van alle onderzochte rassen. Het begrip „alle onderzochte rassen" is in schema (304.1) niet duidelijk. In totaal zijn onderzocht de rassen $k_1 \ldots . k_9$, maar „alle onderzochte rassen" voor het jaar a zijn $k_1 \ldots . k_5$. Wanneer wij de opbrengst van ras k_1 in 't jaar a (dus Ω_{1a}) in verband met (205.10) en (301.4) willen splitsen in

$$\Omega_{1a} = \varrho_1 + M_a \tag{304.2}$$

dan is de grootheid M_a moeilijk te bepalen, wanneer daaronder verstaan moet worden de gemiddelde opbrengst van de negen rassen, die er maar gedeeltelijk gestaan hebben. Omgekeerd is het onderling verband tussen M_a, M_{a+1}, M_{a+2} enz. onduidelijk, wanneer het fictiefras, waarvan dit de opbrengsten zijn, niet steeds hetzelfde is. Wij kunnen zeggen: het is *wenselijk alle* rassen in het fictiefras op te nemen, maar misschien kan het niet. Wij *eisen* daarom. *Neem zoveel mogelijk rassen op in het fictiefras.*

Omdat de theoretische problemen het gemakkelijkst te bespreken zijn wanneer de methode ons duidelijk voor ogen staat, willen wij eerst een geschikte methode uitleggen. Wegens het grillig karakter van de voorkomende schema's kunnen wij daarbij beter met voorbeelden werken dan met formules.

Wij veronderstellen dat wij vier proeven ($p_1 \ldots p_4$) hebben genomen met de rassen $k_1 \ldots k_6$. De opbrengsten $\Omega_{\kappa\pi}$ waren als volgt:

(304.3)

ras / proef	k_1	k_2	k_3	k_4	k_5	k_6
p_1	663	672	819	653	613	
p_2	681	663	690	663	643	643
p_3	556	602	591	301		505
p_4	623	591	623	613	580	580

Men ziet dat op veld p_1 ras k_6 heeft ontbroken, op veld p_3 ras k_5. Het zijn dus de proefvelden p_1 en p_3 die moeilijkheden geven. Voor de velden p_2 en p_4 is het begrip „fictiefras" ondubbelzinnig bepaald. Wij beginnen daarom met de velden p_2 en p_4 samen te vatten. Met behulp van (205.10) berekenen wij voor M_2 en M_4 als afgeronde waarden

$$M_2 = 664$$
$$M_4 = 602.$$

Men ziet dat het milieu p_2 62 eenheden gunstiger was dan het milieu p_4. Om p_2 en p_4 geheel vergelijkbaar te maken kunnen wij b.v. de gunstigheid van p_4 met 62 verhogen door bij alle waarden $\Omega_{\kappa 4}$ het bedrag 62 op te tellen. Wij willen dit aanduiden door te zeggen dat p_4 op p_2 wordt *gestandaardiseerd.* Het op p_2 gestandaardiseerde milieu p_4 willen wij aanduiden door p_4'. Wij vinden dus

(304.4)

ras / proef	k_1	k_2	k_3	k_4	k_5	k_6		M
p_2	681	663	690	663	643	643		664
p_4'	685	653	685	675	642	642		664
som A	1366	1316	1375	1338	1285	1285		
gem. A	683	658	687	669	642	642		664

Om vorenstaand resultaat in de verdere berekening gemakkelijk te kunnen aanhalen, hebben wij de aanduidingen som A en gem(iddelde) A gebruikt. Zo zal straks gesproken worden van som en gem. B, C, D enz.

Men ziet in (304.4) dus driemaal meegedeeld wat de verschillende rassen als opbrengst geven bij de milieugunstigheid $M = 664$. Welke van deze mededelingen is nu het betrouwbaarst? Aanvankelijk zou men kunnen overwegen aan de cijfers van p_2 de voorkeur te geven, omdat deze cijfers inderdaad bij de gunstigheid 664 verkregen zijn. Deze voorkeur zou evenwel in strijd zijn met 202 waar wij van de rekenopbrengsten Ω eisten, dat ze in zulke grootheden werden uitgedrukt, dat de voortreffelijkheidsverschillen van de onderzochte rassen onafhankelijk werden van de gunstigheid van het milieu, waarin het onderzoek plaats vond.

De enige reden, verschil in waarde tussen de drie bovenstaande „mededelingen" te zien, ligt hierin dat gem. A gebaseerd is op meer gegevens dan de cijfers van p_2 of p_4'. Niet alleen is de invloed van de toevallige fout hierdoor verkleind, maar ook de monstername uit het onderzochte gebied *kan* hierdoor verbeterd zijn. Wij willen dus verder werken met gem. A.

Nu is het dus nodig de gegevens van de velden p_1 en p_3 nog in onze conclusie te betrekken. Wanneer wij met p_1 beginnen, staan wij voor de noodzaak als definitie van het fictiefras te gebruiken

$$M_1^* = \frac{\uparrow 5\,[\Omega_{\lfloor 6 \rfloor 1}]}{5}.$$

Wij moeten de gemiddelde opbrengst van de rassen $k_1 \ldots . k_5$ nemen. Dit heeft slechts zin wanneer wij dit gemiddelde vergelijken met het gemiddelde van dezelfde rassen uit gem. A. Wanneer wij dit gemiddelde aanduiden door M_A^* vinden wij

$$M_A^* = 668$$

$$M_1^* = 684.$$

Men ziet dus dat de gunstigheid van p_1 16 hoger ligt dan die van gem. A. Om de getallen vergelijkbaar te maken verlagen wij alle opbrengsten van veld p_1 met 16. Wanneer wij het aldus op gem. A gestandaardiseerde milieu p_1 aanduiden door p_1' vinden wij

(304.5)

ras / proef	k_1	k_2	k_3	k_4	k_5	k_6		M^*
gem. A	683	658	687	669	642	642		668
som A	1366	1316	1375	1338	1285	1285		
p_1'	647	656	803	637	597			668
som B	2013(3)	1972(3)	2178(3)	1975(3)	1882(3)	1285(2)		
gem. B	671	657	726	658	627	642		

Gem. A en p_1' zijn twee reeksen opbrengsten, zoals die bereikt kunnen worden bij de milieugunstigheid $M^* = 668$. Van deze reeksen is gem. A gebaseerd op twee proefvelden, p_1' op een. Wanneer wij nu aannemen dat er geen noodzaak is het milieu te wegen in de zin van 206 kan het gewicht van alle opbrengsten gelijk worden genomen. Het gewicht van gem. A is dan dus tweemaal het gewicht van p_1'. Wij moesten dus *som* A met p_1' samentellen en daaruit het nieuwe gemiddelde B berekenen. Omdat nu niet alle uitkomsten op evenveel gegevens berusten is voor ieder ras het aantal waarnemingen tussen haakjes vermeld.

Aangezien gem. B nu de betrouwbaarste schatting geeft van de rasverschillen, standaardiseren we p_3 op gem. B. Wij vinden

(304.6)

ras / proef	k_1	k_2	k_3	k_4	k_5	k_6		M^{**}
gem. B	671	657	726	658	627	642		671
som B	2013(3)	1972(3)	2178(3)	1975(3)	1882(3)	1285(2)		
p_3'	716	762	751	461		665		671
som C	2729(4)	2734(4)	2929(4)	2436(4)	1882(3)	1950(3)		
gem. C	682	683	732	609	627	650		

Zo zien wij dus dat gem. C een schatting geeft van de onderlinge rasverschillen, die gebaseerd is op alle gegevens. Toch is het misschien mogelijk een betere schatting te verkrijgen. Bij het standaardiseren van veld p_1 is geen rekening gehouden met de gegevens van p_3. Deze onvolkomenheid kan hersteld worden door nu alle

vier proefvelden op gem. C te standaardiseren. Eventuele vergissingen worden langs deze weg tevens automatisch weer hersteld. Wij vinden

(304.7)

ras / proef	k_1	k_2	k_3	k_4	k_5	k_6
gem. C	682	683	732	609	627	650
p_1''	646	655	802	636	596	
p_2''	681	663	690	663	643	643
p_3''	716	762	751	461		665
p_4''	685	653	685	675	642	642
som D	2728(4)	2733(3)	2928(4)	2435(4)	1881(3)	1950(3)
gem. D	682	683	732	609	627	650

Men ziet dat gem. C en gem. D gelijk zijn. Dit gemiddelde geeft nu de betrouwbaarste schatting van de rasverschillen.

Het bovenstaande voorbeeld was erg eenvoudig. Wij geven nu nog een ingewikkelder (304.8). De velden zijn gemakshalve aangeduid met de letters a .... k, de rassen met l .... w.

(304.8)

ras / veld	l	m	n	p	q	r	s	t	u	w
a	732	728		668						
b	372	428		320		408	422		408	310
c	525		513	539		538	568		597	
d	712		733	713		662	706			718
e	662		640	609				651		
f	517		528	531					552	503
g	656		667		742	678	693		736	
h	695		568		702	660	680		690	639
i	618		677		669					632
k	659		655		688				649	695

Wij zoeken eerst weer die velden bij elkaar die de meeste rassen gemeenschappelijk hebben, dit zijn b en h. Als standaardisatieniveau nemen wij $M_b + 300$. Dit getal 300 wordt bij M_b opgeteld, omdat dan ongeveer het niveau van de meeste velden wordt

(304.9)

	l	*m*	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*w*	Fictiefras	[*M*]
b'	672	728		620		708	722		708	610	*l, r, s, u, w*	3420
h'	706		579		713	671	691		701	650		3419
som	1378(2)	728(1)	579(1)	620(1)	713(1)	1379(2)	1413(2)		1409(2)	1260(2)	*l,n,p,r,s,u*	3987
gem.	689	728	579	620	713	689	706		704	630		3988
c'	643		631	657		656	686		715			
som	2021(3)	728(1)	1210(2)	1277(2)	713(1)	2035(3)	2099(3)		2124(3)	1260(2)	*l,n,p,r,s,w*	3925
gem.	674	728	605	638	713	678	700		708	630		3926
d'	659		680	660		609	653			665		
som	2680(4)	728(1)	1890(3)	1937(3)	713(1)	2644(4)	2752(4)		2124(3)	1925(3)	*l,n,q,r,s,u*	4070
gem.	670	728	630	646	713	661	688		708	642		4070
g'	639		650		725	661	676		719			
som	3319(5)	728(1)	2540(4)	1937(3)	1438(2)	3305(5)	3428(5)		2843(4)	1925(3)	*l, n, p, u, w*	3298
gem.	664	728	635	646	719	661	686		711	642		3296
f'	650		661	664					685	636		
som	3969(6)	728(1)	3201(5)	2601(4)	1438(2)	3305(5)	3428(5)		3528(5)	2561(4)	*l, n, q, u, w*	3366
gem.	661	728	640	650	719	661	686		706	640		3366
k'	663		659		692				653	699		
som	4632(7)	728(1)	3860(6)	2601(4)	2130(3)	3305(5)	3428(5)		4181(6)	3260(5)	*l, n, q, w*	2667
gem.	662	728	643	650	710	661	686		697	652		2668
i'	636		695		687					650		
som	5268(8)	728(1)	4555(7)	2601(4)	2817(4)	3305(5)	3428(5)		4181(6)	3910(6)	*l, n, p*	1959
gem.	658	728	651	650	704	661	686		697	652		1959
e'	678		656	625				667				
som	5946(9)	728(1)	5211(8)	3226(5)	2817(4)	3305(5)	3428(5)	667(1)	4181(6)	3910(6)	*l, m, p*	2034
gem.	661	728	651	645	704	661	686	667	697	652		2035
a'	701	697		637								
som	6647(10)	1425(2)	5211(8)	3863(6)	2817(4)	3305(5)	3428(5)	667(1)	4181(6)	3910(6)		
gem.	665	712	651	644	704	661	686	667	697	652		

bereikt; het heeft geen theoretische betekenis. Slechts wordt de onbelangrijkheid van de hoogte van het standaardisatieniveau hierdoor onderstreept.

In (304.9) geven wij de berekening zonder verdere verklaring, zodat duidelijk naar voren komt, hoe snel het resultaat bereikt is. Naast de berekening is vermeld uit welke rassen telkens het fictief-ras is opgebouwd, en wat de som is van de opbrengst van die rassen ([*M*]).

In (304.9) is dus ons eerste gemiddelde verkregen. Nu worden alle proeven ter contrôle op dit gemiddelde gestandaardiseerd. Daarbij moeten wij bedenken dat het gemiddelde van *t* volledig afhankelijk is van proefveld *e*, zodat de hoogte van dit cijfer niets zegt over de juistheid van de standaardisatie van *e*. Wij moeten het gemiddelde van *t* dus negeren.

Wij geven in (304.10) nu de tweede standaardisatie.

(304.10)

	l	*m*	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*w*
1e gem.	665	712	651	644	704	661	686		697	652
a''	696	692		632						
b''	665	721		613		701	715		701	603
c''	646		634	660		659	689		718	
d''	665		686	666		615	659			671
e''	678		656	625				667		
f''	653		664	667					688	639
g''	638		649		724	660	675		718	
h''	707		580		714	672	692		702	651
i''	637		696		688					651
k''	664		600		693				654	700
2e gem.	665	706	653	644	705	661	686	667	697	652

Bij vergelijking van het eerste en het tweede gemiddelde valt het op dat er drie uitkomsten anders zijn geworden. Wij zullen daarom nog eens op het tweede gemiddelde moeten standaardiseren om volledige stabiliteit te bereiken (zie 304.11).

(304.11)

	l	*m*	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*w*
2e gem.	665	706	653	644	705	661	686		697	652
a'''	694	690		630						
b'''	664	720		612		700	714		700	602
c'''	646		634	660		659	689		718	
d'''	665		686	666		615	659			671
e'''	679		757	626				668		
f'''	653		664	667					688	639
g'''	638		649		724	660	675		718	
h'''	707		580		714	672	692		702	651
i'''	638		697		689					652
k'''	664		660		693				654	700
3e gem.	665	705	653	643	705	661	686	668	697	652

Aangezien de uitkomsten nog niet geheel stabiel zijn standaarverdiseren wij nogmaals (304.12)

(304.12)

	l	*m*	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*w*
3e gem.	665	705	653	643	705	661	686		697	652
a^{IV}	694	690		630						
b^{IV}	664	720		612		700	714		700	602
c^{IV}	646		634	660		659	689		718	
d^{IV}	665		686	666		615	659			671
e^{IV}	679		657	626				668		
f^{IV}	653		664	667					688	639
g^{IV}	638		649		724	660	675		718	
h^{IV}	707		580		714	672	692		702	651
i^{IV}	638		697		689					652
k^{IV}	664		660		693				654	700
4e gem.	665	705	653	643	705	661	686	668	697	652

De cijfers van het 4e gem. zijn gelijk aan die van het 3e gem. zodat de berekening is afgelopen.

305 – Contrôle op de gevolgde methode

De gegevens, zoals die zijn weergegeven in tabel (304.8) laten zich ook verwerken met de methode van de kleinste kwadraten

De techniek hiervan wordt door VAN UVEN besproken in hfdst. XII van zijn leerboek.

Bovenbedoelde methode wordt als volgt toegepast. In formule (202.6) laat men C samensmelten met M_p. De 53 gegevens uit tabel (304.8) kunnen nu als volgt in 53 vergelijkingen worden geschreven

(305.1)
$$\begin{aligned} \Omega_{la} &= R_l + M_a = 732 \\ \Omega_{ma} &= R_m + M_a = 728 \\ \Omega_{pa} &= R_p + M_a = 668 \\ \Omega_{lb} &= R_l + M_b = 372 \\ \Omega_{mb} &= R_m + M_b = 428 \\ &\text{enz.} \end{aligned}$$

Omdat het uitsluitend om ras*verschillen* begonnen is, kan het nulpunt van R willekeurig worden gekozen. Wij stellen daarom

(305.2)
$$R_l = 0.$$

Dit geeft dan 10 onbekenden M en 9 onbekenden R. Uit de 53 vergelijkingen van (305.1) moeten deze 19 onbekenden worden gevonden. Wij hebben al deze waarden berekend, en daarna de voortreffelijkheidsverschillen tussen ras l en de overige rassen uitgerekend. Uit tabel (304.12) laten zich dezelfde afwijkingen berekenen. Beide uitkomsten zijn vermeld in (305.3).

(305.3)

Afwijkingen van ras l	l	m	n	p	q	r	s	t	u	w
volgens (304.12)	0	40	–12	–22	40	–4	21	3	32	–13
volgens methode der kleinste kwadraten	0	40	–11	–22	41	–4	21	3	32	–12

Het blijkt dat de verschillen die verkregen worden met onze standaardisatiemethode gelijk zijn aan die, verkregen volgens de methode der kleinste kwadraten. De geringe afwijkingen die (305.3) te zien geeft zijn het gevolg van afrondingsfouten.

306 – HET TOEKENNEN VAN GEWICHTEN

In 304 hebben wij een methode uiteengezet voor het samenvatten van de proefvelden. Bij het ontwikkelen van deze methode zijn wij er van uitgegaan dat alle waarnemingen gelijk gewicht

hadden. In 206 hebben wij er echter nog eens nadrukkelijk op gewezen dat het vaak nodig zal zijn aan de cijfers gewichten toe te kennen op grond van de monsterfout.

Bij het toekennen van gewichten zijn twee gevallen te onderscheiden. In de eerste plaats kan het ene *proefveld* een ander gewicht krijgen dan het andere. In de tweede plaats is het mogelijk dat men aan de opbrengst van een bepaald *ras* op een bepaald proefveld een eigen gewicht toekent. Immers, wanneer niet op alle proefvelden dezelfde rassen staan, zullen de monsterfouten voor ieder ras anders zijn. Het gewicht dat een bepaald proefveld moet krijgen om de monsterfouten van dat ras te herstellen zal dan ook telkens anders zijn, d.w.z. iedere Ω_{kp} zal eventueel zijn eigen gewicht moeten krijgen.

Wij moeten dus nagaan op welke wijze deze gewichten het best kunnen worden toegekend. In de practijk is gebleken, dat het standaardiseren niet prettig verloopt, wanneer iedere waarde een ander gewicht heeft. Het is beter de waarden met wisselend gewicht te vervangen door andere waarden, waaraan het gewicht 1 kan worden toegekend.

Dit gaat gemakkelijk door eerst de proefvelden ongewogen op elkaar te standaardiseren, en vervolgens de afwijkingen u_{kp} te bepalen volgens de formule

(306.1) $$u_{kp} = \Omega^*_{kp} - \overline{\Omega}_k.$$

In deze formule duidt Ω^*_{kp} aan de gestandaardiseerde opbrengst van ras k op veld p, en $\overline{\Omega}_k$ de verkregen gemiddelde opbrengst van ras k.

Aan de opbrengst Ω^*_{kp} kan nu het gewicht g worden toegekend door de waarde

$$\Omega^*_{kp} = \overline{\Omega}_k + u_{kp} \text{ (met gewicht } g)$$

te vervangen door

(306.2) $$\Omega^*_{kp} = \overline{\Omega}_k + a u_{kp} \text{ (met gewicht 1)}$$

en vervolgens weer te standaardiseren.

Voor het berekenen van de juiste waarde van a willen wij een algemener formulering kiezen. Stel dat er n waarnemingen van x zijn (ieder met gewicht 1). Het gemiddelde der waarnemingen is $\overline{x}_{(n)}$. Hieraan wordt toegevoegd een nieuwe waarneming

$$x_{n+1} = \overline{x}_{(n)} + u_{n+1}$$

(met gewicht g). Het nieuwe gemiddelde zal zijn

$$(306.3)\quad \bar{x}_{(n+1)} = \frac{n\bar{x}_{(n)} + g\bar{x}_{(n)} + gu_{n+1}}{n+g} = \bar{x}_{(n)} + \frac{g}{n+g}\, u_{n+1}.$$

Wanneer wij

$$x_{n+1} = x_{(n)} + u_{n+1} \text{ (met gewicht } g)$$

vervangen door

$$x'_{n+1} = \bar{x}_{(n)} + au_{n+1} \text{ (met gewicht 1)},$$

kunnen wij als nieuw gemiddelde berekenen

$$(306.4)\quad \bar{x}'_{(n+1)} = \frac{n\bar{x}_{(n)} + \bar{x}_{(n)} + au_{n+1}}{n+1} = \bar{x}_{(n)} + \frac{a}{n+1}\, u_{n+1}.$$

Uit (306.3) en (306.4) laat zich berekenen

$$(306.5)\quad \bar{x}_{n+1)} - \bar{x}'_{(n+1)} = \frac{gn + g - an - ag}{(n+g)(n+1)}\, u_{n+1},$$

Het verschil tussen de juiste waarde $\bar{x}_{(n+1)}$ en de benaderde waarde $\bar{x}'_{(n+1)}$ is nul, wanneer

$$gn + g - an - ag = 0$$

of

$$(306.6)\qquad a = \frac{n+1}{n+g} \cdot g.$$

Wanneer g klein is ten opzichte van n kunnen we bij benadering stellen

$$(306.7)\qquad a = g.$$

Wij willen nu controleren hoe lang deze benadering geoorloofd is. Daartoe vervangen wij in (306.5) de a door g zodat wij krijgen

$$(306.8)\quad \bar{x}_{(n+1)} - \bar{x}'_{(n+1)} = \frac{g(1-g)}{(n+g)(n+1)}\, u_{n+1}.$$

Deze formule geeft dus de fout, die men maakt wanneer men van (306.7) uitgaat. Haar waarde is nul voor $g = 1$ en voor $g = 0$.

Voor het gebied $0 < g < 1$ wordt ongeveer de grootste waarde bereikt wanneer $g = \frac{1}{2}$ en $n = 1$. Men vindt dan $^1/_{12}\, u_{n+1}$. Bij $g = \frac{1}{2}$; $n = 2$ vindt men nog slechts $^1/_{30}\, u_{n+1}$. Wanneer men deze fouten accepteert kaumen dus zeggen dat (306.8) zeker geldt voor $0 < g < 1$.

Maar ook wanneer g groter is dan 1, is de fout vaak nog gering. Om een indruk te krijgen willen wij nu nagaan hoe groot g mag worden wanneer wij een fout van $-\,^1/_{20}\, u_{n+1}$ toelaten. Wij moeten de g dan oplossen uit de vergelijking

$$\frac{g(1-g)}{(n+g)(n+1)} = -\tfrac{1}{20}.$$

Hieruit wordt gevonden als positieve wortel

$$g = \frac{n + 21 + \sqrt{81n^2 + 122n + (21)^2}}{40}$$

of

$$g = \frac{n + 21 + \sqrt{(9n + 6.777)^2 + 21^2 - 6.777^2}}{40}.$$

Wanneer wij voor de eenvoudigheid de eis verzwaren geeft dit

$$g = \frac{10n + 27.7}{40}$$

of

$$g = {}^1/_4\, n + 0.69.$$

Een vollediger beeld krijgt men wanneer men stelt

$$\frac{g(1-g)}{(n+g)(n+1)} = -\delta$$

en voor enkele waarden van δ en n de maximumwaarde van g berekent. Men vindt dan de waarden van onderstaand tabelletje.

δ \ n	1	5	10	20	50	100
$\frac{1}{50}$	1.07	1.52	2.21	3.69	8.15	15.8
$\frac{1}{20}$	1.18	2.04	3.24	5.72	13.2	25.7
$\frac{1}{10}$	1.35	2.71	4.53	8.2	19.3	37.8
$\frac{1}{5}$	1.64	3.78	6.56	12.2	28.9	56.8

Samenvattend kunnen wij dus concluderen:

(306.9) *Aan een „afwijking van een gemiddelde" u kan men bij benadering een bepaald gewicht g toekennen door u door gu te vervangen met gewicht* 1, *mits g niet te groot wordt. Anders moet u worden vervangen door* $\frac{n+1}{n+g}\,gu$ (zie 306.6).

Deze stelling willen wij aanduiden als de *„gewichtenstelling"*.

Het is wenselijk even uitdrukkelijk bij de inhoud van deze gewichtenstelling stil te staan, aangezien op het eerste gezicht de inhoud tegenstrijdig is met de gangbare gewichtentheorieën. Volgens de gangbare opinie hoort een groot gewicht bij een *kleine* fout, volgens onze gewichtenstelling maakt een groot gewicht de fout *groot* (immers *gu* wordt groot).

De tegensteiling is slechts schijn. De gangbare gewichtentheorieën gaan over de vraag, *waarom* iets gedaan wordt, de gewichtenstelling over de vraag *wat* er gedaan wordt. De gewichtentheorieën zeggen hoe de fout *geschat* wordt, de gewichtenstelling, hoe hij als 't ware *gemaakt* wordt. Wanneer een fout *groot geschat* wordt, krijgt hij een klein gewicht en wordt dan als 't ware *klein gemaakt.*

Wanneer men de conclusie uit (306.9) wil toepassen op (306.2) moet men bedenken dat $\overline{\Omega}_k$ berekend moet zijn uit de opbrengsten $\Omega^*_{k[p]}$, en dat de waarde van n niet het aantal proefvelden aanduidt, maar het aantal (n_k) waarop ras k gestaan heeft min 1 dus

$$\overline{\Omega}_k = \frac{\uparrow(n_k - 1)\,[\Omega^*_{k[p]}]}{n_k - 1}. \tag{306.10}$$

$$n = n_k - 1.$$

In verband met de bestaande fouten zal het in de practijk meestal niet belangrijk zijn hierop te letten, behalve dan dat de n ongeveer de n_k aanduidt.

Tot dusver hebben wij gesproken over het toekennen van een bepaald gewicht aan een bepaalde opbrengst Ω^*_{kp}. Wanneer men aan een geheel proefveld een gewicht wil toekennen gaat het veel eenvoudiger. Men kan zich daarbij houden aan de volgende regel: Wil men aan een proefveld p het gewicht g toekennen (g niet te groot) dan is dit mogelijk door alle opbrengsten Ω_{kp} met g te vermenigvuldigen en daarna aan de cijfers het gewicht 1 toe te kennen. Is g te groot dan moet vermenigvuldigd worden met $\frac{n+1}{n+g} \cdot g$.

Met deze cijfers laat zich natuurlijk geen M_p uitrekenen. Dit dient te geschieden met behulp van de ongewijzigde cijfers.

Met behulp van deze gewogen cijfers laat zich dus het beste gemiddelde berekenen voor alle rassen. Om een juist inzicht in de verhoudingen binnen het proefveld te verkrijgen moeten ongewogen cijfers worden vermeld. Deze zijn te verkrijgen door op de *definitieve* rasgemiddelden de ongewogen proefveldcijfers te standaardiseren. Dan wordt tevens de best bereikbare waarde voor M_p gevonden.

307 – Het schatten van de fouten van de berekende rasverschillen

In 305 hebben wij laten zien dat de standaardisatiemethode dezelfde resultaten geeft als de methode der kleinste kwadraten, en als voordeel heeft dat er veel sneller gewerkt kan worden. Als nadeel staat hiertegenover dat geen exacte foutenberekening mogelijk is. Van onderstaande methode zal in 308 worden aangetoond dat de bereikte resultaten voldoende nauwkeurig zijn. Wij zullen de methode alleen uitwerken voor het geval alle gewichten 1 zijn.

Wanneer er geen interpretatie- of andere fouten gemaakt werden, zouden in tabel (304.12) alle gestandaardiseerde opbrengsten gelijk moeten zijn aan de rasgemiddelden. Door de fouten zijn er afwijkingen u_{kp} (zie 306.1). Het aantal gegevens ($\Omega_{\kappa\pi}$) in deze proef is 53, het aantal onbekenden is 19, nl. 10 waarden M_p en 9 waarden R_k (zie ook 305.2).

De formule voor de varians ψ^2 is dus

$$\psi^2 = \frac{\uparrow 53\,[uu]}{53 - 19}.$$

Indien uit de blokken van ieder proefveld σ^2 is berekend, kan nu worden nagegaan of ψ^2 zuiver aan toevallige fouten moet worden geweten, dan wel of er interpretatiefouten zijn gemaakt. In deze paragraaf willen wij veronderstellen dat

(307.1) $$\psi^2 = \sigma^2.$$

Hoe kan nu de middelbare fout van de verschillen der rasgemiddelden uit σ^2 worden afgeleid?

Met de methode der kleinste kwadraten is dit niet moeilijk. Wij gaan eerst na de verschillen tussen R_l en $R_{[l]}$. Aangezien wij in (305.2) hebben gesteld $R_l \equiv 0$, zijn de berekende waarden $\overline{R}_{[l]}$ tegelijk de waarden van het verschil $\overline{R}_{[l]} - \overline{R}_l$. Wanneer wij als voorbeeld van $[l]$ ras m nemen is

$$\overline{R}_m = \overline{R}_m - \overline{R}_l.$$

De fout van het verschil is de fout van $\overline{R}_m$. Nu geldt volgens Van Uven

(307.2) $$\sigma_{\overline{R}_m}{}^2 = r_{mm}\sigma^2.$$

Voor de andere rassen $[l]$ gelden soortgelijke formules.

De fout van een verschil $\overline{P} = \overline{R}_m - \overline{R}_n$ is volgens van Uven (zie hfdst. XII form. 35 van zijn leerboek)

(307.3) $$\sigma_{\bar{P}}^2 = (r_{mm} - 2r_{mn} + r_{nn})\,\sigma^2$$

Aan de hand van form. (307.2) en (307.3) zijn de waarden van de coëfficienten van σ voor alle verschillen uit te rekenen. De waarden zijn weergegeven in onderstaande tabel. Daar is in de eerste kolom vermeld op welke rassen de fout van het verschil betrekking heeft. In kolom 2 is de coefficient, als in (307.2) en (307.3) bedoeld, aangegeven in gewone breuken. De noemer is steeds 6.901.234.530; de teller is voor ieder geval afzonderlijk vermeld. In kolom 3 is de coëfficient uitgedeeld.

Verschil tussen de rassen	Coëfficient × 6.901.234.530	Coëfficient uitgedeeld	Verschil tussen de rassen	Coëfficient × 6.901.234.530	Coëfficient uitgedeeld
l, m	5.029.116.186	0.7287	*p, q*	3.676.024.800	0.5327
l, n	1.608.301.236	0.2330	*p, r*	2.862.171.919	0.4147
l, p	1.968.410.874	0.2852	*p, t*	9.967.530.496	1.4443
l, q	2.703.615.834	0.3918	*p, u*	2.639.988.856	0.3825
l, r	2.245.044.679	0.3253	*p, w*	2.663.588.536	0.3860
l, t	9.742.890.976	1.4118			
l, u	1.963.276.936	0.2845	*q, r*	3.524.968.699	0.5108
l, w	1.959.668.536	0.2840	*q, t*	11.611.546.576	1.6825
			q, u	3.153.823.936	0.4570
m, n	5.701.358.250	0.8261	*q, w*	3.109.403.416	0.4506
m, p	5.274.261.786	0.7642			
m, q	6.867.012.666	0.9950	*r, s*	2.760.493.812	0.4
m, r	5.992.172.143	0.8683	*r, t*	11.069.237.293	1.6040
m, t	13.885.565.680	2.0120	*r, u*	2.672.766.753	0.3873
m, u	5.795.337.400	0.8398	*r, w*	2.789.685.393	0.4042
m, w	5.823.506.920	0.8438			
			t, u	10.800.253.360	1.5650
n, p	2.282.219.796	0.3307	*t, w*	10.803.066.160	1.5654
n, q	2.803.038.276	0.4062			
n, r	2.448.534.463	0.3548	*u, w*	2.457.677.880	0.3561
n, t	9.847.493.950	1.4269			
n, u	2.145.533.470	0.3109			
n, w	2.133.980.590	0.3092			

Aangezien ras *r* en ras *s* steeds op dezelfde velden voorkomen zijn de fouten van de verschillen met ras *s* gelijk aan die van de verschillen met ras *r*. In bovenstaande tabel is ras *s* daarom slechts genoemd voor het verschil tussen *r* en *s*.

Het is nu de vraag, hoe de coëfficienten die in deze tabel zijn vermeld kunnen worden verkregen, wanneer de uitkomsten volgens de standaardisatiemethode zijn uitgerekend. Empirisch is een methode bruikbaar gebleken, die wij voor de eenvoudigheid willen uitleggen aan de hand van een schema als in (304.7). Het aantal rassen willen wij voor het gemak tot vier terugbrengen. Wij veronderstellen dus dat er vier proefvelden waren $p_1 \ldots . p_4$. Op veld p_1 ontbrak ras k_4, op veld p_3 ontbrak ras k_3. Na een paar keer standaardiseren bleek het laatste gemiddelde (D te noemen) gelijk te zijn aan het voorlaatste gem. (C), zodat gem. C dus reeds de juiste uitkomsten aangaf.

Voor onze empirische foutbepaling *nemen* wij nu *aan* dat de cijfers van gem. C geheel foutloos zijn. *Deze aanname is natuurlijk onjuist,* maar voert tot goede resultaten. De gegevens die op dit gemiddelde C gestandaardiseerd worden, hebben allen een toevallige fout τ_{kp}. (Krachtens (307.1) hoeven wij niet te rekenen met een systematische fout). Wegens deze toevallige fouten zal ieder proefveld foutief op gem. C worden gestandaardiseerd. Wij stellen de standaardisatiefout van proefveld p voor door T_p.

De gestandaardiseerde waarde, die wij hadden moeten krijgen als alles foutloos was, noemen we $\hat{\Omega}_k$. Als werkelijke waarde voor de gestandaardiseerde opbrengsten (Ω^*_{kp}) vinden we dus

(307.4) $$\Omega^*_{kp} = \hat{\Omega}_k + T_p + \tau_{kp}.$$

De fouten T en τ zijn met elkaar gecorreleerd. *Empirisch is het mogelijk gebleken ze als onafhankelijk te beschouwen.* Naar analogie van (304.7) kunnen wij ons schema nu als volgt in formules brengen:

(307.5)

ras / veld	k_1	k_2	k_3	k_4
p_1	$\hat{\Omega}_1+T_1+\tau_{11}$	$\hat{\Omega}_2+T_1+\tau_{21}$	$\hat{\Omega}_3+T_1+\tau_{31}$	
p_2	$\hat{\Omega}_1+T_2+\tau_{12}$	$\hat{\Omega}_2+T_2+\tau_{22}$	$\Omega_3+T_2+\tau_{32}$	$\Omega_4+T_2+\tau_{42}$
p_3	$\hat{\Omega}_1+T_3+\tau_{13}$	$\hat{\Omega}_2+T_3+\tau_{23}$		$\hat{\Omega}_4+T_3+\tau_{43}$
p_4	$\hat{\Omega}_1+T_4+\tau_{14}$	$\hat{\Omega}_2+T_4+\tau_{24}$	$\hat{\Omega}_3+T_4+\tau_{34}$	$\hat{\Omega}_4+T_4+\tau_{44}$
gem. D	$\hat{\Omega}_1+\frac{T_1+T_2+T_3+T_4}{4}+\frac{\tau_{11}+\tau_{12}+\tau_{13}+\tau_{14}}{4}$	$\hat{\Omega}_2+\frac{T_1+T_2+T_3+T_4}{4}+\frac{\tau_{21}+\tau_{22}+\tau_{23}+\tau_{24}}{4}$	$\hat{\Omega}_3+\frac{T_1+T_2+T_4}{3}+\frac{\tau_{31}+\tau_{32}+\tau_{34}}{3}$	$\hat{\Omega}_4+\frac{T_2+T_3+T_4}{3}+\frac{\tau_{42}+\tau_{43}+\tau_{44}}{3}$

Het blijkt dat in gem. D allerlei fouten worden genoemd, die verondersteld worden niet in gem. C aanwezig te zijn, terwijl toch gem. C = gem. D. Ook in deze tegenstrijdigheid komt het benaderingskarakter nogmaals tot uiting.

Het verschil tussen de gemiddelde opbrengst van k_1 en k_2 zoals weergegeven in gem. D (tabel 307.5) is

$$\bar{\Omega}_1 - \bar{\Omega}_2 = (\hat{\Omega}_1 - \hat{\Omega}_2) + \\ + \left(\frac{T_1 + T_2 + T_3 + T_4}{4} - \frac{T_1 + T_2 + T_3 + T_4}{4}\right) + \\ + \left(\frac{\tau_{11} + \tau_{12} + \tau_{13} + \tau_{14}}{4} - \frac{\tau_{21} + \tau_{22} + \tau_{23} + \tau_{24}}{4}\right).$$

Van deze vorm is $(\hat{\Omega}_1 - \hat{\Omega}_2)$ het gezochte verschil. De termen met T vallen tegen elkaar weg. De fout van het verschil wordt dus uitsluitend veroorzaakt door de toevallige fouten τ. Volgens bekende wetten is de toevalsvarians van het verschil $2 \times \frac{\sigma^2}{4}$.

Wij kunnen dus dit zeggen: *Het berekend opbrengstverschil tussen twee rassen, die op dezelfde proefvelden voorkomen, wordt niet beïnvloed door de standaardisatiefouten. Wanneer de beide rassen op n_k velden voorkomen is de toevalsvarians van het verschil* $\frac{2\sigma^2}{n_k}$.

Het verschil tussen de gemiddelde opbrengst van k_1 en k_3, zoals weergegeven in gem. D (tabel 307.5) is

$$(307.6) \quad \bar{\Omega}_1 - \bar{\Omega}_3 = (\hat{\Omega}_1 - \hat{\Omega}_3) + \\ + \left(\frac{T_1 + T_2 + T_3 + T_4}{4} - \frac{T_1 + T_2 + T_4}{3}\right) + \\ + \left(\frac{\tau_{11} + \tau_{12} + \tau_{13} + \tau_{14}}{4} - \frac{\tau_{31} + \tau_{32} + \tau_{34}}{3}\right).$$

Van deze vorm vallen de termen met T niet geheel tegen elkaar weg. Men kan deze termen samenvatten volgens (307.7)

$$(307.7) \quad \frac{T_1 + T_2 + T_3 + T_4}{4} - \frac{T_1 + T_2 + T_4}{3} = \\ = (\tfrac{1}{4} - \tfrac{1}{3})(T_1 + T_2 + T_4) + \tfrac{1}{4}T_3.$$

Wij willen nu de asymptotische waarde berekenen van het kwadraat van de totale fout van $(\hat{\Omega}_1 - \hat{\Omega}_3)$. Daartoe maken wij

gebruik van onze (onjuiste) aanname dat de standaardisatiefouten onafhankelijk zijn van de toevallige fouten en van elkaar. Op grond van deze aanname kunnen wij dus veronderstellen dat *alle* dubbele producten asymptotisch nul worden. We vinden dan voor $\sigma^2_{\bar{\Omega}_1-\bar{\Omega}_3}$ (zie ook 307.6 en 307.7)

$$(307.8)\quad \sigma^2_{\bar{\Omega}_1-\bar{\Omega}_3} = (\tfrac{1}{4}-\tfrac{1}{3})^2(<T_1^2> + <T_2^2> + <T_4^2>) + \\ + (\tfrac{1}{4})^2 <T_3^2> + \frac{\sigma^2}{4} + \frac{\sigma^2}{3}.$$

Om de waarden $<T^2>$ in σ^2 uit te drukken nemen wij aan dat de waarde $<T_p^2>$ van een proefveld p met m_p rassen gelijk is aan $\frac{\sigma^2}{m_p}$ dus

$$(307.9)\qquad <T_p^2> = \frac{\sigma^2}{m_p}.$$

In ons voorbeeld geldt dus

$$<T_1^2> = \frac{\sigma^2}{3}$$

$$<T_2^2> = \frac{\sigma^2}{4}$$

$$<T_3^2> = \frac{\sigma^2}{3}$$

$$<T_4^2> = \frac{\sigma^2}{4}.$$

Zodat (307.8) wordt

$$(307.10)\qquad \sigma^2_{\bar{\Omega}_1-\bar{\Omega}_2} = (\tfrac{1}{4}-\tfrac{1}{3})^2\,\sigma^2\,(\tfrac{1}{3}+\tfrac{1}{4}+\tfrac{1}{4}) + \\ + (\tfrac{1}{4})^2\frac{\sigma^2}{3} + \frac{\sigma^2}{4} + \frac{\sigma^2}{3} = 0.61\,\sigma^2.$$

Omdat k_1 en k_2 in het schema onderling verwisselbaar zijn en k_3 en k_4 ook (zowel k_3 als k_4 ontbreekt op een veld met $m_p = 3$) geeft (307.10) tevens de toevalsvarians van $\bar{\Omega}_2 - \bar{\Omega}_3$; $\bar{\Omega}_1 - \bar{\Omega}_4$ en $\bar{\Omega}_2 - \bar{\Omega}_4$.

Tenslotte het verschil tussen k_3 en k_4. Dit bedraagt

$$\bar{\Omega}_3 - \bar{\Omega}_4 = (\hat{\Omega}_3 - \hat{\Omega}_4) + \\ + \left(\frac{T_1+T_2+T_4}{3} - \frac{T_2+T_3+T_4}{3}\right) + \\ + \left(\frac{\tau_{31}+\tau_{32}+\tau_{34}}{3} - \frac{\tau_{42}+\tau_{43}+\tau_{44}}{3}\right).$$

Van de termen met T vallen T_2 en T_4 tegen elkaar weg; er blijft $\frac{T_1}{3} - \frac{T_3}{3}$.

Kwadrateren van alle fouten geeft weer

$$\sigma^2_{\overline{\Omega}_3 - \overline{\Omega}_4} = \tfrac{1}{9}\left(<T_1^2> + <T_2^2>\right) + \frac{\sigma^2}{3} + \frac{\sigma^2}{3}.$$

Daar $<T_1^2>$ en $<T_3^2>$ beide gelijk aan $\frac{\sigma^2}{3}$ kunnen worden genomen, staat hier

$$\sigma^2_{\overline{\Omega}_3 - \overline{\Omega}_4} = \tfrac{2}{27}\,\sigma^2 + \tfrac{2}{3}\,\sigma^2 = 0.741\,\sigma^2.$$

Samenvattend kunnen wij de volgende regels opstellen voor het berekenen van de toevalsvarians:

(307.11)

1e Wanneer men de fout van het opbrengstverschil tussen twee rassen k en l bestudeert, kan men met vrucht onderscheid maken tussen de zuiver toevallige fouten τ en de standaardisatiefouten T. Deze fouten mogen als onafhankelijk worden beschouwd.

2e De asymptotische waarde $<\tau^2> = \sigma^2$.

3e Voor een proefveld p met m_p rassen is de asymptotische waarde

$$<T_p^2> = \frac{\sigma^2}{m_p}.$$

Voor het vaststellen van m_p tellen niet mee de rassen die slechts op proefveld p voorkwamen en dus geen invloed op de standaardisatie hadden.

4e Wanneer ras k op n_k velden voorkomt en ras l op n_l velden, is de waarde van de toevalsvarians van het verschil, voor zover voortgekomen uit de toevallige fouten τ gelijk aan

$$\left(\frac{1}{n_k} + \frac{1}{n_l}\right)\sigma^2.$$

5e Om het effect van de standaardisatiefouten na te gaan moet men aannemen dat er a velden zijn waarop k en l beide voorkwamen; op b_k velden stond wel ras k en niet ras l; op b_l velden stond wel ras l en niet ras k.
Verder is

$$a + b_k = n_k$$
$$a + b_l = n_l.$$

6e De varians veroorzaakt door de a velden samen is

$$\left(\frac{1}{n_k}-\frac{1}{n_l}\right)^2 \uparrow a\,[<T^2>] = \left(\frac{n_k-n_l}{n_k n_l}\right)^2 \uparrow a\,[<T^2>].$$

Onder $\uparrow a\,[<T^2>]$ wordt verstaan de som van de a waarden $<T^2>$, die volgens (307.11 3e) berekend worden voor de a velden waarop k en l beide voorkomen.
Wanneer $n_k = n_l$ wordt deze vorm $=0$.

7e De varians veroorzaakt door de b_k velden samen is

$$\frac{1}{n_k^2} \uparrow b_k\,[<T^2>],$$

waarbij $\uparrow b_k\,[<T^2>]$ op soortgelijke wijze wordt gedefinieerd als $\uparrow a\,[<T^2>]$.

8e De varians veroorzaakt door de b_l velden samen is

$$\frac{1}{n_l^2} \uparrow b_l\,[<T^2>].$$

308 – Contrôle op de empirische methode

In tabel (304.8) hebben we een voorbeeld gegeven hoe een schema eruit zou kunnen zien, wanneer op verschillende proefvelden telkens gedeeltelijk andere rassen beproefd werden. We hebben toen verder nagegaan hoe we de gemiddelde opbrengst van de verschillende rassen konden bepalen, terwijl op pag. 112 is aangegeven hoe de toevalsvarians is van de diverse rasverschillen. In die tabel zijn nl. steeds de coëfficienten van σ^2 vermeld, die bij een bepaald rasverschil horen. Met behulp van de empirische methode die samengevat is in (307.11), zijn deze coëfficienten te schatten. Om na te gaan hoe nauwkeurig de schatting is zijn in tabel (308.1) naast elkaar vermeld de juiste coëfficienten en de geschatte coëfficienten volgens (307.11). In de 4e kolom zijn de geschatte waarden uitgedrukt in procenten van de juiste. Het blijkt dat dit percentage nooit boven 100 komt.

Het blijkt dus dat wij de coëfficienten, en bijgevolg ook de fouten van de verschillen, bijna steeds te laag schatten.

De grootste afwijking vinden wij in ons voorbeeld bij het verschil $\bar{\Omega}_m - \bar{\Omega}_q$, waar $\sigma^2_{\bar{\Omega}_m-\bar{\Omega}_q}$ 8% te laag is geschat. Globaal wordt $\sigma_{\bar{\Omega}_m-\bar{\Omega}_q}$ dus 4% te laag geschat. Om na te gaan of dit percentage te groot is moge men bedenken, dat een berekend verschil vaak als betrouwbaar wordt aangenomen, wanneer het in absolute

(308.1)

Verschil tussen de rassen	Juiste coëfficient volgens pag. 112	Geschatte coëfficient volgens (307.11)	% age van juist	Geschat volgens (307.11, 4e)	% age van juist	Volgens gepaarde gegevens	% age van juist
1	2	3	4	5	6	7	8
l, m	0.7287	0.6925	95	0.6	82	1	137
l, n	0.2330	0.2308	99	0.225	97	0.25	107
l, p	0.2852	0.2802	98	0.2667	94	0.3333	117
l, q	0.3918	0.3805	97	0.35	89	0.5	128
l, r	0.3253	0.3210	99	0.3	92	0.4	123
l, t	1.4118	1.3877	98	1.1	78	2	142
l, u	0.2845	0.2820	99	0.2667	94	0.3333	117
l, w	0.2840	0.2816	99	0.2667	94	0.3333	117
m, n	0.8261	0.7695	93	0.625	76	—	—
m, p	0.7642	0.7437	97	0.6667	87	1	131
m, q	0.9950	0.9165	92	0.75	75	—	—
m, r	0.8683	0.8219	95	0.7	81	2	230
m, t	2.0120	1.9324	96	1.5	75	—	—
m, u	0.8398	0.7902	94	0.6667	79	2	238
m, w	0.8438	0.7925	94	0.6667	79	2	237
n, p	0.3307	0.3183	96	0.2917	88	0.5	151
n, q	0.4062	0.4004	99	0.375	92	0.5	123
n, r	0.3548	0.3497	99	0.325	92	0.5	141
n, t	1.4269	1.4004	98	1.125	79	2	140
n, u	0.3109	0.3089	99	0.2917	94	0.4	129
n, w	0.3092	0.3077	100	0.2917	94	0.4	129
p, q	0.5327	0.5014	94	0.4167	78	—	—
p, r	0.4147	0.4037	97	0.3667	88	0.6667	161
p, t	1.4443	1.4262	99	1.1667	81	2	138
p, u	0.3825	0.3706	97	0.3333	87	0.6667	174
p, w	0.3860	0.3729	97	0.3333	86	0.6667	173
q, r	0.5108	0.4979	97	0.45	88	1	196
q, t	1.6825	1.6308	97	1.25	74	—	—
q, u	0.4570	0.4500	98	0.4167	91	0.6667	146
q, w	0.4506	0.4454	99	0.4167	92	0.6667	148
r, s	0.4	0.4	100	0.4	100	0.4	100
r, t	1.6040	1.5647	98	1.2	75	—	—

Vervolg (308.1)

r, u	0.3873	0.3851	99	0.3667	95	0.5	129
r, w	0.4042	0.3986	99	0.3667	91	0.6667	165
t, u	1.5650	1.5283	98	1.1667	75	—	—
t, w	1.5654	1.5306	98	1.1667	75	—	—
u, w	0.3561	0.3542	99	0.3333	94	0.5	140

waarde minstens 2× zijn middelbare fout bedraagt. Dit cijfer 2 is tamelijk willekeurig gekozen. Het zou evengoed 1.9 of 2.1 kunnen zijn. Nu bedraagt het verschil tussen 2.0 en 2.1 5%. De „willekeur" van de factor 2 weegt dus zwaarder dan onze grootste schattingsfout.

Gezien de bovenaangeduide onzekerheid in het *gebruik* van de σ kunnen wij wel zeggen, dat een schattingsfout van 10% voor σ vaak toelaatbaar moet worden geacht. Dit zou dus een fout van 20% voor σ^2 betekenen.

Aangezien onze empirische methode veel groter nauwkeurigheid toestaat, zou men kunnen overwegen de schatting nog sterk te vereenvoudigen door alleen rekening te houden met (307.11 4e). De aldus geschatte coëfficienten zijn weergegeven in kolom 5 van tabel (308.1). In de 6e kolom zijn ze in procenten van de juiste waarden uitgedrukt. Het blijkt dat langs deze weg coëfficienten worden gevonden, die over het algemeen minder dan 20% te klein zijn. Slechts in de gevallen dat een van de beide vergeleken rassen slechts op 1 of 2 velden voorkwam zijn de schattingen soms beduidend slechter.

Wij kunnen dus zeggen dat in het algemeen de coëfficient van σ^2 geschat kan worden volgens (307.11 4e). Blijkt bij het interpreteren van de uitkomsten, dat de verhouding

$$\frac{\text{verschil}}{\text{fout van het verschil}}$$

in de buurt van een belangrijk geachte waarde ligt (b.v. 5% point), dan kan met de volledige methode van (307.11) een betere schatting van de fout worden verkregen. Meestal zal die schatting nauwkeurig genoeg zijn. Desnoods kan men het gehele probleem oplossen met de methode der kleinste kwadraten, waarbij de *coëfficienten* van σ^2 juist bepaald worden. De waarde die voor σ^2 zelf berekend wordt, blijft echter ook dan een schatting. Ook dit laatste

is een reden, waarom aan de nauwkeurigheid van de coëfficienten geen al te hoge eisen moeten worden gesteld.

Het spreekt vanzelf, dat één uitgewerkt voorbeeld niet voldoende bewijs is voor de bruikbaarheid van een empirische methode. Toch zullen wij geen nieuwe voorbeelden uitwerken, omdat ieder die meer zekerheid wil, nieuwe voorbeelden kan maken, die meer gelijken op de problemen, die hij bewerkt.

In de practijk hoort men soms de bewering, dat twee rassen alleen mogen worden vergeleken, door voor beide rassen het gemiddelde te nemen van de (gestandaardiseerde) opbrengsten van die velden, waarop beide rassen verbouwd werden. In tabel (308.1) kolom 7 is vermeld, welke coëfficient σ^2 dan moet hebben. Rassen die op a velden gezamelijk voorkwamen hebben als coëfficient $2/a$. Rassen die op geen enkel veld gezamenlijk voorkwamen, kunnen dan natuurlijk niet worden vergeleken, zodat hier ook een foutenbepaling ontbreekt.

In kolom 8 zijn de cijfers van kolom 7 uitgedrukt in procenten van de waarden, die volgens kolom 2 bereikbaar zijn. Het blijkt dat dit percentage in de regel ver boven 100 ligt, zodat wij mogen concluderen, dat de wijze van samenvatten, zoals wij die in 304 uiteen zetten, voor *onafhankelijke rassen* beter is.

309 – Het bepalen van de correlatie tussen de rassen

In 302 hebben wij gevonden dat er bij onafhankelijke rassen een correlatie moet bestaan tussen de schijnbare fouten u_{kp}. *Wij willen er nog eens zeer nadrukkelijk op wijzen dat het dus gaat over correlaties van gestandaardiseerde opbrengstcijfers.* De correlaties van de niet-gestandaardiseerde opbrengsten zijn veel groter, maar blijven hier buiten beschouwing. Wanneer er interpretatiefouten zijn gemaakt kan er volgens (302.10) een „correlatiecoëfficient" $-1/m$ worden verwacht.

Wanneer op ieder proefveld p telkens een ander aantal rassen m_p voorkomt, zal die correlatiecoëfficient voor ieder veld p asymptotisch ongeveer $-\frac{1}{m_p}$ zijn. Dit is geen stabiele grootheid. Bij het toepassen van een formule die op (302.10) gelijkt valt daarom moeilijk te voorspellen, welke waarde precies verwacht moet worden. Het lijkt ons te ver te gaan deze waarde theoretisch uit te zoeken.

Een eenvoudiger theorie laat zich afleiden, wanneer men bedenkt dat de bovengenoemde correlatie veroorzaakt wordt, doordat de op correlatie onderzochte rassen k en l beide in het fictiefras zijn opgenomen. Daardoor is de som van de interpretatiefouten van een proefveld nul (zie 301.23 en 't begin van 302) en daardoor is ook de som van de schijnbare toevallige fouten voor ieder proefveld 0.

Wanneer men een van de rassen k en l of beide buiten de definitie van het fictiefras houdt, wordt het kunstmatig verband van (301.23) verbroken, zodat onafhankelijke rassen dan een correlatie 0 te zien moeten geven tussen hun interpretatiefouten en hun toevallige fouten.

Om deze correlatie empirisch te berekenen kan men *bij het op elkaar standaardiseren* van de proefvelden de gegevens van ras k of ras l, eventueel beide, buiten beschouwing laten. Na het standaardiseren berekent men dan voor ieder ras, dus ook voor en l, de afwijkingen van het gemiddelde. Van deze afwijkingen ordt de correlatie op de gebruikelijke manier onderzocht.

Men moet dan eerst beslissen of men één ras buiten het fictiefras zal houden, of beide. Wanneer er in totaal m rassen zijn, zijn er $\frac{m(m-1)}{2}$ combinaties (k, l) die onderzocht kunnen worden. Zouden telkens beide rassen uitgeschakeld worden, dan moest het standaardiseren $\frac{m(m-1)}{2}$ maal worden overgedaan. Wanneer één ras wordt uitgeschakeld, behoeft het standaardiseren slechts m maal te worden herhaald. Het is daarom beter slechts één ras uit te schakelen. Hierdoor verkrijgt men bovendien het voordeel, dat het fictiefras zo weinig mogelijk verandert.

Wij willen nu bij wijze van voorbeeld de correlatiecoëfficient van ras l en ras n (zie 304.8) langs twee wegen berekenen. Eerst sluiten we ras l buiten het fictiefras, daarna ras n.

Bij de berekening kunnen wij aansluiten bij het resultaat van (304.12). Het feit dat l buiten het fictiefras wordt gesloten zal wel niet zo ingrijpend zijn, dat de uitkomsten veel veranderen. Wij standaardiseren alle proefvelden dus op de reeds verkregen uitkomsten voor de rassen $\mathbb{U}$. Ras t doet natuurlijk weer niet mee, daar het slechts eenmaal voorkomt.

Wanneer wij in (304.12) de gestandaardiseerde cijfers van veld a met het 4e gemiddelde vergelijken, zien wij dat l daar 29 eenheden

boven het gemiddelde ligt. De andere rassen liggen er gezamenlijk dus 29 onder. Er zijn 2 rassen [l], deze liggen gemiddeld dus 15 te laag. De opbrengsten van *a* moeten dus met 15 worden verhoogd. Op dezelfde manier vinden wij dat de opbrengsten van *b* gelijk blijven. Op *c* ligt ras *l* 19 eenheden beneden het gemiddelde, de 5 rassen [l] dus gemiddeld 4 er boven. De cijfers van veld *c* moeten 4 worden verlaagd. Langs deze weg vinden wij zeer snel de cijfers van (309.1).

(309.1)

ras / veld	*l*	*m*	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*w*
4e gem.	665	705	653	643	705	661	686		697	652
a		705		645						
b		720		612		700	714		700	602
c			629	655		654	684		713	
d			686	666		615	659			671
e			664	633						
f			661	664					685	636
g			644		719	655	670		713	
h			587		721	679	699		709	658
i			688		680					643
k			660		693				654	700
gem.		713	652	646	703	661	685		696	652

Het blijkt dat de grootste verandering in het gemiddelde van ras *m* is opgetreden. Door deze verandering zal met name veld *a* op een hoger niveau moeten worden gestandaardiseerd, waardoor het gemiddelde van *m* weer stijgt. Dit uitbalanceren is aanmerkelijk te bekorten, door te schatten wat er uit moet komen, en op dat schattingscijfer te standaardiseren. Wij willen deze standaardisatie niet verder op de voet volgen, doch het eindresultaat geven in (309.2).

Uit de omlijste cijfers laat zich berekenen als correlatiecoëfficient voor de rassen *l* en *n* de waarde

(309.3) $$\gamma_{l/n} = -0.53$$

Bij het standaardiseren hadden wij ook ras *n* kunnen uitschakelen, in plaats van *l*, dan hadden wij de cijfers van (309.4) gevonden.

(309.2)

ras / veld	*l*	*m*	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*w*
a	715	711		651						
b	665	721		613		701	715		701	603
c	642		630	656		655	685		714	
d	665		686	666		615	659			671
e	687		665	634				676		
f	650		661	664					685	636
g	632		643		718	654	669		712	
h	713		586		720	678	698		708	657
i	627		686		678					641
k	662		658		691				652	698
gem.	666	716	652	647	702	661	685	676	695	651

(309.4)

ras / veld	*l*	*m*	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*w*
a	695	691		631						
b	663	719		611		699	713		699	601
c	641		629	655		654	684		713	
d	671		692	672		621	665			677
e	682		660	629				671		
f	657		668	671					692	643
g	637		648		723	659	674		717	
h	695		568		702	660	680		690	639
i	654		713		705					668
k	667		663		696				657	703
gem.	666	705	654	645	706	659	683	671	695	655

Uit de omlijste cijfers berekenen wij nu als waarde van de correlatiecoëfficient

(309.5) $$\gamma_{n/l} = -0.37$$

Dit is een geheel andere waarde dan in (309.3).

Om dit verschijnsel te verklaren denke men zich een aantal proefvelden *n*, met op ieder veld dezelfde *m* rassen. Men heeft dus een schema zonder hiaten. Nu laat zich volgens (302.10) een negatieve correlatie verwachten bij onafhankelijke rassen. Evenals bij onvolledige schema's laat zich ook deze correlatiecoëfficient nul maken, door bij het standaardiseren een of beide van de ge-

correleerde rassen buiten het fictiefras te houden. Veronderstel dat beide rassen (k en l) buiten het fictiefras worden gehouden. Dan zal er maar één correlatiecoëfficient voor het paar (k, l) nodig zijn.

Om niet te abstract te spreken stelle men zich voor, dat de oorzaak van vruchtbaarheidsverschillen in de grondwaterstand is gelegen. Wanneer de grondwaterspiegel 1 cm daalt, zal de opbrengst van het fictiefras (veronderstel: het gemiddelde van 4 rassen $\lfloor kl \rfloor$) 5 eenheden stijgen. De opbrengst van ras k stijgt 20 eenheden, die van ras l 6 eenheden. Wanneer er nu geen toevallige fouten zijn gemaakt en er geen andere oorzaken van vruchtbaarheidsverschillen zijn, zal er een positieve correlatie 1 optreden. Immers wanneer de grondwaterstand a cm beneden zijn gemiddelde ligt zal ras k $15a$ eenheden te hoog zijn en ras l a eenheden. De afwijking van ras k is steeds $15\times$ die van ras l.

Wanneer de opbrengst van ras l per cm lagere grondwaterstand niet 6 doch 5 eenheden steeg, zou ras l in zijn reactie steeds gelijk zijn aan het fictiefras, zodat er geen afwijking voor l bestond; dan zou de correlatie met k dus 0 zijn. Wanneer de stijging van ras l kleiner was dan 5 zou de afwijking tussen ras l zijn gestandaardiseerd gemiddelde negatief zijn bij dalende grondwaterstand; de correlatie met ras k zou dan -1 zijn. Dat slechts de waarden $+1$, 0 en -1 voorkomen, vindt zijn oorzaak in de aanname dat er geen andere vruchtbaarheidsinvloeden en geen toevallige fouten zijn.

Wij willen nu nagaan welke correlaties op zouden treden, wanneer ras k of ras l wel in het fictiefras waren opgenomen.
Wanneer de opbrengst van ras l 6 eenheden stijgt per cm lagere grondwaterstand en het gemiddelde van 4 rassen $\lfloor kl \rfloor$ 5, dan zal het gemiddelde van de 5 rassen $\lfloor k \rfloor$ $\frac{4 \times 5 + 6}{4 + 1} = 5^1/_5$ stijgen. Dit is ons nieuwe fictiefras. Zowel ras k als ras l stijgen meer, dus de correlatie blijft $+1$.

Zou ras k in het fictiefras worden opgenomen, dan steeg dit nieuwe fictiefras $\frac{4 \times 5 + 20}{4 + 1} = 8$ eenheden per cm lagere grondwaterstand. Ras k zou dan meer stijgen en ras l minder; dan zou de correlatie -1 zijn.

Uit deze discussie blijkt, dat de berekende correlatiecoëfficienten de samenhang tussen de rassen steeds beschouwen tegen de achtergrond

van het fictiefras. Verandering van fictiefras geeft dus verandering van samenhang. Hiermee is het gevonden verschil tussen $\gamma_{l/n}$ en $\gamma_{n/l}$ (zie 309.3 en 309.5) voldoende verklaard.

310 – ANALYSE VAN DE CORRELATIEVERSCHIJNSELEN

In de voorgaande paragraaf hebben wij de correlatie tussen de schijnbare fouten u van de rassen l en n onderzocht. Nu bestaan de schijnbare fouten uit een toevallig element en een systematisch element. Het toevallig deel kan niet de oorzaak zijn van een belangrijke correlatie; immers een correlatie tussen toevallige fouten kan ook zelf slechts toevallig zijn. Doordat slechts een van beide rassen in het fictiefras was opgenomen is gezorgd dat er niet kunstmatig een correlatie tussen deze toevallige fouten te voorschijn is geroepen.

Het belang van de correlatie ligt in het systematisch bestanddeel. Men moet weten hoe de samenhang is tussen de interpretatiefouten van de verschillende rassen. A priori kunnen hierbij drie gevallen worden onderscheiden.

In de eerste plaats is er het geval dat de correlatiecoëfficient van de interpretatiefouten van twee rassen $+1$ is, doordat de rassen voor de onderzochte eigenschappen gelijk zijn of slechts een constante verschillen. Dit zou bijv. het geval kunnen zijn, wanneer het ene ras een mutant is van het andere, of er uit geselecteerd is. In zo'n geval zal de enige oorzaak voor variaties van het gevonden verschil tussen beide rassen in de toevallige fouten moeten zijn gelegen. Of dit geval zich voordoet laat zich het gemakkelijkst bepalen door op alle proefvelden het rasverschil rechtstreeks te bepalen, en uit de verkregen cijfers de middelbare fout van deze grootheid te berekenen. Deze m.f. moet $\sigma\sqrt{2}$ zijn. Is de fout betrouwbaar groter, dan heeft men blijkbaar niet met een geval als boven aangeduid te maken.

In de tweede plaats moet men letten op het geval, dat van twee rassen bekend is of stellig verwacht wordt, dat ze gecorreleerd zijn, maar dat de mate van correlatie onbekend is. In de vorige paragraaf is een methode gegeven om in dat geval de correlatiecoëfficient te berekenen.

Wij bespraken reeds dat de daar gegeven grootheid zijn ontstaan te danken heeft aan de samenwerking van interpretatie- en toevallige fouten. Het is in zulke omstandigheden ook gewenst de correlatie tussen de interpretatiefouten alleen te meten. Om

na te gaan hoe men deze kan bepalen, willen wij eerst de berekening bij een volledig schema (op alle velden dezelfde rassen) bestuderen.

Wij stellen dus, dat er n velden zijn met op ieder m rassen, o.a. de rassen k en l. Ras k is niet in het fictiefras opgenomen, ras l wel. Het fictiefras is dus het gemiddelde van $(m-1)$ rassen. Wij willen dit aantal aanduiden door m'.

Uit (301.27) laat zich nu afleiden (wanneer men j^2 weer vervangt door $<i^2_{l\pi}>$)

$$\uparrow n\,[<u^2_{l\pi}>] = \frac{n-1}{n}\,\uparrow n\,[<i^2_{l\pi}>] + (n-1)\,\frac{m'-1}{m'}\,\sigma^2$$

of

$$(310.1)\qquad \frac{n-1}{n}\,\uparrow n\,[<i^2_{l\pi}>] = \uparrow n\,[<u^2_{l\pi}>] - (n-1)\left(1-\frac{1}{m'}\right)\sigma^2.$$

Voor ras k laat zich naar analogie van (307.4) schrijven

$$(310.2)\qquad \Omega^*_{kp} = \hat{\Omega}_k + i_{kp} + \tau_{kp} + T_p = \bar{\Omega}_k + u_{kp}$$

Middelen van deze vorm over p geeft

$$(310.3)\qquad \frac{\uparrow n\,[\Omega^*_{k\pi}]}{n} = \hat{\Omega}_k + \frac{\uparrow n\,[i_{k\pi}]}{n} + \frac{\uparrow n\,[\tau_{k\pi}]}{n} + \frac{\uparrow n\,[T_\pi]}{n} = \bar{\Omega}_k.$$

Substitueren van $\bar{\Omega}_k$ uit (310.3) in (310.2) geeft

$$\hat{\Omega}_k + i_{kp} + \tau_{kp} + T_p = \hat{\Omega}_k + \\ + \frac{\uparrow n\,[i_{k\pi}]}{n} + \frac{\uparrow n\,[\tau_{k\pi}]}{n} + \frac{\uparrow n\,[T_\pi]}{n} + u_{kp}$$

of

$$(310.4)\qquad u_{kp} = \frac{n-1}{n}\,i_{kp} + \frac{n-1}{n}\,\tau_{kp} + \frac{n-1}{n}\,T_p - \\ - \frac{\uparrow n-1\,[i_{k\lfloor p\rfloor}]}{n} - \frac{\uparrow n-1\,[\tau_{k\lfloor p\rfloor}]}{n} - \frac{\uparrow (n-1)\,[T_{\lfloor p\rfloor}]}{n},$$

Aangezien ras k niet in het fictiefras was opgenomen, zijn de fouten i, τ en T onderling onafhankelijk. Wanneer men aanneemt dat ook de verschillende waarden T onderling onafhankelijk zijn en bovendien volgens (307.11 3e) de waarde van $<T_p^2>$ weer $\frac{\sigma^2}{m'}$ stelt, berekent men gemakkelijk uit (310.4)

$$\uparrow n\,[<u^2_{k\pi}>] = \left(\frac{n-1}{n}\right)^2 \uparrow n\,[<i^2_{k\pi}>] + \frac{(n-1)^2}{n}\,\sigma^2 + \\ + \frac{(n-1)^2}{n}\,\frac{\sigma^2}{m'} + \frac{n-1}{n^2}\,\uparrow n\,[<i^2_{k\pi}>] + \frac{n-1}{n}\,\sigma^2 + \frac{n-1}{n}\,\frac{\sigma^2}{m'}$$

of

$$\uparrow n\,[<u^2_{k\pi}>] = \frac{n-1}{n}\uparrow n\,[<i^2_{k\pi}>] + (n-1)\left(\frac{m'+1}{m'}\right)\sigma^2$$

of

$$(310.5)\quad \frac{n-1}{n}\uparrow n\,[<i^2_{k\pi}>] = \uparrow n\,[<u^2_{k\pi}>] - (n-1)\left(1+\frac{1}{m'}\right)\sigma^2.$$

Om de asymptotische waarde $<u_{kp}\,u_{lp}>$ uit te rekenen moet men de u_{lp} op dezelfde manier uitwerken als de u_{kp} in (310.4). Aangezien bij de afleiding van u_{kp} niet tot uiting is gekomen, dat ras k niet in het fictiefras was opgenomen, zal de formule voor u_{lp} gelijk zijn aan (310.4) mits de k door een l wordt vervangen.

In verband met (301.23) is gemakkelijk in te zien, dat de T nu ook nog onafhankelijk is van de interpretatiefout i. Immers, wanneer krachtens definitie de som van de interpretatiefouten van de rassen, die in het fictiefras zijn opgenomen, voor ieder proefveld 0 is, kan hier geen oorzaak van foutieve standaardisatie liggen.

In tegenstelling hiermee zijn τ_{lp} en T_p niet onafhankelijk. Wanneer weer dezelfde veronderstelling wordt gemaakt, die aan (307.9) ten grondslag ligt, kunnen wij schrijven

$$(310.6)\qquad T_p = -\frac{\tau_{lp}}{m'} - \frac{\uparrow(m'-1)\,[\tau_{\lfloor kl\rfloor\,p}]}{m'},$$

waarbij wij door $\lfloor kl\rfloor$ aanduiden dat zowel ras k als ras l uit de som zijn uitgesloten.

Het minteken is nodig om aan te geven dat de standaardisatiefout de tendens heeft de toevallige fouten weg te verklaren. (Met „weg verklaren" wordt bedoeld, dat de berekende waarden zodanig van de ware waarden afwijken, dat de schijnbare fouten middelbaar kleiner zijn dan de ware fouten. Door de berekende waarden, die de waarnemingen moeten verklaren, verdwijnt een deel der fouten dus ogenschijnlijk).

Substitutie van T_p uit (310.6) in (310.4) geeft

$$(310.7)\quad u_{kp} = \frac{n-1}{n}\,i_{kp} + \frac{n-1}{n}\,\tau_{kp} - \frac{n-1}{n}\cdot\frac{\tau_{lp}}{m'} -$$
$$-\frac{n-1}{n}\cdot\frac{\uparrow(m'-1)\,[\tau_{\lfloor kl\rfloor p}]}{m'} - \frac{\uparrow(n-1)\,[i_{k\lfloor p\rfloor}]}{n} - \frac{\uparrow(n-1)\,[\tau_{k\lfloor p\rfloor}]}{n} +$$
$$+\frac{\uparrow(n-1)\,[\tau_{l\lfloor p\rfloor}]}{m'n} + \frac{\uparrow\uparrow(m'-1)(n-1)\,[[\tau_{\lfloor kl\rfloor\lfloor p\rfloor}]]}{m'n}.$$

De formule voor u_{lp} is hieruit af te leiden door de index k door

l te vervangen. De l mag niet door k worden vervangen, omdat l in het fictiefras blijft. Wij vinden dan

$$(310.8)\quad u_{lp} = \frac{n-1}{n}\, i_{lp} + \frac{n-1}{n} \cdot \frac{m'-1}{m'} \cdot \tau_{lp} -$$
$$- \frac{n-1}{n} \cdot \frac{\uparrow(m'-1)\,[\tau_{\lfloor kl \rfloor p}]}{m'} - \frac{\uparrow(n-1)\,[i_{l\lfloor p \rfloor}]}{n} -$$
$$- \frac{m'-1}{m'} \cdot \frac{\uparrow(n-1)\,[\tau_{l\lfloor p \rfloor}]}{n} + \frac{\uparrow\uparrow(m'-1)(n-1)\,[[\tau_{\lfloor kl \rfloor \lfloor p \rfloor}]]}{m'n}.$$

Bij het berekenen van de asymptotische waarde $<u_{kp}\, u_{lp}>$ vallen alle dubbele producten weg behalve die met i. Dit geeft dus

(310.9)

$$<u_{kp}\, u_{lp}> = \frac{(n-1)^2}{n^2} <i_{kp}\, i_{lp}> - \frac{(n-1)^2}{n^2} \cdot \frac{m'-1}{m'^2} <\tau^2{}_{lp}> +$$
$$+ \frac{(n-1)^2}{n^2} \frac{\uparrow m'-1\,[<\tau^2{}_{\lfloor kl \rfloor p}>]}{m'^2} + \frac{\uparrow n-1\,[<i_{k\lfloor p \rfloor}\, i_{l\lfloor p \rfloor}>]}{n^2} -$$
$$- \frac{m'-1}{m^{12}} \cdot \frac{\uparrow n-1\,[<\tau^2{}_{l\lfloor p \rfloor}>]}{n^2} + \frac{\uparrow\uparrow(m'-1)(n-1)[[<\tau_{\lfloor kl \rfloor \lfloor p \rfloor}>]]}{m'^2 n^2}.$$

Wanneer wij de asymptotische waarden $<i_{kp}\, i_{lp}>$ en $<i_{k\lfloor p \rfloor}\, i_{l\lfloor p \rfloor}>$ voorstellen door $<i_{k\pi}\, i_{l\pi}>$ en de asymptotische waarden $<\tau^2>$ door σ^2, wordt dit

$$<u_{kp}\, u_{lp}> = \frac{n-1}{n} <i_{k\pi}\, i_{l\pi}>$$

of

$$(310.10)\qquad \frac{n-1}{n} \uparrow n\,[<i_{k\pi}\, i_{l\pi}>] = \uparrow n\,[<u_{k\pi}\, u_{l\pi}>].$$

Uit (310.1), (310.5) en (310.10) vinden wij als bruikbare schatting voor de correlatiecoëfficient $\Gamma_{k/l}$ van de correlatie tussen de interpretatiefouten van ras k en ras l (waarbij l in het fictiefras is opgenomen):

(310.11)

$$\Gamma_{k/l} = \frac{\uparrow n\,[u_{k\pi}\, u_{l\pi}]}{\sqrt{\left\{\uparrow n[u^2{}_{k\pi}] - (n-1)\left(1 + \frac{1}{m'}\right)\sigma^2\right\}\left\{\uparrow n[u^2{}_{l\pi}] - (n-1)\left(1 - \frac{1}{m'}\right)\sigma^2\right\}}},$$

waarbij σ^2 weer verondersteld wordt uit het verschil tussen de blokken binnen de proefvelden te zijn berekend (zie verder opmerking naar aanleiding van 311.19).

Bij onvolledige schema's, zoals besproken in 304, zal m' geen constante grootheid zijn, maar voor ieder proefveld anders kunnen uitvallen. Dan staat trouwens de gehele formule (310.11) min of meer op losse schroeven, omdat wij in 309 aantoonden dat de berekende correlatiecoëfficient beide rassen beschouwt tegen de achtergrond van het fictiefras. Bij onvolledige schema's is het fictiefras niet constant en de correlatiecoëfficient dus min of meer onbepaald. Toch kan men (310.11) gebruiken om een globale indruk te krijgen. Om complicaties te voorkomen kan men dan het best een waarde voor $\frac{1}{m'}$ berekenen als gemiddelde van de waarden $\frac{1}{m'_p}$, die voor de onderzochte correlatiecoëfficient van belang zijn.

Tenslotte moet nog worden gesproken over het geval, dat er geen correlatie verwacht wordt, doch dat de berekening een bepaalde waarde voor de correlatiecoëfficient oplevert. Zulk een correlatie kan toevallig optreden tussen de toevallige fouten, maar evenzeer tussen de interpretatiefouten. Om te beoordelen of de correlatie als een toevallig verschijnsel mag worden opgevat zal men daarom geen onderscheid behoeven te maken tussen beide soorten fouten. Blijkt de correlatie niet aan toeval te kunnen worden geweten dan moet er een systematische correlatie tussen de interpretatiefouten optreden, aangezien tussen toevallige fouten geen systematische correlatie op kan treden. Zo nodig kan de correlatie tussen de interpretatiefouten weer volgens (310.11) worden vastgelegd.

De vraag is nu: wanneer moet een onverwachte correlatie als reëel worden erkend. Daartoe moet worden nagegaan of de gevonden correlatie betrouwbaar van 0 afwijkt. Dit gaat het gemakkelijkst met behulp van het door R. A. FISHER gevonden resultaat dat de middelbare fout van de grootheid $t = \frac{\gamma}{\sqrt{1-\gamma^2}}$ de waarde $\sqrt{\frac{1}{n-4}}$ heeft, indien er in werkelijkheid geen correlatie is. Door γ is de correlatiecoëfficient aangeduid. Wanneer men als eis van betrouwbaarheid stelt, dat t tweemaal zijn middelbare fout moet zijn, vindt men dus:

$$\left|\frac{\gamma}{\sqrt{1-\gamma^2}}\right| > \frac{2}{\sqrt{n-4}} \quad \text{of}$$

$$\frac{\gamma^2}{1-\gamma^2} > \frac{4}{n-4} \quad \text{of}$$

$$n\gamma^2 - 4\gamma^2 > 4 - 4\gamma^2 \quad \text{of}$$

$$\gamma^2 < \frac{4}{n}. \tag{310.12}$$

Uit (310.12) blijkt wel dat een onverwachte correlatie zeer hoog moet zijn, om bij een gering aantal waarnemingen betrouwbaar te zijn. Formule (310.12) geeft ongeveer het „5% point", het „1% point" wordt bij benadering gegeven door

$$\frac{\gamma^2}{1-\gamma^2} > \frac{6.5}{n-4} \quad \text{of}$$

$$\gamma^2 > \frac{6.5}{n+2.5}. \tag{310.13}$$

Door n wordt in (310.12) en (310.13) aangeduid het aantal proefvelden waarop ras k en ras l beide voorkomen. Het blijkt dat dit aantal minstens 5 moet zijn om een toevallig gevonden correlatie als belangrijk te kunnen aanvaarden. Verder moet er rekening mee worden gehouden dat het standaardiseren misschien ook nog een ongedachte invloed heeft op de berekende correlatiecoëfficient. Enige voorzichtigheid bij het aanvaarden van een correlatie als systematisch is dus wel gewenst.

In het voorgaande hebben wij steeds een positieve correlatie voor ogen gehad. Als materiële grondslag daarvoor kan een extra aantal gemeenschappelijke genen worden aangenomen. Het is natuurlijk niet ondenkbaar dat er ook negatieve correlaties worden waargenomen. In (303.2) toonden wij reeds aan dat een positieve correlatie tussen twee rassen hun correlatie met de andere rassen verlaagt. Het wil ons voorkomen dat hierin vrijwel de enige oorzaak van negatieve correlatie gelegen kan zijn. Meestal zal met een negatieve correlatie daarom voldoende zijn gerekend wanneer met alle positieve op een passende manier rekening is gehouden. Op de enkele uitzonderingen zullen wij hier niet ingaan.

311 – Het samenvatten van proeven met wisselend aantal gedeeltelijk afhankelijke rassen

In 304 is nagegaan hoe proeven met een wisselend aantal onafhankelijke rassen moeten worden samengevat. Als werkvoorbeeld

hebben wij daar de cijfers van (304.3) genomen, die blijkens (304.7) na standaardisatie het volgend resultaat geven.

(311.1)

ras / veld	k_1	k_2	k_3	k_4	k_5	k_6
p_1	646	655	802	636	596	
p_2	681	663	690	663	643	643
p_3	716	762	751	461		665
p_4	685	653	685	675	642	642
gem. D	682	683	732	609	627	650

Uit de cijfers van gem. D blijkt dat ras k_6 veel beter is dan ras k_5. Deze conclusie is in tegenspraak met de waarnemingen op p_2 en p_4 dat beide rassen precies gelijk zijn. Wat is nu de juiste conclusie?

Wanneer inderdaad uit p_2 en p_4 mocht worden geconcludeerd, dat k_5 en k_6 identiek waren, zou hier sprake zijn van sterk gecorreleerde rassen, zodat de vooronderstellingen van 304 dan niet van toepassing zouden zijn. Wanneer daarentegen ras k_5 en ras k_6 in werkelijkheid onafhankelijk zijn, moet het als een speling van het toeval worden gezien, dat beide rassen op beide proefvelden gelijke opbrengsten gaven. Uit de cijfers kan geen keuze tussen beide zienswijzen worden gedaan, omdat blijkens (310.12) het aantal proefvelden te gering is om een onverwachte correlatiecoëfficient + 1 als zeker te aanvaarden. Wel leent bovenstaand voorbeeld zich er toe om het wezen van beide veronderstellingen duidelijk te maken.

Indien men wil volhouden dat de rassen onafhankelijk zijn, zijn de cijfers van gem. D het geschiktst om het verschil tussen ras k_5 en ras k_6 weer te geven (zie eind van 308). Weliswaar hebben beide rassen het evengoed gedaan op p_2 en p_4, doch van ras k_5 is op veld p_1 aangetoond, dat het ook een lagere opbrengst kan geven in vergelijking met het fictiefras. Wegens de onafhankelijkheid mag hieruit niet geconcludeerd worden, dat k_6 het daar ook slecht zou hebben gedaan. Zo heeft ras k_6 op veld p_3 getoond ook hogere opbrengsten te kunnen geven, zonder dat dit een conclusie toelaat over k_5.

Men kan slechts zeggen, dat voor beide rassen drie monsters uit het onderzochte gebied zijn getrokken om de rasvoortreffelijkheid binnen het gebied vast te stellen. Bij gebrek aan beter zal men voor

beide monsters aan moeten nemen, dat ze zo goed mogelijk genomen zijn. Het kan zijn dat k_5 op p_1 te weinig opbracht en k_6 op p_3 te veel; maar evenzeer kan k_5 en p_2 en p_4 te veel opgebracht hebben en k_6 daar te weinig. Als de rassen onafhankelijk zijn valt er niets van te zeggen.

Mocht men er van uit willen gaan, dat ras k_5 en k_6 identiek waren, dan zouden de cijfers van gem. D (zie 311.1) niet de meest geschikte zijn om de onderlinge verschillen tussen alle rassen weer te geven. De eindcijfers van de rassen k_5 en k_6 zouden gelijk moeten zijn. De vraag is nu: Hoe bereikt men dat het rasverschil van k_5 en k_6 zo goed mogelijk wordt vastgelegd, terwijl tevens het verschil met de andere vier rassen op de juiste wijze wordt weergegeven?

In overeenstemming met 310 moet men onderscheid maken tussen rassen die gelijk zijn of slechts een constante verschillen, en rassen die in meerdere of mindere mate gecorreleerd zijn. In 310 werd de laatste groep nog gesplitst al naar er een correlatie werd verwacht of niet. Nu behoeft dit onderscheid niet te worden gemaakt. Wanneer een correlatie eenmaal is aanvaard moet er op een bepaalde manier mee gewerkt worden.

Indien de verwante rassen slechts een constante (eventueel 0) verschillen voor de onderzochte eigenschap, valt er over deze constante niets te leren uit de velden waarop niet beide rassen voorkwamen. De waarde van de constante moet rechtstreeks worden bepaald op de velden, waarop beide rassen wel stonden. Voor ras k_5 en ras k_6 is deze constante in ons voorbeeld dus 0. Hiermee is het verschil tussen k_5 en k_6 zo goed mogelijk aangegeven.

Nu moet deze gelijkheid nog in gem. D van (311.1) tot uitdrukking worden gebracht. Bij het bepalen van de methode willen wij tevens trachten tegemoet te komen aan de wens, die wij bij de discussie van (304.2) uitspraken, dat zoveel mogelijk rassen in het fictiefras moeten worden opgenomen. Wij gaan daarom de ontbrekende opbrengsten schatten. Wegens de gelijkheid van k_5 en k_6 kunnen wij in (311.1) de opbrengsten van beide rassen op de velden p_1 en p_3 ook aan elkaar gelijk achten. Vullen we de geschatte opbrengsten in, dan vinden wij (311.2)

(311.2)

ras / veld	k_1	k_2	k_3	k_4	k_5	k_6
p_1	646	655	802	636	596	596
p_2	681	663	690	663	643	643
p_3	716	762	751	461	665	665
p_4	685	653	685	675	642	642
gem. E	682	683	732	609	637	637

Omdat het schema nu volledig is zal gem. E ongevoelig zijn voor een nieuwe standaardisering. Wel zal het niveau van de verschillende proefvelden daardoor kunnen worden beïnvloed.

Bij bovenstaande methode hebben wij dus de opbrengsten van k_5 en k_6 overal gelijk geschat; wij kunnen evengoed zeggen, dat wij overal de afwijking u_{5p} gelijk hebben geschat aan de afwijking u_{6p}. Wanneer wij onder ε_k verstaan

$$\varepsilon_k = \sqrt{\frac{\uparrow n\,[u^2_{k\pi}]}{n}}$$

kunnen we zelfs zeggen dat we de schatting hebben verricht volgens de formule

(311.3) $$\frac{u_{6p}}{\varepsilon_6} = \frac{u_{5p}}{\varepsilon_5}.$$

Voor rassen die *overal* slechts een constante verschillen moeten de waarden van ε immers gelijk zijn

Ogenschijnlijk is het gebruik van (311.3) in strijd met 124. De afwijkingen u_{6p} en u_{5p} zijn beide normaal verdeeld, niet alleen wat hun toevallige fout betreft, maar — naar wij in 302 aannamen — ook wat hun interpretatiefout betreft. Dit moet dus een geval zijn waar de correlatierekening moet worden toegepast, en niet de lijnvereffening.

De waarden u_{6p} en u_{5p} zijn niet absoluut met elkaar gecorreleerd; beide grootheden zijn gedeeltelijk opgebouwd uit toevallige fouten, die niet gecorreleerd zijn. Volgens 124 hadden inplaats van de lange as (311.3) dus gebruikt moeten worden de regressielijnen

(311.4) $$\frac{u_{61}}{\varepsilon_6} = \gamma\,\frac{u_{51}}{\varepsilon_5}$$

$$\frac{u_{53}}{\varepsilon_5} = \gamma\,\frac{u_{63}}{\varepsilon_6}.$$

Toch bevredigt het gebruik van de regressielijnen in dit geval niet. Wanneer de rassen k_5 en k_6 in wezen als identiek worden gezien, moet de opbrengst van de een zonder meer als schatting van de opbrengst van de ander kunnen dienen. Naar het lijkt hebben wij hier een gevoelsargument tegenover een exact argument. Maar om te zien of het exacte argument inderdaad de doorslag moet geven, willen wij dit argument voor het onderhavige geval nog eens op de voet volgen.

De opbrengst van ras k_5 op veld p_1 is abnormaal laag. In absolute waarde is u_{51} dus groot. Omdat de toevallige fout τ_{5p} normaal verdeeld is, is het onwaarschijnlijk dat τ_{51} zo groot is. En omdat i_{5p} ook normaal verdeeld is, is het eveneens onwaarschijnlijk dat i_{51} zo groot is. De grootste waarschijnlijkheid bereiken wij wanneer we aannemen dat u_{51} en i_{51} in dit geval hetzelfde teken hebben. In absolute waarde schatten wij i_{51} dus kleiner dan u_{51}.

Wanneer wij de opbrengst van ras k_6 schatten, is het niet juist daar de toevallige fout τ_{51} in op te nemen. Wij moeten de systematische interpretatiefout i_{61} schatten. Omdat i_{61} kleiner is in absolute waarde dan u_{51} moet de opbrengst van ras k_6 op veld p_1 hoger geschat worden dan de opbrengst van ras k_5 op dat veld was. Door de regressielijn wordt uitgemaakt hoeveel hoger. (In ons voorbeeld is dit onmogelijk doordat de toevallige fouten toevallig absoluut gecorreleerd zijn (zie 310.12))

Bovenstaande uiteenzetting is slechts dan goed, wanneer wij bereid zijn de consequenties eruit te trekken, die er in 124 ook uitgetrokken werden. Aan het slot van die paragraaf kwamen wij tot een conclusie die hierop neerkomt: De grootte van u_{5p} wordt naar evenredigheid verdeeld over i_{5p} en τ_{5p}. *Wanneer er een grote absolute waarde van de afwijking u_{5p} wordt gevonden nemen we op grond daarvan ook een grote toevallige fout τ_{5p} van.* Welnu deze consequentie kunnen wij hier niet aanvaarden.

Tenslotte is een interpretatiefout een grootheid die men niet kent. Wanneer het proefveldgemiddelde van een ras sterk afwijkt van het cijfer dat men verwacht, dan kan dit gemiddelde blijkens zijn middelbare fout wel zeer nauwkeurig zijn. Men zal dan niet concluderen dat de toevallige fout groot zal zijn, maar alle schuld op de interpretatiefout werpen.

Wij zullen dus geen gebruik maken van formule (311.4) maar van (311.3). Hiermee is evenwel niet gezegd dat wij nu de methoden der lijnvereffening gaan toepassen. Ongetwijfeld hebben wij hier

te maken met een geval van correlatie. Wij weigeren slechts iets over de toevallige fout te zeggen op grond van de interpretatiefout. Daarom mag niet gewerkt worden met de totale correlatie γ, maar moet gebruik worden gemaakt van de *correlatie tussen de interpretatiefouten*: Γ (zie 310.11). Bij identieke rassen is die correlatie 1. Dit is de motivering voor het gebruik van (311.3).

Wij zijn dus tot de conclusie gekomen dat men in het algemeen gebruik moet maken van de correlatiecoëfficient Γ. Maar de vraag blijft nog over, welke? Volgens (310.11) laten zich een $\Gamma_{k/l}$ en een $\Gamma_{l/}$ berekenen, alnaar ras k of ras l buiten het fictiefras wordt gelaten.

Om de keuze te bepalen stelle men zich voor dat men u_{lp} wil schatten aan de hand van u_{kp}. Proefveld p moet dan zo goed mogelijk op de andere proefvelden zijn gestandaardiseerd omdat de standaardisatiefout de grootte van u beïnvloedt. Het „fictiefras" van p moet dus zoveel mogelijk rassen bevatten. Nu kan ras k wel in het fictiefras worden opgenomen, maar ras l niet, omdat het niet op het veld p voorkwam. Het is om deze reden wenselijk k in het fictiefras te handhaven, en dus l uit te sluiten, zodat met de $\Gamma_{l/k}$ moet worden gewerkt voor het schatten van een opbrengst van l.

Wij vinden dus als formule voor het schatten van een onbekende afwijking u_{lp}.

(311.5) $$\frac{u_{lp}}{\varepsilon_l} = \Gamma_{l/k} \frac{u_{kp}}{\varepsilon_k}.$$

Als voorbeeld van het gebruik van deze formule willen wij aannemen, dat wij verwachten, dat ras k_6 gecorreleerd is met ras k_3. Wij willen nu trachten de gestandaardiseerde opbrengst van ras k_6 op veld p_1 te schatten aan de hand van de gestandaardiseerde opbrengst Ω^*_{31} van ras k_3 op dat veld.

Omdat ras k_6 van het fictiefras wordt uitgesloten moeten de gegevens van (311.1) opnieuw gestandaardiseerd worden. Dit geeft

(311.6)

ras / veld	k_1	k_2	k_3	k_4	k_5	k_6
p_1	646	655	802	636	596	
p_2	680	662	689	662	642	642
p_3	721	767	756	466		670
p_4	684	652	684	674	641	641
gem. F	683	684	733	610	626	651

Nu kan de opbrengst van ras k_6 op veld p_1 worden berekend. Daartoe moet eerst worden bepaald de gemiddelde opbrengst van ras k_6 op de velden p_2, p_3 en p_4, en de gemiddelde opbrengst van ras k_3 *op dezelfde velden*. We vinden

(311.7) Gem. van ras $k_6 = 651$
,, ,, ,, $k_3 = 710$.

De afwijkingen u van k_3 en k_6 op deze velden zijn dus

(311.8)

ras / veld	k_3	k_6
p_2	-21	- 9
p_3	46	19
p_4	-26	-10

Hieruit laat zich berekenen

(311.9)
$$[u_3^2] = 3233$$
$$[u_6^2] = 542$$
$$[u_3 u_6] = 1323$$
$$\varepsilon_3 = 32.9$$
$$\varepsilon_6 = 13.5.$$

Om $\Gamma_{6/3}$ te kunnen berekenen moet σ^2 nog bekend zijn.

Wij willen aannemen dat uit het verschil tussen de blokken binnen de proefvelden is berekend dat

(311.10) $$\sigma^2 = 25.$$

De grootheid $\frac{1}{m'}$ is voor p_2, p_3 en p_4 resp. $^1/_5$, $^1/_4$ en $^1/_5$. De gemiddelde waarde is

(311.11) $$\frac{1}{m'} = 0.22.$$

Uit de cijfers van (311.(9, 10, 11)) laat zich berekenen met behulp van (310.11) ($n = 3$)

(311.12) $$\Gamma_{6/3} = \frac{1323}{\sqrt{(542-61)(3233-39)}} = 1.07.$$

Wij vinden hier dat de correlatie tussen de interpretatiefouten groter is dan 1. Dit is natuurlijk onmogelijk. Bij normale correlatie is een dergelijke *reken*resultaat zelfs onmogelijk omdat de berekeningswijze dan geen grotere waarde dan 1 toelaat. Dit hangt samen met de speciale foutenwetten van de correlatiecoëfficient.

Van de hier gebruikte correlatiecoëfficient $\Gamma_{k/l}$ (zie form. 310.11) hebben wij nog geen foutenwet kunnen opstellen. Doordat in de noemer gecorrigeerd wordt met de gevonden waarde van σ^2, zijn de foutenwetten evenwel anders dan van een gewone correlatiecoëfficient. Waarschijnlijk wordt de middelbare fout er groter van, maar dit zegt niets over de vraag of de asymptotische waarde van de $\Gamma_{k/l}$ hierdoor ongunstig wordt beïnvloed. Het is mij ook niet bekend. Zo lang er niet een betere oplossing is, is het in ieder geval beter met (310.11) te werken dan de correlaties tussen de rassen te negeren.

Aangezien wij weten dat $\Gamma_{k/l}$ slechts groter dan 1 berekend kan worden wegens schattingsfouten bij het gebruiken van een op schattingen berustende formule, zullen wij de gevonden waarde terugbrengen tot 1.

Hiermee is het voorbeeld niet teruggevoerd tot het probleem dat in (311.2) werd besproken, nl. dat van de rassen, die gelijk waren of slechts een constante verschilden. Ook in dat geval is $\Gamma_{k/l} = 1$, maar daarbij is tevens $\varepsilon_k = \varepsilon_l$, wat in dit voorbeeld niet opgaat.

In ons onderhavig voorbeeld kan men zeggen dat ras k_6 op dezelfde wijze op de omstandigheden reageert als ras k_3, slechts is de reactie minder heftig. Om het gemiddeld *verschil* tussen beide rassen te bepalen moeten nu wel de hiaten in het schema worden opgevuld. Op veld p_1 waar de opbrengst van k_3 abnormaal hoog was, zal het *verschil* tussen k_3 en k_6 ook abnormaal groot zijn. Dit zal het gemiddeld verschil beïnvloeden.

De vermoedelijke gestandaardiseerde opbrengst van ras k_6 op veld p_1 berekenen wij nu alsvolgt: Blijkens (311.6) bracht ras k_3 op veld p_1 92 meer op dan het gemiddelde in (311.7). Volgens (311.9) is dit $2.8\,\varepsilon_3$. De vermoedelijke opbrengst van ras k_6 op p_1 moet dus $2.8\,\varepsilon_6 = 38$ boven het gemiddelde in (311.7) liggen. Wij vinden dus als waarde van deze opbrengst 689. Vullen wij dit bedrag in (311.6) in dan vinden wij

(311.13)

ras / veld	k_1	k_2	k_3	k_4	k_5	k_6
p_1	646	655	802	636	596	689
p_2	680	662	689	662	642	642
p_3	721	767	756	466		670
p_4	684	652	684	674	641	641
gem. G	683	684	733	610	626	661

De cijfers van (311.13) zullen nu nog op elkaar moeten worden gestandaardiseerd met alle rassen in het fictiefras. Dit geeft

(311.14)

ras / veld	k_1	k_2	k_3	k_4	k_5	k_6
p_1	641	650	797	631	591	684
p_2	682	664	691	664	644	644
p_3	718	764	753	463		667
p_4	687	655	687	677	644	644
gem. H	682	683	732	609	626	660

Hiermee zijn de berekeningen van dit voorbeeld afgelopen.

Wij hebben nu dus tweemaal de opbrengst van ras k_6 op veld p_1 geschat. Eenmaal in de veronderstelling dat ras k_6 identiek is met k_5 (zie 311.2) en eenmaal in de veronderstelling dat ras k_6 gecorreleerd was met k_3.

Door (311.2) met (311.13) te vergelijken ziet men dat de schattingen sterk uiteenlopen. Wij zullen nu het probleem moeten bespreken, hoe deze schattingen op elkaar zouden kunnen worden vereffend, wanneer men de beide gemaakte veronderstellingen even waarschijnlijk achtte. In het onderhavig geval waren de veronderstellingen *kwalitatief* ongelijk. Eerst werd gelijkheid van rassen verondersteld, vervolgens een nauwe correlatie.

Wij willen liever een voorbeeld geven waar de veronderstellingen kwalitatief gelijk zijn, daarom gaan wij uit van de opbrengsten van (311.5), die op elkaar gestandaardiseerd zijn, zonder dat k_6 in het fictiefras was opgenomen.

(311.15)

ras / veld	k_1	k_2	k_3	k_4	k_5	k_6
p_1	646	655	802	636	596	
p_2	680	662	689	662	642	642
p_3	719	765	754	464		668
p_4	684	652	684	674	641	651
p_5	669	669	735	605	656	700
gem.	680	681	733	608	634	665

Wij stellen verder

(311.16) $$\sigma^2 = 25.$$

Nu willen wij de opbrengst van ras k_6 op veld p_1 schatten in de veronderstelling dat k_6 gecorreleerd is met k_3 en k_5.

Voor de correlatie van k_6 met k_3 kunnen wij gebruik maken van de velden p_2 t/m p_5. Op dezelfde wijze als in (311.9 enz.) vinden wij

(311.17) Gem. opbrengst van $k_3 = 715$

,, ,, ,, $k_6 = 665$

$$\varepsilon_3 = 29.8$$

$$\varepsilon_6 = 22.1$$

$$\Gamma_{6/3} = \frac{1849}{\sqrt{(3558 - 59)(1959 - 91)}} = 0.723.$$

Hieruit laat zich berekenen, als vermoedelijke waarde van de opbrengst van ras k_6 op veld p_1, de waarde 711.

Bovenstaande waarde is natuurlijk berekend met behulp van formule (311.5). Voor het onderhavige probleem heeft het evenwel zijn voordelen (zoals uit 312 zal blijken) om de waarde van u_{lp} met behulp van de lange as (zie 311.3) te schatten, en de $\Gamma_{l/k}$ te beschouwen als het gewicht van de schatting.

Volgens (311.5) is gevonden $u_{61} = 46$ (met gewicht 1), volgens (311.3) zou gevonden worden $u'_{61} = 64$ (met gewicht 0.723). Tussen deze uitkomsten is geen verschil volgens de gewichtenstelling (306.9).

(311.18) Op grond van deze gelijkheid kunnen wij als geschatte opbrengst van ras k_6 op veld p_1 ook geven de waarde 729 (met gewicht 0.723).

Voor de correlatie van k_5 en k_6 kunnen wij gebruik maken van de velden p_2, p_4 en p_5. De berekening levert:

(311.19) Gem. opbrengst van $k_5 = 646$

,, ,, ,, $k_6 = 664$

$$\varepsilon_5 = 6.86$$

$$\varepsilon_6 = 25.5$$

$$\Gamma_{6/5} = 1.17.$$

In dit voorbeeld ziet men dat ε_5 nauwelijks groter is dan σ (Volgens 311.16: $\sigma = 5$). Hoewel wij bij wijze van illustratie de berekening verder zullen voortzetten, moeten wij er toch op wijzen, dat dit resultaat het twijfelachtig maakt of ras k_5 wel systematisch van het fictiefras afwijkt. Wanneer ε_5 geheel aan het toeval is te wijten, is het zoeken van correlaties met systematische bestanddelen van ε_5 natuurlijk onjuist. In het algemeen kunnen

wij zeggen dat het gebruik van (310.11) bedenkelijk wordt, wanneer de daar genoemde $\uparrow n\,[u^2_{k\pi}]$ en $\uparrow n\,[u^2_{l\pi}]$ niet „betrouwbaar" groter zijn dan $(n-1)\,\sigma^2$. Deze beperking geeft ook de garantie dat de noemer van (310.11) steeds positief is. Om de uiteenzetting verder voort te kunnen zetten negeren wij ditmaal dit bezwaar.

Wij beginnen weer met $\Gamma^{5}/_{5}$ tot 1 terug te brengen, evenals bij (311.12).

(311.20) Uit (311.19) berekenen wij nu, als vermoedelijke waarde van de opbrengst van ras k_6 op veld p_1, de waarde 478 (met gewicht 1).

Om de beste schatting van de opbrengst van ras k_6 op veld p_1 te krijgen gaan wij de uitkomsten van (311.18) en (311.20) middelen met behulp van de formule (311.21)

$$\overline{x} = \frac{[gx]}{[g]}.$$

Dit geeft $\overline{x} = 583$. Blijkens de formule

$$\sigma_{\overline{x}}^2 = \frac{\sigma^2}{[g]}$$

is het gewicht van deze uitkomst $[g] = 1.723$. Het zou echter niet juist zijn dit gewicht aan de uitkomst toe te kennen. Men mag niet hoger gewicht toekennen dan 1. De reden hiervan zal in 312 worden besproken.

Wij nemen dus als geschatte opbrengst van ras k_6 op veld p_1 de waarde 583 met gewicht 1. Wanneer wij dit getal in (311.15) invullen en vervolgens standaardiseren met alle rassen in het fictiefras, vinden we (311.22)

(311.22)

ras / veld	k_1	k_2	k_3	k_4	k_5	k_6
p_1	657	666	813	647	607	594
p_2	681	663	690	663	643	643
p_3	715	761	750	460		664
p_4	684	652	684	674	641	651
p_5	661	661	727	597	648	692
gem.	680	681	733	608	635	649

Het is natuurlijk mogelijk de opbrengst van k_5 op veld p_3 ook te schatten aan de hand van de correlatie tussen k_5 en k_6. Omdat op p_3 ras k_6 wel in het fictiefras kan worden opgenomen en k_5 niet, is het wenselijk in dit geval met $\Gamma^{5}/_{5}$ te gaan werken.

312 – NADER ONDERZOEK VAN DE ONTWIKKELDE METHODEN

Tot dusver hebben wij in dit hoofdstuk verschillende methoden besproken, die gebruikt kunnen worden bij het samenvatten der proefveldresultaten. Deze methoden willen wij nu in onderlinge samenhang nader bespreken.

Men moet steeds beginnen met de proeven samen te vatten alsof alle rassen van elkaar onafhankelijk zijn (zie 301 en 304). De methode, die in 311 is gegeven voor afhankelijke rassen, doet niet anders dan de methode van 304 corrigeren. Bij afhankelijke rassen heeft men dus een samenvatting volgens 304 nodig als uitgangspunt van de verdere berekening.

Wanneer de proeven op elkaar gestandaardiseerd zijn, kan men controleren of er afhankelijkheid tussen sommige rassen optreedt. Bij schema's die geheel of bijna orthogonaal zijn, zal dit het best gaan aan de hand van de stellingen die uit de formules (302.4) en (303.4) zijn afgeleid. Wanneer er geen afhankelijkheid is moeten de asymptotische waarden van de interpretatievarians van de verschillende rassen gelijk zijn; is er wel afhankelijkheid, dan is de asymptotische waarde van de interpretatievarians voor de afhankelijke rassen kleiner. Aangezien de middelbare toevallige fout voor alle rassen vaak ook gelijk is, gelden bovenstaande stellingen ook voor de asymptotische waarden van $[u^2]$ van ieder ras.

Wanneer de schema's onvolledig zijn geldt deze contrôle niet meer. Bij het standaardiseren zal een proefveld met weinig rassen meer van de toevallige fouten door de milieugunstigheid weg kunnen verklaren (zie voor dit begrip tussen (310.6) en (310.7)) dan een met veel rassen. Zo zal ook het gemiddelde van een ras dat weinig voorkomt, de fouten beter weg kunnen verklaren dan het gemiddelde van een ras dat vaak voorkomt. Uit (301.16) blijkt duidelijk dat er een tendens is dat bij een gegeven i_{kp} de u_{kp} kleiner wordt naarmate m en n kleiner zijn. Uit (301.24) blijkt eveneens dat, bij een gegeven i_{kp}, u_{kp} kleiner wordt naarmate n kleiner wordt. Wegens de definitie van het fictiefras zal een kleine m vaak een verkleining van i_{kp} zelf tot gevolg hebben.

Bovenstaande tendens is min of meer te vermijden door het onderzochte ras k buiten het fictiefras te houden. Dan hebben de fouten van k geen invloed op de standaardisatie en worden dus ook niet gedeeltelijk daardoor weg verklaard. De invloed van het rasgemiddelde blijft natuurlijk bestaan, zodat niet zonder meer

de vormen $[u^2]$ met elkaar kunnen worden vergeleken. Voor ieder ras moet de vorm

$$\psi_k^2 = \frac{\uparrow n_k [u_k^2]}{n_k - 1},$$

bepaald worden, waarbij n_k het aantal velden aangeeft waarop ras k voorkwam.

Het is dus aan te bevelen bij onvolledige schema's de standaardisatie m maal over te doen met telkens een ander ras buiten het fictiefras gesloten. In 309 is besproken hoe deze m standaardisaties gemakkelijk uit de standaardisatie met volledig fictiefras kunnen worden afgeleid. Uit deze m standaardisaties zijn nu voor telkens een ander ras m schattingen van ψ^2 te berekenen, die ongeveer gelijk moeten zijn. Vindt men nu rassen met een kleine ψ^2 dan zijn die vermoedelijk onderling gecorreleerd.

Het is misschien gewenst er op te wijzen, dat bovenstaande methode slechts een globale indruk kan geven. Wanneer er b.v. twee rassen zijn, die precies gelijk zijn aan het fictiefras, dan zal hun interpretatiefout steeds 0 zijn. Dit houdt in dat ze onderling niet gecorreleerd zijn in hun interpretatiefout, en toch een kleine ψ^2 hebben. Bovenstaande contrôle kan dus niet meer dan een aanwijzing geven.

Waarschijnlijk is het meestal nog beter eerst grafisch te controleren of er een correlatie verwacht kan worden. Dit zal meestal direct kunnen na de berekening van 304.

Nauwkeuriger contrôle bereiken we door gebruik te maken van (302.10) of van de methoden, die in 310 genoemd zijn. Deze contrôle is dan ook steeds gewenst wanneer men in onzekerheid verkeert. Komt men tot de conclusie, dat er inderdaad afhankelijkheid van rassen optreedt, dan moeten onvolledige schema's worden samengevat volgens de principes van 311. Daar spraken wij er reeds over dat dit beslist noodzakelijk is om goede uitkomsten te krijgen. Op deze noodzaak willen wij nu iets dieper ingaan.

Bij het trekken van conclusies moet gelet worden op de zin van het proefveld, zoals wij die in 201 bespraken. Daar hebben wij met elkaar vergeleken de mogelijkheid het rassenonderzoek te verrichten door middel van enquêtes en door middel van een proefveldonderzoek. Het bezwaar van de enquêtes is, dat het onderzochte gebied niet homogeen is, zodat er groot gevaar is dat het ene ras steeds gunstige omstandigheden treft en het andere

niet. Het is nu de taak van het proefveldonderzoek de relatieve gevoeligheid van de rassen voor de omstandigheden te verminderen, door alle rassen onder dezelfde omstandigheden te vergelijken. Het proefveldonderzoek leidt tot het begrip „fictiefras" en het begrip „interpretatiefout", wat hier het systematisch verschil tussen opbrengst van ras en fictiefras op een bepaald proefveld betekent.

Het bovenstaande is scherp gedefinieerd zolang alle onderzochte rassen op alle proefvelden hebben gestaan, dus zolang het schema orthogonaal is. Met niet-orthogonale schema's is dit niet meer het geval. Wanneer het ene ras gedeeltelijk op andere proefvelden wordt beproefd dan het andere, verkeert men in een situatie, die het midden houdt tussen een enquête en een proefveldonderzoek. Immers, nu is de kans in zekere mate aanwezig dat het ene ras betere omstandigheden treft dan het andere. De principiële vraag is nu: hoe kan men bereiken, dat het cijfermateriaal zoveel mogelijk aan een proefveldonderzoek doet denken en zo weinig mogelijk aan een enquête.

Men zal zijn materiaal slechts dan als een normaal proefveldonderzoek kunnen benutten, wanneer men kans ziet alle ontbrekende opbrengsten te berekenen uit de aanwezige cijfers. Dit is het wat we in 304 hebben gedaan. Wanneer men b.v. in (304.12) voor alle ontbrekende opbrengsten het berekende rasgemiddelde invult, zullen er geen andere gemiddelde opbrengsten worden berekend, wanneer ook deze geschatte opbrengsten in het gemiddelde worden opgenomen. Men zou (304.12) dus ook alsvolgt weer kunnen geven

(312.1)

ras / veld	*l*	*m*	*n*	*p*	*q*	*r*	*s*	*t*	*u*	*w*
a	694	690	653	630	705	661	686	668	697	652
b	664	720	653	612	705	700	714	668	700	602
c	646	705	634	660	705	659	689	668	718	652
d	665	705	686	666	705	615	659	668	697	671
e	679	705	657	626	705	661	686	668	697	652
f	653	705	664	667	705	661	686	668	688	639
g	638	705	649	643	724	660	675	668	718	652
h	707	705	580	643	714	672	692	668	702	651
i	638	705	697	643	689	661	686	668	697	652
k	664	705	660	643	693	661	686	668	654	700
	665	705	653	643	705	661	686	668	697	652

Wij kunnen dus zeggen dat 304 een bepaalde techniek geeft om opbrengsten te schatten van onafhankelijke rassen. Deze techniek komt hierop neer dat men voor een bepaald proefveld nagaat hoeveel de opbrengsten van de verbouwde rassen gemiddeld boven of beneden hun resp. rasgemiddelden zijn gebleven. Men veronderstelt nu dat de niet verbouwde rassen ook zoveel boven of beneden hun resp. gemiddelden zouden zijn gekomen, als ze er ook verbouwd waren. Door de standaardisatie worden deze veronderstelde opbrengsten dan gelijk aan het bijbehorend gemiddelde.

Het is onwaarschijnlijk dat die veronderstelling voor een bepaald proefveld juist is. Waren de rassen verbouwd dan hadden ze ook een interpretatiefout gegeven. Toch is de veronderstelling niet zinloos. Statistisch is zij juist, d.w.z. wanneer men een aantal opbrengsten van een ras schat, zullen er ongeveer evenveel te hoge als te lage schattingen bij zijn; wanneer men één opbrengst schat, is er evenveel kans dat hij te hoog is als te laag.

Die gelijkheid van kans berust niet op de omstandigheden van het proefveld. Deze omstandigheden kunnen zeer wel maken, dat het ras er een abnormaal hoge of een abnormaal lage opbrengst zou geven. Maar over het aanwezig zijn van die omstandigheden weet men gewoonlijk niets, anders rekende men er wel mee. In zijn kansenwaardering houdt men zich dus bezig met de kansverdeling van de onbekende omstandigheden, die oorzaak zijn van de interpretatiefout. Van deze interpretatiefouten zal men vaak aan moeten nemen dat ze tamelijk normaal om het berekende gemiddelde verdeeld zijn, uit gebrek aan betere gegevens.

Om met deze globale aanname te kunnen werken, moet men dus weten dat er inderdaad geen betere gegevens zijn. Dit is de achtergrond van het onderzoek van de correlaties. Wanneer van het onderzochte ras de interpretatiefout gecorreleerd blijkt met de interpretatiefout van een ander ras, is er meer over de kansverdeling van deze interpretatiefout bekend dan boven en in 304 verondersteld werd.

Wanneer b.v. bekend is dat ras l in zijn interpretatiefouten volledig gecorreleerd is met een ander ras k ($\Gamma^l/_k = 1$), dan zal, al mogen de omstandigheden van een proefveld p niet met name bekend zijn, de reactie van ras l daarop toch wel volledig bepaald zijn. Dit geeft de mogelijkheid de ontbrekende opbrengst van ras l op veld p nader te berekenen. Van deze mogelijkheid *moet*

gebruik worden gemaakt. Immers de berekening van 304 geeft blijkens (312.1) ook een schatting van de opbrengst en wel een foutieve.

De opbrengst van ras l, die men aan de hand van de correlatie berekent, behoeft niet gelijk te zijn aan de opbrengst die in werkelijkheid verkregen zou zijn, als het ras inderdaad verbouwd was. De geschatte opbrengst moet statistisch juist zijn, d.w.z. de kans op een positief verschil tussen schatting en werkelijkheid moet even groot zijn als de kans op een negatief verschil.

Ondertussen kan dit verschil wel erg groot zijn. Immers alle beschikbare gegevens zijn behept met een toevallige fout. Dit geldt met name van de opbrengst van ras k op veld p. Deze toevallige fouten kunnen sterk vergroot worden bij het berekenen van de geschatte opbrengst. Zo is bij het berekenen van (311.20) niet alleen de interpretatiefout van Ω^*_{51}, maar ook de toevallige fout vermenigvuldigd met $\frac{25.5}{6.86}$. Men mag dus verwachten dat de toevallige fout van de geschatte Ω^*_{61} minstens viermaal zo groot is als die van de bestaande opbrengstcijfers.

Dit mag evenwel geen reden zijn aan de schatting een laag gewicht, b.v. $\frac{1}{16}$, toe te kennen. Blijkens de gewichtenstelling, (met uitvoerige motivering bij 306.9), zou dit er op neerkomen dat de interpretatiefout door 16 gedeeld werd. Daarmee zou de schatting zijn statistische juistheid verliezen, omdat het *teken* van de interpretatiefout niet toevallig is. Niettegenstaande de grote verschillen in toevallige fout moeten alle gewichten dus gelijk worden genomen.

In het algemeen kunnen wij als regel stellen, dat de toe te kennen gewichten uitsluitend af mogen hangen van de interpretatiefouten, wanneer er in het materiaal interpretatiefouten en toevallige fouten beide voorkomen. Deze regel vindt slechts hierin zijn beperking dat het soms beter kan zijn een kleine systematische fout te maken dan een grote toevallige. Het is hier niet de plaats uit te zoeken wanneer met deze beperking rekening moet worden gehouden.

Wij hebben tot dusver in deze paragraaf twee uitersten genoemd: onafhankelijkheid van de interpretatiefouten en absolute correlatie. In het laatste geval was alles over de ontbrekende interpretatiefout te berekenen. Aan deze berekende waarde kon dus het gewicht 1 worden toegekend. In geval van onafhankelijk-

heid is niets over de interpretatiefout bekend, d.w.z. dat in dit geval aan het rekenresultaat de waarde 0 moet worden gegeven. Blijkens de gewichtenstelling behoeft dit niet in te houden dat $g = 0$ wordt gesteld, het is voldoende dat gu de waarde 0 krijgt. Men kan nu twee wegen bewandelen om $gu = 0$ te bereiken. In de eerste plaats kan men $g = 0$ stellen, en de u onbepaald laten; dit deden wij in 304, waar wij geen geschatte opbrengsten vermeldden omdat het gewicht toch 0 was. Men kan evenwel ook $g = 1$ stellen en $u = 0$. Dit deden wij in (312.1), waar wij wel opbrengsten vermeldden met gewicht 1, maar waar de u steeds 0 was.

Wanneer de $\Gamma_{l/k}$ in absolute waarde tussen 0 en 1 ligt, is er natuurlijk een tussenpositie tussen beide uitersten. Er bestaat wel een band, maar die band is niet absoluut. Hoe groter de correlatiecoëfficient, des te steviger de band.

Voor dit geval hebben wij in (311.5) een formule gevonden, die een statistisch juiste schatting toelaat van de ontbrekende opbrengsten. Dat deze schatting statistisch juist is betekent niet dat de *interpretatie*fout juist wordt geschat. Voor de interpretatiefout bestaan er drie mogelijkheden: De schatting volgens (311.5) kan te klein zijn, goed, of te groot. Dat de schatting statistisch juist is betekent, dat niet alleen de toevallige fout asymptotisch 0 wordt, maar dat ook de te grote en de te kleine schattingen van het systematisch deel van u_{lp} elkaar asymptotisch in evenwicht houden.

Omdat de schatting statistisch juist is, moet ze in de berekening het gewicht 1 hebben, zoals wij zo juist op grond van de gewichtenstelling nog eens mochten formuleren. *Een ander gewicht verandert immers feitelijk de grootte van de geschatte interpretatiefout.*

Wanneer ras l alleen met één bepaald ras k is gecorreleerd, valt er over de vraag, of in de berekende u_{lp} de interpretatiefout te groot of te klein geschat is, niets te zeggen. Anders wordt dit, wanneer ras l gecorreleerd is met twee rassen k_1 en k_2. Wanneer de u_{lp} die aan de hand van k_2 wordt berekend tegengesteld is aan de u_{lp} die aan de hand van k_1 is berekend, dan maken beide schattingen elkaar dubieus.

Wanneer u_{lp} aan de hand van beide correlaties gelijk berekend wordt zijn er nog weer twee mogelijkheden: Door de correlatie met k_1 en die met k_2 zijn *dezelfde* vruchtbaarheidselementen aan-

gewezen, die interpretatiefouten geven, of er zijn *ongelijke* vruchtbaarheidselementen aangewezen.

Wanneer *ongelijke* vruchtbaarheidselementen zijn aangewezen moeten de beide gevonden waarden van u_{lp} bij elkaar worden *opgeteld*, immers beide groepen vruchtbaarheidselementen drijven de u_{lp} in de gevonden richting, min of meer onafhankelijk van elkaar.

Wanneer *dezelfde* vruchtbaarheidselementen zijn aangewezen, moeten beide waarden u_{lp} als twee berekeningen van dezelfde grootheid worden gezien. In dit geval moeten de gevonden waarden van u_{lp} dus worden *gemiddeld*. Dit laatste is stellig het geval wanneer de k_1 en k_2 onderling volledig zijn gecorreleerd.

Wanneer k_1 en k_2 onderling niet volledig zijn gecorreleerd is minder gemakkelijk te zeggen in welke geval men verkeert. Toch vallen hier wel enkele regels op te stellen. Wanneer namelijk ras *l volledig* met ras k_1 is gecorreleerd, dan is het onmogelijk, dat de correlatie met k_2 vruchtbaarheidselementen aanwijst, die niet door de correlatie met k_1 zijn aangewezen. In dit geval is *optelling* van de gevonden waarden van u_{lp} dus *ongeoorloofd*. Naarmate de correlaties zwakker worden is er meer *kans* dat er ongelijke vruchtbaarheidselementen door de verschillende correlaties worden aangewezen. Er moet dus een samenvattingswijze zijn, die naar het *additieve* helt, wanneer de correlaties klein zijn, en naar het *middelen*, wanneer de correlaties groot zijn.

Een dergelijke berekeningswijze is te vinden, wanneer we de afwijking u_{lp} schatten volgens de formule

$$\frac{u_{lp}}{\varepsilon_l} = \frac{u_{kp}}{\varepsilon_k}$$

en aan deze schatting het gewicht $\Gamma_{l/k}$ toekennen. Deze methode hebben wij gevolgd bij de bespreking van (311.15), zoals in de tekst onder (311.17) is uiteengezet.

Bij (311.21) hebben wij gezegd dat men de *opbrengsten*, die zo uit de verschillende correlaties zijn berekend, moet middelen met de formule

$$\bar{x} = \frac{[gx]}{[g]} \tag{311.21}$$

en dat aan de aldus berekende $\bar{x}$ het gewicht $[g]$ moet worden toegekend, *mits* $[g]$ *niet groter is dan* 1. In het licht van bovenstaande overwegingen worden deze voorschriften duidelijk. Voor zover

de berekende afwijkingen van het oorspronkelijke gestandaardiseerde gemiddelde elkaar tegenspreken heffen ze elkaar bij de toepassing van (311.21) op. Aangezien blijkens de gewichtenstelling de waarde gu feitelijk de invloed van de afwijking meet (dus het *effectief* gewicht is), kunnen wij zeggen dat het effectief gewicht van tegenstrijdige schattingen ongeveer 0 wordt.

Voor zover de berekende afwijkingen overeenstemmen, worden ze opgeteld zolang $[g] < 1$, dus bij lage correlaties; en ze worden steeds meer gemiddeld, zodra $[g] > 1$, terwijl er zuiver van middelen sprake is zodra iedere g, dus iedere correlatie 1 is.

Tenslotte willen wij er nog op wijzen, dat het voor kan komen dat waarden u_{lp} die op grond van hoge correlaties zijn berekend elkaar tegenspreken. In dit geval zijn de correlatiecoëfficienten ten onrechte te hoog geschat, meestal wegens het toeval. Ook hier geldt dat het gemiddelde van de geschatte waarden u_{lp} dan naar verhouding klein wordt, zodat ook hier het *effectief* gewicht uiteindelijk laag is.

In deze paragrafen hebben wij dus een aantal voorschriften gegeven, die beter zijn dan het negeren van de correlatieverschijnselen. Maar de voorschriften zijn niet streng mathematisch afgeleid. Het is: bij gebrek aan beter.

313 – Onderzoek van de interpretatiefout

Tot dusver hebben wij veel gesproken over de interpretatiefout als functie van de proefveldinvloeden, maar wij hebben nog niet nagegaan, hoe het verband tussen omstandigheden en interpretatiefout was. Wel hebben wij door een correlatierekening soms trachten op te sporen hoe een interpretatiefout op een bepaald veld af kan hangen van onbekende omstandigheden, die een bekende uitwerking hebben op een ander ras.

Wanneer de omstandigheden van de proefvelden met name bekend zijn, kan getracht worden de interpretatiefout als functie van de met name bekende invloeden te zien. Hierbij moet het probleem besproken worden op welke wijze de interpretatiefout het gemakkelijkst kan worden uitgedrukt.

Wanneer met een orthogonaal schema kan worden gewerkt is de interpretatiefout eenvoudig te benaderen. Blijkens (301.24) is het systematisch deel van

$$u_{kp} = i_{kp} - \frac{[i_{k\pi}]}{n}.$$

Aangezien de vorm $\frac{[i_{k\pi}]}{n}$ gedurende het onderzoek van ras k constant is, en dus geen invloed heeft op de reactie van u_{kp} op de omstandigheden, kan men u_{kp} als beste benadering van i_{kp} zien. De asymptotische waarde van de toevallige fouten is natuurlijk ook 0. Men kan de waarden $u_{k\pi}$ dus gewoon bestuderen aan de hand van de bekende omstandigheden.

Bij onvolledige schema's is de werkwijze minder van zelfsprekend. Bij het standaardiseren zullen alle geschatte opbrengsten het best mee kunnen tellen. Het schatten heeft juist ten doel het samenvatten der proefvelden, en dus ook het standaardiseren te verbeteren.

Voor het verder onderzoek zijn de geschatte cijfers misschien niet altijd even waardevol. Het probleem is nu immers niet, hoe een aantal rassen onder dezelfde omstandigheden kunnen worden vergeleken, waarbij nergens een hiaat mag zijn; maar nu is de vraag, hoe de ene grootheid een functie is van de ander. Hierbij is het vaak niet belangrijk of er een gegeven meer of minder is.

Veronderstel b.v. dat men het verband tussen interpretatiefout en pH grafisch wil weergeven. Dan is het ideaal een nauwe puntenbundel te vinden. Een cijfer moet dus niet alleen statistisch juist zijn, maar ook een kleine toevallige fout hebben. Het is duidelijk dat de gewichten nu volgens geheel andere principes moeten worden toegekend, dan bij het samenvatten der proefvelden, en dat met name die schattingen, die een gewicht $g = 1$ kregen, omdat de u toch 0 was, bij een onderzoek, zoals nu bedoeld, vaak het gewicht 0 zullen moeten krijgen. Zo zullen de meeste schattingen waarschijnlijk waardeloos zijn.

Maar ook aan de waargenomen cijfers kleven moeilijkheden. Hoe minder rassen op een proefveld, hoe beter kan het fictiefras zich aan de voorhanden rassen aanpassen, dus des te kleiner interpretatiefout moet worden verwacht. Dit ongemak kan worden voorkomen door het onderzochte ras buiten het fictiefras te sluiten, een middel dat wij reeds vaker aanbevalen. Dit kan evenwel weer het bezwaar opleveren, dat ieder ras in verband met een ander fictiefras wordt beschouwd. Hier moet dus even op gelet worden.

Op welke wijze het verband tussen de interpretatiefouten en de omstandigheden nader moet worden onderzocht, willen we niet nagaan. Het is voldoende er op te wijzen dat de interpretatiefouten continue grootheden zijn, evenals de omstandigheden

meestal. Over de methodiek die voor dit soort problemen gewenst is zijn reeds vele onderzoekingen verricht, o.a. door Ir W. C. VISSER.

Het spreekt van zelf dat de milieugunstigheid M_p op dezelfde wijze tegen de omstandigheden kan worden afgezet als de interpretatiefouten. Als resultaat hiervan zal men een nader inzicht in de reacties van het *gewas* kunnen krijgen.

Voor zover de interpretatiefouten als functie van verschillende invloeden zijn begrepen, kunnen ze voor de omstandigheden worden gecorrigeerd. Het is niet uitgesloten dat na alle toe te passen correcties blijkt dat de interpretatiefouten nog niet geheel zijn verklaard. Bij het beoordelen van deze kwestie moet men bedenken, dat iedere correctie met een fout behept is en dus de toevallige fout van het nog verder te corrigeren cijfer vergroot. Men moet zich dus steeds afvragen, hoe groot de toevallige fout langzamerhand geworden is. Wanneer zeer zeker de afwijkingen nog groter zijn dan de vergrote toevallige fout, is op bepaalde invloeden niet gelet. Men staat dan voor de taak uit te zoeken welke invloeden dat geweest kunnen zijn. Bij dit zoeken kan het steun geven, wanneer men weet welke proefvelden op elkaar gelijken, wat de reactie van de rassen betreft. Van zulke proefvelden kan men nagaan welke eigenschappen ze gemeen hebben.

Om n proefvelden met ieder m rassen in een aantal groepen verwante velden in te delen, zou men als volgt te werk kunnen gaan. Men rangschikt de velden volgens opklimmende waarde van u_{1p} (de afwijking van ras k_1 op de diverse velden). De n proefvelden laten zich zo in a klassen verdelen met ieder ongeveer evenveel velden.

Binnen iedere klasse rangschikt men dan de velden naar opklimmende waarde van u_{2p}. Zo laat iedere klasse zich weer in a klassen onderverdelen. Men heeft dan a^2 klassen. Met de m rassen vindt men zo a^{m-1} klassen. (De interpretatiefouten van het m^e ras zijn volledig afhankelijk van de eerste $(m - 1)$).

In iedere klasse moeten minstens 2 velden voorkomen, het hoogst toelaatbare aantal klassen is dus $\frac{n}{2}$. Uit de formule

$$a^{m-1} < \frac{n}{2}$$

laat zich nu berekenen hoe groot a hoogstens gekozen mag

worden. Het spreekt van zelf dat deze formule slechts als oriënterend is bedoeld.

Misschien is het soms ook mogelijk op dezelfde manier als in (305.1) de *gestandaardiseerde* (zie begin van 309) opbrengsten te verwerken en met behulp van de gevonden grootheden r_{aa}, r_{ab} enz. correlatiecijfers tussen de verschillende proefvelden te berekenen. Wanneer iemand deze mogelijkheid practisch wil toepassen moet hij vooraf bedenken dat het aantal onbekenden gelijk is aan de som van het aantal proefvelden en het aantal rassen, min 1. Wanneer hij niet over benaderingsmethoden beschikt om zulke grote vergelijkingen op te lossen is het meestal practisch onuitvoerbaar.

IV – HET BEREKENEN VAN VRUCHTBAARHEIDS-CORRECTIES OP AFZONDERLIJKE PROEFVELDEN

401 – De aard van de vruchtbaarheidscorrectie

In het vorig hoofdstuk hebben wij gesproken over het samenvatten van de resultaten van verschillende proefvelden. Deze proefvelden konden in milieugunstigheid M_p aanzienlijk uit elkaar lopen. Door deze verschillen had de rassenvergelijking erg kunnen worden bemoeilijkt, wanneer wij niet in 202 reeds het begrip „rekenopbrengst" hadden ingevoerd, dat deze moeilijkheden grotendeels wegnam (zie formule 202.6). Dank zij dit begrip kan de conclusie van hfdst. III alsvolgt in het kort worden weergegeven: Bij niet-orthogonale schema's moeten de proefvelden op elkaar worden gestandaardiseerd, om de juiste rasgemiddelden te berekenen (zie b.v. 304); bij orthogonale schema's is dit niet nodig (zie 301.5). Wanneer de *juiste reken*opbrengsten worden gebruikt, komt bij orthogonale schema's een extra gunstig milieu alle rassen evenveel ten goede, een extra ongunstig milieu geeft alle rassen evenveel nadeel. Het heeft in dat geval dus bij orthogonale schema's geen invloed op de berekende rasverschillen of op de verschillen in milieugunstigheid wordt gelet.

Toch worden deze verschillen gewoonlijk in het onderzoek betrokken. Hun aanwezigheid veroorzaakt dat er een grote toevallige fout wordt berekend, wanneer niet nagegaan wordt in hoeverre deze berekende fout aan milieu-omstandigheden moet worden geweten. Het rekenen met de milieugunstigheid heeft in dit geval dus niet het doel, de rasverschillen beter te leren kennen, maar de berekende fout kleiner te maken, en zo een gunstiger indruk te krijgen van de betrouwbaarheid van het resultaat.

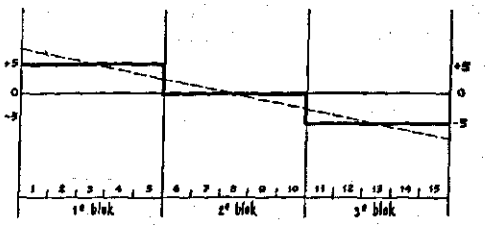
De methode, die aan het gehele proefveld één gunstigheid M_p toekent, is in dit geval ook zeer voor de hand liggend. De verschillen in milieugunstigheid worden veroorzaakt, doordat het ene proefveld in het Noorden lag, een ander in het Zuiden; het ene werd vroeg gezaaid, een ander laat; het ene werd goed bemest, een ander matig, het ene had een goede ontwatering en een ander een slechte. Kortom, de omstandigheden van de verschillende proefvelden kunnen zozeer uiteenlopen, dat het vanzelfsprekend is, een proefveld als een eenheid te zien.

Dezelfde rekentechniek, die gebruikt wordt voor het samenvatten van proefvelden, wordt vaak aangewend bij het verwerken van de resultaten van één proefveld, wanneer dat proefveld van alle objecten enkele herhalingen heeft. Een zeer gebruikelijke aanleg verdeelt het proefveld in *n* blokken en plaatst in ieder blok de *m* objecten (b.v. rassen) eenmaal. Wanneer er gunstigheidsverschillen in het veld zijn, zal het ene blok het waarschijnlijk iets gunstiger treffen dan het andere. De meerdere of mindere gunstigheid van een blok kan nu op dezelfde wijze in rekening worden gebracht als de meerdere of mindere gunstigheid van een bepaald proefveld. En ook nu zal het resultaat zijn, dat de berekende rasverschillen niet veranderen, doch dat de berekende fout kleiner wordt.

Voor het verwerken van één proefveld is bovenstaande methode ongetwijfeld toelaatbaar, maar niet vanzelfsprekend. Binnen een proefveld zijn er maar weinig oorzaken van verschillen in milieugunstigheid. Het zijn de oorzaken die samengevat worden in het begrip *vruchtbaarheid*. Om de gedachten te bepalen neme men aan dat men een strokenproef heeft met 3 blokken en 5 rassen. In zo'n proef komen dus 15 veldjes voor, die op één rij liggen. De eerste vijf veldjes vormen een blok, de tweede en derde vijf eveneens. Wanneer nu uit de berekening blijkt dat het eerste blok per veldje 5 eenheden vruchtbaarder is dan het gemiddelde, dat het tweede blok de gemiddelde vruchtbaarheid heeft, en het derde blok 5 eenheden beneden het gemiddelde ligt, dan kan inderdaad met deze cijfers als „blokgunstigheid" worden gerekend.

Het ligt evenwel meer voor de hand uit bovenstaande gegevens af te leiden, dat de vruchtbaarheid regelmatig 5 eenheden daalt per 5 veldjes en dus 1 per veldje. Dit geeft de mogelijkheid, aan de hand van *dezelfde* rekenresultaten een vruchtbaarheidscorrectie binnen de blokken toe te passen. In figuur (401.1) zijn beide correcties in beeld gebracht.

(401.1)



Wanneer men voor ieder blok een bepaald vruchtbaarheidsniveau aanneemt, krijgt men een vruchtbaarheidsverloop, zoals door de trapjeslijn is aangegeven. Wanneer men ook binnen de blokken corrigeert vindt men een vruchtbaarheidsverloop, zoals de stippellijn aangeeft.

Voor de proefveldverwerking zijn beide methoden geoorloofd, maar wij kunnen ons toch niet aan de indruk onttrekken, dat de methode, die binnen de blokken corrigeert, natuurlijker is. De indeling in blokken heeft een kunstmatig karakter, dat niet door de proefveldomstandigheden wordt gerechtvaardigd. Het is onwaarschijnlijk dat veldje no. 5 even vruchtbaar is als veldje no. 1, doch veel vruchtbaarder dan veldje no. 6. Een blok is slechts een kunstmatige eenheid, en wanneer men bij een vruchtbaarheidscorrectie met zulke kunstmatige eenheden werkt, kan men niet anders dan tamelijk ruwe correcties toepassen.

In de termen van R. A. FISHER uitgedrukt kunnen wij zeggen dat een vruchtbaarheidscorrectie, die slechts met blokverschillen (subblokverschillen) werkt, wel „consistent", maar niet „efficient" is.

Dit gebrek aan „efficiency" uit zich niet alleen hierin, dat de vruchtbaarheid slechts globaal wordt uitgeschakeld. Een tweede bezwaar is, dat er vaak te veel vrijheidsgraden worden gebruikt. In voorbeeld (401.1) zijn er 2 vrijheidsgraden nodig om met de blokverschillen te rekenen, terwijl met 1 vrijheidsgraad kan worden volstaan om vast te leggen dat de vruchtbaarheid 1 eenheid per veldje daalt.

Zo heeft men b.v. bij een Latijns vierkant met 8 objecten (en dus 8 rijen en 8 kolommen) 14 vrijheidsgraden nodig om de vruchtbaarheidsverschillen van de rijen en de kolommen vast te leggen, dat is meer dan één per vijf veldjes. Het is duidelijk dat de vruchtbaarheidscorrectie niet veel zou veranderen, wanneer de vruchtbaarheid van de 1e, de 3e, de 5e en de 7e kolom berekend werd, terwijl de andere vruchtbaarheden hiertussen werden geïnterpoleerd. Wanneer men hetzelfde voor de rijen deed, zou in totaal met 8 vrijheidsgraden kunnen worden volstaan. Nu willen wij niet graag beweren dat de vrijheidsgraden dan wel efficient waren gebruikt, maar het zou al een hele besparing zijn.

Wij willen nu dus nagaan hoe we een vruchtbaarheidscorrectie aan moeten vatten om zo natuurlijk en dus zo efficient mogelijk

te werken *). Daartoe moet men allereerst letten op het veld, waarop de proef genomen wordt. Dit veld heeft bijna altijd een geologische wordingsgeschiedenis achter de rug, die maakt dat er allerlei banen en uitzonderlijke hoeken in aanwezig zijn, meestal in de ondergrond. Er kunnen b.v. door het veld sporen lopen van allerlei kreken met hun oeverwallen. Dit kan op korte afstand soms grote verschillen in vruchtbaarheid geven. Vruchtbaarheidsverschillen van 10% op 20 meter afstand zijn geen zeldzaam verschijnsel.

Wanneer men een dergelijk veld als proefveld gebruiken, moet men op twee dingen letten:

In de eerste plaats zijn de veranderingen van vruchtbaarheid in de regel continu, er kan een vruchtbaarheidskaart getekend worden zonder plotselinge overgangen. Op de enkele uitzonderingen op deze regel willen wij niet ingaan.

In de tweede plaats zijn de vormen van de kreken en zandruggen zelden in wiskundige formules te vangen. Het verloop van een zandrug of andere eigenaardigheid moet aan zekere wetten beantwoorden, maar die wetten zijn er niet op gericht, met weinig constanten beschreven te kunnen worden.

Wanneer men b.v. ergens een strokenproef zou aanleggen met 100 veldjes in één rij, dan zou de vruchtbaarheidskaart hoogstwaarschijnlijk gekenmerkt worden door een aantal maxima en minima. Laten wij nu aannemen dat de minima worden veroorzaakt door een kreek en de maxima door een zandrug, dan is het geologisch duidelijk dat het vruchtbaarheidsverloop van de ene kant van een kreek samenhangt met het verloop aan de andere kant. Een kreek moet nu eenmaal twee zijden hebben. Maar het is evenzeer duidelijk dat het verband tussen de beide oevers tamelijk los is. Wanneer iemand de ene oever kent, weet hij nog de breedte van de kreek niet, en dus ook niet de plaats van de andere oever. Zo is ook de gedetailleerde vorm van die oever niet af te leiden.

Wij kunnen zo verder gaan en zeggen dat de plaats van de ene kreek geen conclusies toelaat over de plaats van de volgende. Kortom, uit de vruchtbaarheid op een bepaald veldje valt weinig af te leiden over de vruchtbaarheid van een ander veldje dat b.v. 30 m verder ligt.

*) Aan het eind van 404 wordt er op gewezen dat de efficiëntie van de hier gegeven methode ook tegen kan vallen.

Wij kunnen dus zeggen: De vruchtbaarheid van dicht bij elkaar gelegen veldjes hangt ten nauwste samen, omdat de vruchtbaarheid continu verloopt, maar de vruchtbaarheid van verder uit elkaar liggende veldjes heeft weinig samenhang, omdat een geologische detailkaart weinig extrapolaties toelaat.

Bovengenoemde eigenaardigheden van het veld zijn bepalend voor de wijze van vruchtbaarheidscorrectie. Wij wezen er reeds op dat die correctiesystemen, die de vruchtbaarheid als een discontinue grootheid zien, gewoonlijk niet de meest natuurlijke zijn. Maar er zijn ook bezwaren verbonden aan die methoden, die de continuiteit van de vruchtbaarheid erkennen, en daarom de vruchtbaarheidskaart in de vorm van een formule geven, die de vruchtbaarheid geeft als functie van de plaats op het veld. Het is een wezenlijke eigenschap van een formule, dat hij geëxtrapoleerd kan worden; het is even wezenlijk voor het proefveld, dat extrapoleren slechts op korte afstand toelaatbaar is.

Wanneer men een formule berekent, die drie maxima en minima moet beschrijven, dan zullen de constanten, die het eerste maximum beschrijven, dezelfde zijn als diegene die het derde vastleggen. Door het gebruiken van een formule wordt het eerste maximum dus geëxtrapoleerd naar het derde en omgekeerd. Hiermee wordt het probleem geweld aangedaan.

De enige oplossing, die natuurlijk is, is een grafische bewerking. Bij het tekenen van grafieken wordt op korte afstand de continuiteit gehandhaafd, terwijl er geen noodzaak is over grotere afstand te extrapoleren. Grafische bewerkingsmethoden worden dan ook reeds lang toegepast.

Als bezwaar van de grafische verwerking wordt vaak gevoeld, dat het moeilijk is, het overzicht over de vrijheidsgraden en fouten te houden. Om deze reden willen wij in het vervolg van dit hoofdstuk nagaan, welke mogelijkheden er zijn om bij grafische correctie bedoeld overzicht te houden. Om dit te bereiken worden de grafieken voorlopig in een bepaalde wiskundige vorm gegoten, nl. die van het verschuivend gemiddelde.

402 – Nadere bepaling van de gewenste methode

In 202 zijn wij gekomen tot de formule

$$\Omega_{kp} = R_k + M_p + C. \tag{202.6}$$

In 301 hebben wij deze formule uitgebreid tot de vorm

(301.8) $$\Omega_{kp} = <\bar{\bar{\Omega}}> + \hat{\varrho}_k + \hat{\mu}_p + t_{kp}.$$

De vermelde t_{kp} werd in (301.18) nog gesplitst in $i_{kp} + \tau_{kp}$. Wij zouden (301.8) dus kunnen schrijven in de vorm

(402.1) $$\Omega_{kp} = <\bar{\bar{\Omega}}> + \hat{\varrho}_k + \hat{\mu}_p + i_{kp} + \tau_{kp}.$$

Wanneer we de p nu niet laten slaan op een proefveld, doch op een bepaald blok van het onderzochte proefveld, dan moet $\hat{\mu}_p$ aanduiden, hoeveel dit blok in vruchtbaarheid van de gemiddelde vruchtbaarheid afwijkt, terwijl i_{kp} aangeeft, in hoeverre de opbrengst Ω_{kp} van ras k in blok p systematisch van de som $<\bar{\bar{\Omega}}> + \hat{\varrho}_k + \hat{\mu}_p$ verschilt. Aangezien binnen het proefveld verondersteld mag worden, dat de enige systematische invloed in de vruchtbaarheidsverschillen ligt, duidt i_{kp} dus aan, in hoeverre de vruchtbaarheid van het veldje waarop ras k verbouwd werd van het blokgemiddelde afwijkt.

Omdat i_{kp} hier een veel beperkter betekenis heeft dan in het vorig hoofdstuk, willen wij hem vervangen door het symbool λ_{kp}. Het veldje, waarop λ_{kp} betrekking heeft, noemen wij in het vervolg veldje (kp).

Het is niet noodzakelijk λ_{kp} beslist als een *fout* te beschouwen. Wanneer men er naar streeft de vruchtbaarheid in rekening te brengen door de waarden μ en λ gezamenlijk, dan is de λ een deel van de benodigde formule. Aangezien deze formule in principe in staat is het vruchtbaarheidsprobleem op te lossen, is het niet meer nodig van een *noodformule* te spreken. Hier is veeleer sprake van een *ideale formule* (zie 106).

Slechts wanneer de vruchtbaarheidskaart voor het ene ras anders zou uitvallen, dan voor het andere, zou men van een noodformule moeten spreken, wanneer men toch maar één kaart tekende. Dit komt evenwel weinig voor; tengevolge van ons begrip „rekenopbrengst". Daarom willen wij in het vervolg aannemen, dat er met een ideale formule gewerkt wordt.

Dit heeft tot gevolg dat de asymptotische waarden $\hat{\varrho}_k$ en $\hat{\mu}_p$ als „ware waarden" mogen worden opgevat. Wij willen ze daarom gewoon aanduiden als ϱ_k en μ_p. Daarentegen moet de schrijfwijze $<\bar{\bar{\Omega}}>$ worden gehandhaafd omdat de vorm $\bar{\bar{\Omega}}$ steeds besmet zal zijn met toevallige fouten.

Formule (402.1) lezen wij nu dus als volgt

(402.2) $$\Omega_{kp} = <\bar{\bar{\Omega}}> + \varrho_k + \mu_p + \lambda_{kp} + \tau_{kp}.$$

Op grond van (301.23) en de discussie na (301.9) geldt nu, in verband met het feit dat er geen monsterfouten mogelijk zijn, bij een ideale formule,

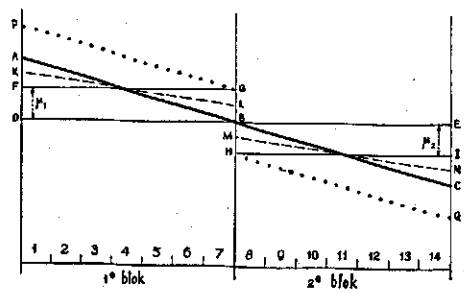
(402.3)
$$[\varrho_\kappa] = 0$$
$$[\mu_\pi] = 0$$
$$[\lambda_{\kappa p}] = 0$$
$$[[\lambda_{\kappa\pi}]] = 0;$$

terwijl $[\lambda_{k\pi}]$ niet 0 behoeft te zijn.

Van de constanten die in (402.2) genoemd worden heeft ϱ_k betrekking op de rasinvloed, τ_{kp} slaat op de toevallige fout van veldje (kp), terwijl zijn vruchtbaarheid wordt vertegenwoordigd door de waarden $<\bar{\bar{\Omega}}>$, μ_p en λ_{kp}.

Wij willen de betekenis van deze symbolen door een tekening duidelijk maken. Daartoe stellen wij dat wij een strokenproef bestuderen, waarbij de veldjes gelegen zijn op de manier als in (402.4) in beeld gebracht. Alle blokken zijn 1 veldje breed en liggen achter elkaar. Wij willen nu met name veronderstellen, dat wij twee blokken hebben, met in ieder blok 7 rassen. Het vruchtbaarheidsverloop is in fig. (402.4) door de lijn AC weergegeven.

(402.4)



Wanneer wij de opbrengst van het fictiefras op een bepaald veldje als de vruchtbaarheid van dat veldje beschouwen, kunnen wij in verband met (402.3) zeggen dat $<\bar{\bar{\Omega}}>$ de gemiddelde vruchtbaarheid van het gehele veld is. Deze gemiddelde vruchtbaarheid is weer te geven door de lijn DE. De waarde μ_1 geeft aan hoeveel de gemiddelde vruchtbaarheid van het eerste blok van het totaal gemiddelde afwijkt. Laten wij deze gemiddelde vruchtbaarheid voorstellen door de lijn FG, dan is de loodrechte afstand tussen FG en DE gelijk aan μ_1. Zo kunnen wij de gemiddelde vruchtbaarheid van het tweede blok weergeven door HI, die op de af-

stand μ_2 van DE is verwijderd. De waarden λ_{kp} geven voor ieder veldje de vertikale afstand tussen AB en FG in het eerste blok, en tussen BC en HI in het tweede.

Wanneer wij de opbrengst Ω_{kp} van een bepaald veldje grafisch voor willen stellen in fig. (402.4) dan zal deze voorstelling niet op de vruchtbaarheidslijn AC liggen, doch daar een bedrag ϱ_k van verwijderd zijn. (Omdat we in deze paragraaf de methodiek bestuderen willen wij aannemen dat er geen toevallige fouten zijn).

Hoe kan nu uit de opbrengstcijfers Ω_{kp} de lijn AC worden berekend? Het is duidelijk dat daartoe de invloed van ϱ_k moet worden uitgeschakeld. Om dit te bereiken middelen wij (402.2) over p. Wanneer wij tevens de toevallige fouten onvermeld laten, geeft dit

$$\frac{\uparrow n\,[\Omega_{k\pi}]}{n} = <\bar{\bar{\Omega}}> + \varrho_k + \frac{\uparrow n\,[\mu_\pi]}{n} + \frac{\uparrow n\,[\lambda_{k\pi}]}{n}. \tag{402.5}$$

Substitutie van ϱ_k uit (402.5) in (402.2) geeft in verband met (402.3), en onder weglating van de toevallige fout,

(402.6)

$$\Omega_{kp} - \frac{\uparrow n\,[\Omega_{k\pi}]}{n} = \mu_p + \lambda_{kp} - \frac{\uparrow n\,[\lambda_{k\pi}]}{n} = \mu_p + \frac{n-1}{n}\lambda_{kp} - \frac{\uparrow (n-1)[\lambda_{k[p]}]}{n}.$$

Het is de bedoeling van (402.6) de waarde van $\mu_p + \lambda_{kp}$ te berekenen, maar er ontbreekt nog het een en ander aan. Allereerst is daar de term

$$\frac{\uparrow (n-1)\,[\lambda_{k[p]}]}{n},$$

die storend werkt. Van deze term zullen wij de waarde moeten bepalen.

Aangezien volgens (402.3) voor ieder blok geldt $[\lambda_{\kappa p}] = 0$, moet de waarde van λ in ieder blok om 0 schommelen. Dit schommelen kan zeer systematisch plaats vinden. In fig. (402.2) is λ aan het begin van ieder blok sterk positief, met het opklimmen van de veldjesnummers daalt de waarde van λ tot hij sterk negatief is.

Om nu toch de waarde van

$$\frac{\uparrow (n-1)\,[\lambda_{k[p]}]}{n}$$

te kunnen bepalen moet dit systematisch verloop worden omzeild. Dit kan bereikt worden door in ieder blok de rasvolgorde door het toeval te laten bepalen. Het verloop van λ mag dan

systematisch zijn binnen een blok, maar het is toevallig welke λ uit dit blok in de vorm

$$\frac{\uparrow(n-1)[\lambda_{k\lfloor p\rfloor}]}{n}$$

wordt opgenomen. Uit het ene blok zal een positieve λ worden genomen en uit het ander een negatief. Bij gevolg mag de asymptotische waarde van de onderzochte vorm gelijk aan 0 worden gesteld wanneer aan de eis van toevallige verdeling is voldaan. Op grond hiervan kunnen wij (402.6) dus bij benadering schrijven als

$$\Omega_{kp} - \frac{\uparrow n\,[\Omega_{k\pi}]}{n} = \mu_p + \frac{n-1}{n}\,\lambda_{kp}. \tag{402.7}$$

De waarden die volgens (402.7) worden berekend zijn in figuur (402.4) weergegeven door de stippellijnen KL en MN. Door dit resultaat met ABC te vergelijken blijkt dat de gevonden lijn twee gebreken heeft: de helling is verkeerd en er is een plotselinge overgang op de blokgrens.

Van deze gebreken is de helling het gemakkelijkst te verbeteren. Daartoe schrijven wij (402.7) in de vorm

$$\Omega_{kp} - \frac{\uparrow n\,[\Omega_{k\pi}]}{n} = \frac{n-1}{n}\,\Omega_{kp} - \frac{\uparrow(n-1)[\Omega_{k\lfloor p\rfloor}]}{n} = \mu_p + \frac{n-1}{n}\,\lambda_{kp}.$$

Door beide leden met $\frac{n}{n-1}$ te vermenigvuldigen vinden wij

$$\Omega_{kp} - \frac{\uparrow(n-1)[\Omega_{k\lfloor p\rfloor}]}{n-1} = \frac{n}{n-1}\,\mu_p + \lambda_{kp}. \tag{402.8}$$

De waarden die volgens (402.8) worden berekend, zijn in fig. (402.4) aangegeven door de kruisjeslijnen PG en HQ. Het blijkt dat de helling van de lijnen nu goed is, maar dat de plotselinge overgang op de blokgrens nog erger is geworden.

De moeilijkheid waarvoor wij staan komt scherp naar voren door (402.7) met (402.8) te vergelijken. Volgens de eerste formule wordt de μ_p goed berekend, volgens de tweede de λ_{kp}. De μ_p en de λ_{kp} gehoorzamen dus niet aan dezelfde wetten.

Zowel (402.7) als (402.8) vergelijken de opbrengst Ω_{kp} met de gemiddelde opbrengst van ras k. Om μ_p goed te berekenen *moet* Ω_{kp} zelf in dit gemiddelde zijn opgenomen; om λ_{kp} goed te berekenen *mag* Ω_{kp} *niet* in dit gemiddelde opgenomen zijn. Wat is de oorzaak hiervan?

Bij het bespreken van (402.6) is gebleken dat de vorm

$$\frac{\uparrow(n-1)[\lambda_{k\lfloor p\rfloor}]}{n}$$

asymptotisch de waarde 0 heeft. Dit is het gevolg van een kansverdeling waaraan iedere $\lambda_{k\lfloor p\rfloor}$ is gebonden. Wij zouden kunnen zeggen dat iedere $\lambda_{k\lfloor p\rfloor}$ krachtens zijn aard een inhaerente kansverdeling heeft. Op grond daarvan kunnen wij ook zeggen, dat iedere $\lambda_{k\lfloor p\rfloor}$ asymptotisch 0 wordt. Wanneer wij aannemen dat ras k in fig. (401.1) of fig. (402.4) op veldje no. 2 heeft gestaan, en wanneer wij in verband met deze aanname de vruchtbaarheid van veldje 2 onderzoeken, dan weten wij nog niets over de waarde van $\lambda_{k\lfloor 1\rfloor}$ in de andere blokken, omdat de ligging van ras k in die andere blokken onafhankelijk is van de ligging van ras k op veldje 2. Wij weten slechts dat $\lambda_{k\lfloor 1\rfloor}$ in waarde om 0 schommelt en asymptotisch 0 is.

Daarentegen mogen wij λ_{k1} niet asymptotisch 0 stellen omdat bekend is dat λ_{k1} op de vruchtbaarheid van veldje 2 betrekking heeft. Om de vruchtbaarheid van veldje 2 te kunnen berekenen moeten wij dus λ_{k1}, belichaamd in Ω_{k1}, buiten het rasgemiddelde laten (zie 402.8).

Met μ_p ligt de zaak heel anders. Blijkens (402.3) is $[\mu_\pi] = 0$. Hieruit volgt

(402.9) $$\mu_p = -\uparrow(n-1)[\mu_{\lfloor p\rfloor}].$$

Wanneer μ_p niet 0 is, is de vorm $-\uparrow n(-1)[\mu_{\lfloor p\rfloor}]$ het ook niet. Om μ_p goed te berekenen moeten dus alle waarden μ_π, belichaamd in $\Omega_{k\pi}$, in het rasgemiddelde opgenomen worden (zie 402.7). Bij het bestuderen van (402.5) hebben wij van de eigenschap $[\mu_\pi] = 0$ gebruik gemaakt.

(402.10) Eenvoudig uitgedrukt kunnen wij zeggen, *dat een afwijking van een gemiddelde slechts asymptotisch* 0 *kan worden, wanneer zij bewegingsvrijheid heeft over het gebied, waarvoor het gemiddelde berekend is.*

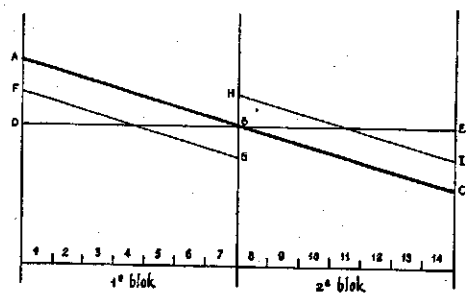
Zo geven de waarden μ de afwijkingen aan van het proefveldgemiddelde. Wanneer met formule (402.8) gewerkt wordt hebben deze waarden geen bewegingsvrijheid over het gehele proefveld, omdat ze bij het bepalen van het gemiddelde in de blokken $\lfloor p\rfloor$ moeten blijven. Formule (402.8) is voor het bepalen van μ_p dus ongeschikt.

Daarentegen zijn de waarden λ afwijkingen van het blokge-

middelde. Omdat de plaatsen van een bepaald ras in de diverse blokken onderling onafhankelijk zijn, heeft iedere λ volledige bewegingsvrijheid, behalve de λ in het blok dat onderzocht wordt. Wanneer de λ van veldje 2 wordt onderzocht, is λ_{k1} aan veldje 2 gebonden. Deze λ moet dus buiten het rasgemiddelde worden gehouden. Nu is (402.8) dus wel goed.

Uit het bovenstaande is duidelijk dat we μ en λ tijdens de berekening van elkaar moeten scheiden. Hoe dit het best kan, willen wij uitleggen aan de hand van fig. (402.11), waarin enkele gegevens van (402.4) zijn overgenomen.

(402.11)



De lijn AC geeft weer het vruchtbaarheidsverloop met het gemiddelde DE. Omdat μ van λ moet worden gescheiden, hebben wij de lijnstukken AB voor μ_1 en BC voor μ_2 gecorrigeerd. Dit gaf als resultaat FG en HI. Wanneer wij nu formule (402.8) toepassen omdat λ nog gezocht moet worden, zullen we de lijnstukken FG en HI berekenen.

Hier blijft dus een kunstmatige plotselinge overgang op de blokgrens voorhanden. Deze is gemakkelijk te verwijderen, door bij de berekende waarden de correctie voor μ weer ongedaan te maken. Dan wordt de lijn AC teruggevonden.

Dit summiere overzicht zal in de volgende paragraaf worden uitgewerkt en verduidelijkt, terwijl dan tevens de gemaakte fouten worden bestudeerd.

403 – Het berekenen van de vruchtbaarheidslijn

In de vorige paragraaf hebben wij gesproken over de formule

(402.2) $$\Omega_{kp} = <\bar{\bar{\Omega}}> + \varrho_k + \mu_p + \lambda_{kp} + \tau_{kp}.$$

Van deze formule moeten de waarden $<\bar{\bar{\Omega}}>$, μ_p en λ_{kp} de vruchtbaarheidskaart omschrijven. Wij willen dus trachten deze waarden te berekenen.

Het spreekt vanzelf dat de asymptotische waarde van de gemiddelde opbrengst niet is te vinden. Wij moeten volstaan met de empirische waarde ervan. Deze kan verkregen worden door (402.2) over k en p te middelen. Dit geeft

$$\bar{\bar{\Omega}} = \frac{\uparrow\uparrow mn\,[[\Omega_{\kappa\pi}]]}{mn} = <\bar{\bar{\Omega}}> + \frac{\uparrow m\,[\varrho_\kappa]}{m} + \frac{\uparrow n\,[\mu_\pi]}{n} + \\ + \frac{\uparrow\uparrow mn\,[[\lambda_{\kappa\pi}]]}{mn} + \frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

In verband met (402.3) kan dit vereenvoudigd worden tot

$$\bar{\bar{\Omega}} = <\bar{\bar{\Omega}}> + \frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}. \tag{403.1}$$

Substitutie van $<\bar{\bar{\Omega}}>$ uit (403.1) in (402.2) geeft

$$\Omega_{kp} - \bar{\bar{\Omega}} = \varrho_k + \mu_p + \lambda_{kp} + \tau_{kp} - \frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}. \tag{403.2}$$

De tweede constante die berekend moet worden is μ_p. Het spreekt vanzelf dat ook hier niet de ware waarde is te vinden, doch slechts de empirische $\bar{\mu}_p$. Uit (403.2) laat zich deze vorm berekenen door over k te middelen. Dit geeft

$$\bar{\mu}_p = \frac{\uparrow m\,[\Omega_{\kappa p}]}{m} - \bar{\bar{\Omega}} = \frac{[\varrho_\kappa]}{m} + \mu_p + \frac{[\lambda_{\kappa p}]}{m} + \frac{\uparrow m\,[\tau_{\kappa p}]}{m} - \frac{\uparrow\uparrow mn[[\tau_{\kappa\pi}]]}{mn}.$$

In verband met (402.3) kan dit worden vereenvoudigd tot

$$\bar{\mu}_p = \mu_p + \frac{\uparrow m\,[\tau_{\kappa p}]}{m} - \frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}. \tag{403.3}$$

Substitutie van μ_p uit (403.3) in (403.2) geeft

$$\Omega_{kp} - \bar{\bar{\Omega}} - \bar{\mu}_p = \varrho_k + \lambda_{kp} + \tau_{kp} - \frac{\uparrow m\,[\tau_{\kappa p}]}{m}. \tag{403.4}$$

In het vervolg willen wij stellen

$$\begin{aligned} \omega_{kp} &= \Omega_{kp} - \bar{\bar{\Omega}} \\ \omega'_{kp} &= \omega_{kp} - \bar{\mu}_p; \end{aligned} \tag{403.5}$$

zodat (403.4) de vorm krijgt

$$\omega'_{kp} = \varrho_k + \lambda_{kp} + \tau_{kp} - \frac{\uparrow m\,[\tau_{\kappa p}]}{m}. \tag{403.6}$$

Nu moet λ_{kp} worden berekend. Naar analogie van (402.5) en (402.6) schakelen wij daartoe eerst ϱ_k uit. Daartoe middelen wij (403.6) over p. Dit geeft

$$(403.7)\quad \frac{\uparrow n\,[\omega'_{k\pi}]}{n} = \varrho_k + \frac{\uparrow n\,[\lambda_{k\pi}]}{n} + \frac{\uparrow n\,[\tau_{k\pi}]}{n} - \frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

Substitutie van ϱ_k uit (403.7) in (403.6), geeft

$$(403.8)\quad V'_{kp} = \omega'_{kp} - \frac{\uparrow n\,[\omega'_{k\pi}]}{n} = \lambda_{kp} - \frac{\uparrow n\,[\lambda_{k\pi}]}{n} + \\ + \tau_{kp} - \frac{\uparrow m\,[\tau_{\kappa p}]}{m} - \frac{\uparrow n\,[\tau_{k\pi}]}{n} + \frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

In bovenstaande formule zijn in het rechterlid de vormen μ uitgeschakeld, en de vormen λ aanwezig. In overeenstemming met formule (402.8) moeten beide leden van de vergelijking nu met $\frac{n}{n-1}$ worden vermenigvuldigd. Wij schrijven daartoe (403.8) eerst in de vorm

$$V'_{kp} = \frac{n-1}{n}\lambda_{kp} - \frac{\uparrow(n-1)\,[\lambda_{k\lfloor p\rfloor}]}{n} + \frac{n-1}{n}\tau_{kp} - \\ - \frac{\uparrow(n-1)\,[\tau_{k\lfloor p\rfloor}]}{n} - \frac{\uparrow m\,[\tau_{\kappa p}]}{m} + \frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

Door beide leden met $\frac{n}{n-1}$ te vermenigvuldigen vinden wij nu

$$(403.9)\quad F'_{kp} = \frac{n}{n-1}V'_{kp} = \lambda_{kp} - \frac{\uparrow(n-1)\,[\lambda_{k\lfloor p\rfloor}]}{n-1} + \tau_{kp} - \\ - \frac{\uparrow(n-1)\,[\tau_{k\lfloor p\rfloor}]}{n-1} - \frac{n}{n-1}\,\frac{\uparrow m\,[\tau_{\kappa p}]}{m} + \frac{n}{n-1}\,\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

Blijkens het voorschrift aan het eind van 402 moet de waarde $\bar{\mu}_p$, die in (403.4) van het linkerlid van de vergelijking werd afgetrokken, er nu weer bij worden opgeteld. Uit (403.3) en (403.9) berekenen wij

$$(403.10)\quad F_{kp} = F'_{kp} + \bar{\mu}_p = \mu_p + \lambda_{kp} - \frac{\uparrow n-1\,[\lambda_{k\lfloor p\rfloor}]}{n-1} + \tau_{kp} - \\ - \frac{\uparrow n-1\,[\tau_{k\lfloor p\rfloor}]}{n-1} - \frac{1}{n-1}\,\frac{\uparrow m\,[\tau_{\kappa p}]}{m} + \frac{1}{n-1}\,\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

In deze formule komen μ_p en λ_{kp} op de juiste wijze voor, maar ze zijn nog niet gescheiden van de toevallige fout τ_{kp}. Men weet dus nog niet of de grootheid F_{kp}, die men voor veldje (kp) berekent, toegeschreven moet worden aan toeval of aan vruchtbaarheidsinvloeden.

Over deze vraag zijn inlichtingen te verkrijgen door gebruik te maken van de eigenschap dat vruchtbaarheidsverschillen continu veranderen. Wanneer wij ons, om de gedachten te bepalen, weer voorstellen dat wij te maken hebben met een strokenproef, kunnen wij de continuïteit in formule brengen, door aan te nemen dat het vruchtbaarheidsverloop over de korte afstand van g veldjes als lineair mag worden beschouwd. Door deze aanname wordt onze formule een noodformule. Wij zullen de g steeds oneven nemen, zodat de gemiddelde vruchtbaarheid van de g veldjes gelijk is aan de vruchtbaarheid van het middelste veldje.

Wanneer het middelste veldje (kp) is, willen wij de g veldjes aanduiden als (k^*p). De verbouwde rassen noemen we ${}_pk^*$, met bijbehorende lopende index ${}_p\varkappa^*$. De p in het symbool herinnert eraan, dat de rassen ${}_p\varkappa^*$ slechts in het blok p naast elkaar verbouwd werden op een rijtje van g veldjes (k^*p); in de andere blokken zijn deze rassen volgens toeval verspreid.

Op grond van onze aanname dat het vruchtbaarheidsverloop op korte afstand lineair is kunnen wij nu stellen

$$\frac{\uparrow g\,[\lambda_{p\varkappa^*p}]}{g} = \lambda_{kp}. \tag{403.11}$$

Wanneer wij (403.10) over de g veldjes (k^*p) middelen, vinden wij (in verband met (403.11))

$$\begin{aligned} P_{kp} = \frac{\uparrow g\,[F_{p\varkappa^*p}]}{g} &= \mu_p + \lambda_{kp} - \\ &- \frac{\uparrow\uparrow g\,(n-1)\,[[\lambda_{p\varkappa^*\lfloor p\rfloor}]]}{g\,(n-1)} + \frac{\uparrow g\,[\tau_{p\varkappa^*p}]}{g} - \\ &- \frac{\uparrow\uparrow g\,(n-1)\,[[\tau_{p\varkappa^*\lfloor p\rfloor}]]}{g\,(n-1)} - \frac{\uparrow m\,[\tau_{\varkappa p}]}{m(n-1)} + \frac{\uparrow\uparrow mn\,[[\tau_{\varkappa\pi}]]}{mn\,(n-1)}. \end{aligned} \tag{403.12}$$

De waarde P_{kp} kan nu worden beschouwd als een eerste benadering van de vruchtbaarheid $\mu_p + \lambda_{kp}$. In de volgende paragrafen zal er aandacht aan worden besteed of deze benadering bruikbaar is en of zij verbeterd kan worden.

Tot slot moet nog worden opgemerkt, dat formule (403.11) niet meer bruikbaar is aan de grens van twee achter elkaar liggende blokken. Dan kan wel de gemiddelde vruchtbaarheid van g veldjes gelijk worden geacht aan de vruchtbaarheid (λ_{kp}) van het middelste veldje, maar de g veldjes liggen niet meer alle in blok p. Door berekening is ons gebleken, dat dit de geldigheid van de

hierna volgende conclusies en formules niet ongunstig beïnvloedt. Omdat de berekeningen erg ingewikkeld zijn laten wij ze liever hier achterwege.

404 – De toelaatbaarheid van een algemene vruchtbaarheidscorrectie

In formule (403.12) hebben wij een waarde gegeven voor de vruchtbaarheid van het veldje (kp). Uit de formule blijkt duidelijk dat de berekende vruchtbaarheid met een fout behept is. Dit brengt het gevaar mee, dat door een vruchtbaarheidscorrectie de gemaakte toevallige fouten zoveel groter worden, dat het nut van de vruchtbaarheidscorrectie hier niet tegen opweegt. Hierdoor zou een vruchtbaarheidscorrectie de conclusies onbetrouwbaarder kunnen maken.

Wij willen nu onderzoeken, wanneer op dit gevaar gelet moet worden. Daartoe berekenen wij onze conclusies eerst met, en dan zonder vruchtbaarheidscorrectie binnen de blokken. De aanwezige fouten zullen wij met elkaar vergelijken.

De schatting $(\varrho_k)_p$ van ϱ_k, die met behulp van de vruchtbaarheidscorrectie P_{kp} voor ras k op veldje (kp) kan worden berekend, is als volgt te omschrijven:

$$(\varrho_k)_p = \Omega_{kp} - \bar{\bar{\Omega}} - P_{kp}. \tag{404.1}$$

Deze waarde is het gemakkelijkst te berekenen aan de hand van formules (403.2) en (403.12). Dit geeft

$$(\varrho_k)_p = \varrho_k + \frac{\uparrow\uparrow g\,(n-1)\,[[\lambda_{p\kappa^*\lfloor p\rfloor}]]}{g\,(n-1)} + \tau_{kp} - \frac{\uparrow g\,[\tau_{p\kappa^* p}]}{g} + \tag{404.2}$$
$$+ \frac{\uparrow\uparrow g\,(n-1)\,[[\tau_{p\kappa^*\lfloor p\rfloor}]]}{g\,(n-1)} + \frac{\uparrow m\,[\tau_{\kappa p}]}{m(n-1)} - \frac{n}{n-1}\,\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

Uit ieder blok laat zich met (404.2) een waarde voor ϱ_k berekenen. Dit geeft in totaal dus n waarden. De beste schatting $\bar{\varrho}_k$ van ϱ_k krijgen wij door deze n waarden te middelen. Om dit gemakkelijk te kunnen doen, zullen wij de rassen $_p\varkappa^*$ splitsen in ras k en de overige rassen, die wij aan zullen duiden als rassen $_p\varkappa^\circ$. Wanneer wij deze scheiding aanbrengen in (404.2) vinden wij

$$(404.3)\quad (\varrho_k)_p = \varrho_k + \frac{\uparrow(n-1)[\lambda_{k\lfloor p\rfloor}]}{g(n-1)} + \frac{\uparrow\uparrow(g-1)(n-1)[[\lambda_{p\kappa^\circ\lfloor p\rfloor}]]}{g(n-1)} +$$

$$+ \frac{g-1}{g}\tau_{kp} - \frac{\uparrow g-1\,[\tau_{p\kappa^\circ p}]}{g} + \frac{\uparrow(n-1)[\tau_{k\lfloor p\rfloor}]}{g(n-1)} +$$

$$+ \frac{\uparrow\uparrow\,(g-1)(n-1)[[\tau_{p\kappa^\circ\lfloor p\rfloor}]]}{g(n-1)} + \frac{\uparrow m\,[\tau_{\kappa p}]}{m(n-1)} - \frac{n}{n-1}\,\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

Formule (404.3) moet nu dus over p gemiddeld worden. Voor de overzichtelijkheid zullen wij de negen termen van het rechterlid een voor een middelen. Wij vinden dan

1e De eerste term geeft als gemiddelde ϱ_k.

2e Om de tweede term te middelen schrijven wij hem in de vorm

$$\frac{\uparrow n-1\,[\lambda_{k\lfloor p\rfloor}]}{g(n-1)} = \frac{\uparrow n\,[\lambda_{k\pi}]}{g(n-1)} - \frac{\lambda_{kp}}{g(n-1)}.$$

Middelen wij dit over p, dan blijft de eerste term van het rechterlid onveranderd, de tweede term wordt $-\dfrac{\uparrow n\,[\lambda_{k\pi}]}{gn(n-1)}$. De som van beide termen wordt dus

$$\frac{n\uparrow n\,[\lambda_{k\pi}]}{gn(n-1)} - \frac{\uparrow n\,[\lambda_{k\pi}]}{gn(n-1)} = \frac{\uparrow n\,[\lambda_{k\pi}]}{gn}.$$

3e Het gemiddelde van de derde term willen wij voorlopig aanduiden door

$$\frac{\uparrow\uparrow\uparrow n\,(g-1)(n-1)\,[[[\lambda_{\pi\kappa^\circ\lfloor\pi\rfloor}]]]}{gn(n-1)}.$$

4e Het gemiddelde van de vierde term wordt

$$\frac{g-1}{g}\,\frac{\uparrow n\,[\tau_{k\pi}]}{n}.$$

5e Het gemiddelde van de vijfde term wordt

$$-\frac{\uparrow\uparrow n\,(g-1)\,[[\tau_{\pi\kappa^\circ\pi}]]}{gn}.$$

6e Het gemiddelde van de zesde term is te vinden naar analogie van de tweede. Het wordt

$$\frac{\uparrow n\,[\tau_{k\pi}]}{gn}.$$

7e Het gemiddelde van de zevende term is te schrijven naar analogie van de derde term. Wij schrijven

$$\frac{\uparrow\uparrow\uparrow n\,(g-1)\,(n-1)\,[[[\tau_{\pi\kappa^\circ[\pi]}]]]}{gn\,(n-1)}.$$

8e Het gemiddelde van de achtste term wordt

$$\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn\,(n-1)}.$$

9e Het gemiddelde van de laatste term is

$$-\frac{n}{n-1}\,\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

Van bovenstaand overzicht kunnen de vierde en de zesde term worden samengenomen, evenals de achtste en de negende. Wij kunnen dus de formule opstellen

$$(404.4)\quad \bar{\varrho}_k = \varrho_k + \frac{\uparrow n\,[\lambda_{k\pi}]}{gn} + \frac{\uparrow\uparrow\uparrow n\,(g-1)\,(n-1)\,[[[\lambda_{\pi\kappa^\circ[\pi]}]]]}{gn\,(n-1)} + \frac{\uparrow n\,[\tau_{k\pi}]}{n} - \frac{\uparrow\uparrow n\,(g-1)\,[[\tau_{\pi\kappa^\circ\pi}]]}{gn} + \frac{\uparrow\uparrow\uparrow n\,(g-1)\,(n-1)\,[[[\tau_{\pi\kappa^\circ[\pi]}]]]}{gn\,(n-1)} - \frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

Alles wat hier na de eerste term van het tweede lid staat, moet worden beschouwd als de fout van $\bar{\varrho}_k$, wanneer er binnen de blokken gecorrigeerd wordt.

Wanneer er geen correcties worden toegepast binnen de blokken doch wel met blokverschillen wordt rekening gehouden, is de schatting $(\varrho_k)_p'$ van ϱ_k, die wij voor ras k in blok p kunnen berekenen, als volgt te omschrijven

$$(404.5)\qquad (\varrho_k)_p' = \Omega_{kp} - \bar{\bar{\Omega}} - \bar{\mu}_p.$$

Deze waarde is gegeven in (403.4). In verband met (404.5) lezen wij daar

$$(404.6)\qquad (\varrho_k)_p' = \varrho_k + \lambda_{kp} + \tau_{kp} - \frac{\uparrow m\,[\tau_{\kappa p}]}{m}.$$

De beste schatting $\bar{\varrho}_k'$ van ϱ_k vinden wij nu door (404.6) over p te middelen. Dit geeft

$$(404.7)\qquad \bar{\varrho}_k' = \varrho_k + \frac{\uparrow n\,[\lambda_{k\pi}]}{n} + \frac{\uparrow n\,[\tau_{k\pi}]}{n} - \frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}.$$

Alles wat hier na de eerste term van het tweede lid staat, moet worden beschouwd als fout van $\bar{\varrho}_k'$.

Een vruchtbaarheidscorrectie binnen de blokken is nu gewenst, wanneer de fout van $\bar{\varrho}_k$ (zie 404.4) kleiner is in absolute waarde dan de fout van $\bar{\varrho}_k'$ (zie 404.7).

Hierbij moet evenwel nog op een bizonderheid worden gelet. Zowel in (404.4) als in (404.7) komt als laatste de term

$$-\frac{\uparrow\uparrow mn\,[[\tau_{k\pi}]]}{mn}$$

voor. Het is duidelijk dat deze term voor alle rassen gelijk uitvalt en dus op de ras*verschillen* geen invloed heeft. Om deze reden kunnen wij deze term beter weglaten.

Door (404.4) en (404.7) te vergelijken komen wij dus tot de volgende voorwaarde voor een vruchtbaarheidscorrectie binnen de blokken

(404.8)

$$\left|\frac{\uparrow n\,[\lambda_{k\pi}]}{n}+\frac{\uparrow n\,[\tau_{k\pi}]}{n}\right| > \left|\frac{\uparrow n\,[\lambda_{k\pi}]}{gn}+\frac{\uparrow\uparrow\uparrow n(g-1)(n-1)[[[[\lambda_{\pi\varkappa^\circ[\pi]}]]]]}{gn\,(n-1)}+\right.$$
$$\left.+\frac{\uparrow n\,[\tau_{k\pi}]}{n}-\frac{\uparrow\uparrow n\,(g-1)\,[[\tau_{\pi\varkappa^\circ\pi}]]}{gn}+\frac{\uparrow\uparrow\uparrow n\,(g-1)\,(n-1)[[[[\tau_{\pi\varkappa^\circ[\pi]}]]]]}{gn\,(n-1)}\right|.$$

Deze ongelijkheid is overzichtelijker wanneer wij beide leden in het kwadraat verheffen.

Omdat (404.8) een algemene voorwaarde moet uitdrukken, die geldt voor alle rassen en het gehele proefveld, hoeven wij slechts de asymptotische waarde van de kwadraten vast te stellen. Dit heeft het voordeel dat er gebruik van kan worden gemaakt, dat het dubbele product van onafhankelijke factoren asymptotisch $= 0$ is.

Bij het kwadrateren van (404.8) vindt deze regel echter slechts beperkte toepassing, aangezien er in het rechterlid veel grootheden voorkomen die wel afhankelijk zijn.

In de eerste plaats wordt het begrip $\varkappa^\circ$ in ieder blok opnieuw gedefinieerd. Er zijn dus n definities ${}_\pi\varkappa^\circ$. Het is waarschijnlijk dat sommige rassen verschillende malen in een groep $\varkappa^\circ$. betrokken worden zodat een aantal dubbele producten in werkelijkheid zuivere kwadraten worden.

In de tweede plaats treden er tussen de waarden λ dezelfde correlaties op, die wij in 302 voor de i bestudeerden.

Wanneer wij met al deze correlaties rekening houden, worden de berekeningen zo ingewikkeld, dat wij ze hier liever niet uitwerken. Wij willen ons daarom beperken tot de zuivere kwadraten.

De asymptotische waarde van het kwadraat van alle toevallige fouten duiden wij daarbij aan door σ^2, de asymptotische waarde van het gemiddeld kwadraat van alle waarden λ door λ^2 (zonder indices).

Uit (404.8) leiden wij op deze manier af

$$\frac{n}{n^2}\lambda^2 + \frac{n}{n^2}\sigma^2 > \frac{n}{g^2n^2}\lambda^2 + \frac{n(g-1)(n-1)}{g^2n^2(n-1)^2}\lambda^2 +$$
$$+ \frac{n}{n^2}\sigma^2 + \frac{n(g-1)}{g^2n^2}\sigma^2 + \frac{n(g-1)(n-1)}{g^2n^2(n-1)^2}\sigma^2.$$

Wanneer wij alle termen met λ^2 naar het linkerlid brengen, en alle termen met σ^2 naar het rechter, wordt dit

$$\left\{\frac{1}{n} - \frac{1}{g^2n} - \frac{g-1}{g^2n(n-1)}\right\}\lambda^2 > \left\{\frac{g-1}{g^2n} + \frac{g-1}{g^2n(n-1)}\right\}\sigma^2.$$

Door beide leden met het positieve getal $g^2n(n-1)$ te vermenigvuldigen wordt dit

$$\left\{g^2(n-1) - (n-1) - (g-1)\right\}\lambda^2 > \left\{(n-1)(g-1) + (g-1)\right\}\sigma^2$$

of

$$\frac{\lambda^2}{\sigma^2} > \frac{n(g-1)}{(g-1)\{(g+1)(n-1)-1\}}.$$

Een kleine vereenvoudiging geeft

$$\frac{\lambda^2}{\sigma^2} > \frac{n}{g(n-1)+(n-2)}. \tag{404.9}$$

Wanneer wij bij het kwadrateren van (404.8) met de afhankelijkheid van de termen van het rechterlid volledig rekening hadden gehouden, zou het rechterlid van (404.9) kleiner zijn geworden. Voorwaarde (404.9) blijft dus aan de veilige kant.

Het vinden van deze voorwaarde heeft een principiële betekenis. In 401 hebben wij op aprioristische gronden aangenomen, dat een grafische vruchtbaarheidscorrectie de meest efficiënte was. Nu blijkt dat een grafische correctie aan grenzen is gebonden, is deze efficiëntie dubieus. Dit maakt ons onderzoek evenwel niet waardeloos. Soms moet er grafisch gecorrigeerd worden. Maar in andere gevallen kan dit onderzoek er toe leiden de proeven toch volgens een FISHER-schema op te zetten.

405 – Over de voorwaarden voor correctie in een blok

In voorwaarde (404.9) is weergegeven, hoe groot de vruchtbaarheidsschommelingen in verhouding tot de toevallige fout moeten zijn, opdat een vruchtbaarheidscorrectie voordeel zal geven. Hierbij is rekening gehouden met de gemiddelde vruchtbaarheidsschommelingen van het gehele proefveld.

Het laat zich denken, dat in het algemeen een vruchtbaarheidscorrectie niet gewenst is, doch dat op een bepaalde plaats er wel aanleiding toe bestaat. Wij zullen nu de voorwaarden voor vruchtbaarheidscorrectie in één blok onderzoeken. Wij nemen daartoe aan, dat in de blokken $\lfloor p \rfloor$ de vruchtbaarheidscorrectie door middel van de waarden $\bar{\mu}_{\lfloor p \rfloor}$ afdoende is.

Indien wij bij het corrigeren van blok p de methode van 403 volgen, zal de berekende rasvoortreffelijkheid $(\varrho_k)_p$ in formule (404.3) zijn uitgedrukt. De $n - 1$ waarden $(\varrho_k)_{\lfloor p \rfloor}$ zijn weergegeven in (404.6).

De beste schatting $\bar{\varrho}_k$ van de ware ϱ_k vinden wij nu met de formule

$$\bar{\varrho}_k = \frac{(\varrho_k)_p + \uparrow (n-1)\,[\varrho_{k\lfloor p \rfloor}]}{n}. \tag{405.1}$$

Uit (404.3) en (404.6) vinden wij

$$\begin{aligned}
\bar{\varrho}_k = \frac{1}{n}\Bigg\{ & \varrho_k + \frac{\uparrow (n-1)\,[\lambda_{k\lfloor p \rfloor}]}{g\,(n-1)} + \frac{\uparrow\uparrow (g-1)\,(n-1)\,[[\lambda_{p\kappa^\circ\lfloor p \rfloor}]]}{g\,(n-1)} + \\
& + \frac{g-1}{g}\,\tau_{kp} - \frac{\uparrow (g-1)\,[\tau_{p\kappa^\circ p}]}{g} + \frac{\uparrow (n-1)\,[\tau_{k\lfloor p \rfloor}]}{g\,(n-1)} + \\
& + \frac{\uparrow\uparrow (g-1)\,(n-1)\,[[\tau_{p\kappa^\circ\lfloor p \rfloor}]]}{g\,(n-1)} + \frac{\uparrow m\,[\tau_{\kappa p}]}{m\,(n-1)} - \\
& - \frac{n}{n-1}\,\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}\Bigg\} + \\
& + \frac{1}{n}\Bigg\{\uparrow (n-1)\left[\varrho_k + \lambda_{k\lfloor p \rfloor} + \tau_{k\lfloor p \rfloor} - \frac{\uparrow m\,[\tau_{\kappa\lfloor p \rfloor}]}{m}\right]\Bigg\}
\end{aligned}$$

of

$$\begin{aligned}
\bar{\varrho}_k = \varrho_k & + \frac{\uparrow (n-1)\,[\lambda_{k\lfloor p \rfloor}]}{gn\,(n-1)} + \frac{\uparrow (n-1)\,[\lambda_{k\lfloor p \rfloor}]}{n} + \\
& + \frac{\uparrow\uparrow (g-1)\,(n-1)\,[[\lambda_{p\kappa^\circ\lfloor p \rfloor}]]}{gn\,(n-1)} + \frac{g-1}{gn}\,\tau_{kp} - \frac{\uparrow (g-1)\,[\tau_{p\kappa^\circ p}]}{gn} +
\end{aligned} \tag{405.2}$$

$$+\frac{\uparrow(n-1)\,[\tau_{k\lfloor p\rfloor}]}{gn\,(n-1)}+\frac{\uparrow(n-1)\,[\tau_{k\lfloor p\rfloor}]}{n}+\frac{\uparrow\uparrow\,(g-1)(n-1)[[\tau_{p\kappa^\circ\lfloor p\rfloor}]]}{gn\,(n-1)}+$$

$$+\left\{\frac{\uparrow m\,[\tau_{\kappa p}]}{mn\,(n-1)}-\frac{\uparrow\uparrow m\,(n-1)[[\tau_{\kappa\lfloor p\rfloor}]]}{mn}-\frac{1}{n-1}\,\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}\right\}.$$

Het deel van de formule, dat tussen $\{\}$ is geplaatst, is voor alle rassen gelijk, en kan dus bij het bestuderen van de ras*verschillen* worden weggelaten.

Wanneer er niet gerekend wordt met vruchtbaarheidscorrecties binnen de blokken is de beste schatting $\bar{\varrho}_k'$ van ϱ_k te vinden in formule (404.7) waaruit de term

$$-\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}$$

weer kan worden weggelaten.

Uit (405.2) en (404.7) zien wij dat een vruchtbaarheidscorrectie zin heeft, wanneer

$$(405.3)\quad \left|\frac{\uparrow n\,[\lambda_{k\pi}]}{n}+\frac{\uparrow n\,[\tau_{k\pi}]}{n}\right|>\left|\frac{\{g\,(n-1)+1\}\uparrow(n-1)\,[\lambda_{k\lfloor p\rfloor}]}{gn\,(n-1)}+\right.$$

$$+\frac{\uparrow\uparrow\,(g-1)\,(n-1)\,[[\lambda_{p\kappa^\circ\lfloor p\rfloor}]]}{gn\,(n-1)}+\frac{g-1}{gn}\,\tau_{kp}-\frac{\uparrow(g-1)\,[\tau_{p\kappa^\circ p}]}{gn}+$$

$$\left.+\frac{\{g\,(n-1)+1\}\uparrow(n-1)\,[\tau_{k\lfloor p\rfloor}]}{gn\,(n-1)}+\frac{\uparrow\uparrow\,(g-1)\,(n-1)\,[[\tau_{p\kappa^\circ\lfloor p\rfloor}]]}{gn\,(n-1)}\right|.$$

Evenals bij (404.8) zullen wij deze ongelijkheid in het kwadraat verheffen. Hierbij moeten wij bedenken dat wij slechts in blok p een vruchtbaarheidscorrectie nodig achten. In blok p zijn de waarden van λ naar onze mening dus groter dan in de blokken $\lfloor p\rfloor$. In verband hiermee zullen wij de asymptotische waarde van de kwadraten van de grootheden λ in blok p aanduiden met λ_p^2, in de andere blokken met $\lambda_{\lfloor p\rfloor}^2$.

Bij het kwadrateren van (404.8) hebben wij met de aanwezige correlaties van de verschillende termen geen rekening gehouden. Dat zullen wij nu wel doen. Bij (405.3) is dit vrij gemakkelijk daar alle grootheden van het linkerlid onderling onafhankelijk zijn, terwijl er in het rechterlid geen mogelijkheid is dat een bepaald veldje tweemaal is genoemd. De toevallige fouten zijn dus allen onderling onafhankelijk, terwijl bij de waarden λ slechts met correlaties in de zin van 302 behoeft te worden gerekend.

Wij willen deze correlaties eerst nagaan, en beginnen daartoe met het kwadraat te zoeken van de eerste twee termen van het rechterlid van (405.3), dus van

$$(405.4)\quad \frac{\{g(n-1)+1\}\uparrow(n-1)[\lambda_{k\lfloor p\rfloor}]}{gn(n-1)}+\frac{\uparrow\uparrow(g-1)(n-1)[[\lambda_{p\kappa^0\lfloor p\rfloor}]]}{gn(n-1)}.$$

De waarden λ zijn genomen uit $(n-1)$ blokken. Volgens de bespreking van formule (402.6) treden er geen correlaties op tussen waarden λ uit verschillende blokken. Wij kunnen dus beginnen met een willekeurig blok q te bestuderen. Blok q heeft tot (405.4) het volgende bijgedragen:

$$(405.5)\quad \frac{\{g(n-1)+1\}\lambda_{kq}}{gn(n-1)}+\frac{\uparrow(g-1)[\lambda_{p\kappa^0 q}]}{gn(n-1)}.$$

Het kwadraat van (405.5) is nu

$$(405.6)\quad \frac{\{g(n-1)+1\}^2\lambda_{kq}^2}{g^2n^2(n-1)^2}+\frac{\uparrow(g-1)[\lambda_{p\kappa^0 q}^2]}{g^2n^2(n-1)^2}+$$

$$+\frac{2\{g(n-1)+1\}\uparrow(g-1)[\lambda_{kq}\lambda_{p\kappa^0 q}]}{g^2n^2(n-1)^2}+\frac{\uparrow\uparrow(g-1)(g-2)[[\lambda_{lq}\lambda_{l'q}]]}{g^2n^2(n-1)^2},$$

waarbij in de laatste term door l en l' twee willekeurige rassen ${}_pk^0$ worden aangeduid, die niet dezelfde mogen zijn.

De asymptotische waarde van de kwadraten uit de eerste twee termen is $\lambda_{\lfloor p\rfloor}^2$; de asymptotische waarde van de dubbele producten uit de laatste twee termen is volgens (302.4) gelijk aan $-\frac{\lambda_{\lfloor p\rfloor}^2}{m-1}$. De asymptotische waarde van (405.6) is dus

$$\frac{\{g(n-1)+1\}^2\lambda_{\lfloor p\rfloor}^2}{g^2n^2(n-1)^2}+\frac{(g-1)\lambda_{\lfloor p\rfloor}^2}{g^2n^2(n-1)^2}-$$

$$-\frac{2\{g(n-1)+1\}(g-1)\lambda_{\lfloor p\rfloor}^2}{g^2n^2(n-1)^2(m-1)}-\frac{(g-1)(g-2)\lambda_{\lfloor p\rfloor}^2}{g^2n^2(n-1)^2(m-1)}.$$

Na enige vereenvoudiging wordt dit

$$\frac{g^2(n-1)^2}{g^2n^2(n-1)^2}\lambda_{\lfloor p\rfloor}^2+\frac{2g(n-1)+g}{g^2n^2(n-1)^2}\lambda_{\lfloor p\rfloor}^2-$$

$$-\frac{g-1}{m-1}\cdot\frac{2g(n-1)+g}{g^2n^2(n-1)^2}\lambda_{\lfloor p\rfloor}^2$$

of

$$\frac{1}{n^2}\lambda_{\lfloor p\rfloor}^2+\frac{m-g}{m-1}\cdot\frac{2(n-1)+1}{gn^2(n-1)^2}\lambda_{\lfloor p\rfloor}^2.$$

Dit is dus het kwadraat van (405.5). Het kwadraat van (405.4) is $(n-1)$ maal zo groot omdat er $(n-1)$ blokken $\lfloor p \rfloor$ zijn. Dit kwadraat is dus

$$\frac{n-1}{n^2}\lambda_{\lfloor p \rfloor}^2 + \frac{m-g}{m-1} \cdot \frac{2(n-1)+1}{gn^2(n-1)}\lambda_{\lfloor p \rfloor}^2.$$

Met behulp van deze uitkomst brengen wij (405.3) in de kwadratische vorm (zie ook de tekst onder (405.3)):

$$\frac{1}{n^2}\lambda_p^2 + \frac{n-1}{n^2}\lambda_{\lfloor p \rfloor}^2 + \frac{n}{n^2}\sigma^2 > \frac{n-1}{n^2}\lambda_{\lfloor p \rfloor}^2 +$$
$$+ \frac{m-g}{m-1} \cdot \frac{2(n-1)+1}{gn^2(n-1)}\lambda_{\lfloor p \rfloor}^2 + \frac{(g-1)^2}{g^2n^2}\sigma^2 + \frac{g-1}{g^2n^2}\sigma^2 +$$
$$+ \frac{\{g(n-1)+1\}^2(n-1)}{g^2n^2(n-1)^2}\sigma^2 + \frac{(g-1)(n-1)}{g^2n^2(n-1)^2}\sigma^2$$

of, na enige vereenvoudiging,

$$\lambda_p^2 > \frac{m-g}{m-1} \cdot \frac{2(n-1)+1}{g(n-1)}\lambda_{\lfloor p \rfloor}^2 + \frac{n}{g(n-1)}\sigma^2. \tag{405.7}$$

Wij willen voorwaarde (405.7) op twee manieren uitwerken. In de eerste plaats willen wij veronderstellen

$$\lambda_{\lfloor p \rfloor}^2 = 0. \tag{405.8}$$

Formule (405.7) wordt dan

$$\lambda_p^2 > \frac{n}{g(n-1)}\sigma^2$$

of

$$\frac{\lambda_p^2}{\sigma^2} > \frac{n}{g(n-1)}. \tag{405.9}$$

Wanneer wij ons herinneren dat (404.9) eist

$$\frac{\lambda^2}{\sigma^2} > \frac{n}{g(n-1)+(n-2)},$$

dan zien wij dat de eis van (405.9) iets zwaarder is.

Wij willen nu (405.7) nogmaals bestuderen in de veronderstelling dat $\lambda_{\lfloor p \rfloor}^2$ niet gelijk is aan 0. Immers, wanneer $\lambda_{\lfloor p \rfloor}^2$ te klein is om aan voorwaarde (404.9) of (405.9) te voldoen, zal in de blokken $\lfloor p \rfloor$ geen correctie geoorloofd zijn. De waarde van $\lambda_{\lfloor p \rfloor}^2$ zal dan invloed hebben op de gemaakte fouten, en dus ook op de eis die wij aan een vruchtbaarheidscorrectie stellen.

Volgens (405.9) is de grootste waarde van $\lambda_{[p]}$, die geen correctie toelaat,

$$\lambda_{[p]}^2 = \frac{n}{g(n-1)}\sigma^2. \tag{405.10}$$

Deze waarde kan $\lambda_{[p]}^2$ dus ook in (405.7) hebben. Wij zullen die formule uitwerken in de veronderstelling dat dit inderdaad het geval is, en vinden dan

$$\lambda_p^2 > \frac{m-g}{m-1} \cdot \frac{2(n-1)+1}{g(n-1)} \cdot \frac{n}{g(n-1)}\sigma^2 + \frac{n}{g(n-1)}\sigma^2$$

of, na enige uitwerking,

$$\frac{\lambda_p^2}{\sigma^2} > \frac{n}{g(n-1)} \cdot \frac{g + \frac{m-g}{m-1}\left(2 + \frac{1}{n-1}\right)}{g}. \tag{405.11}$$

406 – Samenvatting van de voorwaarden voor toelaatbaarheid van een vruchtbaarheidscorrectie

In de voorgaande paragrafen hebben wij enkele voorwaarden gevonden voor de toelaatbaarheid van een vruchtbaarheidscorrectie. In de eerste plaats zij herinnerd aan de formules

$$\frac{\lambda^2}{\sigma^2} > \frac{n}{g(n-1)+(n-2)}. \tag{404.9}$$

$$\frac{\lambda_p^2}{\sigma^2} > \frac{n}{g(n-1)}. \tag{405.9}$$

$$\frac{\lambda_p^2}{\sigma^2} > \frac{n}{g(n-1)} \cdot \frac{g + \frac{m-g}{m-1}\left(2 + \frac{1}{n-1}\right)}{g}. \tag{405.11}$$

Deze drie formules geven minimumeisen. Wanneer de verhouding $\frac{\lambda^2}{\sigma^2}$ kleiner is dan de waarde van het rechterlid is een correctie gewoonlijk nadelig. Is de verhouding $\frac{\lambda^2}{\sigma^2}$ gelijk aan de waarde van het rechterlid dan geeft een correctie even vaak nadeel als voordeel. Is het linkerlid groter dan het rechterlid dan geeft een correctie meestal voordeel, terwijl het gemiddeld voordeel stijgt naarmate het linkerlid meer in waarde verschilt van het rechterlid.

Nu is het duidelijk dat niet tot correctie zal worden overgegaan wanneer het linkerlid en rechterlid precies gelijk zijn. Maar ook

wanneer het linkerlid slechts weinig groter is dan het rechterlid, heeft een correctie weinig zin. Dan zal de proeffout niet zoveel verlaagd worden, dat de uitkomsten er veel betrouwbaarder door worden.

Het is niet onze bedoeling na te gaan hoeveel het linkerlid groter moet zijn dan het rechter-, om een correctie zinvol te maken. Slechts deze conclusie willen wij trekken, dat het geen zin heeft bovenstaande drie formules in de practijk te onderscheiden. Wij kunnen in alle gevallen werken met formule (405.11), die de zwaarste eisen stelt.

De betekenis van (404.9) is dan deze, dat hieruit blijkt dat een vruchtbaarheidscorrectie over het gehele proefveld wenselijk is, wanneer met formule (405.11) voor iedere plaats aangetoond kan worden, dat een correctie daar nodig is. Deze conclusie lijkt vanzelfsprekend, maar wegens de correlaties die tussen allerlei fouten optreden gedurende de berekening, zou toch aan de juistheid hiervan getwijfeld kunnen worden, als het niet rechtstreeks was aangetoond.

Nu zou men nog het bezwaar kunnen voelen, dat formule (405.11) nog niet streng genoeg is. Deze formule is ontstaan door (405.9) (in de vorm van (405.10)) in (405.7) te substitueren. Men zou ook nog (405.11) in (405.7) kunnen substitueren en zo doorgaan tot een limiet bereikt werd.

Toch menen wij dat dit zelden noodzakelijk is. Het is onwaarschijnlijk dat de vruchtbaarheidsafwijking λ op alle niet-gecorrigeerde plaatsen juist zo groot is, dat een correctie bijna toelaatbaar is; en het is ook onwaarschijnlijk, dat tot correctie wordt besloten, wanneer aan (405.11) juist wordt voldaan. Mochten deze uitzonderlijke gevallen zich voordoen, dan valt bovenbedoelde limietwaarde ook nog gemakkelijk af te leiden.

Wij willen dus steeds werken met formule (405.11). Deze formule heeft evenwel nog geen bruikbare vorm, omdat λ_p^2, het kwadraat van de *ware* vruchtbaarheidsafwijking, ons onbekend is. Wij kunnen niet verder komen dan $(P_{kp} - \bar{\mu}_p)$, een schatting van λ_p. Deze schatting laat zich berekenen met behulp van de formules (403.3) en (403.12), waaruit we afleiden

(406.1)

$$P_{kp} - \bar{\mu}_p = \lambda_{kp} - \frac{\uparrow\uparrow g\,(n-1)\,[[\lambda_{p\kappa^*|p}]]}{g\,(n-1)} + \frac{\uparrow g\,[\tau_{p\kappa^*p}]}{g} -$$

$$-\frac{\uparrow\uparrow g(n-1)\,[[\tau_{p\kappa^*\lfloor p\rfloor}]]}{g(n-1)}+$$

$$+\left\{-\frac{n}{n-1}\,\frac{\uparrow m\,[\tau_{\kappa p}]}{m}+\frac{n}{n-1}\,\frac{\uparrow\uparrow mn\,[[\tau_{\kappa\pi}]]}{mn}\right\}.$$

We zien naast λ_{kp} in de formule allerlei fouten vermeld. Hiervan hebben de eersten invloed op de schatting van de vruchtbaarheidsverschillen in blok p; de fouten die tussen $\{\}$ zijn geplaatst hebben daarop geen invloed en kunnen dus verder worden genegeerd.

Uit (406.1) laat zich nu een waarde van λ_p^2 afleiden door beide leden te kwadrateren. Hierbij moet er opgelet worden dat de waarden onderling afhankelijk zijn voorzover ze in hetzelfde blok liggen. Met sommige dubbele producten moet dus worden gerekend als in de formules (405.4) tot (405.6). Verder moet er weer onderscheid worden gemaakt tussen λ_p^2 en $\lambda_{\lfloor p\rfloor}^2$. Wij streven immers naar een bruikbare vorm van (405.11), die ook op dat onderscheid is gebaseerd.

Door (406.1) te kwadrateren vinden wij dus als asymptotische waarde

$$<(P_{kp}-\bar{\mu})^2> = \lambda_p^2+\frac{g(n-1)}{g^2(n-1)^2}\lambda_{\lfloor p\rfloor}^2-$$

$$-\frac{(n-1)\,g\,(g-1)}{g^2(n-1)^2}\cdot\frac{\lambda_{\lfloor p\rfloor}^2}{m-1}+\frac{g}{g^2}\sigma^2+\frac{g(n-1)}{g^2(n-1)^2}\sigma^2$$

of

$$<(P_{kp}-\bar{\mu}_p)^2> = \lambda_p^2+\frac{m-g}{m-1}\cdot\frac{1}{g(n-1)}\lambda_{\lfloor p\rfloor}^2+\frac{n}{g(n-1)}\sigma^2$$

of

$$(406.2)\quad \lambda_p^2=<(P_{kp}-\bar{\mu}_p)^2>-\frac{m-g}{m-1}\cdot\frac{1}{g(n-1)}\lambda_{\lfloor p\rfloor}^2-\frac{n}{g(n-1)}\sigma^2.$$

Wanneer wij λ_p^2 uit (406.2) in (405.7) substitueren vinden wij

$$<(P_{kp}-\bar{\mu}_p)^2>-\frac{m-g}{m-1}\cdot\frac{1}{g(n-1)}\lambda_{\lfloor p\rfloor}^2-$$

$$-\frac{n}{g(n-1)}\sigma^2>\frac{m-g}{m-1}\cdot\frac{2(n-1)+1}{g(n-1)}\lambda_{\lfloor p\rfloor}^2+\frac{n}{g(n-1)}\sigma^2$$

of

$$(406.3)\quad <(P_{kp}-\bar{\mu}_p)^2>>\frac{m-g}{m-1}\cdot\frac{2n}{g(n-1)}\lambda_{\lfloor p\rfloor}^2+\frac{2n}{g(n-1)}\sigma^2.$$

Deze formule is analoog aan (405.7). Door nu in deze formule de waarde van $\lambda_{[p]}$ uit (405.10) te substitueren zullen wij een formule vinden, die analoog is aan formule (405.11). Deze luidt

$$<(P_{kp}-\bar{\mu}_p)^2> > \frac{m-g}{m-1} \cdot \frac{2n}{g(n-1)} \cdot \frac{n}{g(n-1)} \sigma^2 + \frac{2n}{g(n-1)} \sigma^2$$

of

$$\text{(406.4)} \quad \frac{<(P_{kp}-\bar{\mu}_p)^2>}{\sigma^2} > \frac{2n}{g(n-1)} \left(1 + \frac{m-g}{m-1} \cdot \frac{n}{n-1} \cdot \frac{1}{g}\right).$$

Deze formule moet dus worden gebruikt om te controleren of een vruchtbaarheidscorrectie, die men in de practijk vindt, wenselijk is. Maar ook deze formule is nog niet goed bruikbaar. Zowel de teller als de noemer van het linkerlid zijn nog onbekend. Van de teller is evenwel een schatting te verkrijgen door de asymptotische waarde te vervangen door de empirische. Hierover wordt in 407 en 408 nader gesproken. Een manier om de σ^2 te vinden wordt gegeven in 411.

Wij willen nu nog even letten op de vorm $\frac{m-g}{m-1}$ uit het rechterlid. Uit deze vorm blijkt dat de verhouding tussen het aantal rassen g, dat gebruikt wordt om de vruchtbaarheid van een bepaald veldje te schatten, en het totaal aantal rassen m, invloed heeft op de toelaatbaarheid van de vruchtbaarheidscorrectie op een bepaalde plaats. Naarmate die verhouding groter wordt, wordt de eis van toelaatbaarheid minder zwaar.

Dit verschijnsel vindt zijn verklaring in de vruchtbaarheidsafwijkingen in de niet gecorrigeerde blokken $[p]$. Immers, wanneer wij deze afwijkingen $=0$ hadden gesteld (zie 405.8), dan kwam de uitdrukking $\frac{m-g}{m-1}$ niet in de formule voor (zie 405.9).

Wanneer nu $g=m$, dan hebben deze afwijkingen geen invloed. In dit geval immers heeft men te maken met (402.3)

$$\text{(402.3)} \qquad [\lambda_{\kappa p}] = 0$$

Deze regel geldt voor ieder blok, zodat de invloed van $\lambda_{[p]}$ zichzelf opheft. Dit komt hierin tot uiting dat $\frac{m-g}{m-1}=0$ wordt.

Een ander uiterste wordt gevonden wanneer $g=1$. Dan bereikt de vorm $\frac{m-g}{m-1}$ de waarde 1. In de practijk zal de waarde steeds tussen 0 en 1 liggen. In verband met het feit, dat wij niet te snel

tot vruchtbaarheidscorrectie willen overgaan is het gemakkelijk de waarde van deze breuk gelijk aan 1 te nemen.

Wij willen nu de voorwaarden, die door formule (406.4) worden gesteld, in tabelvorm brengen (zie 406.5). Voor het gemak geven wij niet de verhouding van de kwadraten, maar van de absolute grootten.

(406.5) *Minimumwaarde voor de verhouding* $\frac{|P_{kp} - \bar{\mu}_p|}{\sigma}$ *bij een toelaatbare vruchtbaarheidscorrectie, bij verschillende waarden van n en g (wanneer* $m = \infty$*)*

g \ n	2	3	4	5	10	∞
1	3.47	2.74	2.50	2.38	2.17	2.—
3	1.50	1.23	1.14	1.09	1.01	0.94
5	1.06	0.89	0.83	0.79	0.74	0.70
7	0.86	0.72	0.67	0.65	0.61	0.57
9	0.74	0.62	0.58	0.56	0.53	0.50
11	0.66	0.56	0.52	0.50	0.47	0.45
13	0.60	0.51	0.48	0.46	0.43	0.41
15	0.55	0.47	0.44	0.43	0.40	0.38

Uit formule (406.4) en ook uit tabel (406.5) blijkt, dat het aantal blokken n een vrij geringe invloed heeft op de toelaatbaarheid van de vruchtbaarheidscorrectie. Immers komt n steeds in de vorm $\frac{n}{n-1}$ voor. Vanuit het oogpunt van vruchtbaarheidscorrectie heeft het daarom meestal geen zin n groter dan 3 te nemen.

Daarentegen heeft g een grote invloed op de toelaatbaarheid. In (406.4) komt g uitsluitend in de noemer voor. Hieruit kunnen wij de regel afleiden: *Hoe smaller de vruchtbaarheidstop is, des te hoger moet hij zijn om weggecorrigeerd te mogen worden.*

Wanneer men formule (406.4) of tabel (406.5) nauwkeurig bekijkt, ziet men, dat er ook een voorwaarde is gevonden voor een correctie over de breedte van 1 veldje ($g = 1$). Zulk een correctie kan er op neerkomen dat het veldje wordt uitgeschakeld. Het zou ons te ver voeren het probleem van het uitschakelen van gegevens hier in zijn geheel te bespreken.

Tenslotte nog dit: In het bovenstaande hebben wij speciaal gelet op de toelaatbaarheid van een correctie op een bepaalde plaats. Hierbij hebben wij de conclusie getrokken, dat men steeds

af kon gaan op de voorwaarden van (406.5). Dit mag ons evenwel niet doen vergeten dat wij in 401 reeds over een ander criterium hebben gesproken.

Bij het bespreken van fig. (401.1) hebben wij op grond van de continuïteit van het vruchtbaarheidsverloop gebruik gemaakt van de stelling *dat de vruchtbaarheid over korte afstanden geïnterpoleerd mag worden*. Van deze stelling kan ook gebruik worden gemaakt, wanneer wij met het criterium van (406.5) werken. Aan de hand van dit criterium zal voor bepaalde plaatsen geconcludeerd kunnen worden, hoe de vruchtbaarheid daar in rekening moet worden gebracht. De vruchtbaarheidscorrectie van de tussenliggende plaatsen kan gevonden worden door interpolatie.

407 – Voorbeeld van het toetsen van een vruchtbaarheidslijn

Het gebruik van het criterium van (406.5) kan het best worden toegelicht aan de hand van een voorbeeld. Daartoe hebben wij in

(407.1)

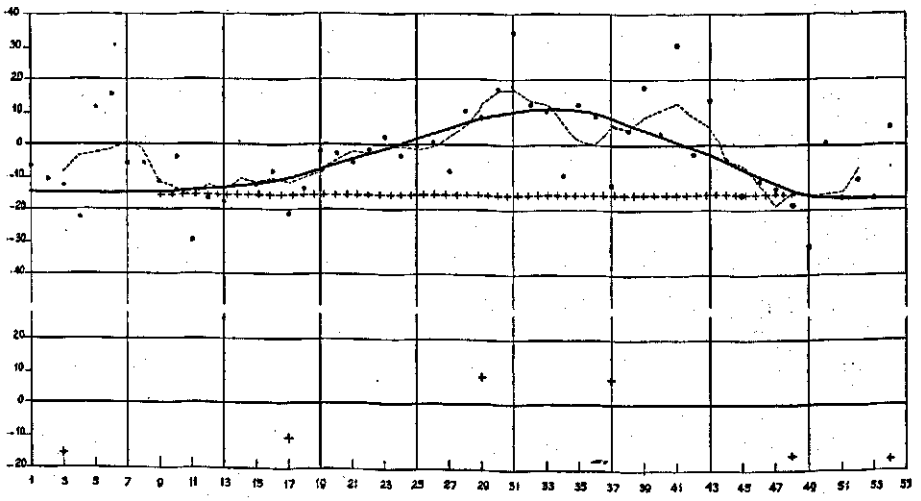


fig. (407.1) een deel van een proefveld in beeld gebracht. Van dit proefveld kunnen wij aannemen dat

(407.2) $\sigma = 12$

$n = 2$

$m = 70$ (de veldjes 1—55 liggen allen in blok p)

$\bar{\mu}_p = 0.$

De punten in grafiek (407.1) geven voor ieder veldje de waarde van F_{kp} (zie 403.10). De rasinvloeden zijn dus reeds uitgeschakeld. Om nu met behulp van formule (403.12) waarden P_{kp} te kunnen berekenen moeten wij nog de waarde van g kiezen.

Wij kiezen

(407.3) $$g = 5$$

Met behulp van deze keuze hebben wij voor ieder veldje de waarde van P_{kp} berekend. Deze waarden zijn in de grafiek weergegeven, verbonden door een stippellijn.

Het blijkt duidelijk dat de stippellijn nog kleine fluctuaties vertoont tengevolge van de toevallige fout, maar deze fluctuaties zijn toch zeer gering. Wanneer alle punten F door een lijn verbonden werden, zouden de fluctuaties veel groter zijn.

Het is wellicht nuttig er op te wijzen *dat de fluctuaties geen maat zijn voor de toevallige fout.* De waarden P, die voor twee naast elkaar liggende veldjes zijn berekend, zijn in hoge mate gecorreleerd omdat ze grotendeels uit dezelfde waarden F zijn afgeleid. Bij gevolg zijn de fluctuaties abnormaal sterk onderdrukt in verhouding tot de schommelingen van de waarden F. De stippellijn is dus minder betrouwbaar dan hij op het oog lijkt. De waarde van de stippellijn is alleen te beoordelen met behulp van (406.5).

Uit deze tabel lezen wij af dat de afwijking tussen de stippellijn en de nullijn $(P_{kp} - \bar{\mu}_p)$ bij $n = 2$ en $g = 5$ minstens $1.06\,\sigma$ moet zijn, dus minstens 1.06×12, of 13. Het blijkt, dat deze afwijkingen voorkomen op vier plaatsen nl. de veldjes (10, 11, 13), (29—32), (41), en (47—51).

Wij zien dus dat er een viertal plaatsen zijn, die voor correctie in aanmerking komen. Maar nergens is de wenselijkheid overtuigend. Alle vier plaatsen zijn misschien ontstaan doordat telkens de toevallige fout van één veldje „toevallig" erg groot was, nl. van de veldjes 11, 31, 41 en 49. Asymptotisch is het wenselijk te corrigeren, maar in het licht van deze mogelijkheid kunnen wij het in dit geval misschien beter nalaten. Wij kunnen het ook zo zeggen: In dit geval betwijfelen wij of het mogelijk is, de *berekende* waarde van $(P_{kp} - \bar{\mu}_p)$ te zien als een voldoend nauwkeurige schatting van de *asymptotische* waarde, die in (406.4) genoemd wordt.

Er zijn verschillende wegen, waarlangs wij een betere schatting van de asymptotische waarde kunnen vinden. Wanneer wij b.v.

de grafiek bezien voor de veldjes 10—20 dan is het kennelijk dat de kleine fluctuaties toevallig zijn. Wanneer deze fluctuaties worden genegeerd door het trekken van een strakke lijn, wordt de asymptotische afwijking beter benaderd. Ook op andere plaatsen zouden wij dit kunnen proberen.

Een tweede weg ligt in het vergroten van g. Wanneer wij niet stellen $g = 5$, maar b.v. $g = 9$ of $g = 15$, dan zullen de kleine fluctuaties steeds kleiner worden. Wij kunnen er dus naar streven g zo groot mogelijk te kiezen. Dit geeft tevens het voordeel dat aan de eis van (406.5) gemakkelijker kan worden voldaan, omdat de eis bij toenemende g minder zwaar wordt.

Het vergroten van g heeft echter een grens. Wij hebben formule (403.11) en de lateren opgebouwd op de veronderstelling dat het vruchtbaarheidsverloop over een afstand van g veldjes als lineair mag worden beschouwd. In ons voorbeeld is deze veronderstelling zeker niet meer te handhaven, wanneer $g > 20$ wordt.

Wij zouden nu als methode kunnen aanbevelen:

1e kies de g zo groot mogelijk;

2e bereken het verschuivend g-voudig gemiddelde;

3e trek gevonden slingerlijn uit de hand strak.

De lijn die men langs deze weg krijgt, zal evenwel weinig verschillen van de lijn die men vindt als men de stippellijn recht trekt, die wij vonden in de veronderstelling dat $g = 5$ was. Iemand die een zekere oefening heeft, behoeft in de regel zelfs de steun van het verschuivend gemiddelde (de stippellijn) niet.

De lijn, die in fig. (407.1) strak is getrokken, is op bovengenoemde wijze gevonden. De afwijkingen tussen deze lijn en de nullijn willen wij aanvaarden als betere benadering van de asymptotische waarde $<P_{kp} - \bar{\mu}_p>$ dan de gevonden afwijkingen. Wij zullen dus de strakke lijn op toelaatbaarheid moeten controleren.

Men ziet dat deze lijn drie extremen heeft: bij veldje 10, bij veldje 34 en bij veldje 50. Van iedere top moeten wij nagaan of hij toelaatbaar is. Hierbij moeten wij bedenken, dat aanname (407.3) zijn zin heeft verloren. Wanneer de strakke lijn goed is getrokken, dan zouden wij dezelfde lijn hebben gevonden, wanneer de stippellijn was berekend in de veronderstelling dat $g = 9$.

Toch moeten wij de beschikking hebben over een bepaalde waarde van g, omdat anders tabel (406.5) onbruikbaar is. Wij zouden deze regel willen volgen:

(407.4) *Bij het beoordelen van de toelaatbaarheid van de vruchtbaarheidscorrectie op een bepaalde plaats moet g de grootste afstand aangeven waarover de vruchtbaarheid als voldoende lineair kan worden beschouwd.*

Wanneer men een strakke lijn uit de hand trekt, bestaat de mogelijkheid dat men een verkeerde kijk op de puntenbundel heeft. Zo is bij het trekken van de strakke lijn voor de veldjes 1—9 alleen maar aansluiting gezocht bij het traject 10—20, zonder te letten op het verloop van de stippellijn ter plaatse. Men zou nu ten onrechte voor dit gebied de strakke lijn als de beste schatting van de asymptotische waarde $<P_{kp} - \bar{\mu}_p>$ kunnen beschouwen. Het is mogelijk dat de stippellijn hier een betere schatting geeft. In verband hiermee zouden wij nog de volgende regel willen formuleren:

(407.5) *Bij het beoordelen van de toelaatbaarheid van een vruchtbaarheidscorrectie moet niet alleen die waarde toelaatbaar blijken, die gegeven wordt door de „strakke lijn", maar ook de waarde van het verschil, dat berekend wordt volgens formule* (406.1) *met de waarde van g, die in* (407.4) *wordt genoemd.*

Wanneer wij nu de eerste top in de practijk toetsen, zien wij dat volgens de strakke lijn het vruchtbaarheidsverloop over een afstand van 15 veldjes zeker voldoende lineair is te noemen. Daar in (407.2) is aangenomen dat $n = 2$, is volgens tabel (406.5) een vruchtbaarheidscorrectie toelaatbaar wanneer de afwijking $0.55\,\sigma$ of ongeveer 7 is. Blijkens de grafiek voldoet de strakke lijn ruimschoots aan deze eis; blijkens berekening het gemiddelde van de eerste 15 veldjes maar nauwelijks. De strakke lijn is in het begin lager getrokken dan met de gegevens overeenkomt.

Wanneer wij alleen het traject van veldje 9—18 in ogenschouw nemen is $g = 9$. De afwijking moet dan $0.74\,\sigma$ of 9 zijn. Het is duidelijk dat nu zowel de strakke lijn als het berekend cijfer aan de voorwaarde voldoet. In dit traject is dus een vruchtbaarheidscorrectie gewenst.

Bij het laatste extreem kan de vruchtbaarheidslijn rondom veldje 49 over een afstand van 5 veldjes als voldoende lineair worden beschouwd. Volgens (406.5) moet de afwijking $1.06\,\sigma$ of 13 zijn. Dit is zowel voor de berekende waarde als voor de strakke lijn het geval, dus ook hier is de vruchtbaarheidscorrectie gewenst.

Met de top bij 34 is het een eigenaardig geval. Hier is de

strakke vruchtbaarheidslijn kennelijk krom. Ook reeds over een afstand van 3 veldjes. Wanneer wij nu $g = 3$ nemen, moet de afwijking 18 zijn, wat niet het geval is; de maximale afwijking is slechts 11. Indien wij nu toch g groter nemen, en dus het krom zijn negeren, zien wij dat de eis snel daalt. Zo moet bij $g = 7$ de afwijking 10 zijn, waaraan de veldjes 30—36 alle voldoen en hun gemiddelde dus ook. *Door een geschikte keuze van g kan ook hier nog de toelaatbaarheid van de correctie worden aangetoond.*

Willen wij nu nog de contrôle van (407.5) toepassen, dan blijkt ook hieruit dat de correctie toelaatbaar is. Toch zien we aan de grafiek dat het mogelijk is dat op de top, waar de grootste vruchtbaarheid wordt aangenomen, „toevallig" een paar slechte veldjes liggen, zodat de contrôle (407.5) negatief uit zou vallen. *In zulke gevallen dient de uitslag als positief te worden beschouwd, wanneer het mogelijk is met een grotere g dan in* (407.4) *bedoeld, een positieve uitslag te krijgen.* Hierbij moeten natuurlijk spitsvondigheden worden vermeden.

408 – Precisering van de vruchtbaarheidslijn

In de vorige paragraaf hebben wij aan de hand van een practisch voorbeeld nagegaan hoe de ontwikkelde correctiemethode in de practijk werkt. Hierbij bleek dat het mogelijk was op drie plaatsen de noodzaak van een vruchtbaarheidscorrectie aannemelijk te maken, nl. rondom de veldjes 10, 34 en 50.

Wanneer wij nader over de gebruikte criteria nadenken, is het resultaat evenwel niet bevredigend. Dit laat zich als volgt aantonen. In grafiek (407.1) zijn 55 veldjes afgebeeld. Volgens (407.2) liggen er 70 veldjes in dit blok p. Wanneer wij nu eens aannemen dat de veldjes 56—70 veel vruchtbaarder waren dan tot dusver is aangenomen, dan verandert er aan de veldjes 1—55 niets. De nullijn $\bar{\mu}_p$ moet evenwel hoger komen te liggen, omdat deze nullijn het gemiddelde van de 70 veldjes weergeeft. Hoewel de top bij veldje 34 hierdoor niets verandert, wordt hij numeriek toch l ager. Onze conclusie aan het eind van de vorige paragraaf, dat de correctie toelaatbaar was, komt hierdoor dan zwak te staan. Dit is natuurlijk onbevredigend.

Deze moeilijkheid ontstaat doordat de betekenis van $\bar{\mu}_p$ is overschat. Wanneer in de vorige paragraaf is bewezen dat rondom de veldjes 10 en 50 betrekkelijk slechte plekken voorkomen, dan is voor het tussenliggend gebied de enige natuurlijke vraag, of de

vruchtbaarheid daar betrouwbaar hoger is. M.a.w. men moet weten of de tussenliggende top betrouwbaar afwijkt van de kruisjeslijn, die in fig. (407.1) beide slechte plekken verbindt. Dit is met de top bij 34 kennelijk het geval.

Omdat $\bar{\mu}_p$ dus blijkbaar niet de natuurlijkste vergelijkingsbasis is, moeten wij trachten zijn taak aan een andere grootheid over te dragen. Daartoe moeten wij eerst nagaan welke taak $\bar{\mu}_p$ gedurende de berekening had te vervullen.

In 402 hebben wij de onderlinge verhouding tussen μ_p en λ_{kp} nagegaan. Uit (402.7) en (402.8) bleek dat μ en λ ongelijke wetten (402.10) volgden, zodat ze gescheiden moesten worden. Blijkens is daarbij wezenlijk voor het gedrag van λ_{kp} dat

(402.3) $$[\lambda_{kp}] = 0$$

a. Wanneer $\bar{\mu}_p$ door een andere grootheid wordt vervangen moet deze stelling, dat $[\lambda_{kp}] = 0$, blijven gehandhaafd. Onder λ_{kp} moet n dit verband worden verstaan: Het verschil tussen de ware vruchtbaarheidslijn, en een schatting er van zoals $\bar{\mu}_p$ er een is.

In 403 is de functie van $\bar{\mu}_p$ nader bestudeerd. Volgens (403.3) is het een schatting die van μ_p is te verkrijgen. Volgens (403.4) wordt hiermee de waarde $(\Omega_{kp} - \bar{\bar{\Omega}})$ gecorrigeerd. Maar deze correctie wordt in (403.10) weer ongedaan gemaakt, om te bereiken dat er geen plotselinge sprongen in de vruchtbaarheidslijn ontstaan op de grens van twee blokken.

b. Wanneer $\bar{\mu}_p$ door een andere grootheid wordt vervangen, moet er voor gezorgd worden dat ook dan geen sprongen ontstaan.

Doordat de correctie van (403.4) in (403.10) weer ongedaan is gemaakt is de invloed van de fout van $\bar{\mu}_p$ (als schatting van μ_p) slechts gering. Bovendien zagen wij bij het bespreken van (405.2) dat deze fout ook nog mag worden verwaarloosd. Als derde eis kunnen wij dus stellen

c. Wanneer $\bar{\mu}_p$ door een andere grootheid wordt vervangen, moet deze grootheid practisch onafhankelijk zijn van de toevallige fouten van de proef.

Aan al deze eisen kan de strakke vruchtbaarheidslijn die wij in de vorige paragraaf trokken bevredigend voldoen. De waarde die deze lijn als eerste schatting voor de vruchtbaarheid van veldje (kp) aangeeft, willen we aanduiden met ς_{kp}; de lijn zelf als ς-lijn.

Wij zullen nu de voorwaarden achtereenvolgens bespreken:

ad a. Wanneer uit de vrije hand door de puntenzwerm een vruchtbaarheidslijn wordt getrokken, is het niet zeker dat $[\lambda_{\kappa p}] = 0$ is. Na contrôle kan evenwel de lijn dusdanig gewijzigd worden dat aan de voorwaarde bevredigend wordt voldaan.

ad b. Wanneer er verschillende blokken in het verlengde van elkaar liggen, en in een grafiek alle waarden F afgezet zijn, dan zal er geen gevaar voor plotselinge sprongen in de ς-lijn ontstaan, wanneer uit de vrije hand een vruchtbaarheidslijn door alle blokken wordt getrokken. Slechts moet er op gelet worden dat de aansluiting bewaard blijft, wanneer er achteraf wijzigingen in de lijn worden aangebracht om te zorgen dat aan voorwaarde *a* wordt voldaan.

ad c. Wanneer uit de vrije hand een vruchtbaarheidslijn is geconstrueerd zal deze lijn des te meer onafhankelijk zijn van de toevallige fouten, naarmate hij eenvoudiger is. In verband hiermee zouden wij dan ook de eis willen stellen: Kies de vruchtbaarheidslijn zo eenvoudig mogelijk. In 409 wordt dit nader uitgewerkt.

Het blijkt dus dat inderdaad de ς-lijn de functie van de $\bar{\mu}_p$-lijn mag overnemen. Het criterium van (406.5) geldt dus niet alleen voor afwijkingen van de $\bar{\mu}_p$-lijn, maar ook voor afwijkingen van de ς-lijn.

409 – Practisch voorschrift voor de vruchtbaarheidscorrectie

Bij een vruchtbaarheidscorrectie kunnen wij verschillende etappes onderscheiden.

Allereerst kan men aannemen dat er geen vruchtbaarheidsverschillen zijn. De vruchtbaarheidslijn loopt dan horizontaal. Deze lijn willen wij aanduiden als ς_0.

Vervolgens kan men voor ieder blok de waarde $\bar{\mu}_p$ berekenen. Volgens de methode die in fig. (401.1) is toegepast kan hieruit een tweede ς-lijn worden geconstrueerd, die zodanig moet lopen, dat voor ieder blok bij benadering geldt (zie 408, ad *a*):

$$\frac{[\varsigma_{\kappa p}]}{n} = \bar{\mu}_p. \tag{409.1}$$

Deze ς-lijn willen wij aanduiden als ς_1. Wanneer door „Fisheren" of anderszins is gebleken dat de blokverschillen „betrouwbaar" zijn, is ook deze ς_1 voldoende gemotiveerd. Er moet wel op gelet

worden dat er niet angstvallig kleine fluctuaties in worden aangebracht om aan (409.1) te voldoen. In verband met de toevallige fouten hoeft (409.1) slechts bij benadering te gelden. Wanneer er geen betrouwbare blokverschillen zijn, is ς_1 gelijk aan ς_0.

Bij een proefveld met kleine blokken kan een nadere precisering worden bereikt door per veldje af te zetten de waarde van het verschil

$$\Omega_{kp} - \frac{\uparrow n\,[\Omega_{k\pi}]}{n}.$$

Volgens (402.7) is dit verschil weliswaar behept met fouten, zodat asymptotisch een lijn als $KLMN$ uit fig. (402.4) wordt verkregen, maar wanneer men zich dit bewust is, zal men vaak uit de vrije hand toch een goede ς-lijn kunnen trekken. Uit de ontwikkelde theorie is immers wel te schatten in welke zin en in welke mate de ς-lijn van de getekende puntenzwerm af moet wijken.

Wanneer er veel objecten onderzocht worden, zodat de blokken zeer groot worden, en ook wanneer de blokken zonder onderling verband liggen, zal het nodig zijn de waarden F_{kp} uit (403.10) te berekenen en deze in grafiek te brengen. Door deze puntenzwerm zal dan meestal een ς-lijn uit de hand getrokken kunnen worden. Desnoods moet eerst met een willekeurige waarde g voor ieder veldje een waarde P_{kp} worden berekend met behulp van (403.12). De lijn die zo bij kleine of grote velden ontstaat, willen wij aanduiden als ς_2.

Bij het controleren van ς_2 zal men gebruik moeten maken van het criterium van (406.5). Van ieder maximum en minimum moet men controleren of het voldoende van ς_1 (eventueel ς_0) afwijkt.

Wanneer een aantal maxima of minima aanvaard moeten worden, wordt voor de overigen gecontroleerd of ze voldoende afwijken van de eenvoudigste lijn die de opeenvolgende erkende maxima en minima verbindt. Als hieruit blijkt dat er nog enkele „uitstulpingen" erkend moeten worden, wordt er tenslotte een nieuwe ς-lijn getrokken, die alle erkende uitstulpingen verbindt, en bovendien globaal moet voldoen aan (409.1). Deze ς-lijn is dus gelijk aan ς_2 of eenvoudiger. Wij willen hem aanduiden als ς_3.

Het is vaak nog mogelijk, vooral bij grote blokken, de correctie ς_3 ook nog te verbeteren. Dit blijkt wanneer men let op formule (403.10). Daar ziet men dat de fout van F_{kp} beïnvloed wordt door

de waarden $\lambda_{k[p]}$. Zou men deze waarden kleiner kunnen maken dan zou ook F_{kp} nauwkeuriger bepaald kunnen worden.

Nu hebben wij in 408 besproken dat wij λ_{kp} niet alleen kunnen definiëren als het verschil tussen ware vruchtbaarheid van het veldje (kp) en het ware vruchtbaarheidsgemiddelde van blok p, maar ook als het verschil tussen de ware vruchtbaarheid van veldje (kp) en de waarde ς_{kp} die door de ς_3-lijn wordt gegeven. Wanneer ς_3 als correctielijn nuttig is zal λ_{kp} volgens de tweede definitie als regel kleiner zijn dan volgens de eerste. Het is dan dus nuttig bij het berekenen van F_{kp} volgens (403.10) de λ_{kp} in verband met ς_3 te definiëren.

Wij willen nu nog even weer wijzen op formule (402.7). Wanneer men λ_{kp} op de laatstgenoemde wijze definieert moet hier worden gelezen

$$\Omega_{kp} - \frac{\uparrow n\,[\Omega_{k\pi}]}{n} = \varsigma_{kp} + \frac{n-1}{n}\,\lambda_{kp}. \tag{409.2}$$

In verband met (405.10) mag aangenomen worden dat λ niet veel groter zal worden dan $\sigma\sqrt{\frac{n}{n-1}\cdot\frac{1}{g}}$, omdat anders de waarde van ς_{kp} anders zou zijn gekozen. Het verschil $\frac{1}{n}\,\lambda_{kp}$ tussen de ware vruchtbaarheid $\varsigma_{kp} + \lambda_{kp}$ en de waarde van het rechterlid van (409.2) bedraagt in zijn systematisch deel dus niet veel meer dan

$$\sigma\sqrt{\frac{1}{n\,(n-1)\,g}}.$$

Dit zal over het algemeen te verwaarlozen zijn. De volgende methode spreekt nu verder voor zich zelf.

(409.3)

1e Bereken de lijnen ς_0, ς_1, ς_2 en ς_3 als boven omschreven.

2e Lees de waarden ς_{kp} uit de lijn ς_3 af.

3e Bereken de waarden

$$\Omega^*_{kp} = \Omega_{kp} - \varsigma_{kp}$$

4e Bereken het rasgemiddelde

$$\overline{\Omega}^*_k = \frac{\uparrow n\,[\Omega^*_{k\pi}]}{n}.$$

Dit is nu de beste schatting van het rasgemiddelde waarover men beschikt.

5e Bepaal voor ieder veldje het verschil tussen de ongecorrigeerde opbrengst en het gecorrigeerde rasgemiddelde, dus

$$V^*_{kp} = \Omega_{kp} - \overline{\Omega}^*_k$$

6e Breng de waarden V^*_{kp} in grafiek, en teken een nieuwe vruchtbaarheidslijn.

7e Controleer de nieuwe ς-lijn (ς_4) op dezelfde manier als ς_2. De lijn die na contrôle gevonden wordt noemen we ς_5.

Omdat volgens (409.5e) het verschil wordt bepaald tussen *ongecorrigeerde* opbrengst en gecorrigeerd gemiddelde moet de lijn ς_4 worden gezien als een *vervanging* van ς_2. Bij de contrôle van ς_4 moet dus niet worden nagegaan of hij voldoende van ς_2 of ς_3 afwijkt, maar *of hij voldoende van* ς_1 *of* ς_0 *verschilt.*

Verder moet er opgelet worden dat het criterium van (406.5) is afgeleid ten behoeve van het controleren van ς_4 en *niet* ten behoeve van het controleren van ς_2. Immers (406.5) is er opgebaseerd dat (405.10) juist is. Deze veronderstelling is slechts te motiveren, wanneer door een voorlopige ς_3 lijn de grotere waarden van λ zijn gecorrigeerd. Het is dus theoretisch altijd noodzakelijk een ς_4-lijn te tekenen. Wanneer echter λ_{kp} overal klein is kan soms met het tekenen van een ς_3-lijn worden volstaan.

Wanneer daarentegen ς_5 sterk van ς_3 afwijkt, kan betwijfeld worden of ς_3 wel heeft bewerkt dat formule (405.10) juist was. In dit geval kan men van ς_5 uitgaan en zo tot een nieuwe ς_6 en ς_7 komen. Het eindpunt is bereikt wanneer de lijn waarvan men uitgaat en de lijn waartoe men komt practisch gelijk zijn.

Aan de hand van de definitieve vruchtbaarheidslijn $\overline{\varsigma}$ vindt men dan als gecorrigeerde opbrengst van ieder veldje

(409.4) $$\Omega_{kp}^{**} = \Omega_{kp} - \overline{\varsigma}_{kp}$$

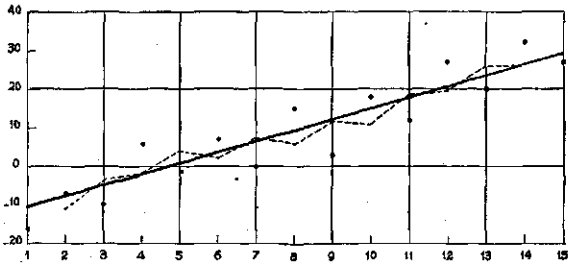
en als waarde van de rasconstante ϱ_k

(409.5) $$\overline{\varrho_k}^{**} = \frac{\uparrow n\,[\Omega_{k\pi}^{**}]}{n} - \overline{\overline{\Omega}}^{**}.$$

410 – Het verbruikt aantal vrijheidsgraden

Het grote bezwaar dat tegen een grafische vruchtbaarheidscorrectie wordt gevoeld is dat het niet gemakkelijk is een overzicht te houden over het gebruik van de vrijheidsgraden, zodat zich geen toevallige fout laat berekenen. Wij willen nu nagaan of dit bezwaar kan worden ondervangen.

(410.1)



Om de gedachten te bepalen willen wij beginnen met een voorbeeld te geven. In figuur (410.1) zijn een aantal waarden F afgezet. Volgens formule (403.12) zijn met een verschuivend drievoudig gemiddelde ($g = 3$) de waarden P berekend. Deze zijn in de grafiek getekend, verbonden door een stippellijn. Wij kunnen de stippellijn nu dus als een vruchtbaarheidslijn beschouwen. Hoeveel vrijheidsgraden heeft deze lijn geëist? Langs verschillende wegen kunnen wij trachten hierop een antwoord te vinden.

In de eerste plaats kunnen wij trachten een formule voor de gevonden lijn op te stellen en te zien hoeveel constanten deze formule bevat. Het aantal verbruikte vrijheidsgraden is gelijk aan het aantal constanten. Het is duidelijk dat het maar zelden gelukt een passende formule te vinden.

In de tweede plaats kan getracht worden na te gaan door hoeveel *punten* de lijn bepaald wordt. Dit aantal is gelijk aan het aantal constanten uit de formule, maar is gemakkelijker te vinden. (Denk b.v. aan de lijn $y = mx + q$, die door twee punten bepaald is en $y = ax^2 + bx + c$ die door drie punten bepaald is. Zo is $x^2 + bxy + cy^2 + dx + ey + f = 0$ door 5 punten bepaald).

Het aantal punten dat nodig is om onze gebroken lijn te bepalen is gelijk aan het aantal knikpunten van de lijn en dus practisch gelijk aan het aantal veldjes. Wanneer dit aantal vrijheidsgraden inderdaad benodigd was voor het berekenen van de vruchtbaarheidslijn zou het totaal aantal verbruikte vrijheidsgraden groter worden dan het beschikbare aantal, daar het ook nog mogelijk blijkt een aantal rasgemiddelden te berekenen. Dit is in strijd met de stelling aan het eind van 101. Bovendien zou verwacht mogen worden dat de kwadraatsom der afwijkingen de waarde 0 aannam, wat blijkens grafiek (410.1) ook niet het geval is. Dit is dus niet de juiste methode voor het vaststellen van het verbruikte aantal vrijheidsgraden.

De fout van onze poging is dat wij iedere waarde P_{kp} die in fig.

(410.1) in beeld is gebracht, als een „uitkomst" hebben opgevat. In 101 vonden wij echter dat men slechts dan van „uitkomsten" kan spreken als ze onderling onafhankelijk zijn. Dit is met de waarden P geenszins het geval. De waarden P die voor de veldjes 5 en 6 zijn berekend berusten beide op de waarden F van de veldjes 5 en 6. Pas de waarde P van veldje 8 is onafhankelijk van die van veldje 5. De waarden P van de veldjes 6 en 7 zijn er op een min of meer omslachtige manier tussen geïnterpoleerd, zonder nieuwe uitkomsten te verschaffen.

In verband met het feit dat we $g = 3$ gekozen hebben zouden wij dus kunnen zeggen dat er hoogstens $^{15}/_{3} = 5$ vrijheidsgraden verbruikt kunnen zijn, omdat er niet meer onafhankelijke waarden P zijn berekend.

Maar ook dit aantal is nog te hoog. In 409 bespraken wij dat wij na een ς_3 nog een ς_5 en misschien een ς_7 zochten tot we de definitieve vruchtbaarheidslijn hadden gevonden. Bij dit zoeken van de definitieve lijn wordt de natuurlijke verwantschap van de P van veldje 5 en van veldje 8 in het oog gehouden. Over deze afstand wordt naar een zekere continuïteit gestreefd in het vruchtbaarheidsverloop, waardoor ook de waarden P van veldje 5 en veldje 8 onderling afhankelijk worden.

Bij het bestuderen van het verbruikt aantal vrijheidsgraden moeten wij ons dus baseren op de strakke ς-lijn, die wij als definitieve vruchtbaarheidslijn beschouwen.

Van deze lijn is soms een formule te vinden zoals in voorbeeld (410.1) waar het een rechte lijn is. In dit geval is het verbruikte aantal vrijheidsgraden hoogstens gelijk aan het aantal parameters der lijn. Het verbruikt aantal *kan* lager zijn, wanneer de definitieve vruchtbaarheidslijn ingewikkelder is gekozen dan uit de gegevens valt af te leiden. Dit zou het geval geweest zijn wanneer de stippellijn uit (410.1) als definitieve vruchtbaarheidslijn was beschouwd.

Van de ς-lijn van figuur (407.1) is geen formule bekend. Wij hebben getracht voldoende aansluiting te vinden bij een algebraïsche functie van de $(a + 1)^e$ graad wanneer er a maxima en minima waren. Voor ons lijnstuk met 1 maximum en 2 minima is $a = 3$, zodat de formule zou moeten worden

$$y = Ax^4 + Bx^3 + Cx^2 + Dx + E;$$

er zouden dan dus $(a + 2)$ vrijheidsgraden verbruikt zijn. Meestal

is de aansluiting slecht, wat niet betekent dat er meer vrijheidsgraden verbruikt zijn, slechts is de contrôle-methode verkeerd.

Meer succes heeft men wanneer men tracht na te gaan door hoeveel punten de lijn bepaald wordt. Onder grafiek (407.1) hebben wij 6 kruisjes geplaatst. Wanneer iemand gevraagd wordt door deze kruisjes de eenvoudigste vloeiende lijn te trekken die mogelijk is, zal er vermoedelijk een lijn ontstaan, die niet veel van de ς-lijn van (407.1) afwijkt. De gevonden ς-lijn is dus door ongeveer 6 kruisjes wel bepaald.

Als regel voor het vaststellen van het verbruikt aantal vrijheidsgraden kunnen wij dus geven:

(410.2) *Ga na hoeveel punten minstens gegeven moeten zijn om de lijn dragelijk te reconstrueren.* Met dragelijk reconstrueren bedoelen wij, dat men zoveel punten kiest dat de verschillen tussen de lijnen, die men er op 't oog door trekt, klein zijn in verhouding tot de toevallige fouten. Het minimum aantal benodigde punten willen wij b noemen.

Feitelijk is het getal b steeds 1 te hoog omdat de absolute hoogte van de vruchtbaarheidslijn onbelangrijk is. In verband met het feit dat b geschat is, kan men dit echter verwaarlozen. Het getal b geeft nu aan hoeveel vrijheidsgraden *hoogstens* verbruikt zijn.

411 – De bepaling van de toevallige fout

Wanneer men de formules met elkaar vergelijkt die Van Uven geeft voor de bepaling van de varians σ^2, ziet men dat ze kunnen worden samengevat in de algemene formule

$$\sigma^2 = \frac{[u^2]}{N - M} \tag{411.1}$$

waarbij N het beschikbaar aantal vrijheidsgraden aangeeft, en M het reeds verbruikte aantal. Uit het feit dat deze formule vaak voorkomt mag niet worden afgeleid dat hij algemeen geldig is. Wij willen op twee beperkingen wijzen.

In 128 hebben wij nagegaan volgens welk principe de formule voor een geval van lijnvereffening moest worden gekozen. Wij zagen dat de richting van middelen niet steeds doorslaggevend was. Dit neemt niet weg, dat de door Van Uven gegeven richting van middelen bij regressielijnen wel beslissend is voor de richting waarin de grootte van de fout moet worden gemeten. Wanneer

men nu de voorkeur geeft aan een andere lijn dan die met de methode der *kleinste* kwadraten voor die richting van middelen wordt berekend, dan houdt dit in dat men een grotere kwadraatsom moet vinden. Uit het feit dat men toch de voorkeur aan de lijn met de kleinste kwadraten onthoudt, blijkt niet dat men σ^2 groter wil schatten. De noemer van (411.1) moet dus wel groter worden genomen.

Dit ligt ook voor de hand. Een berekening volgens de methode der kleinste kwadraten is erop gespitst de fouten weg te verklaren. Men zoekt immers de kleinste kwadraten. Blijkbaar geeft iedere uitkomst die berekend wordt, een mogelijkheid meer de schijnbare fout te verkleinen. Vandaar dat het aantal verbruikte vrijheidsgraden (M) dan voor compensatie in de noemer staat.

Wanneer men een ander principe volgt, eventueel grafisch werkt, is men er niet zo op gespitst. Dan vervalt ook de reden, M in zijn geheel van het beschikbaar aantal vrijheidsgraden N af te trekken. *Dus wanneer volgens een ander principe wordt vereffend dan volgens de methode der kleinste kwadraten is er de mogelijkheid dat de schatting van σ^2 in* (411.1) *te hoog is.*

Een tweede beperking van de geldigheid van (411.1) houdt verband met formule (116.12). In deze formule wordt gezegd dat de totale varians ψ^2 bestaat uit de toevalsvarians σ^2 en nog een systematisch deel dat niet kleiner dan 0 kan worden.

Asymptotisch is dit juist, d.w.z. wanneer men een bepaalde proef steeds weer herhaalt, zal er steeds weer een varians berekend kunnen worden. Het gemiddelde van al deze variansen voldoet aan formule (116.12). Ook het voorschrift, dat wij aan het eind van 118 gaven, heeft slechts asymptotische geldigheid. Daar stelden wij als regel voor het kiezen van een formule: *Kies die (nood)-formule, die de kleinste varians geeft.* Wij wezen er daar op dat volgens dit principe de ideale formule gevonden moet worden als die zich onder de beproefde formules bevindt. Ook dit is slechts asymptotisch waar. D.w.z. als men steeds weer hetzelfde probleem met dezelfde formule tracht op te lossen, zal men op den duur met de ideale formule het best uitkomen.

Bij vruchtbaarheidscorrecties kan men niet asymptotisch werken. Een vruchtbaarheidskaart doet zich maar eenmaal voor. Hetzelfde perceel geeft het volgend jaar onder andere weersomstandigheden een ander vruchtbaarheidsbeeld. Bij vruchtbaarheidscorrecties mag men daarom niet veronderstellen dat de

dubbele producten tussen toevallige en systematische fouten nul zullen zijn. Er is een tendens dat ze negatief zullen zijn, zodat ψ^2 kleiner wordt dan σ^2 is.

In verband met deze tweede beperking is het ook niet mogelijk de doelmatigheid van een vruchtbaarheidscorrectie te beoordelen aan de hand van de resterende varians. Dit is de reden waarom het criterium (406.5) nodig is.

Aangezien door dit criterium het wegverklaren van toevallige afwijkingen vaak wordt tegengegaan (zie 404.9; 405.9; 405.11) willen wij aannemen dat ook het zoeken van negatieve dubbele producten binnen redelijke grenzen blijft.

Bij grafische vruchtbaarheidscorrecties heeft men dus met twee tendensen te maken. Er is een tendens om de schatting van de varians te groot te laten zijn en er is een tendens om diezelfde schatting te verkleinen.

Wanneer wij aannemen, dat die beide tendensen elkaar in het evenwicht houden, kunnen wij als formule voor de varians geven

(411.2) $$\psi^2 = \frac{\Sigma u^2}{mn - m - b}.$$

waarbij u aanduidt het verschil tussen definitief gecorrigeerde opbrengst en definitief rasgemiddelde.

In (411.2) hebben wij de varians aangeduid door ψ^2 en niet door σ^2. Wij moeten niet vergeten dat de varians niet alleen is opgebouwd uit zuivere toevallige fouten, maar ook uit geringe vruchtbaarheidsverschillen die niet konden worden weggecorrigeerd (zie 406.5). In formule (301.28) is aangetoond dat dergelijke bestanddelen soms andere wetten volgen dan σ^2. Het lijkt ons waarschijnlijk dat dit verschil in dit geval mag worden verwaarloosd.

De waarde ψ^2 van formule (410.2) moet nu voor tweeërlei doel gebruikt worden. In de eerste plaats moet de nauwkeurigheid van de gecorrigeerde rasgemiddelden er door worden aangegeven. Die nauwkeurigheid zal niet alleen beïnvloed worden door de toevallige fouten (σ), maar ook door het te kort schieten van de vruchtbaarheidcorrectie (λ). De waarde ψ, die beide elementen bevat, mag wel als een geschikte grootheid worden beschouwd om de onzekerheid weer te geven. Wij kunnen dus globaal van de middelbare fout van het gecorrigeerde rasgemiddelde, en dus ook van de berekende waarde $\overline{\varrho_k{}^{**}}$ zeggen

(411.3) $$\sigma_{\overline{\varrho_k{}^{**}}} = \frac{\psi}{\sqrt{n}}.$$

In de tweede plaats moet ψ^2 dienen om na te gaan of een correctie geoorloofd is. In tabel (406.5) is een criterium gegeven voor de toelaatbaarheid, maar dit criterium hangt af van de waarde van σ, die niet bekend is. De waarde ψ kan als een schatting van σ worden genomen, waarbij men mag bedenken dat ψ voor dit doel vermoedelijk te groot is, zodat *niet te snel* tot correctie wordt overgegaan. Bij het toetsen van de ς_4-lijn (zie 409) moet dan een schatting van ψ worden berekend met behulp van de ς_3-lijn.

412 – VRUCHTBAARHEIDSCORRECTIES BIJ PROEVEN IN VIERKANTSVERBAND

Tot dusver hebben wij steeds verondersteld dat wij een strokenproef bewerkten, dus een proef, waarbij alle veldjes achter elkaar in een strook lagen. Deze aanname diende alleen om iets gemakkelijker de gedachten weer te kunnen geven; geen enkele berekening is er op gebaseerd. De voorgaande beschouwingen zijn dus gelijkelijk toepasselijk wanneer men te maken heeft met een proef in vierkantsverband; waarbij de veldjes dus gerangschikt zijn als de vakjes van een dambord.

Er zou dan ook geen enkele reden zijn speciaal over deze proeven te spreken wanneer er niet de moeilijkheid was, dat er feitelijk drie dimensionaal gewerkt moet worden. Er zijn twee assen nodig voor de plaatsbepaling van het veldje en een voor de vruchtbaarheid.

Omdat er geen principiële punten bij aan de orde komen, willen wij hier gewoon een voorschrift geven. Dit voorschrift is er op gebaseerd dat een proef in vierkantsverband volgens twee richtingen (evenwijdig aan de coördinaatassen) in stroken kan worden verdeeld. De stroken evenwijdig aan de ene as kunnen als „rijen" worden aangeduid, de stroken evenwijdig aan de andere as als „kolommen".

Een vruchtbaarheidscorrectie is nu in de eerste plaats mogelijk door het proefveld in rijen of kolommen in te delen, en de gevormde stroken gewoon als strokenproef te verwerken. De nauwkeurigheid is dan even groot als die van een normale strokenproef met stroken van dezelfde lengte.

Het ligt evenwel voor de hand de nauwkeurigheid op te voeren door de vruchtbaarheidslijnen van naast elkaar liggende stroken op elkaar te vereffenen. Aan het Landbouwproefstation in Gro-

ningen werd mij een methode gedemonstreerd die op het volgende neer kwam:

1e Het proefveld werd in rijen verdeeld.
2e In iedere ontstane strook werd een vruchtbaarheidslijn getekend.
3e Voor ieder veldje werd genoteerd welke vruchtbaarheid daaraan moest worden toegekend op grond van de vruchtbaarheidslijnen in de rijen.
4e Nu werd het proefveld in kolommen verdeeld.
5e In deze kolommen werden de waarden die volgens de rijen waren gevonden op elkaar vereffend.
6e De waarden die nu werden gevonden werden nog eens volgens de rijen vereffend.
7e Hetzelfde werd gedaan door de oorspronkelijke afwijkingen eerst volgens kolommen te vereffenen, dan volgens rijen en tenslotte weer volgens kolommen.
8e De waarden die langs beide wegen waren gevonden werden tenslotte gemiddeld.

Deze methode heeft als nadeel dat het achteraf moeilijk is te schatten hoeveel vrijheidsgraden verbruikt zijn. Verder kan het onderling verband tussen naastliggende rijen of kolommen misschien gemakkelijker langs een andere weg worden bereikt.

In de practijk is onderstaande methode goed bevallen:

1e Vul op een plattegrond van het proefveld voor ieder veldje de waarde F_{kp} (zie 403.10) in.
2e Vorm een blokje van g veldjes (meestal negen, dus 3×3). Bereken uit de waarden F de waarde P_{kp} (zie 403.12). Dit is dan een schatting van de vruchtbaarheid van het middelste veldje. Bereken zo voor ieder veldje de waarde P.
3e Voor de buitenste rijen en buitenste kolommen kan zo geen waarde P worden berekend. Dit kan desnoods door deze als strokenproef op te vatten maar mag meestal worden nagelaten.
4e Teken voor iedere rij en iedere kolom een grafiek, waarin de waarden P op elkaar worden vereffend. Hierbij moet het verloop der punten wel tamelijk goed gevolgd worden.
5e Lees uit deze grafieken af op welke plaatsen in iedere rij en iedere kolom de vruchtbaarheid een bepaalde waarde A heeft.

6e Geef deze plaatsen door een pijltje aan op een plattegrond van het proefveld. Laat daarbij het pijltje aanwijzen in welke richting de vruchtbaarheid lager is.

7e Vereffen deze pijltjes op elkaar door er een niveaulijn tussen door te trekken.

8e Construeer zo het benodigd aantal niveaulijnen. Voordat met een volgende lijn begonnen wordt moeten alle pijltjes van de voorgaande zijn uitgegomd, omdat anders verwarring ontstaat.

9e Tracht de gevonden kaart zoveel mogelijk te vereenvoudigen.

10e Controleer met behulp van het criterium van (406.5) of de vereenvoudigde kaart toelaatbaar is, en of de niet-vereenvoudigde kaart ook teveel van de vereenvoudigde afwijkt. Maak in de geest van 409 een definitieve kaart.

Wanneer men langs deze weg een vruchtbaarheidskaart heeft geconstrueerd, moet men nog trachten na te gaan hoeveel vrijheidsgraden er door zijn verbruikt. Hierbij staan dezelfde methoden ten dienste die in 410 zijn besproken. In de eerste plaats kan men trachten een formule voor de gevonden kaart te vinden. Dit gelukt in de practijk bijna nooit.

In de tweede plaats kan men nagaan hoeveel „punten" moeten worden vastgelegd om de kaart voldoende te reconstrueren. Het begrip „punten" hoeft dan niet zeer letterlijk te worden genomen. Meestal is het 't gemakkelijkst één of enkele niveaulijnen door een aantal punten vast te leggen, en van de anderen na te gaan op welke afstand ze evenwijdig aan de eerste lopen. Wil men deze methode toepassen, dan moet men er bij het tekenen van de kaart naar streven de lijnen evenwijdig te trekken. Behalve naar evenwijdigheid kan natuurlijk ook naar andere eenvoudige mathematische samenhangen worden gestreefd.

In de derde plaats kan geconstateerd worden, dat het aantal vrijheidsgraden nooit groter is dan het aantal veldjes, gedeeld door g. Dit is bij deze aanleg meestal de gemakkelijkste contrôle. Geschat moet dan worden hoe groot g genomen mag worden om toch dezelfde vruchtbaarheidskaart te krijgen.

STELLINGEN

I

Variansen, die bij een variansanalyse betrouwbaar groter blijken dan de kleinste varians die gevonden is, bevatten waarschijnlijk systematische elementen, en kunnen daarom niet zonder meer als „toetsvarians" dienen.

II

Groene erwten wijken in hun reactie op uitwendige omstandigheden zo zeer af van capucijners en grauwe erwten, dat ze als een afzonderlijk landbouwgewas moeten worden beschouwd.

III

De rechtszekerheid, welke het Kwekersbesluit 1941 aan de kwekers wil geven, zal slechts bereikt kunnen worden, indien bij een geding over het plaatsen van een ras op de rassenlijst de bewijslast van de bijzondere waarde van het betreffende ras op de kweker wordt gelegd.

IV

Het duurzaam samenleven van twee of meer volken op hetzelfde grondgebied is bijna uitsluitend mogelijk, wanneer deze volken bereid zijn samen te smelten.

Wanneer ons volk de banden met Indonesië wil handhaven, zal het hier volledig rekening mee moeten houden.

V

Bij grafieken, die het verloop weergeven van indexcijfers of koersen, is het wenselijk op de as voor deze grootheden een logarithmische schaalverdeling te gebruiken.

VI

Bij het oplossen van het kleine boerenvraagstuk is niet maatgevend, hoeveel personen in de landbouw werkzaam zullen zijn, maar wel, hoe doelmatig er gewerkt zal worden.

G. HAMMING