MATHEMATISCH CENTRUM
AMSTERDAM

STATISTISCHE AFDELING

21 November 1949.

Vijfde verslag betreffende groeiproeven met Wistar-ratten. 1)

I

1. De correlatie- en regressieberekeningen hadden tot nog toe betrekking op de grootheden g_k (k = 1,...,4), zijnde de gemiddelde gewichtsvermeerdering binnen een zekere groep gedurende de eerste k weken tezamen. De correlatie, die tussen g_k en $g_{k'}$ (k,k' = 1,...,4) bestaat, wordt echter gedeeltelijk veroorzaakt door het feit, dat beide variabelen deels dezelfde elementen bevatten; immers, als $k < k'$ is, is $g_{k'}$ gelijk aan g_k plus de gewichtsvermeerderingen gedurende de (k+1)ste,...,k'-de week.
Daarom verzocht Drs. Thomasson deze berekeningen op enkele essentiële punten te herhalen voor de grootheden h_k (k = 1,...,4), zijnde de gemiddelde gewichtsvermeerdering van een groep gedurende de k-de week alleen. Deze berekeningen, waarvan het resultaat sub 2 wordt gegeven, zijn door onze Rekenafdeling verricht.

2. De grootheden g_k en h_k hangen dus als volgt samen:

$$g_1 = h_1$$
$$g_2 = h_1 + h_2$$
$$g_3 = h_1 + h_2 + h_3$$
$$g_4 = h_1 + h_2 + h_3 + h_4.$$

In onderstaande Tabellen I, II en III zijn op grond van deze relaties berekend:
tabel I: de correlatie-coefficienten $r_{kk'}$, die het verband aangeven tussen de groei gedurende de k-de week (h_k) en die gedurende de k'-de week ($h_{k'}$);
tabel II: de regressiecoefficienten $b_{kkk'}$, gedefinieerd volgens de regressie-vergelijking

$$h_k = b_{kk'}h_{k'} + \text{constante};$$

tabel III: de regressiecoefficienten $b_{k'k}$, die met de bovengenoemde grootheden samenhangen volgens de identiteit

$$b_{kk'}b_{k'k} = r^2_{kk'}$$

Tabel I: $r_{kk'}$

k \ k'	1	2	3	4
1	1	0,903	0,689	0,405
2		1	0,826	0,567
3			1	0,726
4				1

1) Door H. Theil en J. Hemelrijk.

Tabel II: $b_{kk'}$

k \ k'	1	2	3	4
1	1	0,722	0,627	0,467
2		1	0,941	0,819
3			1	0,919
4				1

Tabel III: $b_{k'k}$

k \ k'	1	2	3	4
1	1	1,141	0,757	0,352
2		1	0,725	0,394
3			1	0,574
4				1

3. Zoals te verwachten is, zijn de correlatiecoefficienten $r_{kk'}$ tussen de h's weliswaar positief, maar lager dan de corresponderende correlatiecoefficienten $r'_{kk'}$ tussen de g's. Ter vergelijking volgt hieronder een tabel voor $r'_{kk'}$.

Tabel IV: $r'_{kk'}$

k \ k'	1	2	3	4
1	1	0,97	0,92	0,87
2		1	0,98	0,94
3			1	0,98
4				1

II

4. Drs. Thomasson maakt voor zijn significantieberekeningen gebruik van de grootheid G_4, hieronder te bespreken. Ons werd verzocht de spreiding van deze grootheid te berekenen. Deze berekening stuitte op moeilijkheden, want de gevraagde spreiding bleek afhankelijk te zijn van enige onbekende correlatiecoefficienten, waarvan de berekening wegens de grote uitgebreidheid van het materiaal, vrij kostbaar zou zijn. Daarom werd voorlopig volstaan met een - zij het ruwe - schatting van deze spreiding aan de veilige kant, welke wèl zonder omvangrijke berekeningen te geven bleek.

5. Van de volgende notatie wordt gebruik gemaakt:
   x_{ijk} : de gewichtsvermeerdering gedurende de eerste k weken van rat j in groep i.
   N_i : het aantal ratten in groep i. Aangenomen is, dat dit aantal constant is gedurende alle weken. In werkelijkheid neemt het door sterfte e.d. af, hetgeen echter

terwille van de eenvoud is verwaarloosd.

x_{ik} : de gemiddelde gewichtsvermeerdering gedurende de eerste k weken in groep i:

$$x_{ik} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ijk}$$

g_{ik} : dito, echter naar boven of naar beneden afgerond:

$$g_{ik} = x_{ik} + \xi_{ik} \quad ,$$

waarbij

ξ_{ik} beschouwd wordt als een stochastische variabele, die homogeen verdeeld is in het interval

$$-\frac{1}{2} < \xi_{ik} \leq \frac{1}{2}$$

(zie het Vierde Verslag).

Dan wordt de G_4-waarde van groep i (kortheidshalve geschreven als G_i) gedefinieerd volgens

$$\begin{aligned} 4G_i &= \sum_{k=1}^{4} \frac{G_{ik}}{k} \\ &= \sum_{k=1}^{4} \frac{x_{ik} + \xi_{ik}}{k} \\ &= A_i + \sum_{k=1}^{4} \frac{\xi_{ik}}{k} \end{aligned} \qquad (1)$$

waarbij

$$\begin{aligned} A_i &= \sum_{k=1}^{4} \frac{x_{ik}}{k} \\ &= \sum_{k=1}^{4} \frac{1}{k} \cdot \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ijk} \\ &= \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{x_{ij1}}{1} + \frac{x_{ij2}}{2} + \frac{x_{ij3}}{3} + \frac{x_{ij4}}{4} \right) \end{aligned}$$

Nu is de spreiding van x_{ijk}/k ongeveer gelijk aan $\sigma = 3{,}8$ (zie het Vierde Verslag). Nu geldt:

$$\begin{aligned} \operatorname{var}(N_i A_i) = \sum_{j=1}^{N_i} \Big[& \operatorname{var} \frac{x_{ij1}}{1} + \ldots + \operatorname{var} \frac{x_{ij4}}{4} + \\ & + 2 \operatorname{cov} \left(\frac{x_{ij1}}{1}, \frac{x_{ij2}}{2} \right) + 2 \operatorname{cov} \left(\frac{x_{ij1}}{1}, \frac{x_{ij3}}{3} \right) + \\ & + \ldots + 2 \operatorname{cov} \left(\frac{x_{ij2}}{2}, \frac{x_{ij3}}{3} \right) + \ldots \Big] \\ = N_i \Big[& 4\sigma^2 + 2\sigma^2 (\rho^i_{12} + \rho^i_{13} + \rho^i_{14} + \rho^i_{23} + \rho^i_{24} + \rho^i_{34}) \Big] , \end{aligned}$$

waarbij $\rho_{kk'}$ $(k,k' = 1,\ldots,4;\ k \neq k')$ de correlatiecoefficient is,

die het verband aangeeft tussen de gewichtsvermeerderingen van de ratten in groep i gedurende de eerste k weken en die van dezelfde ratten gedurende de eerste k' weken. Deze correlatiecoefficienten zijn echter niet bekend: de correlatieberekeningen van onze Rekenafdeling hadden immers betrekking op ongeveer 200 groepsgemiddelden (terwijl het hier gaat om afzonderlijke ratten) en beperkten zich evenals de berekeningen van drs. Thomasson niet tot afzonderlijke groepjes, maar hadden betrekking op de grote collectie (van N~/300 waarnemingen). Een bovengrens van de spreiding van A_i kan echter gegeven worden, immers geldt:

$$\begin{aligned} \text{var } A_i &= \frac{1}{N_i} \text{ var } (N_i A_i) \\ &= \frac{\sigma^2}{N_i}\left[4 + 2\sum_{k=1}^{3}\sum_{k'=k+1}^{4} \rho^{i}_{kk'}\right] \\ &\leq \frac{16\,\sigma^2}{N_i} \end{aligned}$$

Voor de variantie van $\sum_i'(\xi_{ik}/k)$ in (1) kan geschreven worden (zie het Vierde Verslag):

$$\begin{aligned} \text{var } \left(\sum_{k=1}^{4}\frac{\xi_{ik}}{k}\right) &= \frac{1}{12}\left(1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16}\right) \\ &= 0{,}119\,. \end{aligned}$$

Dan is (zie (1)):

$$\begin{aligned} \text{var } G_i &= \frac{1}{16} \text{ var } \left(A_i + \sum_{k=1}^{4}\frac{\xi_{ik}}{k}\right) \\ &\leq \frac{\sigma^2}{N_i} + 0{,}007\,. \end{aligned} \qquad (2)$$

Hieruit kan een bovengrens voor σ_{g_i} (de spreiding van G_i) berekend worden. Deze wordt gegeven in Tabel V, voor $N_i = 7, 8, \ldots, 12$; voor σ is de waarde 3,8 gekozen.
De term 0,007 kan verwaarloosd worden; de afronding van de gemiddelden heeft ook hier geen invloed van betekenis.

Tabel V: bovengrens voor σ_{g_i}

N_i	Bovengrens voor σ_{g_i}
7	1,45
8	1,35
9	1,27
10	1,20
11	1,15
12	1,10

6. Met behulp van deze grootheid σ_{g_i} kan men nu, afwijkingen van normaliteit van de verdeling van G_i verwaarlozende en, evenals bij de G_i, werkende met een spreiding, die slechts afhangt van de grootte van de steekproef (en niet van het voedsel) een significantietest opstellen voor het verschil van twee G-waarden. Indien G_i en G_j de gevonden waarden zijn

in groepjes met N_i en N_j ratten, is de grootheid

$$\frac{G_i - G_j}{\sigma} \sqrt{\frac{N_i N_j}{N_i + N_j}}$$

(met $\sigma = 3{,}8$) normaal verdeeld met gemiddelde nul en een spreiding ≤ 1. Hierbij is de term 0,007 van (2) verwaarloosd. Indien toepassing van deze significantie-test een overschrijdingskans van bijv. 0,05 oplevert, is de werkelijke overschrijdingskans $\leq 0{,}05$. In het bijzonder als $N_i = N_j = 12$ is,

$$0{,}64(G_i - G_j)$$

normaal verdeeld met gemiddelde 0 en een spreiding ≤ 1.

III

7. Drs. Thomasson verzocht ons tenslotte een kritiek te geven op het gebruik van de G_4-waarden, en verder enige suggesties te verstrekken ter vervanging van deze grootheden door doelmatiger. De kritiek volgt sub 8; het slot van dit verslag is gewijd aan een methode, die wij aan de opdrachtgever willen voorleggen.

8. Uit de vier vergelijkingen, vermeld sub 2, volgt:

$$G = \frac{25h_1 + 13h_2 + 7h_3 + 3h_4}{48}$$

Hieruit blijkt, dat G in hoofdzaak bepaald wordt door h_1, terwijl de gewichtstoenamen gedurende de latere weken, in het bijzonder de vierde, weinig gewicht in de schaal leggen. De Heer Thomasson deelde ons echter mede, dat de wijziging van het voedsel van grote invloed is op h_1; deze wijziging nu staat geheel los van het doel van het onderzoek: de bepaling van de groeisnelheid. Doordat G als maat van de groeisnelheid zo sterk door h_1 beinvloed wordt, ondervindt zij ipso facto grote invloed van de voedselwijziging. Daarom moet G beschouwd worden als een onzuivere maatstaf van de groeisnelheid. Aan dit principiële bezwaar moet het practische nadeel worden toegevoegd, dat de verdeling van G onbekend is en dat haar spreiding slechts geschat kan worden.

9. Een algemeen principe, waaruit men tot een keuze zou kunnen komen van een grootheid, die G_4 zou kunnen vervangen, wordt hieronder beschreven.
De bedoeling van drs. Thomasson is de groei van een rattengroepje in één cijfer uit te drukken, daarbij gebruik makende van de grootheden $g_1, \ldots, g_4$. Prof. van Dantzig gaf, naar analogie van de methoden der "discriminant functions" het volgende algemene principe voor het bepalen van een keuze

uit de zeer vele mogelijkheden aan:
Beperken wij ons tot lineaire combinaties van de $g_1,\ldots,g_4$, dus tot grootheden van de vorm

$$G^* = \lambda_1 g_1 + \lambda_2 g_2 + \lambda_3 g_3 + \lambda_4 g_4, \qquad \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$$

waarin de λ's nog gekozen kunnen worden [1], dan kan men nu trachten deze keuze zo uit te voeren, dat de G^*, wanneer men deze voor alle ratten van een groep apart zou berekenen, een spreiding zou vertonen, die ten opzichte van de spreiding van G^* over de verschillende groepen zo klein mogelijk is. Dit leidt dan tot de volgende berekeningen:
Zij wederom x_{ijk} de voedseltoename van de j^{e} rat uit de i^{e} groep gedurende de eerste k weken en zij verder

$$g_{ik} \approx \frac{1}{N_i} \sum_{j=1}^{N} x_{ijk}$$ (eigenlijk is g_{ik} de afgeronde waarde van het rechterlid). Kent men nu aan de weekreeksen de gewichten $\lambda_1,\ldots,\lambda_4$ toe, dan wordt de spreiding van de met G^* voor iedere rat overeenkomende grootheid gemiddeld genomen over alle groepen, gegeven door:

$$A = \frac{1}{n} \sum_i \frac{1}{N_i} \sum_j \Big\{ \sum_k \lambda_k x_{ijk} - \sum_k \lambda_k g_{ik} \Big\}^2 =$$

$$= \frac{1}{n} \sum_i \frac{1}{N_i} \sum_j \sum_k \sum_{k'} \lambda_k \lambda_{k'} (x_{ijk} - g_{ik})(x_{ijk'} - g_{ik'}) =$$

$$= \sum_k \sum_{k'} \lambda_k \lambda_{k'} C_{kk'}$$

met $C_{kk'} = \frac{1}{n} \sum_i \frac{1}{N_i} \sum_j (x_{ijk} - g_{ik})(x_{ijk'} - g_{ik'})$.
De spreiding tussen de groepen wordt, als we nog

$$u_k = \frac{1}{n} \sum_i g_{ik}$$

stellen, gegeven door

$$B = \frac{1}{n} \sum_i \Big\{ \sum_k \lambda_k g_{ik} - \sum_k \lambda_k u_k \Big\}^2 =$$

$$= \sum_k \sum_{k'} \lambda_k \lambda_{k'} R_{kk'}$$

met $R_{kk'} = \frac{1}{n} \sum_i (g_{ik} - u_k)(g_{ik'} - u_{k'})$

Indien de covariantiematrices $(C_{kk'})$ en $(R_{kk'})$ van de kwadratische vormen A en B bekend zijn, kan men de waarden $\lambda_1,\ldots,\lambda_4$ waarvoor de breuk $\frac{A}{B}$ minimaal wordt berekenen. Deze gewichten hebben dan het voordeel, dat het resulterende getal G^* het onderscheid tussen de verschillende groepen zo duidelijk mogelijk doet uitkomen in vergelijking met de variabiliteit binnen

---

1) Keuze van G_4 betekent $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \frac{1}{4}$; keuze van g_4 (d.w.z. van gelijke gewichten voor de minder afhankelijke h's) betekent $\lambda_1 = \lambda_2 = \lambda_3 = 0$ en $\lambda_4 = 1$.

de afzonderlijke groepen.

10. De covarianties $C_{kk'}$, zijn echter tot nog toe niet berekend. Aangezien berekening hiervan aanzienlijke tijd zou vergen, daar het totaal aantal waarnemingen groot is, hebben wij getracht uit de reeds berekende covarianties voorlopig een schatting van de $C_{kk'}$ en met behulp daarvan van de λ_k te verkrijgen. Daar men voor $\lambda_1,\ldots,\lambda_4$ toch eenvoudige waarden zal willen gebruiken, kan reeds een vrij ruwe schatting voldoende zijn.
Bekend zijn: $R_{k,k'}$ voor iedere k en k', berekend door de Rekenafdeling van het Mathematisch Centrum. Voorts zijn van de matrix $(T_{k,k'})$, waarin

$$T_{k,k'} = \frac{1}{n}\sum_i \frac{1}{N_i}\sum_j (x_{ijk} - u_k)(x_{ijk} - u_{k'}) \quad \text{is,}$$

de grootheden T_{11}, T_{22}, T_{33}, T_{44}, $T_{12}= T_{21}$, $T_{23}= T_{32}$ en $T_{34}= T_{43}$ bekend, de grootheden $T_{13}= T_{31}$, $T_{14}= T_{41}$ en $T_{24}= T_{42}$ echter niet. Eerstgenoemde grootheden kunnen n.l. bij benadering verkregen worden uit tabel 4, na deling door de uitgebreidheden, pag.8 van "Een statistische analyse van groeiproeven met Wistarratten" van drs. Thomasson.
Nu geldt:

$$\begin{aligned} C_{kk'} &= \frac{1}{n}\sum_i \frac{1}{N_i}\sum_j (x_{ijk} - g_{ik})(x_{ijk'} - g_{ik'}) = \\ &= \frac{1}{n}\sum_i \frac{1}{N_i}\sum_j \{(x_{ijk} - u_k) - (g_{ik} - u_k)\} \cdot \\ &\qquad \cdot \{(x_{ijk'} - u_{k'}) - (g_{ik'} - u_{k'})\} \\ &= \frac{1}{n}\sum_i \frac{1}{N_i}\sum_j (x_{ijk} - u_k)(x_{ijk'} - u_{k'}) + \\ &+ \frac{1}{n}\sum_i \frac{1}{N_i}\sum_j (g_{ik} - u_k)(g_{ik'} - u_{k'}) + \\ &- \frac{1}{n}\sum_i \frac{1}{N_i}\sum_j (x_{ijk} - u_k)(g_{ik'} - u_{k'}) - \\ &- \frac{1}{n}\sum_i \frac{1}{N_i}\sum_j (x_{ijk'} - u_{k'})(g_{ik} - u_k) \end{aligned}$$

Hiervan is

1^{e} term $= T_{kk'}$

2^{e} term $= \frac{1}{n}\sum_i (g_{ik} - u_k)(g_{ik'} - u_{k'}) = R_{kk'}$

3^{e} term $= -\frac{1}{n}\sum_i (g_{ik'} - u_{k'})\frac{1}{N_i}\sum_j (x_{ijk} - u_k) =$

$$= -\frac{1}{n}\sum_i (g_{ik'} - u_{k'})(g_{ik} - u_k) = -T_{kk'}$$

4^{e} term $= -T_{kk'}$, evenals de derde.

Dus

$$C_{kk'} = R_{kk'} - T_{kk'}$$

Van de matrix $(C_{kk'})$ kan dus op deze wijze een schatting verkregen worden van de elementen van de hoofddiagonaal en van de aangrenzende, daarmee evenwijdige lijnen; dat de $R_{kk'}$ berekend zijn met noemers $\frac{1}{n-1}$ in plaats van $\frac{1}{n}$ heeft, wegens de grootte van n, geen invloed van betekenis. In genoemde tabel 4 van "Een statistische analyse..." staan de waarden

$$\sum_i \sum_j (x_{ijk} - u_k)(x_{ijk'} - u_{k'}) \quad \text{voor} \quad k-k' = \pm 1 \quad \text{en}$$

$$\sum_i N_i = N$$

vermeld. Indien N_i voor iedere i dezelfde is, zijn dit de waarden van $T_{kk'}$ voor $k-k' = \pm 1$ en dan kunnen de overeenkomstige $C_{kk'}$ zonder meer berekend worden. Voor het geval de N_i verschillend zijn, kan de aldus berekende waarde als een benadering van de juiste gelden.

11. Bij het berekenen van $C_{11},\ldots,C_{34}$ op deze wijze deed zich echter een anomslie voor: $C_{33} = T_{33} - R_{33} = 86$, terwijl in tabel I van "Een statistische analyse ..." dezelfde grootheid berekend als $\frac{1}{N_i}\sum(x-x)^2$ de waarde 110 bezit. Deze discrepantie maakt het uitvoeren van de bovengenoemde methode vooralsnog practisch onmogelijk, daar de willekeur van de bepaling van de C-matrix, waarvan bovendien de elementen C_{13}, C_{14} en C_{24} geschat moeten worden met behulp van de overige elementen, te groot wordt om een vertrouwen in de uitkomst te rechtvaardigen.
Wellicht kan drs. Thomasson, beter bekend met het materiaal waaruit tabel 1 en 4 berekend zijn dan wij, een middel aan de hand doen om deze discrepantie op te heffen. In dat geval zou getracht kunnen worden met boven beschreven methode de schatting van de gewichten $\lambda_1,\ldots,\lambda_4$ uit te voeren. Tevens zou dan vermoedelijk de schatting van de spreiding van G_4 kunnen worden verbeterd.