STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

S 4

Dantzig, D. van, Theil, H.

Eerste verslag betreffende groeiproeven met Wistar-ratten.

1948

Den Heer Thomassen
Unilever Laboratories
<u>ZWIJNDRECHT</u>.

Geachte Heer Thomassen,

Naar aanleiding van ons gesprek van heden kan ik U nog het volgende mededelen.

Laat gegeven zijn de E_0 waarden $x_1, \ldots, x_N$
" E_{2026} " $y_1, \ldots, y_N$

N is het aantal groepen (24)
Elke x_i resp. y_i is het groepsgemiddelde van één groep van ± 12 dieren en heeft een spreiding $\frac{\sigma}{\sqrt{n}}$ (n = 12), waarin σ de spreiding in de gewichtstoename per dier is (alles gedurende een bepaalde week). Ook de spreiding van elke y_i is = $\frac{\sigma}{\sqrt{n}}$ aangenomen; de correlatiecoefficient van elke x_i met de y_i van <u>dezelfde</u> groep is ϱ. De x onderling, de y_i onderling en de x_i en y_j voor j ≠ i zijn <u>ongecorreleerd</u> ondersteld; de verwachting der x_i en y_i zijn = 0 ondersteld.

Hieruit is de regressievergelijking $y' = a + bx$ opgemaakt, die de waarden y_i: $y_1', \ldots, y_N'$ geeft. Tenslotte worden de gemiddelden $z_i = \frac{y_i + y_i'}{2}$ $z_1, \ldots, z_N$ bepaald.

De spreidingen der z_i kunnen bij benadering bepaald worden, in de onderstelling dat de N paren (x_i, y_i) beschouwd kunnen worden als N onafhankelijke trekkingen uit een oneindig grote collectie met regressievergelijking $y' = \alpha + \beta x$, en dat de verschillen van a en b met α resp. β <u>verwaarloosd</u> kunnen worden. Dit houdt in, dat $b = r\frac{s_y}{s_x}$ door $\beta = \varrho\frac{\sigma_y}{\sigma_x} = \varrho$ vervangen wordt, en dat zijn spreiding verwaarloosd wordt. (Griekse letters hebben op de oneindige collectie, latijnse op de steekproef betrekking. Laatstgenoemde zijn dus bekend, eerstgenoemde hypothetisch.) Voorts dat $a = \frac{1}{n}\sum y_i$ door de verwachting daarvan, dus 0 vervangen wordt, en dat ook hiervan de spreiding verwaarloosd wordt.

Dan is
$$\sigma_{y'}^2 = \beta\,\sigma_x^2 = \varrho^2\frac{\sigma^2}{n}$$

$$\varrho_{yy'}\,\sigma_y\,\sigma_{y'} = \text{verwachting van } yy'$$
$$\beta\,.\ \text{verwachting van } yx = \beta\,\varrho\,\sigma_y\,\sigma_x = \varrho^2\sigma^2/n$$

Dus
$$\varrho_{yy'} = \frac{\varrho^2\sigma^2/n}{\sigma_y\,\sigma_{y'}} = \frac{\varrho^2\sigma^2/n}{\frac{\sigma}{\sqrt{n}}\,\varrho\,\frac{\sigma}{\sqrt{n}}} = \varrho$$

Derhalve is wegens $2z = y + y'$
$$\sigma_{2z}^2 = \sigma_y^2 + \sigma_{y'}^2 + 2\varrho_{yy'}\,\sigma_y\,\sigma_{y'} =$$

$$= \frac{1}{n}\left\{\sigma^2 + \rho^2\sigma^2 + 2\sigma.\rho\sigma.\rho\right\}$$

$$= \sigma^2(1+3\rho^2)/n$$

$$\sigma_2^2 = \frac{\sigma^2(1+3\rho^2)}{4n}$$

of

$$\sigma_2 = \frac{\sigma}{\sqrt{4n}}\sqrt{1+3\rho^2}$$

Schrijft men dit in de vorm $\frac{\sigma}{\sqrt{2n}}\cdot\lambda^*$ dan is de "correctiefactor" $\lambda = \sqrt{\frac{1+3\rho^2}{2}}$ dus >1 als $3\rho^2>1$, $|\rho|>\frac{1}{3}\sqrt{3}$ is en alleen dan. Het voordeel van het invoeren van deze grootheid zie ik echter niet in.

Schrijft men echter

$$\sigma_2 = \frac{\sigma}{\sqrt{n}}\cdot\lambda$$

dan is de correctiefactor $\lambda = \frac{1}{2}\sqrt{1+3\rho^2}$ Deze is $\frac{1}{2}$ voor $\rho = 0$ en 1 voor $\rho = \pm 1$.

Bij volledige correlatie vinden we dus de oorspronkelijke $\frac{\sigma}{\sqrt{n}}$ terug. Bij volledige ongecorreleerdheid $\rho = 0$ wordt zij echter gehalveerd. Dat is als volgt in te zien. Indien x en y volledig ongecorreleerd waren, zou de regressielijn van y t.o.v. x een horizontale rechte worden, dus, daar de gemiddelden 0 zijn, de x-as. D.w.z. dan waren alle $y_i' = 0$. Dus waren de $z_i = \frac{y_i+y_i'}{2}$ eenvoudig de helften der y_i, dus ook $\sigma_2 = \frac{1}{2}\sigma_y$

Voor $\rho = 0,8$ wordt de factor λ $\frac{1}{2}\sqrt{1+0,8^2\cdot 3} \approx$
$= 0,85$
dus $\sigma_2 \approx 0,85\frac{\sigma}{\sqrt{n}}$ $\left(\approx 1,19\frac{\sigma}{\sqrt{2n}}\right)$

Voor	$\rho = 0,9$	$\lambda = 0,91$
"	$\frac{1}{3}\sqrt{3} = 0,577..$	0,707
	0,5	0,6
	0,3	0,55
	0,0	0,5

enz. (alle getallen in grove benadering).

Ik herinner er nog even aan, dat altijd als $z = \frac{y+y'}{2}$ is, (onverschillig of y al dan niet op bovenstaande wijze uit y ontstaan is)

$$\sigma_z = \tfrac{1}{2}\sqrt{\sigma_y^2 + \sigma_{y'}^2 + 2\sigma_y\sigma_{y'}\rho_{yy'}}$$

is,dus **steeds**

$$\sigma_z \leq \tfrac{1}{2}\sqrt{\sigma_y^2 + \sigma_{y'}^2 + 2\sigma_y\sigma_{y'}} = \frac{\sigma_y + \sigma_{y'}}{2}$$

Anderzijds is ook steeds

$$\sigma_z \geq \frac{|\sigma_y - \sigma_{y'}|}{2}$$

Als dus $\sigma_y = \sigma_{y'}$ is,is σ_z **steeds** **hoogstens** daaraan gelijk, onverschillig of deze grootheden spreidingen van gemeten grootheden (in het bovenstaande σ),dan wel de spreidingen van daaruit gevonden groepsgemiddelden (in het bovenstaande $\frac{\sigma}{\sqrt{n}}$) zijn.

De hier gevonden correctiefactor $\frac{1}{2}\sqrt{1+3\rho^2}$ behoeft zelf weer een correctie doordat de identificatie van a en b met α resp. β eigenlijk **niet** gerechtvaardigd is, vooral omdat het aantal groepen (N = 34) zo klein is. **Exacte** berekening van deze verdere correctie is niet mogelijk,approximatieve wellicht.Het is evenwel de vraag,of deze de moeite loont.Wanneer ik U de oude waarde σ = 3,8.. aanhoudt,evenwel met een noemer $\sqrt{n}$ en niet $\sqrt{2n}$ blijft U in ieder geval aan de veilige kant.

In de hoop,dat deze gegevens nog van nut kunnen zijn voor Uw rapport en ook overigens verhelderend mogen werken, verblijf ik,

hoogachtend,

(Prof.Dr.D.van Dantzig).

EENIGE VOORLOPIGE AANTEKENINGEN BIJ HET ONDERZOEK VAN

Drs Thomasson

Ad § 2. 1. De afronding van het gemiddelde zonder correctie is een te ruwe benadering, die systematische fouten doet ontstaan. De spreidingen worden n.l. regelmatig overschat.

Ad § 3. 2. Aangezien β_1 een kwadraat is, moeten de mintekens voor de waarden van β_1 vervangen worden door plustekens (Tabel I); de mintekens zouden voor de wortel uit β_1 moeten staan.

3. Voor de berekening van de standaarddeviaties van b_1 en b_2 zijn de Tables for Statisticians and Biometricians in dit geval niet geschikt. Het doel van Tabel I is n.l. te toetsen of de 6 beschouwende verdelingen als normaal beschouwd kunnen worden. Het procédé is dan, dat men uitgaat van de veronderstelling, dat de verdelingen normaal zijn, om vervolgens al of niet tot een tegenspraak te komen. Men moet dus ook voor de berekening van de standaard-deviatie van b_1 en b_2 uitgaan van de hypothese der normaliteit. Deze gedachtegang is bij de constructie van de tabellen van Pearson niet gevolgd; men heeft daar een veel uitgebreider groep van verdelingen op het oog gehad.

4. Het is in dit verband doelmatig, gebruik te maken van het werkje "Tests of normality" van R.C. Geary and E.S. Pearson. Dit verschaft de mogelijkheid om van het werken met standaarddeviaties van β_1 en β_2 af te zien en direct af te lezen, hoe groot de waarschijnlijkheid is, dat als de modelcollectie normaal verdeeld is, de steekproefwaarden van β_1 en β_2 buiten een bepaald gebied rondom de theoretische waarden 0 resp. 3 vallen. Het resultaat is, dat het slechts in 1 op de 100 gevallen zal voorkomen, dat bij een steekproef van 1300 waarnemingen de steekproefwaarden b_1 groter is dan 0.026. Kiezen we een overschrijdingskans van 5% in plaats van 1%, dan is de maximale steekproefwaarde van b_1 hoogstens 0.012; het is dus dubieus, of de collectie van de 3e week nog geacht kan worden normaal verdeeld te zijn.

5. Op analoge wijze vinden we, dat bij een steekproef van 1300 waarnemingen de steekproefwaarde van b_2 slechts in 1 op de 100 gevallen zal liggen buiten het interval (2,71 - 3,36). Uit tabel I blijkt, dat alle steekproefwaarden van β_2 op één na duidelijk rechts van dit interval liggen. Het is duidelijk, dat de 6 beschouwde verdelingen niet normaal geacht kunnen worden.

Ad § 4. 6. Bij een voorlopig onderzoek van de analyse, die de opdrachtgever in deze paragraaf geeft, betreffende relaties tussen spreidingen, leek het ons toe, dat de methode niet geheel correct is. Hierop zal nader worden ingegaan.

Ad § 5 en
§ 6 7. Deze paragrafen moeten nog nader onderzocht worden.

8. De conclusie, opgesloten in de zin: "Uit het feit, dat de standaard-deviatie in de loop van de proef met een constante grootheid vermeerdert, wordt waarschijnlijk, dat tussen de gewichtsvermeerderingen in de achtereenvolgende weken een rechtlijnig verband zal voorkomen", moet zeer gewaagd geacht worden.