' VOOR CULTUURTECHNIEK EN WATERHUISHOUDING
NOTA no. 95 dd. 12 september 1961

Analyse van een samengestelde steekproef

L.P. Kamil

Inleiding

In een omschreven gebied, bestaande uit een groot aantal veehouderijen, worden de kavels grasland per bedrijf ingedeeld in de klassen:

I	kavels	op	afstand	0 - 500	m	van	de	boerderij
II	"	"	"	500 - 1000	m	"	"	"
III	"	"	"	1000 - 1500	m	"	"	"
IV	"	"	"	1500 - 2000	m	"	"	"
V	"	"	"	2000 - 3000	m	"	"	"
VI	"	"	"	> 3000	m	"	"	"

Wil men het aantal weidedagen per klasse weten, dan kan men de veehouders verzoeken te noteren hoe frequent gedurende het weideseizoen de kavels in de verschillende klassen worden gebruikt door melkvee. Indien men deze gegevens verkrijgt kan men een gemiddelde gebruiksfrequentie per klasse voor het gebied maken.

Dit systeem heeft de volgende nadelen:

1e. Men kan niet verwachten, dat alle veehouders aan het verzoek zullen voldoen.

2e. Een onbekend aantal van de veehouders zullen de notities niet systematisch bijhouden, zodat foutieve gegevens niet uitgesloten zijn.

3e. Indien het mogelijk is een aantal veehouders te selecteren, welke wel geacht kunnen worden gegevens te verschaffen, dan behoeft de gemiddelde gebruiks-frequentie van deze groep nog geen juist beeld te geven van de gemiddelde gebruiks-frequentie van het gebied.

Een ander systeem verkrijgt men door een toevalssteekproef te nemen uit alle bedrijven. Van het aantal bedrijven in de steekproef zal men door eigen krachten dagelijks laten noteren waar het vee staat.

Zodoende verkrijgt men een betrouwbaar beeld. Het behoeft echter geen betoog, dat de kosten, verbonden aan een op deze wijze opgezet onderzoek hoog zullen zijn.

Een systeem, dat een zuiver beeld zal geven en waarvan de kosten belangrijk minder zullen zijn, zal in het volgende worden verklaard. Ter vereenvoudiging wordt slechts gewerkt met twee klassen:

I de kavels op afstand 0 - 500 m van de boerderij

II de andere kavels.

In de gegeven afleidingen kan men de klasse 0 - 500 m vervangen door elke der andere klasse. De berekeningen verlopen overigens analoog. Achtereenvolgens worden behandeld de steekproeven uit de weidedagen, de steekproef uit de bedrijven, terwijl tenslotte door samenstelling van de beide soorten steekproeven een oplossing gevonden wordt voor het gestelde probleem.

1. De steekproef uit het aantal weidedagen

Men beschouwe één bedrijf, waarvan men steekproefsgewijs het aantal dagen wil bepalen, dat melkvee op de kavels op afstand 0 - 500 m van de boerderij staat. Men noemt het aantal weidedagen N_1, het totaal aantal weidedagen op genoemde kavels X en het aantal dagen op andere kavels $N_1 - X$. Men neemt een steekproef ter grootte n_1, waarvan op x dagen het vee op genoemde kavels wordt gevonden en op (n_1-x) dagen op andere kavels. Indien de steekproef zo gekozen is, dat alle dagen evenveel kans hebben gehad in de steekproef te worden opgenomen, dan geldt dat $\hat{\underline{X}}$ een zuivere schatter is van X:

$$\hat{\underline{X}} = \frac{x}{n_1} N_1 \qquad (1)$$

(Een schatter $\hat{\underline{X}}$ van X heet zuiver als geldt: $E(\hat{\underline{X}}) = X$)

Een zuivere schatter van de variantie van $\hat{\underline{X}}$ is:

$$\mathrm{var}(\hat{\underline{X}}) = \frac{N_1(N_1 - n_1)}{n_1 - 1} pq \qquad (2),$$

waarin $p = \frac{x}{n_1}$ en $q = \frac{n_1 - x}{n_1}$

Om in § 3 verwarring te voorkomen wordt voor $\hat{\underline{X}}$ het symbool $\underline{Y}$ ingevoerd. Men kan nu schrijven:

$$\underline{Y} = X + \underline{d}$$

Doordat men X schat door een steekproef-uitkomst $\underline{Y}$, zal in het algemeen een afwijking $\underline{d}$ bestaan.

Nu is $E(\underline{d}) = E(\underline{Y}) - X = X - X = 0$ (3)

en $\text{var}(\underline{d}) = E[\underline{Y} - X - E(\underline{Y} - X)]^2$

$= E[\underline{Y} - E(\underline{Y}) - X + E(X)]^2$

$= E[\underline{Y} - E(\underline{Y})]^2 = \text{var}(\underline{Y}) = \text{var}(\hat{\underline{X}})$

dus $\text{var}(\underline{d}) = \dfrac{N_1(N_1 - n_1)}{n_1 - 1} pq$ (4)

2. De steekproef uit het aantal bedrijven

Stel dat men van het gebied van elk bedrijf het juiste aantal weidedagen X_i kent, dat het vee op de kavels op afstand 0 - 500 m van de boerderij staat. Het totaal aantal bedrijven in de populatie is N_2 en men neemt een toevalssteekproef ter grootte n_2, om een schatting te maken van het gemiddelde aantal weidedagen μ op genoemde kavels over alle bedrijven.

$$\mu = \frac{\sum_{i=1}^{N_2} X_i}{N_2} = \bar{X}_{N_2} \qquad \text{(niet stochastisch)}$$

Een zuivere schatter van μ is:

$$\hat{\underline{\mu}} = \bar{\underline{X}}_{n_2} = \frac{\sum_{i=1}^{n_2} X_i}{n_2} \qquad (5)$$

De schatter van de variantie van $\hat{\underline{X}}_{n_2}$ is:

$$\text{var}(\bar{\underline{X}}_{n_2}) = \frac{N_2 - n_2}{N_2} \frac{\Sigma(X_i - \bar{X})^2}{n_2(n_2-1)} \qquad (6)$$

Het deel $\Sigma (X_i-\bar{X})^2/(n_2-1)$ in het rechterlid van (6) is een schatter van S^2, de populatie-variantie.

3. De samengestelde steekproef

Men neemt uit N_2 bedrijven een steekproef ter grootte n_2. Van de

in de steekproef voorkomende bedrijven schat men voor elk bedrijf het aantal weidedagen op de kavels in de klasse 0 - 500 m door $\underline{Y}$ uit steekproeven ter grootte n_1.

Men heeft dus:

N_2 bedrijven in de populatie waarvan n_2 bedrijven in de steekproef, en voor elk van de n_2 in de steekproef voorkomende bedrijven N_1 weidedagen in het seizoen waarvan n_1 weidedagen in de steekproef.

Nu kan geschreven worden:

$$\begin{array}{lll} \underline{Y}_1 = X_1 + \underline{d}_1 & & \underline{Y}_1 - X_1 = \underline{d}_1 \\ \cdot \quad \cdot \quad \cdot & & \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot & \text{of} & \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot & & \cdot \quad \cdot \quad \cdot \\ \underline{Y}_{N_2} = X_{N_2} + \underline{d}_{N_2} & & \underline{Y}_{N_2} - X_{N_2} = \underline{d}_{N_2} \end{array}$$

De stochastieken $\underline{d}_1, \underline{d}_2, \ldots . \underline{d}_{N_2}$ zijn onderling onafhankelijk en onafhankelijk van X.
Verder geldt $E(\underline{d}) = 0$ terwijl alle $\sigma^2_{\underline{d}_i}$ verschillend kunnen zijn.
Volgens (1) en (5) geldt dat $\bar{\underline{Y}}_{n_2}$ een zuivere schatter is van μ, nl. :

$$\bar{\underline{Y}}_{n_2} = \frac{\sum_{i=1}^{n_2} \underline{Y}_i}{n_2} = \hat{\bar{\underline{X}}}_{n_2} = \hat{\underline{\mu}} \qquad (7)$$

Omdat pq in (2) maximaal is als p=q=0.5, volgt uit (4), onder voorwaarde dat N_1 en n_1 voor alle bedrijven gelijk is:

$$\text{var}(\underline{d}_i) \leqslant 0{,}25 \; \frac{N_1(N_1 - n_1)}{n_1 - 1} \qquad (8)$$

en

$$\text{var}(\bar{\underline{d}}_{n_2}) \leqslant 0{,}25 \; \frac{N_1(N_1 - n_1)}{n_2(n_1 - 1)} \qquad (9)$$

(Doordat de steekproefgrootte uit de bedrijven n_2 is, heeft men dus n_2 trekkingen uit de stochastieken $\underline{d}_i$;

$$\text{var}(\bar{\underline{d}}_{n_2}) = \text{var}\left(\frac{\underline{d}_1 + \ldots . + \underline{d}_{n_2}}{n_2}\right) = \frac{N_1(N_1 - n_1)}{n_2(n_1 - 1)} \cdot \frac{\Sigma p_i q_i}{n_2} \leqslant 0.25 \frac{N_1(N_1 - n_1)}{n_2(n_1 - 1)}$$

De variantie van $\underline{Y}_i$ volgt uit:

$$\begin{aligned} \text{var}(Y_i) &= E[(\underline{X}_i + \underline{d}_i) - E(\underline{X}_i + \underline{d}_i)]^2 \\ &= E[\underline{X}_i - E(\underline{X}_i) + \underline{d}_i - E(\underline{d}_i)]^2 \\ &= E[\underline{X}_i - E(\underline{X}_i)]^2 + E[\underline{d}_i - E(\underline{d}_i)]^2 + 2E[\underline{X}_i - E(\underline{X}_i)][\underline{d}_i - E(\underline{d}_i)] \\ &= E[\underline{X}_i - E(\underline{X}_i)]^2 + E[\underline{d}_i - E(\underline{d}_i)]^2 \end{aligned}$$

wegens onafhankelijkheid van $\underline{X}$ en $\underline{d}$.

Dus $\text{var}(\underline{Y}_i) = \text{var}(\underline{X}_i) + \text{var}(\underline{d}_i)$ (10)

Voor de $\text{var}(\bar{\underline{Y}}_{n_2})$ volgt na substitutie van (6) en (9) in (10):

$$\text{var}(\bar{\underline{Y}}_{n_2}) \leq \frac{N_2 - n_2}{N_2} \frac{S^2}{n_2} + 0.25 \frac{N_1(N_1 - n_1)}{n_2(n_1 - 1)} \qquad (11)$$

Indien n_1 en n_2 niet te klein zijn, geldt dat de normale verdeling als benadering toegepast kan worden, zodat met 95% betrouwbaarheid de waarde μ zal liggen binnen de grenzen:

$$\bar{\underline{Y}} - 1.96 \sqrt{\text{var}(\bar{\underline{Y}})} < \mu < \bar{\underline{Y}} + 1.96 \sqrt{\text{var}(\bar{\underline{Y}})} \qquad (12)$$

4. De keuze van n_1 en n_2

Indien men uit gelijksoortig waarnemingsmateriaal een schatting van S^2 kan maken, kan men de grootte van de nieuw te nemen steekproeven zo kiezen, dat het betrouwbaarheidsinterval een van tevoren bepaalde lengte heeft.

Noemt men deze lengte 2L dan is (zie(12)):

$$L = 1.96 \sqrt{\text{var}(\bar{\underline{Y}})} \qquad (13)$$

Stelt men nu $(\frac{L}{1.96})^2 = Q$ dan is: $Q = \text{var}(\bar{\underline{Y}})$

Substitutie van (11) geeft bij benadering:

$$Q = \frac{N_2 - n_2}{N_2} \frac{S^2}{n_2} + 0.25 \frac{N_1(N_1 - n_1)}{n_2(n_1 - 1)} ,$$

hetgeen na uitwerking geeft:

$$n_1 = \frac{(QN_2 + S^2)n_2 + 0.25N_1^2N_2 - N_2S^2}{(QN_2 + S^2)n_2 + 0.25N_1N_2 - N_2S^2} \qquad (14)$$

In (14) zijn N_1, N_2, Q en S^2 constanten, terwijl na keuze van n_2 de grootte van n_1 vaststaat.
De benadering van $\mathrm{var}(\bar{\underline{Y}}_{n_2})$ door $\frac{N_2 - n_2}{N_2} \frac{S^2}{n_2} + 0.25 \frac{N_1(N_1 - n_1)}{n_2(n_1 - 1)}$

heeft tot gevolg dat de grootte van n_1 en n_2 aan de veilige kant gekozen worden, d.w.z. dat de kans dat μ ligt binnen het interval

$$\bar{\underline{Y}} - L < \mu < \bar{\underline{Y}} + L$$

groter is dan 95%.
Indien men verwacht, dat p veel groter of kleiner dan 0,5 zal zijn, (b.v. $p < 0,3$ of $p > 0,7$), vervangt men 0,25 door pq, waarin voor p en $q = (1-p)$ de verwachte waarde van p wordt ingevuld.

5. Steekproef opzet

Omschrijf het proefgebied.
Bepaal het aantal bedrijven: N_2.
Bepaal (schat) het totaal aantal weidedagen: N_1.
Schat S^2 uit gelijksoortig waarnemingsmateriaal.
Bepaal de gewenste nauwkeurigheid 2L bij een gewenste betrouwbaarheid.
Bepaal n_1 en n_2 uit (14), zodat de kosten van het onderzoek minimaal zijn.
Trek een toevalssteekproef ter grootte n_2 uit de bedrijven.
Trek voor elk van de gekozen n_2 bedrijven een toevalssteekproef ter grootte n_1 uit de weidedagen.

6. Voorbeeld

Voor de volgende waarden zijn in enige figuren de verbanden tussen n_1, n_2, S^2, L en het betrouwbaarheidspercentage weergegeven. Gesteld wordt dat het aantal weidedagen 200 zal zijn, terwijl in het gebied 1500 bedrijven liggen.

Nu is dus: $N_1 = 200$ weidedagen
$N_2 = 1500$ bedrijven

Figuur 1 geeft het verband tussen n_1 en n_2 weer bij:

$L = 2$
$S^2 = 25 ; 49 ; 81 ; 121$

De betrouwbaarheid is 95%.

In de figuur is af te lezen dat men bij b.v. S^2 = 121 met de volgende paren n_1 en n_2 de gewenste nauwkeurigheid kan verwachten in 95 op de 100 steekproeven:

n_1: 120 100 80 60 40

n_2: 136 152 164 182 214

Bij kleiner worden van n_1 ziet men, dat de lijnen steeds steiler lopen, zodat de waarde van n_2 sterker toeneemt.

Figuur 2 geeft het verband tussen n_1 en n_2 weer bij

L = 4

S^2= 25 ; 49 ; 81 ; 121

De betrouwbaarheid is 95%.

Bepaalt men thans de paren n_1 en n_2 bij S^2= 121 dan vindt men:

n_1: 120 100 80 60 40

n_2: 36 40 47 57 64

Figuur 3 geeft eenzelfde verband bij:

L = 4

S^2= 25 ; 121

De betrouwbaarheid is 80%

Men kan nu de gewenste nauwkeurigheid (L=4) bij een betrouwbaarheid van 80% bereiken met de paren:

n_1: 120 100 80 60 40 30 20 10

n_2: 16 17 19 23 32 40 65 110

Stelt men dat men op een dag 40 bedrijven kan bereiken, dan zou men in het weideseizoen 30 keer een plaatsbepaling van het vee moeten maken.

Figuur 4 tenslotte geeft het verband tussen L en het percentage betrouwbaarheid bij:

S^2 = 25 ; 121

n_2 = 40

n_1 = 10 ; 20 ; 30

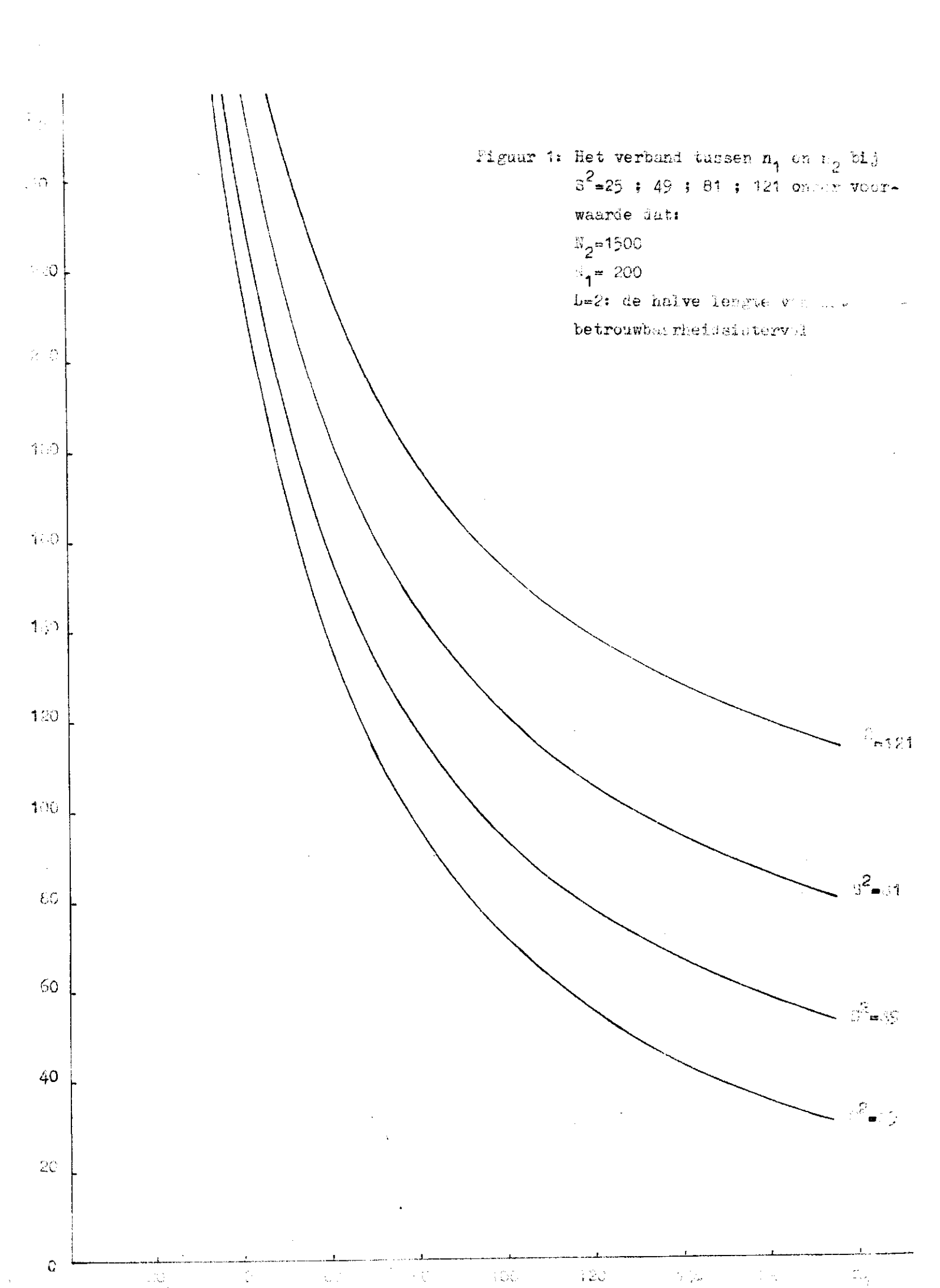
In de figuur is af te lezen dat bij n_1 = 10, 20 of 30 en P = 95%; 90%; 80% men de volgende waarden voor L zal kunnen bereiken:

De waarden van L bij $N_2 = 40$

P \ n_1	10	20	30
95%	11.1	8.0	6.5
90%	9.1	6.5	5.3
80%	7.0	5.0	4.0

Tot slot moet worden opgemerkt, dat een toevalssteekproef over de weidedagen vervangen moet worden door een gelaagde steekproef, indien de gebruiks-frequentie in het seizoen verandert, Men zou dan het weideseizoen in drie klassen kunnen indelen, waarna binnen elke klasse een toevals-steekproef wordt getrokken. Het principe van de steekproef-opzet verandert hierdoor niet, doch de formules voor de steekproef uit de weidedagen, zowel als die voor de samengestelde steekproef, zullen veranderen.

De bewijzen voor de formules welke niet in de tekst zijn bewezen, worden gegeven in: W.G.Cochran "Sampling Techniques"(London 1953).



Figuur 1: Het verband tussen n_1 en n_2 bij s^2=25 ; 49 ; 81 ; 121 onder voorwaarde dat:

N_2=1500

s_1= 200

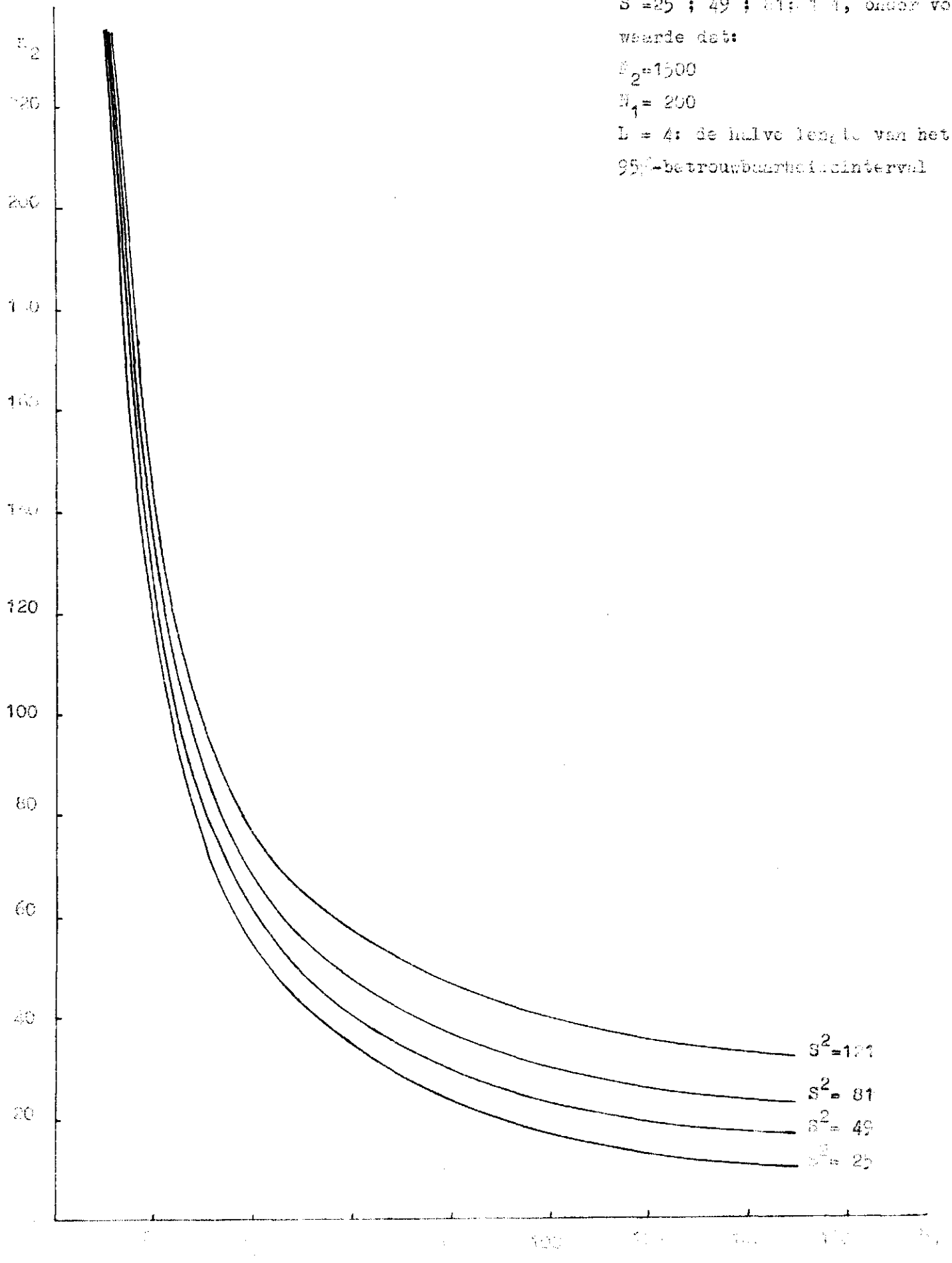
b=2: de halve lengte van het betrouwbaarheidsinterval

Figuur 2: Het verband tussen n_1 en n_2 bij S^2=25 ; 49 ; 81; 121, onder voorwaarde dat:

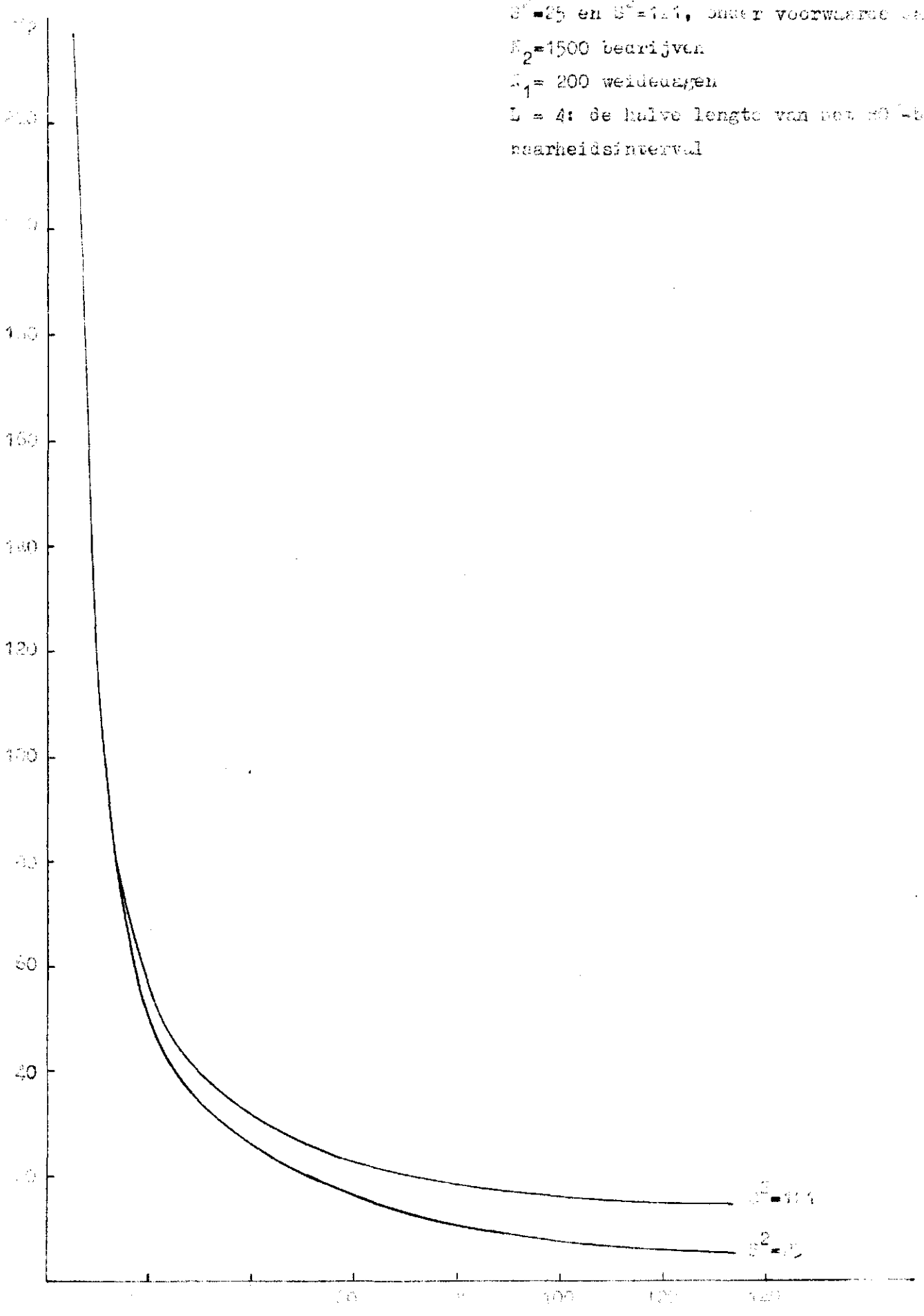
N_2=1500

N_1= 200

L = 4: de halve lengte van het 95%-betrouwbaarheidsinterval



Figuur 3: Het verband tussen n_1 en n_2 bij
$S^2=25$ en $S^2=1.1$, onder voorwaarde van
$N_2=1500$ bedrijven
$N_1=200$ weidedagen
L = 4: de halve lengte van het 95 %-betrouwbaarheidsinterval



Figuur : het verband tussen n_1, de halve lengte van het betrouwbaarheidsinterval en het percentage betrouwbaarheid bij constante $\bar{x}_1, \bar{x}_2, n_2$ en S^2.