Rapport 66

INSTITUUT VOOR CULTUURTECHNIEK EN WATERHUISHOUDING

Over het à priori aannemen van een kansverdeling.

L, Kamil

Inleiding

Wanneer men de beschikking heeft over een verzameling waarnemingsuitkomsten $x_1, x_2, \ldots\ldots, x_i, \ldots\ldots, x_n$ wordt vaak zonder nader onderzoek verondersteld, dat $\underline{x}$ normaal verdeeld is met verwachtingswaarde μ en spreiding σ. Men drukt het normaal zijn van de verdeling uit door te zeggen dat $\underline{x}$ een $N(\mu, \sigma)$-verdeling heeft of door het symbool :

$\underline{x} \approx \mu+\sigma\underline{\varkappa}$ spreek uit: isomoor met , aan welk symbool de volgende definities ten grondslag liggen:

1)Een stochastiek is een variabele met een kansverdeling en wordt aangegeven door een letter met een streepje er onder: $\underline{x}$

2)Hebben twee stochastieken $\underline{u}$ en $\underline{v}$ een zelfde kansverdeling, dan zijn ze isomoor, hetgeen aangegeven wordt door $\underline{u} \approx \underline{v}$.

3)De stochastiek met kansdichtheid $f(x) = \frac{1}{\sqrt{2\pi}} \exp.\ (-\frac{1}{2} x^2)$ is de standaard normale stochastiek $\underline{\varkappa}$ met verwachtingswaarde $\mu = 0$ en spreiding $\sigma = 1$.

Uit de definities volgt dat de grafieken van $\underline{x}$ en $\underline{\varkappa}$ door verschuiving en schaalverandering tot dekking zijn te brengen, waardoor het gebruik van het congruentie symbool $\approx$ verklaard is.

Men schat μ en σ door respectievelijk te berekenen:

$S(\mu) = \frac{\Sigma x}{n} = \overline{x}$

$S(\sigma) = \sqrt{\frac{\Sigma (x - \overline{x})^2}{n - 1}}$, en gebruikt deze schattingen eventueel om nieuwe gevolgtrekkingen te maken of hypothesen te toetsen.
De verleiding om dit te doen is daarom zo groot, omdat de aan de normale verdeling verbonden toetsen in het algemeen scherper zijn dan de parameter-vrije, waarbij de verdeling buiten beschouwing gelaten wordt.
Het gevaar is dat, hoe exact de methode verder is, een verkeerd uitgangspunt noodzakelijk tot een verkeerde conclusie moet leiden.

Het een en ander wordt verklaard aan de hand van een praktijkvoorbeeld.
Gegeven zijn een dertigtal k- waarden van kleigrond, afkomstig van metingen verricht door K.E.Wit.
Het symbool k stelt de doorlaatfactor voor met dimensie m/dag, berekend uit doorlatendheidsmetingen van ongeroerde grondmonsters.
De vraag kan gesteld worden of een waarde van k, bijvoorbeeld k = 0,10,

een vertegenwoordigende waarde kan zijn voor de reeks uitkomsten.
Voor het beantwoorden van deze vraag zullen drie verschillende methoden toegepast en met elkaar vergeleken worden.

I. De veronderstelling dat k normaal verdeeld is.

Het uitgangspunt luidt: $\underline{k} \approx \mu + \sigma \underline{x}$
De hypothese H_o luidt: $\mu = 0{,}10$
Onderzocht wordt of H_o juist kan zijn, waarbij een risico, dat is de kans op een verkeerde uitspraak, kleiner of gelijk aan 5% wordt aanvaard.
In tabel 1 wordt de analyse schematisch uitgewerkt. Met f wordt aangegeven het aantal malen dat een bepaalde waarde voorkomt.

Tabel 1

Analyse van dertig waarnemingen van de doorlaatfaktor k van kleigrond.

k	f	f cumulatief	z=f in % cumulatief	f.k	$f.k^2$
0.004	1	1	3.33	0.004	0.000016
0.007	3	4	13.33	0.021	0.000147
0.008	2	6	20.00	0.016	0.000128
0.010	4	10	33.33	0.040	0.000400
0.020	3	13	43.33	0.060	0.001200
0.030	5	18	60.00	0.150	0.004500
0.040	2	20	66.67	0.080	0.003200
0.050	1	21	70.00	0.050	0.002500
0.080	2	23	76.67	0.160	0.012800
0.090	2	25	83.33	0.180	0.016200
0.140	1	26	86.66	0.140	0.019600
0.160	1	27	90.00	0.160	0.025600
0.250	1	28	93.33	0.250	0.062500
0.340	1	29	96.67	0.340	0.115600
0.460	1	30	100.00	0.460	0.211600
Som	30			2.111	0.475991

Bij de uitwerking van het probleem wordt nog gebruik gemaakt van de uitdrukking

$$S\left(\sigma_{\bar{k}}^2\right) = \frac{S(\sigma_k^2)}{n}$$

hetgeen de schatting van de variantie van het gemiddelde van n onafhankelijke waarnemingsuitkomsten voorstelt. Hieruit wordt de spreiding $\sigma_{\bar{k}}$ van de gemiddelde waarde van $\underline{k}$ geschat.

$$n = \Sigma f = 30$$

$$\Sigma fk = 2.111$$

$$S(\mu) = \bar{k} = \frac{\Sigma fk}{\Sigma f} = 0.0703$$

$$S\left(\sigma_{\bar{k}}^2\right) = \frac{\Sigma f(k - \bar{k})^2}{n - 1} = \frac{\Sigma fk^2 - \frac{(\Sigma fk)^2}{\Sigma f}}{n - 1} = 0.011277$$

$$S\left(\sigma_{\bar{k}}^2\right) = \frac{S(\sigma_k^2)}{n} = 0.00376$$

$$S\left(\sigma_k\right) = 0.1062$$

$$S\left(\sigma_{\bar{k}}\right) = 0.0194$$

Nu heeft $\frac{\bar{k} - \mu}{S(\sigma_k)}$ een zogenaamde t-verdeling van Student bij 29 dimensies. Het 95% betrouwbaarheids-interval voor μ wordt geconstrueerd door in een t-tabel bij 29 dimensie de overschrijdingswaarden bij α_L en α_R op te zoeken. waarbij $\alpha_L = \alpha_R = \frac{1}{2}\alpha = 0.025$(L is linkszijdig, R = rechtszijdig). Men vindt dan de waarden $t_{29} = \pm 2.045$. Hieruit volgt dan met 95% betrouwbaarheid

$$-2.045 < \frac{\bar{k} - \mu}{S(\sigma_{\bar{k}})} < 2.045$$

en na oplossing van de betrekking

$$0.0306 < \mu < 0.1100$$

Nu blijkt dat het 95% betrouwbaarheids-interval de hypothetische waarde k = 0.10 bevat. waar uit volgt dat $\mu = 0.10$ een vertegenwoordigende waarde van de reeks uitkomsten kan zijn (inzet figuur 1 geeft schematisch het toetsen van de hypothese weer).

Conclusie: de hypothese wordt aanvaard.

Gaat men echter de frequentieverdeling van $\underline{k}$ onderzoeken dan kan men de punten (k_i, z_i) waarvan de berekening in tabel 1 is uitgevoerd, op

op kanspapier tekenen; dat wil zeggen, dat men k_i op een lineaire schaal en z_i op een kansschaal uitzet (figuur 1).
Slechts indien $\underline{k}$ normaal verdeeld is moeten de punten op een rechte lijn liggen.
Nu blijkt dat de punten van een rechte afwijken. De verdeling is dus niet normaal. In figuur 1 staat de uit de verdeling berekende lijn

$$\underline{x} \approx 0.0703 + 0.1062\ \underline{z}$$

grafisch weergegeven.

II. De veronderstelling dat log $\underline{k}$ normaal verdeeld is.

Uit het feit dat $\underline{k}$ niet normaal verdeeld blijkt te zijn volgt de vraag of het mogelijk is door een geschikt gekozen transformatie wel tot een normale verdeling te komen.
Worden de gegevens nu getransformeerd met $y_i = \log k_i$ en tekent men de punten (y_i, z_i) op kanspapier, dan blijken de punten een rechte te benaderen.

Men kan dus beter van de veronderstelling uitgaan, dat $\underline{y} = \log \underline{k}$ een normale verdeling heeft (figuur 2).
In tabel 2 wordt met de nieuwe veronderstelling de berekening van de verwachtingswaarde en de spreiding nogmaals uitgevoerd.

Het uitgangspunt luidt: $\underline{y} \approx \nu + \sigma \underline{z}$

$$H_o : \mu = 10^{\nu} = 0.10 \text{ of } \nu = -1.0$$

Het 95% betrouwbaarheidsinterval wordt analoog aan het vorige geval vastgesteld.

Tabel 2

Analyse van y = log k

k	y = log k	f	f cumulatief	z = f in % cumulatief	f.y	f. y^2
0.004	- 2.398	1	1	3.33	- 2.398	5.75040
0.007	- 2.155	3	4	13.33	- 6.465	13.93208
0.008	- 2.097	2	6	20.00	- 4.194	8.79482
0.010	- 2.000	4	10	33.33	- 8.000	16.00000
0.020	- 1.699	3	13	43.33	- 5.097	8.65980
0.030	- 1.523	5	18	60.00	- 7.615	11.59764
0.040	- 1.398	2	20	66,67	- 2.796	3.90881
0.050	- 1.301	1	21	70.00	- 1.301	1.69260
0.080	- 1.097	2	23	76.67	- 2.194	2.40682
0.090	- 1.046	2	25	83.33	- 2.092	2.18823
0.140	- 0.854	1	26	86.66	- 0.854	0.72932
0.160	- 0.796	1	27	90.00	- 0.796	0.63362
0.250	- 0.602	1	28	93.33	- 0.602	0.36240
0.340	- 0.468	1	29	96.67	- 0.468	0.21902
0.460	- 0.337	1	30	100.00	- 0.337	0.11357
Som		30			-45.209	76.98913

$S(\nu) = \bar{y} = -1.5069 \quad S(\mu) = 0.03$

$S(\sigma_y) = 0.5528$

$S(\sigma_{\bar{y}}) = 0.1009$

Het 95% betrouwbaarheidsinterval volgt uit:

$$-2.045 < \frac{\bar{y} - \nu}{S(\sigma_{\bar{y}})} < 2.045$$

en na oplossing van de betrekking

$$-1.713 < \nu < -1.301$$

Substitutie van $\mu = 10^{\nu}$ geeft

$$0.0193 < \mu < 0.0501$$

Nu blijkt dat het 95% betrouwbaarheidsinterval de veronderstelling $\mu = 0.10$ niet bevat

Conclusie: H_o : $\mu = 0.10$ wordt verworpen.

31/0161/100/5

Omgekeerd kan worden vastgesteld, dat een waarde van $\mu > 0.10$ slechts met 1.9% kans als verwachtingswaarde van een zelfde serie $\underline{k}$-uitkomsten kan optreden, welke kans te klein is om een dergelijke μ te aanvaarden

III. Over de verdeling van k wordt niets verondersteld

Een andere mogelijkheid om te toetsen of k = 0.10 een vertegenwoordigende waarde is krijgt men door geen veronderstelling over de verdeling te maken en de hypothese te toetsen, dat de mediaan m = 0.10 is. De mediaan m van een reeks uitkomsten $\underline{k}$ is de waarde waarvoor geldt dat de kans P op het voorkomen van waarden groter dan m , gelijk is aan die op het voorkomen van waarden kleiner dan m.
In formule:

$$P(\underline{k} > m) = P(\underline{k} < m) = \frac{1}{2}$$

Let men alleen op de eigenschap $\underline{k} > m$ en $\underline{k} < m$ dan is de stochastiek $\underline{x}$ het aantal waarden dat kleiner is dan m.
Deze stochastiek is isomoor met de stochastiek $\underline{y}$: het aantal keren dat kruis boven komt bij n keer werpen met een zuivere munt.
Deze kansen zijn dus te berekenen met de binomiale verdeling waarvoor geldt:

$$\mu = np = 15$$
$$\sigma^2 = np(1-p) = 7.5$$
$$\sigma = 2.7386$$

Het construeren van het 95% betrouwbaarheidsinterval geschiedt nu door de overschrijdingswaarden van de binomiale verdeling te bepalen.

Bij 30 waarnemingen is de binomiale verdeling te benaderen door de normale verdeling met continuïteits-correctie. Deze laatste is noodzakelijk doordat de binomiale verdeling discreet is en nu benaderd wordt door een continue verdeling.

Dan is

$$P\left(\underline{x} \geq \frac{x - 15 - \frac{1}{2}}{2.7386}\right) = 0.025 \; ; \; P\left(\underline{x} \leq \frac{x - 15 + \frac{1}{2}}{2.7386}\right) = 0.025$$

Uit de tabel voor normale verdeling volgt:

$$\frac{x - 15 - \frac{1}{2}}{2.7386} = 1.96 \qquad \frac{x - 15 + \frac{1}{2}}{2.7386} = -1.96$$

Bij het oplossen van deze betrekkingen worden de dichtst bijzijnde gehele waarden van x gevonden.

$x = 21$ $\qquad$ $x = 9$

De waarde x is het aantal waarnemingen kleiner dan de grenswaarde voor k. Men vindt de grenswaarden het eenvoudigst door de k uitkomsten te rangschikken naar grootte.
Tabel 3 geeft de rangschikking van de gegevens.

Tabel 3

In volgorde van grootte gerangschikte uitkomsten van 30 k-bepalingen

x	k	x	k	x	k	x	k	x	k	x	k
0		5		10		15		20		25	
	0.004		0.008		0.02		0.03		0.05		0.14
1		6		11		16		21	←	26	
	0.007		0.01		0.02		0.03		0.08		0.16
2		7		12		17		22		27	
	0.007		0.01		0.02		0.03		0.08		0.25
3		8		13		18		23		28	
	0.007		0.01		0.03		0.04		0.09		0.34
4		9	←	14		19		24		29	
	0.008		0.01		0.03		0.04		0.09		0.46

De pijlen in de tabel geven de grenswaarden aan. Het 95% betrouwbaarheidsinterval wordt dus:

$0.01 < m < 0.08$

De beste schatting voor m is : $m = 0.03$

Ook nu blijkt dat de veronderstelling $m = 0.10$ buiten het betrouwbaarheidsinterval ligt.

Conclusie: H_o : $m = 0.10$ wordt verworpen.

Samenvatting en conclusies

De uitkomsten van de uitgevoerde analyses kunnen als volgt samengevat worden.

Geval	S (μ)	Betrouwbaarheidsinterval	H_o	Uitspraak
I	0.07	$0.0306 < \mu < 0.1100$	$\mu = 0.10$	aanvaard
II	0.03	$0.0193 < \mu < 0.0501$	$\mu = 0.10$	verworpen
III	0.03	$0.01 < m < 0.08$	$m = 0.10$	verworpen

Uit de berekeningen blijkt duidelijk tot welke verschillende uitspraken men kan komen door uit te gaan van verschillende veronderstellingen omtrent de kansverdeling van een stochastiek.

De uitspraken van methoden II en III komen overeen, waarbij opgemerkt wordt dat bij III het betrouwbaarheidsinterval wat breder is dan bij II. Methode II is dus scherper.
Methode I leidt tot een afwijkende conclusie. De oorzaak hiervan ligt in het aannemen van de normale kansverdeling voor een stochastiek die niet normaal verdeeld is. Dat juist in dit geval de hypothese niet verworpen wordt vindt zijn oorzaak in het feit dat $\bar{x}$ naar de hypothetische waarde toe wordt getrokken door enkele grote waarnemingen en doordat een grotere $S(\sigma)$ berekend wordt. Aangezien deze twee waarden tezamen het betrouwbaarheidsgebied bepalen komen minder aannemelijke schattingen voor μ binnen het gebied te liggen.

Het is dus raadzaam te onderzoeken of een stochastiek inderdaad normaal verdeeld kan zijn. Is dit niet zo dan is het in vele gevallen mogelijk door een geschikte transformatie de stochastiek te normaliseren.

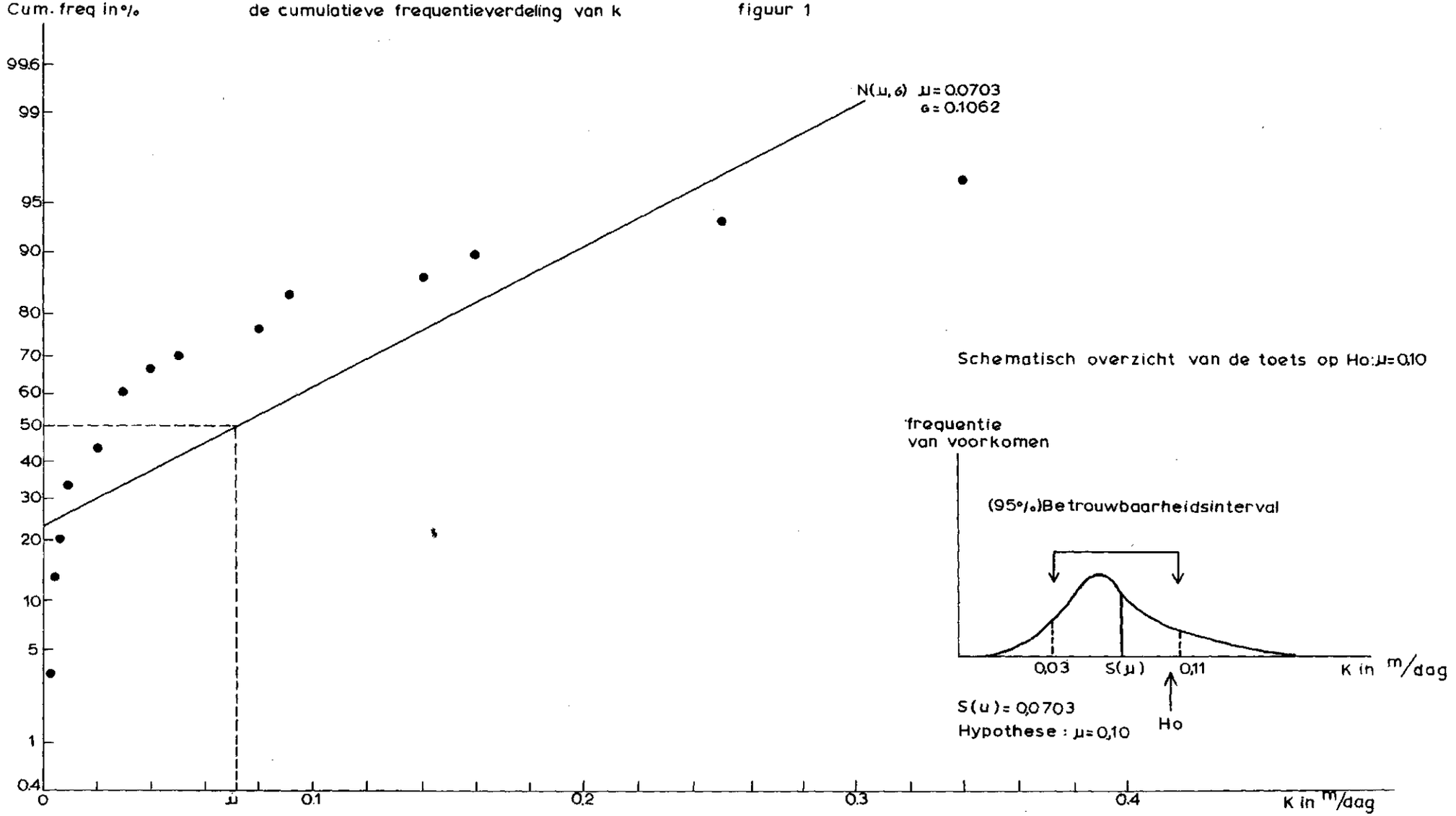
Blijkt het niet mogelijk een dergelijke transformatie te vinden of wil men de verdeling buiten beschouwing laten, dan kan het toetsen van een hypothese slechts met parameter-vrije methoden worden uitgevoerd. (methode III).

Figuur 3 ten slotte geeft een schematische voorstelling van de gevolgde gedachtengang.

Cum. freq in %

de cumulatieve frequentieverdeling van k

figuur 1

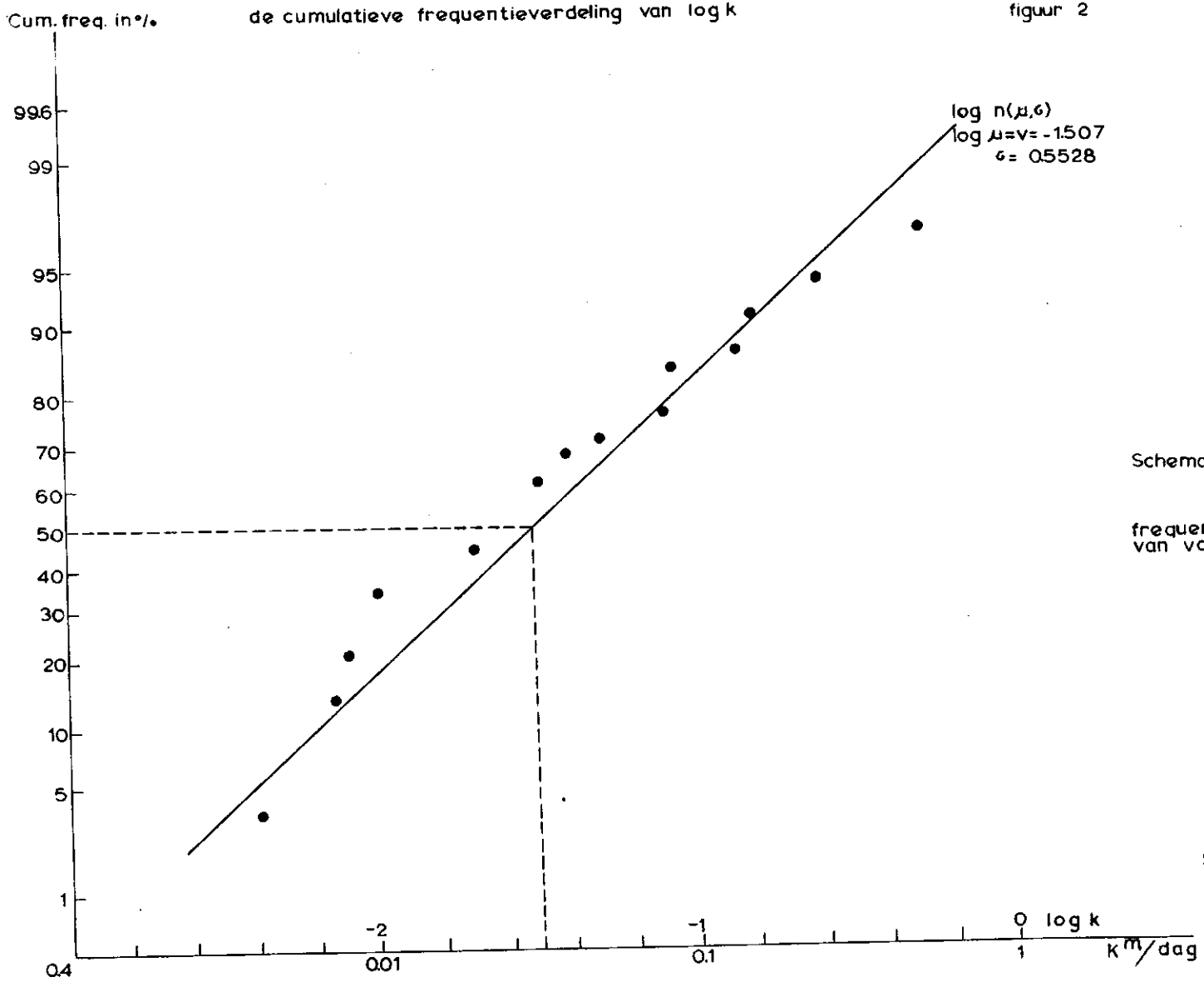


Schematisch overzicht van de toets op Ho: μ = 0,10

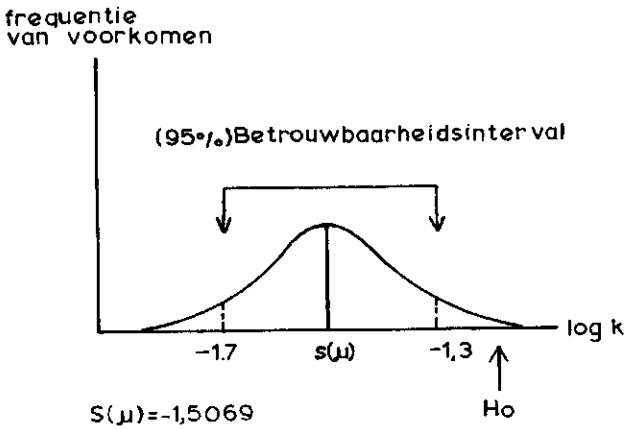
61 C
15-2

de cumulatieve frequentieverdeling van log k

figuur 2



Schematisch overzicht van de toets op Ho: V=-1,0



S(μ)=-1,5069
Hypothese: μ=-1,0

61 C
15-1

figuur 3

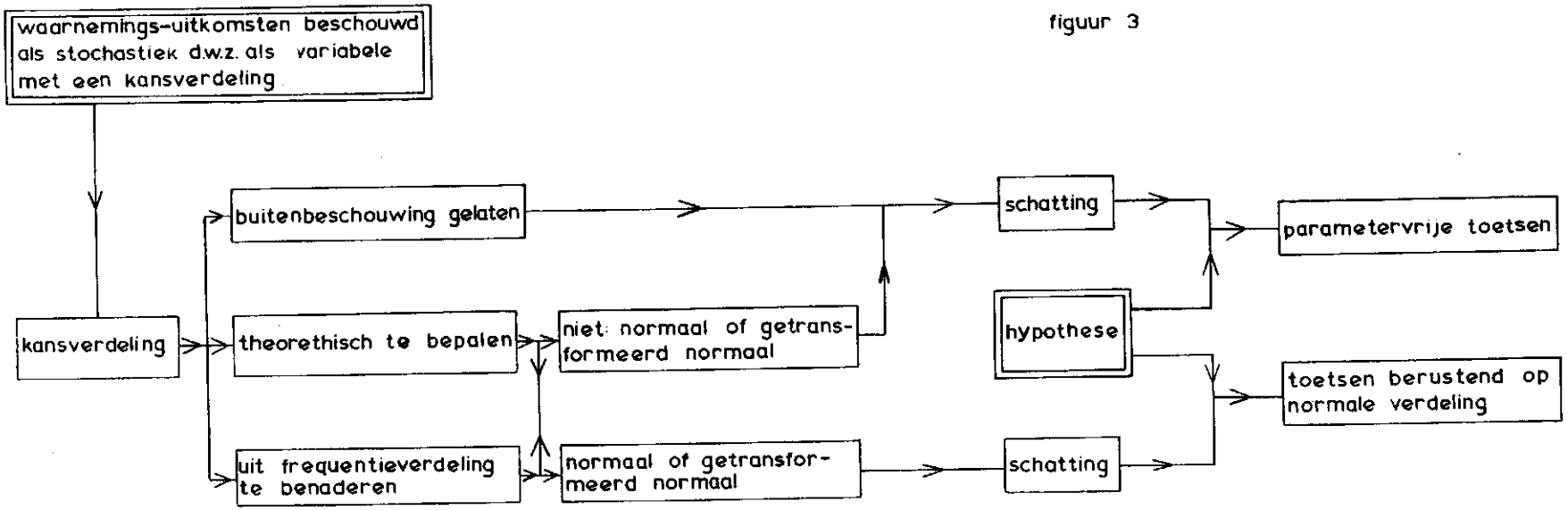


Fig 3 SCHEMA VAN DE MOGELIJKHEDEN TOT HET TOETSEN VAN EEN HYPOTHESE UITGEVOERD OP EEN REEKS WAARNEMINGSUITKOMSTEN.