MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 117

Onderzoek naar het caroteengehalte van palmolie uit de Belgische Congo

door

J.F.van Haastrecht
en
Ph.van Elteren

Vertrouwelijk

1953

INHOUD

1. Inleiding.

Dit rapport bevat de conclusies en uitkomsten van een onderzoek over het caroteengehalte (afkorting: car geh.) van palmolie uit de Belgische Congo

Naast andere vragen wordt in dit rapport in het bijzonder nagegaan of, en zo ja hoe, het car.geh. verandert tijdens het vervoer van de olie van de plaats van herkomst naar de raffinaderijen.

Voorafgaand aan de behandeling van de gestelde vragen zijn de resultaten van een statistisch vooronderzoek opgenomen, waarbij is nagegaan in hoeverre tussen de diverse gegevens in het waarnemingsmateriaal correlaties bestaan, om daarmede bij de verdere analyse rekening te houden.

1.1. Materiaal.

De gegevens betreffen in hoofdzaak het caroteengehalte van palmolie uit de Belgische Congo, bepaald in de verschillende stadia van vervoer van de olie uit de plaats van herkomst naar de raffinaderijen. Zo zijn car gehn. bepaald:

a) Bij de mills en bulking stations in het binnenland.

b) Te Leopoldville, waarheen bijna alle Congo-olie per schip getransporteerd wordt en vanwaar de olie per trein naar Matadi gaat. (Vanuit het aan zee gelegen bulking station Boma wordt de olie rechtstreeks naar Europa verscheept.)

c) Te Matadi, een zeehaven vanwaar de olie per schip naar Antwerpen vervoerd wordt.

d) Te Antwerpen, waar de olie wordt overgeladen in lichters.

e) Bij de raffinaderijen, waar deze lichters aankomen.

Verder is steeds, of alleen bij bepaalde gedeelten in het materiaal gegeven:

a) Datum en plaats van monstername.

b) VD = vertrekdatum van de olie uit mill of bulking station

c) Tonnage van de vervoerde partij.

d) FFA = vrij vetzuurgehalte van de olie.

e) Of de palmolie afkomstig is van de Belgische Unilever-firma H.C.B. (UL) of van derden (T)

f) Of de olie uit het noorden (N) of uit het zuiden (Z) van de Congo afkomstig is.

g) Productiecijfers van een aantal mills en bulking stations van de Unilever.

Te Leopoldville is het car.geh. steeds gemeten op de dag van aankomst van de desbetreffende partij aldaar. Uit die datum en uit de VD van die partij volgt onmiddellijk de vervoerstijd (VT) van de partij van mill of bulking station naar Leopoldville.

1.2 Vragen.

Hieronder volgen vragen, die aan de hand van bovengenoemd materiaal gesteld zijn. Achter elke vraag is tussen haakjes vermeld in welke paragraaf van dit rapport die vraag behandeld is.

1. Wat is het gemiddelde car.geh. van de palmolie die te Leopoldville arriveert (4.1).
Hoeveel procent van deze palmolie bezit een car.geh. boven en hoeveel beneden dit gemiddelde (4.1).
Hoeveel procent van de palmolie, die te Leopoldville aankomt, heeft een car.geh. van 0,60 - 0,65, van 0,65 - 0,70, van 0,70 - 0,75 en boven 0,75 mg/g (4.1)
2. Bestaat er een systematisch verschil, en zo ja hoe groot, tussen het car.geh. van de olie afgeleverd door de mills in de Belgische Congo en het car.geh. van de olie die te Leopoldville arriveert (5.1).
3. Bestaat er een systematisch verschil, en zo ja hoe groot, tussen het car.geh. van de palmolie afkomstig van de H.C.B. en zusterorganisatie (UL) bij aankomst te Leopoldville en het car.geh. van de olie afkomstig van derden (T) bij aankomst te Leopoldville (4.3).
4. Bestaat er een systematisch verschil, en zo ja hoe groot, tussen het car.geh. van de palmolie afkomstig uit het noorden van de Belgische Congo bij aankomst te Leopoldville en het car.geh. van de palmolie afkomstig uit het zuiden bij aankomst te Leopoldville (4.2).
5. Bestaat er een systematisch verschil, en zo ja hoe groot, tussen het car.geh. van de palmolie afgeleverd door de mills in het noorden van de Belgische Congo en het car.geh. van de olie afkomstig van die mills bij aankomst in Leopoldville (5.2).
6. Als vraag 5, maar nu voor de mills in het zuiden van de Belgische Congo (5.3).
7. Indien bij vraag 5 en 6 inderdaad zulke verschillen bestaan, is er dan een systematisch verschil tussen deze verschillen (5.4).
8. Bestaat er een correlatie tussen het car.geh. en het percentage vrij vetzuur (FFA) van de monsters genomen te Leopoldville, Coquilhatville en Boma (6.1).
9. Bestaat er een systematisch verschil, en zo ja hoe groot, tussen het car.geh. van de olie afgeleverd te Coquilhatville en het car.geh. van dezelfde olie bij aankomst te Leopoldville (5.5).
10. Bestaat er een systematisch verschil, en zo ja hoe groot, tussen het car.geh. van de olie die Matadi verlaat (onderzoek 1952) en de olie die Boma verlaat (6.2).

11. Bestaat er een systematisch verschil, en zo ja hoe groot, tussen het car.geh. van de olie die Matadi verlaat en het car. geh. van de olie die te Antwerpen arriveert (onderzoek 1951) (5.6).
12. Hetzelfde als vraag 11, maar nu voor Antwerpen en de raffinaderijen te Baasrode en Merksem (onderzoek 1951) (5.7).
13. Bestaat er een systematisch verschil, en zo ja hoe groot, tussen het car geh. van de olie die Matadi verliet bij het onderzoek in 1951 en bij het onderzoek in 1952 (6.3).

2. Enige opmerkingen over de statistische bewerking.

Deze opmerkingen betreffen de overschrijdingskansen en de kwantitatieve resultaten, die veelal bij de conclusies zijn opgenomen.

2.1. Overschrijdingskansen

Onze conclusies berusten op statistische toetsingsmethoden. Toetsingsmethoden dienen om na te gaan of een gevonden effect (b.v. een correlatie tussen twee reeksen waarnemingen, een verschil van twee gemiddelden etc.) systematisch is dan wel aan toevallige omstandigheden kan worden toegeschreven. Gewoonlijk formuleren wij in een of andere mathematische vorm, nulhypothese genoemd, de onderstelling dat het gevonden effect "toevallig" is. Wij berekenen dan de kans op het gevonden of een nog sterker effect in de onderstelling dat de nulhypothese vervuld is. Deze kans wordt overschrijdingskans genoemd. De nulhypothese wordt verworpen als de overschrijdingskans beneden een bepaalde grens α (gewoonlijk neemt men $\alpha = 0{,}05$) ligt. Wij concluderen dan dat het effect "systematisch" (ook wel "duidelijk" of "aantoonbaar") is. De betekenis van α (de z.g. onbetrouwbaarheidsdrempel) is, dat de kans op ten onrechte verwerpen van de nulhypothese, dus de kans op het verwerpen van deze hypothese, als hij juist is, gelijk is aan α. Voor $\alpha = 0{,}05$ zal dus gemiddeld ongeveer één of de twintig juiste nulhypothesen per ongeluk toch verworpen worden. De conclusie, dat een effect niet op toeval berust, maar systematisch is, is des te hechter naarmate de gevonden overschrijdingskans kleiner is. Ligt de overschrijdingskans boven de gebruikte grens α, dan kunnen wij de nulhypothese niet verwerpen. Het effect kan dan toch wel systematisch zijn, doch het was dan niet duidelijk genoeg (of de toets niet scherp genoeg) om het als zodanig te onderkennen. Een vaak voorkomende oorzaak hiervoor is, dat het aantal verrichte waarnemingen te gering is om effecten, die niet zeer sterk zijn, te kunnen aantonen.

In hoofdzaak is bij dit onderzoek gebruik gemaakt van de toets van WILCOXON voor het probleem van twee steekproeven en van

methoden van rangcorrelatie van KENDALL en van ELFVING en WHITLOCK voor het ontdekken van een verloop. Tussen de verschillende plaatsen van herkomst blijken grote verschillen te bestaan in car.geh , VT en tonnage van de uit die plaatsen afkomstige partijen palmolie. Daarom zijn de analyses vaak eerst voor ieder der mills en bulking stations apart uitgevoerd, terwijl de resultaten daarna werden gecombineerd op dezelfde wijze als door ELFVING en WHITLOCK voor de rangcorrelatiemethode is aangegeven.

Van de voornaamste toetsingsmethoden, die bij het onderzoek zijn gebruikt, zijn memoranda toegevoegd aan dit rapport (de toets van WILCOXON: memorandum S 47 (M 7); rangcorrelatie: S 47 (M 13)). Deze memoranda bevatten korte beschrijvingen van de desbetreffende toetsingsmethoden. Ter oriëntatie is toegevoegd memorandum S 47 (M 6) over de algemene gang van zaken bij het toetsen van hypothesen.

Literatuur

W.J.Dixon and A.M.Mood, The statistical sign test, Journal of American Statistical Association 41 (1946), p. 556-566.

G.Elfving and J.H.Whitlock, A simple trend-test with application to erythrocyte data, Biometrics 6 (1950), p. 282-288.

M.G.Kendall, Rank correlation methods, London 1948, Hoofdstuk 1.

H.B.Mann and D.R.Whitney, On a test of whether one of two random variables is stochastically larger than the other, Annals of Mathematical Statistics 18 (1947), p. 50-60.

F.Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945), p. 80-83.

2.2. Kwantitatieve resultaten.

Voor zover mogelijk zijn bij het eigenlijke onderzoek de gemiddelden van het car.geh. berekend door te wegen met de tonnages van de partijen palmolie, d.w.z. op de volgende manier:

$$\frac{\text{som van: tonnage van een partij x car.geh. van die partij}}{\text{som van tonnages}}.$$

De aldus berekende waarde is een maat voor het desbetreffende gemiddelde in de periode waarin de waarnemingen in de Congo gedaan zijn. Deze bezit een zekere speiding door de schommelingen in het car.geh. van de olie op verschillende tijdstippen en door schommelingen in de bepaling alsmede in de tonnages van de partijen. Daar er volgens opgave grote schommelingen optreden in de productiecijfers van de afzonderlijke mills en daar

het car.geh van de olie bij de verschillende mills verre van constant is, is ook het gemiddelde aan flinke schommelingen onderhevig. Deze zijn ons echter niet bekend, daar de waarnemingen zich slechts over een korte tijd uitstrekken. De resultaten van dit onderzoek kunnen daarom het best gezien worden als een soort momentopname, die een indruk geeft van de algemene toestand zonder ons in staat te stellen nauwkeurige voorspellingen te doen

In sommige gevallen is in plaats van een gewogen gemiddelde een ongewogen gemiddelde gebruikt omdat de grootten van de partijen niet bekend waren; dit is er steeds bij vermeld.

3. Resultaten van het vooronderzoek.

Bij dit vooronderzoek is in het materiaal te Leopoldville gezocht naar eventueel optredende afhankelijkheden tussen de volgende kenmerken van de partijen olie: car.geh., FFA, VD, VT en tonnage.

Steeds is bij dit onderzoek gebruik gemaakt van de rangcorrelatiemethode [1]).

In tabel I is voor elke combinatie van twee kenmerken de overschrijdingskans opgegeven van de toets. De resultaten van de toets zijn eerst gecombineerd voor de partijen van de mills, van de bulking stations en voor de partijen waarvan niet bekend is of zij al dan niet via een bulking station vervoerd zijn. Daarna zijn de resultaten van deze groepen weer gecombineerd

Bij de bepaling van de overschrijdingskans van het correlatieonderzoek tussen car.geh. en VT is aangenomen dat het car. geh. niet kan stijgen door een langere vervoersduur.

Uit het materiaal blijkt dat verschillende keren meerdere lichters op dezelfde dag van een mill of bulking station vertrekken en ook op dezelfde dag te Leopoldville aankomen. Daar het zeer aannemelijk is dat deze lichters gelijk zijn opgevaren, b.v. in een sleep, is het niet gerechtvaardigd de VT's van de in deze schepen vervoerde partijen olie als onderling onafhankelijk te beschouwen. Bij het onderzoek of VD en tonnage de VT beïnvloeden hebben wij daarom elke dergelijke groep partijen beschouwd als één partij, waarvan het tonnage gelijk is aan het gemiddelde van de tonnages van die groep.

Het + of - teken in de tabel betekent dat een positieve resp. negatieve correlatie gevonden is.

---

1) Zie memorandum S 47 (M 13).

Tabel I

Overschrijdingskansen, gevonden bij de toepassing van de rang-correlatiemethode.

onderzoek	mills	bulking stations	onbekend	gecom-bineerd
car geh. - FFA	+ 0,65	+ 0,66	- 0,66	+ 0,65
" " - VD	+ 0,33	+ 0,032	- 0,80	+ 0,040
" " - VT	- 0,40	- 0,021	+ 0,52	- 0,06
" " - tonnage	+ 0,81	0 1,00	- 0,10	- 0,71
FFA - VD	+ 0,034	+ 0,50	+ 0,40	+ 0,036
" - VT	+ 0,89	+ 0,22	- 0,65	+ 0,44
" - tonnage	+ 0,94	+ 0,45	- 0,80	+ 0,61
VD - VT	- 0,71	- 0,43	- 0,43	- 0,29
" - tonnage	+ 0,28	+ 0,76	- 0,96	+ 0,39
VT - "	+ 0,16	- 0,83	- 0,26	+ 0,68

We vinden uit de overschrijdingskansen van de gecombineerde toets een aanwijzing dat:

a) het car.geh bij latere VD in het algemeen hoger is;

b) het car.geh. bij langere VT in het algemeen lager is;

c) FFA bij latere VD in het algemeen hoger is.

De effecten a) en c) zouden wijzen op een zekere seizoen-invloed op car.geh. en FFA; het effect b) zou er op wijzen dat het verlies aan car.geh. tijdens de reis (vervoersverlies) groter wordt bij langere VT.

Door steeds 2 kenmerken, los van de andere, met elkaar te vergelijken verwaarloost men de invloed, die veranderingen in die andere kenmerken op de vergeleken grootheden kunnen hebben.

In enkele gevallen, waarbij een dergelijke invloed van één ander kenmerk aanwezig zou kunnen zijn, hebben wij het materiaal gesplitst in groepen, waarbij vóór de partijen uit één groep de waarde van zo'n kenmerk in eenzelfde interval is gelegen. Een toetsing is dan voor elke groep apart uitgevoerd en daarna zijn de resultaten gecombineerd.

De resultaten hiervan waren:

Tabel II

Overschrijdingskansen, gevonden bij de toepassing van de rang-correlatiemethode, na verdere splitsing van het materiaal.

onderzoek	gesplitst naar gelijke	mills	bulking stations	onbekend	gecomb.
car.geh. - VD	VT	+ 0,61	+ 0,038	+ 0,91	+ 0,046
" " - VD	FFA	+ 0,34	+ 0,024	+ 0,17	+ 0,007
" " - VT	VD	+ 0,97	- 0,12	- 0,31	+ 0,52

De aanwijzing voor een systematisch seizoeneffect op het car.geh. blijft dus bestaan. Dit seizoeneffect, waarvan het bestaan bij de mills niet is aangetoond, maar toch ook daar aanwezig kan zijn, kan gevaar voor schijneffecten geven. Overal waar dit gevaar bestaat is het zoveel mogelijk uitgeschakeld door het materiaal op analoge wijze als hierboven beschreven is te splitsen naar gelijke intervallen voor de VD.

We vinden nu geen systematische correlatie tussen car.geh. en VT. Uit tabel I vinden we eveneens geen systematische correlatie tussen het tonnage van de partij tijdens het vervoer naar Leopoldville en het car.geh. bij aankomst te Leopoldville. Dit zou er op wijzen dat het verlies in het car.geh. van olie tijdens het vervoer naar Leopoldville, waarvan het bestaan in 5.1 nagegaan wordt, niet wordt beïnvloed door VT of tonnage. De waarnemingen zijn voor het onderzoeken van deze kwestie echter slechts weinig geschikt, daar er geen directe metingen van het vervoersverlies beschikbaar waren [2]). De negatieve uitkomst betekent dus geenszins, dat VT en tonnage geen invloed op het caroteengehalte hebben, maar alleen dat het hierdoor (eventueel) tussen de partijen ontstane verschil niet zo groot is, dat het aan het totale caroteengehalte te bespeuren is. Voor ons onderzoek zijn de negatieve conclusies van belang omdat wij dus verder geen gevaren voor schijneffecten van de kant van VT en tonnages te wachten hebben, zodat wij niet voor deze grootheden in homogene groepen behoeven te splitsen.

---

2) Onder een directe meting van dit verlies verstaan wij het verschil tussen een meting van het car.geh. aan het begin en aan het eind van de reis van een bepaald schip.
In 5.1 wordt weliswaar gebruik gemaakt van een bepaling van het vervoersverlies, maar die slaat niet op een bepaalde partij en is niet bruikbaar voor het onderzoek of het vervoersverlies beïnvloed wordt door VT of tonnage. Door de in 2.1 genoemde verschillen tussen de diverse mills en bulking stations moet dit laatste onderzoek namelijk voor elke mill of bulking station apart worden uitgevoerd en bij de in 5.1 gebruikte methode krijgen we per plaats van herkomst slechts één waarneming van het vervoersverlies, zodat een eventueel verband met VT of tonnage niet kan worden nagegaan.

4. Verwerking van de waarnemingen, gedaan te Leopoldville.

4.1. De verdeling van de gemeten caroteengehalten.

In deze paragraaf worden enige vragen beantwoord die een indruk van de frequentieverdeling van de waargenomen car.gehn. te Leopoldville geven.

Voor het gewogen gemiddelde car.geh. van de palmolie, die te Leopoldville arriveert vonden we 0,58 mg/g.

Voor het gewichtspercentage van deze olie met car.geh. hoger en het percentage met car.geh. lager dan het gemiddelde vonden wij resp. 47 en 53%. Statistisch gezien is het verschil tussen deze beide percentages bij een materiaal van deze omvang onbetekenend, zodat men op grond van dit materiaal niet tot een afwijking van de verhouding 50:50 kan besluiten.

Verder vonden wij voor het gewichtspercentage van de olie die te Leopoldville arriveert met car.geh.

0,61 t/m 0,65 mg/g : 15%
0,66 t/m 0,70 mg/g : 11%
0,71 t/m 0,75 mg/g : 1%
0,76 en hoger : 2%

Ook deze cijfers mag men niet als nauwkeurig bepaald beschouwen; alleen de orde van grootte ervan is van betekenis te achten.

4.2. Verschil in het car.geh. van olie uit het zuiden en uit het noorden van de Congo.

Om na te gaan of hier van een systematisch verschil sprake is, hebben wij de toets van WILCOXON (zie memorandum S 47 (M 7)) toegepast op de gemiddelde car.gehn. per maand (i.v.m. seizoen-effect dat bij het vooronderzoek gevonden werd) van palmolie voor de verschillende plaatsen van herkomst.

Wij vonden, gecombineerd over de maanden, dat het car.geh. van de palmolie uit het noorden duidelijk kleiner is dan dat van de palmolie uit het zuiden van de Belgische Congo bij aankomst te Leopoldville ($k = 10^{-5}$). Gesplitst naar de maanden waarin de olie van mill of bulking station vertrokken was, vinden wij:

Juni gemiddeld car.geh. zuid (0,625) - gemiddeld car.geh. noord (0,547) = 0,08 mg/g
Juli gemiddeld car.geh. zuid (0,647) - gemiddeld car.geh. noord (0,569) = 0,08 mg/g
Aug. gemiddeld car.geh. zuid (0,642) - gemiddeld car.geh. noord (0,570) = 0,07 mg/g.

Overal waar dit verschil aanleiding kan geven tot schijn-effecten gaan we, op analoge wijze als in 3 beschreven is, te werk door het materiaal te splitsen naar noord en zuid.

4.3. Verschil in het car.geh. van olie van Unilever en van derden.

De toets van WILCOXON werd hier toegepast op dezelfde gemiddelden als bij de toets van 4.2 gebruikt zijn. De toets werd voor elke maand apart uitgevoerd voor noord en voor zuid. Gecombineerd over deze groepen vonden we geen systematisch verschil tussen het car.geh. van de palmolie afkomstig van de Unilever (gemiddeld 0,[illegible] mg/g) bij aankomst te Leopoldville en het car. geh. van de olie afkomstig van derden (gemiddeld 0,57 mg/g) bij aankomst te Leopoldville ($k = 0,37$).

5. Vervoersverlies in het car.geh van palmolie.

Dit hoofdstuk behandelt enkele vragen over het vervoersverlies in het car.geh. van palmolie gedurende enkele perioden van het vervoer. Daarnaast worden enige vragen behandeld over het vervoersverlies in het car.geh. van palmolie, speciaal afkomstig van het noorden van de Congo, van het zuiden van de Congo en van Coquilhatville.

5.1. Verlies tijdens het vervoer naar Leopoldville.

Hiertoe werden de waarnemingen van het car.geh., gedaan bij een mill of bulking station in het binnenland, vergeleken met die gedaan te Leopoldville van olie afkomstig van dezelfde mill of bulking station.

Dit werd uitgevoerd m.b.v. de toets van WILCOXON. Gecombineerd over alle mills en bulking stations waarbij waarnemingen van het car.geh ter plaatse zijn verricht, vonden we als overschrijdingskans $k = 0,07$. Bij deze methode is de invloed van de weinige (steeds slechts 1 of 2) waarnemingen van het car.geh. bij een mill of bulking station des te groter naarmate meer waarnemingen van het car.geh. van de olie van die mill of bulking station te Leopoldville verricht zijn. Deze invloed van de uitgebreidheid van de steekproeven is te elimineren door gebruik te maken van de tekentoets [3]), toegepast op ongewogen gemiddelden van de 1 of 2 waarnemingen bij de mills en bulking stations en de gewogen gemiddelden van de waarnemingen van die mills en bulking stations te Leopoldville.

Het gebruik van de tekentoets betekent hier dat aan alle plaatsen van herkomst, waarvan ter plaatse waarnemingen van het car.geh. gedaan zijn, gelijk gewicht wordt toegekend. Als overschrijdingskans met de tekentoets vonden we $k = 0,02$. Bij de berekening van de in deze paragraaf voorkomende overschrijdings-

---

3) Zie memorandum S 53 (M 22).

kansen is ondersteld dat het car.geh. niet kan stijgen door het vervoer. (Indien deze onderstelling niet gerechtvaardigd is, zouden de overschrijdingskansen tweemaal zo groot genomen moeten worden.)

Er bestaat dus, ondanks het geringe aantal waarnemingen, gedaan bij de mills en bulking stations, inderdaad een aanwijzing dat het car.geh. van de olie aldaar gemeten groter is dan het car.geh van de olie van dezelfde mills en bulking stations gemeten te Leopoldville.

Het verschil van de gewogen gemiddelden van het car.geh. gemeten bij de bron en te Leopoldville geeft geen juist beeld van het vervoersverlies, daar juist één mill met hoge productie, Elisabetha, een erg laag car.geh. bij de mill te zien gaf, dat echter op slechts twee bepalingen berustte Daarom bepaalden wij het verschil tussen het (ongewogen) gemiddelde car geh. aan de bron en het gewogen gemiddelde car.geh te Leopoldville per mill of bulking station apart. Dit gaf:

<u>zuid</u>

Leverville	:	- 0;10 mg/g
Tango	:	- 0,16 mg/g
Lunungu (Kisia, Putubumba)	:	0,00 mg/g
Pindi (Kunga)	:	- 0,13 mg/g
Brabanta	:	- 0,10 mg/g

<u>noord</u>

Elisabetha	:	+ 0,06 mg/g
Mosite	:	- 0,07 mg/g
Bomaneh	:	- 0,06 mg/g
Yaligimba	:	- 0,01 mg/g

(Hierbij slaat een - teken op vervoersverlies en een + teken op schijnbare vervoerswinst.)

Elk van deze verschillen op zichzelf heeft slechts weinig betekenis, daar er van ieder der mills en bulking stations slechts zeer weinig waarnemingen zijn; tezamen geven zij echter een vrij goede indruk van het vervoersverlies.

Enerzijds constateerden wij dus een vervoersverlies en anderzijds vonden wij bij het vooronderzoek geen correlatie tussen vervoersduur van een partij olie naar Leopoldville en het car.geh. van die partij te Leopoldville. Dit kan geweten worden aan het ontbreken van directe metingen van het vervoersverlies (zie <u>3</u>), maar het kan wellicht ook verklaard worden door aan te nemen dat het verlies optreedt door één of meer van de volgende drie oorzaken: de opslag in de tank van mill of bulking station

(de monsters aldaar zijn genomen voordat de olie in de tank komt), het laden en lossen van de olie en de eerste dagen van het vervoer van de olie. Uitsluitsel hierover kan door de gegevens van dit onderzoek niet worden verkregen.

5.2. Vervoersverlies voor olie uit het noorden van de Congo.

Een systematisch verschil tussen het car.geh. van de olie bij de mills en bulking stations in het noorden en het car.geh van de olie afkomstig van die mills en bulking stations bij aankomst te Leopoldville kan niet apart worden aangetoond (met de toets van WILCOXON: $k = 0,70$ en met de tekentoets: $k = 0,31$, waarbij weer is aangenomen dat het car.geh. tijdens het vervoer niet stijgt). Dit betekent echter niet dat dit verschil niet bestaat; het kan zijn dat het niet gevonden kon worden omdat er slechts zeer weinig waarnemingen bij de noordelijke mills zijn verricht. Bovendien bederven de lage uitkomsten van de waarnemingen van het car.geh. te Elisabetha het algemene beeld van een (eventuele) daling.

Voor een indruk van de gevonden verschillen verwijzen we naar het gedeelte van de bij 5.1 gegeven tabel, dat op het noorden slaat

5.3. Vervoersverlies voor olie uit het zuiden van de Congo.

Bij de mills en bulking stations uit het zuiden is dit verschil wel aan te tonen (met de toets van WILCOXON: $k = 0,005$ en met de tekentoets: $k = 0,03$, met de reeds gemaakte onderstelling).

Voor een indruk van de grootte van dit verschil verwijzen wij naar het gedeelte van de bij 5.1 gegeven tabel, dat op het zuiden slaat.

5.4. Verschil tussen de vervoersverliezen van de olie uit het noorden en uit het zuiden van de Congo.

De tabel bij 5.1 en de uitkomsten van 5.2 en 5.3 houden de suggestie in, dat het vervoersverlies in noord kleiner is dan dat in zuid. Wij konden echter met de toets van WILCOXON aan de hand van de weinige beschikbare gegevens (zie tabel uit 5.1) niet aantonen dat dit het geval is ($k = 0,11$). Ook voor het onderzoeken van deze vraag zou men bij voorkeur over directe metingen van het vervoersverlies moeten beschikken.

5.5. Vervoersverlies bij olie afkomstig van Coquilhatville.

Of er een systematisch verschil bestaat tussen het car.geh. van olie, gemeten te Coquilhatville en het car.geh. van olie afkomstig van Coquilhatville bij aankomst te Leopoldville kan niet

goed onderzocht worden, daar de waarnemingen te Coquilhatville later gedaan werden dan te Leopoldville. Het in 3 gesignaleerde seizoeneffect is dus in dit geval gevaarlijk. Dit klemt te meer, daar de in Leopoldville waargenomen car.gehn. van olie uit Coquilhatville alleen reeds een aanwijzing gaven voor dit seizoeneffect (ondanks het feit, dat dit slechts een vrij klein aantal waarnemingen was, vonden wij met de methode der rangcorrelatie een overschrijdingskans k = 0,06) Een extrapolatie van de reeks in Leopoldville verrichte waarnemingen tot het tijdstip, waarop in Coquilhatville waarnemingen werden verricht - een poging daartoe werd gewaagd met behulp van de covariantieanalyse - bleek niet goed mogelijk, daar er bij de in Coquilhatville verrichte waarnemingen eerder een daling dan een stijging van het car.geh met de tijd gevonden werd. Daardoor wordt lineaire extrapolatie uitgesloten en voor kromlijnige extrapolatie zijn er niet genoeg gegevens.

Voor het gemiddelde car.geh. van de olie vonden we te Coquilhatville 0,55 mg/g (September) en voor dat te Leopoldville van olie afkomstig van Coquilhatville vonden we voor die vertrokken was in Juni 0,47 mg/g, in Juli 0,56 mg/g en in Augustus 0,54 mg/g.

5.6. Verlies tijdens het vervoer over zee.

Uit de gegevens van 1951 vonden we dat het car geh. van olie die Matadi verliet systematisch hoger was dan het car.geh. van olie die te Antwerpen arriveerde (met de toets van WILCOXON: k = 0,03). Hierbij is aangenomen dat stijging van het car.geh. gedurende de zeereis uitgesloten moet worden geacht.

Voor de grootte van het verschil tussen het gemiddelde van het car.geh. gemeten te Matadi (0,51 mg/g) en gemeten te Antwerpen (0,48 mg/g) vinden we 0,03 mg/g.

5.7. Verlies tussen Antwerpen en de raffinaderijen.

Tussen de car.gehn. van palmolie te Antwerpen (gemiddeld car.geh. = 0,48) en van olie te Baasrode of Merksem (gemiddeld car.geh. = 0,47) werd met de toets van WILCOXON geen systematisch verschil gevonden (k = 0,48), waarbij weer is ondersteld dat het car.geh. te Baasrode of Merksem niet hoger kan zijn dan te Antwerpen.

6. Beantwoording van de verder gestelde vragen.

6.1. Correlatie tussen car.geh. en vrij vetzuur gehalte.

Te Coquilhatville en Boma vonden we met behulp van de methode der rangcorrelatie een niet zeer grote, doch door het uitge-

breide materiaal aantoonbare, negatieve correlatie tussen car. geh. en percentage vrij vetzuur van de palmolie ($k = 0{,}002$ resp. 0,009). Dit betekent, dat dus in het algemeen een hoog caroteengehalte gepaard zal gaan met een laag vrij vetzuur gehalte en omgekeerd. Bij de monsters genomen te Leopoldville kon een dergelijke correlatie niet aangetoond worden ($k = 0{,}66$).

6.2. Verschil tussen car.gehn. van partijen olie, die van Boma en die van Matadi vertrekken.

We vonden met de toets van WILCOXON geen systematisch verschil tussen de car.gehn. van olie die Boma en olie die Matadi verliet uit de gegevens van 1952 ($k = 0{,}68$). Voor de gemiddelde car.gehn. vonden we bij beide 0,57 mg/g.

6.3. Verschil tussen car.gehn. van partijen olie, die in 1951 en die in 1952 van Matadi vertrekken.

Het car.geh van olie die Matadi verliet in 1952 is duidelijk groter dan dat van de olie die Matadi verliet in 1951 (met de toets van WILCOXON: $k = 0{,}001$). Voor het verschil tussen de gemiddelde car.gehn. in 1952 (0,57 mg/g) en 1951 (0,51 mg/g) vinden we 0,06 mg/g. Dit kan wellicht geheel of gedeeltelijk verklaard worden doordat de monsters, die te Matadi genomen zijn, in 1951 over zee en in 1952 per luchtpost zijn verzonden, voordat het car.geh. ervan bepaald werd.

7. Overzicht van de conclusies [4]).

a) Uit het materiaal te Leopoldville vinden we dat het car.geh. en het vrij vetzuur gehalte van een partij palmolie systematisch positief gecorreleerd met de vertrekdatum van die partij (3). Dit is een aanwijzing voor een seizoeneffect in car.geh. en FFA. Verder kunnen we o.a. niet concluderen tot een systematisch verband tussen vervoersverlies in het car.geh. en vervoersduur resp. tonnage, hetgeen wij echter onder voorbehoud vermelden, daar het gevonden resultaat wellicht is toe te schrijven aan het ontbreken van directe waarnemingen van het vervoersverlies (3).

b) Het car.geh. van palmolie, afkomstig uit het zuiden van de Congo, is systematisch hoger dan dat van olie uit het noorden van de Congo, beide bij aankomst te Leopoldville. Voor de grootte van dit verschil vonden we 0,08 mg/g (4.2).

c) Tussen de car.gehn. van partijen olie, afkomstig van de Unilever en van derden vinden wij geen systematisch verschil (4.3)

d) We vinden een aanwijzing dat het car.geh. van olie, bij de

---

4) Achter iedere conclusie wordt tussen haakjes de desbetreffende paragraaf van dit rapport vermeld.

bron gemeten, systematisch hoger is dan het car.geh. van die olie bij aankomst te Leopoldville. Het is zeer aannemelijk dat dit effect veel duidelijker naar voren komt wanneer meer waarnemingen in het binnenland of directe waarnemingen van het vervoersverlies beschikbaar waren (5.1). Apart voor palmolie, uit het zuiden afkomstig, is wel een systematisch vervoersverlies aan te tonen; voor palmolie uit het noorden echter niet (5.5 en 5.2). De grootte van de verschillen was van de orde van 0,1 mg/g

Tussen de vervoersverliezen van olie uit het zuiden en uit het noorden vinden we uit de weinige daartoe geschikte bepalingen geen systematisch verschil, al houdt het materiaal wel een geringe aanwijzing daartoe in (5.4).

e) We vinden dat het car.geh. van olie, gemeten te Matadi, systematisch hoger is dan dat van olie bij aankomst te Antwerpen. Voor de grootte van het verschil vonden we 0,03 mg/g (5.6)

f) We kunnen aan de hand van het materiaal niet aantonen, dat het car.geh. van palmolie, gemeten te Antwerpen, systematisch verschilde van dat, gemeten bij de raffinaderijen (5.7).

g) Uit de waarnemingen, gedaan te Coquilhatville en Boma, vinden we een systematisch negatieve correlatie tussen car.geh. en vrij vetzuur gehalte; bij die, gedaan te Leopoldville, kunnen we een dergelijke correlatie niet aantonen (6.1).

h) We vinden geen systematisch verschil tussen het car.geh. van palmolie die vertrekt van Boma en dat van olie die van Matadi vertrekt (6.2)

i) We vinden dat het car.geh. van palmolie, die in 1952 van Matadi vertrok, systematisch hoger is dan dat van olie, die in 1951 van Matadi vertrok. Voor de grootte van het verschil vonden wij 0,06 mg/g (6.3).

Statistische Afdeling.

S47(M6)

Algemene gang van zaken bij het toetsen van een hypothese. [1]

De toetsing van een hypothese H_0 berust steeds op een aantal waarnemingen $x_1, x_2, \ldots, x_n$ van één of meer stochastische grootheden[2], of op enige groepen van waarnemingen (bv. twee steekproeven).

Bij een toets behoort een toetsingsgrootheid u (soms meer dan één), die een functie is van bovengenoemde stochastische grootheden en die, voor de waargenomen waarden $x_1, x_2, \ldots, x_n$ een waarde aanneemt, die berekend kan worden (bv.: het gemiddelde der waarnemingen, of de spreiding, of het verschil van de gemiddelden van twee waarnemingen).

De toetsingsgrootheid wordt steeds zo gekozen, dat men, op grond van de onderstelling, dat H_0 juist is, de waarschijnlijkheidsverdeling van deze grootheid kan berekenen.

Vervolgens kiest men een verzameling Z van mogelijke uitkomsten van u, en wel op zodanige wijze, dat de kans, dat u een in Z gelegen waarde aanneemt, onder de hypothese H_0, gelijk is aan een gegeven getal α, zodat Z dus van α afhankelijk is [3]. Z heet de kritieke zône van de toets, α de onbetrouwbaarheidsdrempel (Engels: level of significance). Voor α neemt men veelal de waarde 0,05 of 0,01.

Men verwerpt nu H_0 op grond van de waarnemingen $x_1, x_2, \ldots, x_n$, indien de bij deze waarnemingen behorende waarde van u in Z ligt. Dit wordt vaak uitgedrukt door te zeggen, dat het resultaat van het experiment "significant" is. De waarde van α moet dan echter worden vermeld. De kans, dat dit zal gebeuren, is, indien H_0 juist is, gelijk aan α. Derhalve is α de kans op ten onrechte verwerping van de juiste hypothese, ook de kans op een fout van de eerste soort genoemd. Indien men deze methode toepast, met $\alpha = 0{,}05$ resp. 0,01, zal men in gemiddeld ongeveer één op 20 resp. op 100 van de gevallen, waarin de hypothese die men toetst juist is, deze toch verwerpen.

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

2) Een stochastische grootheid is een grootheid, die een waarschijnlijkheidsverdeling bezit, of, anders gezegd, een grootheid, die voor de elementen van een collectie(universum,populatie) gedefinieerd is en daarop allerlei waarden aanneemt. Stochastische grootheden worden aangegeven door onderstreepte letters.

3) Soms kan men slechts bereiken, dat deze kans $\leq \alpha$ is.

De toetsingstheorie biedt in het algemeen geen mogelijkheid om tot aanvaarding van een hypothese te komen. Indien een bepaalde hypothese H_0 niet verworpen kan worden, is dit gewoonlijk met een hele verzameling van hypothesen tegelijk het geval. Niet-verwerpen staat dus niet gelijk met aanvaarden.

Wel zal men vaak in de loop van een statistische analyse bepaalde onderstellingen, die plausibel schijnen en voor de verdere analyse van nut zijn, toetsen, alvorens ze bij de verdere bewerking van het materiaal te gebruiken. Worden zij dan op grond van de toets niet verworpen, dan houdt dit in zo verre een rechtvaardiging van die onderstellingen in, dat een grote afwijking door de toets veelal wel zou zijn ontdekt. Indien men dan verder de onderstellingen gebruikt, verwaarloost men eventueel aanwezige afwijkingen van onbekende grootte, die echter niet zo groot zijn, dat zij door de toets zijn ontdekt.

Vele toetsen gelden zelf alleen onder bepaalde onderstellingen omtrent de waarschijnlijkheidsverdelingen der stochastische grootheden, waarvan waarnemingen zijn verricht. Deze nevenvoorwaarden dienen steeds uitdrukkelijk te worden vermeld en, zo mogelijk, zelf te worden getoetst.

In plaats van de onbetrouwbaarheidsdrempel α wordt vaak bij de uitslag van een toetsing de overschrijdingskans k opgegeven; dit is de kleinste waarde van α, waarbij in het betrokken geval, nog tot verwerping van H_0 zou zijn overgegaan; anders gezegd: de kleinste α, waarvoor de gevonden waarde der toetsingsgrootheid nog juist in de (bij α behorende) kritieke zône Z ligt. Wordt dus de waarde k opgegeven en werkt men met onbetrouwbaarheidsdrempel α, dan wordt verworpen, indien $k \leq \alpha$ is.

Voor het onderscheid tussen één- en tweezijdige toetsing en de keuze tussen deze twee mogelijkheden vergelijke men bv. de tweede hieronder gegeven litteratuurplaats. Wij moeten hier volstaan met de opmerking, dat éénzijdige toetsing veelal eerder tot verwerping van H_0 leidt, maar dat deze slechts onder bijzondere omstandigheden kan worden toegepast.

Litteratuur:

J.Neyman, First course in probability and statistics, New York, 1950, Chapter 5.

J.Hemelrijk en H.R. van der Vaart, Het gebruik van één- en tweezijdige overschrijdingskansen voor het toetsen van hypothesen, Statistica 4 (1950) p.54-66.

Mathematisch Centrum,
2de Boerhaavestraat 49,
Amsterdam O.
Statistische Afdeling,
S47 (M7).

Maart, 1952.

De toets van Wilcoxon.[1]

Deze methode dient tot het toetsen van de hypothese H_o, inhoudende, dat twee steekproeven $x_1,\ldots,x_n$ en $y_1,\ldots,y_m$ afkomstig zijn uit één collectie (ook wel populatie of universum genaamd).

Voor het toetsen van de hypothese H_o wordt gebruik gemaakt van een toetsingsgrootheid $\underline{U}$ [2]), die als volgt uit de waarnemingen berekend wordt. Onderstellen we, dat de waarnemingen $x_1,\ldots,x_n$ en $y_1,\ldots,y_m$ naar opklimmende grootte gerangschikt zijn, dan bepalen we eerst het aantal waarnemingen uit de tweede steekproef, dat kleiner is dan de kleinste waarneming x_1 uit de eerste steekproef (bij gelijkheid telle. wij $\frac{1}{2}$ in plaats van 1). Noem dit aantal V_1. Vervolgens wordt het aantal waarnemingen uit de tweede steekproef bepaald, dat kleiner is dan de op één na kleinste waarneming x_2 uit de eerste steekproef (bij gelijkheid wordt weer een $\frac{1}{2}$ in plaats van 1 geteld). Dit aantal noemen we V_2. Evenzo worden met betrekking tot $x_3,x_4,\ldots,x_n$ de aantallen $V_3,V_4,\ldots,V_n$ bepaald. De waarde U van de toetsingsgrootheid $\underline{U}$ wordt voor de twee steekproeven dan gegeven door

$$U=V_1+V_2+\ldots+V_n.$$

Wanneer onder de waarnemingen niet te veel gelijken voorkomen, kan bewezen worden, dat de toetsingsgrootheid $\underline{U}$ onder de hypothese H_o voor grote waarden van n en m (beide $\geq$ 10) bij benadering een normale verdeling bezit. De waarnemingen $x_1,\ldots,x_n$ en $y_1,\ldots,y_m$ tezamen genomen vallen uiteen in een aantal groepen van gelijke waarnemingen. Noem het aantal van deze groepen k, dan is k minstens 1 (als alle waarnemingen gelijk zijn) en hoogstens m+n (als alle waarnemingen verschillend zijn).

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

2) Stochastische grootheden worden door onderstreping aangeduid.

Zijn $t_1,\ldots,t_k$ de aantallen waarnemingen in deze groepen van gelijken, dan worden het gemiddelde μ en de variantie σ^2 van de toetsingsgrootheid $\underline{U}$ gegeven door

$$\mu(\underline{U}) = \tfrac{1}{2}nm,$$

en

$$\sigma^2 = \mathrm{Var}\ (\underline{U}) = \frac{1}{12}\frac{nm}{(n+m)(n+m-1)}\left\{(n+m)^3 + - (t_1^3+t_2^3+\ldots+t_k^3)\right\} \quad 1)$$

De grootheid $\mu(\underline{U})$ is dus onafhankelijk van de waarden vast. Indien de hypothese H_o niet vervuld is, zal de grootheid $\underline{U}$ grote of kleine waarden bezitten, al naar gelang $\underline{y}$ systematisch kleiner of groter is dan $\underline{x}$.
De (tweezijdige) toets bestaat nu daarin, dat men H_o verwerpt indien de gevonden waarde U van $\underline{U}$ te sterk van μ afwijkt, d.w.z. als

$$\frac{|U-\mu|}{\sigma} > \xi_\alpha \ , \quad 2)$$

waarin α de onbetrouwbaarheidsdrempel is en ξ_α volgt uit

$$\frac{1}{\sqrt{2\pi}}\int_{\xi_\alpha}^{\infty} e^{-\frac{1}{2}x^2}\,dx = \tfrac{1}{2}\alpha \ ,$$

en in een tabel van de normale verdeling kan worden opgezocht.
De (tweezijdige) overschrijdingskans k , behorende bij T, is gedefinieerd als

$$k = \frac{2}{\sqrt{2\pi}}\int_{\left|\frac{U-\mu}{\sigma}\right|}^{\infty} e^{-\frac{1}{2}x^2}\,dx \quad 2)$$

en kan ook in eenttabel van de normale verdeling worden gevonden.
Bij eenzijdige toetsing wordt α door 2α vervangen, resp. k gehalveerd.
Een bijzonder geval van het bovenstaande is, dat onder de waarnemingen voor $\underline{x}$ en $\underline{y}$ in 't geheel geen gelijken voorkomen. In dat geval kan de uitdrukking voor de variantie herleid worden tot

$$\sigma^2 = \frac{1}{12}\,nm(n+m+1).$$

---

1) Deze formule is een door T.J.Terpstra gegeven vereenvoudiging van de door J.Hemelrijk([5] en [7]) afgeleide formule. De afleiding van deze vereenvoudigde formule zal nog gepubliceerd worden.

2) Deze formules berusten op de normale benadering van de verdeling van $\underline{U}$.

Indien n en m kleiner zijn dan 10, zijn tabellen beschikbaar voor het berekenen van de overschrijdingskans k voor de uit de steekproef bepaalde waarde U van U (zie [2] en [4]). Dergelijke tabellen bestaan echter niet voor het geval, dat gelijke waarnemingen optreden.

Opmerking. Men kan gemakkelijk bewijzen, dat de variantie van U door het optreden van gelijke waarnemingen vermindert. Het verschil, dat door deze gelijken optreedt, is echter in het algemeen gering. Men kan daarom in eerste instantie deze correctie op σ^2 verwaarlozen. De overschrijdingskansen, die men dan vindt, zijn iets te groot.

Litteratuur:

1. F.Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945), p.80-83.

2 H.B.Mann and D.R.Whitney On a test of whether one of two random variables is stochastically larger than the other, Amer.Math.Stat. 18 (1947),p. 50-60.

3 H.R.van der Vaart Some remarks on the power function of Wilcoxon's test for the problem of two samples, Proceedings van de Kon. Ned.Ak.v.Wet., 53 (1950),p. 494-520.

4 H.R.van der Vaart Gebruiksaanwijzing voor de toets van Wilcoxon, met tabellen voor n en m≤10, Rapport S32 (M4) (1950).

5 H.R.van der Vaart De toets van Wilcoxon voor het probleem van twee steekproeven. (Cursus "Parametervrije Methoden", 1951-'52).

6 D.van Dantzig Kadercursus Mathematische Statistiek, Math. Centrum, Amsterdam (1947-'50), hoofdst. 6, § 3.

7 J.Hemelrijk Note on Wilcoxon's two sample test, when ties are present, Ann.Math.Stat. 23 (1952) no. 2.

MATHEMATISCH CENTRUM
2de Boerhaavestr. 49,
A m s t e r d a m -O.
Statistische Afdeling
S47(M13)

Rangcorrelatie[1)]

1. Beschrijving van de methode.

De door M.G. Kendall ontwikkelde methode der rangcorrelatie is toepasbaar op de volgende situatie:

De stochastische grootheden x en y bezitten een simultane verdeling. Over deze verdeling zelf behoeft niets ondersteld te worden.

(x_i, y_i) $(i = 1,\ldots,n)$, zijn onafhankelijke waarnemingsparen van deze stochastische grootheden

Voorbeeld:

i =	1	2	3	4	5	6
x_i	0,11	0,12	0,10	0,11	0,15	0,13
y_i	3,4	3,0	3,2	3,5	3,5	3,5

Wij zeggen dat de waarnemingsparen (x_i, y_i) en (x_j, y_j) positief gecorreleerd zijn, als de volgorde van x_i en x_j hetzelfde is als die van y_i en y_j (bv. $x_i < x_j$ en $y_i < y_j$); zij zijn negatief gecorreleerd als de volgorde van x_i en x_j tegengesteld is aan de volgorde van y_i en y_j (bv. $x_i > x_j$ en $y_i < y_j$) en zij zijn niet gecorreleerd als $x_i = x_j$ of $y_i = y_j$.

In tabel 1 hebben wij van alle tweetallen (x_i, y_i) en (x_j, y_j) uit ons voorbeeld nagegaan of zij positief, negatief dan wel niet gecorreleerd zijn. Een positieve correlatie is aangeduid met +1, een negatieve met -1, terwijl het ontbreken van correlatie wordt aangegeven door een 0.

De toetsingsgrootheid van de methode van rangcorrelatie is nu het aantal positief gecorreleerde tweetallen verminderd met het aantal negatief gecorreleerde, of wel de som van de getallen, die in tabel 1 in de kolom "correlatie" voorkomen.

De verdeling van S voor het geval dat x en y onafhankelijk zijn is bekend (zie § 2). De hypothese dat x en y onafhankelijk

---

1) Dit memorandum is slechts bedoeld ter orientatie en streeft niet naar volledigheid of volledige exactheid.

Tabel 1

Berekening van S voor het voorbeeld

i	j	Correlatie
1	2	-1
1	3	+1
1	4	0
1	5	+1
1	6	+1
2	3	-1
2	4	-1
2	5	+1
2	6	+1
3	4	+1
3	5	+1
3	6	+1
4	5	0
4	6	0
5	6	0
		S = +5

zijn, kan dus getoetst worden.

Is de hypothese van onafhankelijkheid niet vervuld, dan is de waarschijnlijkheid van grote positieve of grote negatieve waarden van S groter, dan wanneer dit wel het geval is. De kritieke zône is daarom van de vorm $|S| \geqq S_0$, en bij éénzijdige toetsing van de vorm $S \geqq S'_0$ (rechtszijdige toetsing) of $S \leqq S''_0$ (linkszijdige toetsing).

2. Verdeling van S als x en y onafhankelijk zijn.

Als er noch bij de x_i noch bij de y_j gelijke waarden voorkomen kunnen wij gebruik te maken van exacte tabellen, die voorkomen in [1] pg 141 (n = 4 t/m 10) en in [2] (tables I and II, n = 4 t/m 40). Bovendien vindt men in [2] table III de kleinste waarden van S, waarvan de overschrijdingskansen onder de hypothese van onafhankelijkheid hoogstens gelijk zijn aan α voor α = 0,005; 0,01; 0,025; 0,05 en 0,10 en n = 4,5,6,...,40.

Als er bij de x_i óf bij de y_i, doch niet bij beide tweetallen of drietallen gelijken voorkomen, kan men voor $n \leqq 10$ gebruik maken van de tabel van Sillitto [4].

Voor grote waarden van n is de verdeling van $\frac{S}{\sigma_S}$ (waarin σ_S de spreiding van S is, die uit een hieronder op te geven formule berekend kan worden) bij benadering normaal met gemiddelde 0 en spreiding 1. Hiervan kunnen we gebruik maken om de hypothese van onafhankelijkheid te toetsen in de gevallen waar de exacte verdeling niet getabelleerd is. Dit geschiedt dan, door in een tabel van de normale verdeling de

overschrijdingskans op te zoeken, die behoort bij de gevonden waarde van $\frac{S}{\sigma_S}$.

Om σ_S te berekenen, nemen wij in de rij der waarnemingen x_i de gelijke waarnemingen in groepen bij elkaar. De aantallen waarnemingen in die groepen duiden wij aan met t_h, waarin $h = 1,2,\ldots,k_1$. Evenzo doet men in de rij der waarnemingen y_j, waar we de overeenkomstige aantallen aanduiden met u_l waarin $l = 1,2,\ldots,k_2$. σ_S kan dan gevonden worden uit de volgende formule:

$$(1)\quad \sigma^2_S = \frac{1}{18}\Big\{ n(n-1)(2n+5) - \sum_{h=1}^{k_1} t_h(t_h-1)(2t_h+5) - \sum_{l=1}^{k_2} u_l(u_l-1)(2u_l+5)\Big\} + \frac{1}{9n(n-1)(n-2)}\sum_{h=1}^{k_1} t_h(t_h-1)(t_h-2)\sum_{l=1}^{k_2} u_l(u_l-1)(u_l-2) + \frac{1}{2n(n-1)}\sum_{h=1}^{k_1} t_h(t_h-1)\sum_{l=1}^{k_2} u_l(u_l-1).$$

In ons voorbeeld van § 1 komt in de rij x_i één tweetal gelijken (dus $k_1=1$ en $t_1=2$) en in de rij y_j één drietal gelijken ($k_2= 1$, $u_1 = 3$) voor. Dus geldt:

$t_1(t_1- 1)(2t_1+ 5) = 2.1.9 = 18$
$u_1(u_1- 1)(2u_1+ 5) = 3.2.11 = 66$
$t_1(t_1- 1)(t_1- 2) = 0$, $u_1(u_1- 1)(u_1- 2) = 0$
$t_1(t_1- 1) = 2.1 = 2$
$u_1(u_1- 1) = 3.2 = 6$
$n(n-1)(2n+5) = 6.5.17 = 510$
$n(n-1) = 6.5 = 30$

zodat:

$$\sigma^2_S = \frac{1}{18}\{ 510-18-66\} + \frac{1}{60} \times 2 \times 6 = 23,87$$

en $\sigma_S = 4,89$ is.

Als alle t_h en alle u_l gelijk zijn aan 1 en er dus in geen van beide rijen gelijken voorkomen, gaat formule (2) over in:

$$(2)\quad \sigma_S = \sqrt{\frac{1}{18} n(n-1)(2n+5)}$$

Een tabel van deze functie voor $n = 40,41,\ldots,100$ vindt men in [2] (table IV).

3. Rangcorrelatiecoëfficiënt

Als maat voor de correlatie in de rij van waarnemingsparen $(x_1,y_1),\ldots,(x_n,y_n)$ heeft Kendall de coëfficiënt τ gedefinieerd, die + 1 is als de volgorden der waarnemingen in beide rijen $x_1,\ldots,x_n$ en $y_1,\ldots,y_n$ volledig overeenstemmen en -1 is, als deze volgorden volkomen tegengesteld zijn. De definitie van τ is:

$$(3)\quad \tau = \frac{2S}{\left\{n(n-1)-\sum_{h=1}^{k_1} t_h(t_h-1)\right\}^{\frac{1}{2}}\left\{n(n-1)-\sum_{l=1}^{k_2} u_l(u_l-1)\right\}^{\frac{1}{2}}}.$$

Als er in geen van beide rijen gelijke waarnemingen voorkomen wordt deze formule:

$$(4)\quad \tau = \frac{2S}{n(n-1)}.$$

Literatuur:

[1] M.G. Kendall. Rank correlation Methods London 1948, Hoofdstuk 1.

[2] L. Kaarsemaker en A. van Wijngaarden. Tables for use in rank correlation . (1952) Report R 73 of the Computation Department of the Mathematical Centre.

[3] J. Hemelrijk. Kendall's rangcorrelatie-coëfficiënt . Hoofdstuk I der cursus "Parametervrije Methoden" Rapport S 59 (1951) Mathematisch Centrum, blz. 1-17.

[4] G.P. Sillitto. "The Distribution of Kendall's coefficient of rankcorrelation in rankings containing ties. Biometrica 34 (1947) p. 36-40.

MATHEMATISCH CENTRUM,
2de Boerhaavestr. 49,
A m s t e r d a m -O.

Statistische Afdeling
S 53 (M 22)

Tekentoets[1)]

Deze toets dient voor het toetsen van de hypothese H_o, dat een aantal grootheden $\underline{z}_1,\ldots,\underline{z}_n$ alle nul tot mediaan hebben, d.w.z. dat

$$P\left[\underline{z}_i > 0\right] = P\left[\underline{z}_i < 0\right] \qquad i = 1,\ldots,n,$$

is. De toets geldt zonder enige verdere beperking dan de eis, dat de grootheden $\underline{z}_i$ onderling onafhankelijk verdeeld zijn; zij behoeven niet dezelfde waarschijnlijkheidsverdeling te bezitten.

De toets berust op één waarneming van ieder der grootheden $\underline{z}_i$, dus op n waarnemingen $z_1,\ldots,z_n$. De waarnemingen, die de waarde 0 bezitten, laten wij buiten beschouwing[2)]. Als toetsingsgrootheid gebruiken wij nu $\underline{n}_1$, het aantal positieve waarnemingen. Zijn er m waarnemingen $\neq 0$, dan bezit $\underline{n}_1$ een binomiale verdeling, onderstellende, dat H_o juist is:

$$P\left[\underline{n}_1 = n_1 \mid H_o\right] = \binom{m}{n_1} 2^{-m}.$$

Als kritieke zône worden de grote en kleine waarden van n_1 genomen. De kritieke zône is, voor onbetrouwbaarheidsdrempels 0,01; 0,05; 0,10 en 0,25 en m =1 tot 100 getabelleerd door

W.J.Dixon and A.M.Mood, The statistical sign test, Jrn.Am.Stat.Ass. 41 (1946) p. 556-566.

Voor een groter aantal waarnemingen gebruikt men als benadering van de binomiale verdeling de aangepaste normale verdeling.

Opmerking: De toets wordt vaak gebruikt, indien men een aantal grootheden twee maal heeft waargenomen, voor en na een bepaalde gebeurtenis, om na te gaan of deze gebeurtenis invloed op de grootheden heeft uitgeoefend. Noemen wij de waarnemingen vóór het optreden der gebeurtenis $\underline{x}_i$ (i=1,...,n) en erna $\underline{y}_i$, dan hebben de grootheden $\underline{x}_i-\underline{y}_i$ alle o als mediaan, indien $\underline{x}_i$ dezelfde verdeling bezit als $\underline{y}_i$ (dus als de gebeurtenis geen invloed heeft gehad). De toets wordt nu toegepast op $\underline{z}_i=\underline{x}_i-\underline{y}_i$ (i=1,...,n).

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

2) In tegenstelling tot de gewoonte deze waarnemingen voor de helft bij de positieve en voor de helft bij de negatieve te tellen; de door ons gebruikte methode geeft de toets een groter onderscheidingsvermogen.