Projectnr.: 505.0060
Normalisatie monsterneming en analyse
Projectleider: dr W. G. de Ruig

Rapport 90.55 December 1990

Statistiek in de Chemometrie
Deel 1. Theoretisch gedeelte

dr W. G. de Ruig, drs P.H.U. de Vries,
ir A.A.M. Jansen en drs J.H. Oude Voshaar

Afdeling: Coördinatie Chemometrie

Goedgekeurd door dr F.A. Huf

Rijks-Kwaliteitsinstituut voor land- en tuinbouwprodukten (RIKILT)
Bornsesteeg 45, 6708 PD Wageningen
Postbus 230, 6700 AE Wageningen
Telefoon 08370-75400
Telex 75180 RIKIL
Telefax 08370-17717

VERZENDLIJST

INTERN:

directeur
sectorhoofden
coördinatie Chemometrie (5x)
coördinatie Kwaliteit en Veiligheid
afdeling Sensoriek (2x)
deelnemers aan de cursus (A. van Polanen, M.A.H. Tusveld, J.J.M. Driessen, J.H. Slangen, D.P. Venema, M.J.B. Mengelers, W. Haasnoot, H.J. Keukens, G.M. Binnendijk, Th. C. Wolters, P.J. Herben, H.J. Horstman)
programmabeheer en informatieverzorging (2x)
bibliotheek
circulatie

EXTERN:
Dienst Landbouwkundig Onderzoek
Directie Wetenschap en Techniek
Directie Voedings- en Kwaliteitsaangelegenheden
Groep Landbouwwiskunde (ir A.A.M. Jansen, drs J.H. Oude Voshaar)

INHOUD

SAMENVATTING

In april - juni 1990 werd op het RIKILT een interne cursus over de toepassing van statistiek in de chemometrie gegeven. De cursus bestond uit een theoretisch gedeelte en de behandeliong van praktische toepassingen met behulp van SPSS. Cursusleider voor het theoretische deel was dr W.G. de Ruig, voor het praktische deel dr ir A.B. Cramwinckel.
Bij het opzetten en geven van de cursus werd veel medewerking ondervonden van ir A.A.M. Jansen en drs J.A. Oude Voshaar van de Groep Landbouwwiskunde.

Dit rapport omvat het theoretische gedeelte van de cursus.
Hierbij zijn een aantal basisbegrippen uit de statistiek, die in de chemometrie van pas komen, behandeld. Aan de orde zijn gekomen: populaties, steekproeven, eigenschappen en toepassing van de normale verdeling, steekproefgemiddelde, standaardafwijking, betrouwbaarheidsinterval, toetsen van hypothesen, ringonderzoek, proefopzetten, variantieanalyse, lineaire-regressieanalyse, calibratie, multipele lineaire regressie en met deze onderwerpen gerelateerde begrippen.

Hoofdstuk 14 'Interlaboratoriumonderzoek in de praktijk' werd verzorgd door drs P.H.U. de Vries, hoofdstuk 18 'Lineaire-regressieanalyse' en hoofdstuk 20 'Multipele regressie' door drs J.A. Oude Voshaar en hoofdstuk 19 'Calibratie' door ir A.A.M. Jansen.

April 1990

STATISTIEK IN DE CHEMOMETRIE

1 INLEIDING

Een aanzienlijk deel van de werkzaamheden die op het RIKILT uitgevoerd worden betreft analytisch onderzoek van binnenkomende monsters.
Er wordt een analyse uitgevoerd op het monster met als doel inzicht te verkrijgen in een bepaalde eigenschap van dat monster. Met name betreft het een toetsing, of het monster al dan niet aan gestelde eisen voldoet.
Zo'n monster kan afkomstig zijn van een partij. Voor het trekken van juiste conclusies is de wijze van monsterneming van groot belang.
Een andere RIKILT-activiteit is, dat men kennis wil vergaren over een bepaald onderwerp en daartoe een onderzoek opzet, bijvoorbeeld inventariserend onderzoek over een contaminant of een residu in Nederland, maar ook onderzoek naar een nieuwe of verbeterde analysemethode.

Voor het uitvoeren van deze activiteiten kan de statistiek een zinvolle bijdrage leveren.
Statistiek is een onderdeel van de wiskunde, die in allerlei disciplines van nut is. Op het RIKILT interesseert ons de toepassingen op het terrein van de analytische chemie. Gemakshalve wordt hier onder deze term ook verstaan niet-chemische technieken, zoals microbiologie, microscopie en toxicologie. Het onderdeel van de analytische chemie dat kennis verzamelt uit chemische metingen heet chemometrie. De chemometrie gebruikt hiervoor statistische en andere mathematische technieken.
Statistiek wordt veelvuldig toegepast in landbouwkundig onderzoek, met name bij de opzet van experimenten of steekproeven en bij het trekken van conclusies uit de verkregen waarnemingen. Deze conclusies hebben meestal betrekking op modellen waarmee de invloed van variabelen of factoren op responsvariabelen wordt beschreven. Een complicatie daarbij is de variabiliteit van de respons. Vooral als men werkt met levend materiaal, kan deze variatie aanzienlijk zijn. Verder moet men behalve met de te onderzoeken behandelingsfactoren ook rekening houden met andere variatiebronnen. Indien deze vooraf aanwijsbaar zijn, kan men daarmee al tijdens de opzet van het onderzoek rekening houden en aldus voorkomen dat ze verstrengeld raken met de te onderzoeken factoren. Kortom: in de

onderzoeksmethodologie neemt statistiek een belangrijke plaats in, waarbij de proefopzet zeker zo belangrijk is als de statistische analyse van de uiteindelijk verkregen gegevens.
Hoewel deze strategieën vooral gericht zijn op de onderzoekmethoden van DLO instituten die zelf een onderzoek opzetten, en minder afhankelijk zijn van ingezonden monsters, kunnen ze ook voor het RIKILT van nut zijn bij het evalueren van nieuwe bepalingsmethoden en bij het sensorisch en toxicologisch onderzoek. Voor wat betreft het keuringsonderzoek kan de chemometrie een bijdrage leveren door het toekennen van een waardeoordeel aan het analyseresultaat.
Bij het onderzoek, anders dan keuringen, wordt vaak een veelheid van waarnemingsresultaten verkregen, die, in ongeordende vorm weinig inzicht geven in het onderwerp van studie. Het ordenen en reduceren (bewerken en samenvatten) van waarnemingsresultaten is het terrein van de *beschrijvende statistiek*.
Ordenen geschiedt door het maken van frequentietabellen, histogrammen en grafieken. Reductie omvat het weergeven van veel waarnemingen in enkele kengetallen voor de meest karakteristieke aspecten van de waarnemingsresultaten: een centrummaat voor het niveau en een maat voor de spreiding of variabiliteit.
De *mathematische statistiek* doet op grond van een beperkt aantal waarnemingen gegeneraliseerde uitspraken. Zij maakt daartoe gebruik van de *waarschijnlijkheids-* of *kansrekening*, dat is het onderdeel van de wiskunde dat de wetten van het toeval bestudeert. Uitspraken van de mathematische statistiek kunnen zijn:
schattingen van een bepaald kenmerk, aan de hand van de beschikbare waarnemingsresultaten;
toetsingen van bepaalde veronderstellingen, aan de hand van de beschikbare waarnemingsresultaten.
Hiervan wordt gebruik gemaakt bij de keuringen.

Aangezien het RIKILT veelal gebonden is aan officiële keuringseisen wordt in deze cursus veel aandacht besteed aan NEN- en ISO-normen en regels voorgeschreven of voorgesteld door IUPAC, EG of ORA.

2

POPULATIES

Statistiek houdt zich bezig met uitspraken over populaties. Meestal baseert men de uitspraken op een steekproef uit de populatie (zie Hoofdstuk 4). Omdat de uitspraken echter eigenschappen van de populatie betreffen, zullen we eerst bespreken hoe die populatie-eigenschappen gedefinieerd zijn. Hiertoe zullen we in dit hoofdstuk doen alsof de gehele populatie bekend is (hetgeen in de praktijk meestal niet het geval is).

Populaties

Een populatie is de verzameling elementen waarover men een uitspraak wil doen. Bijvoorbeeld alle in Nederland aangevoerde slachtdieren in 1989, alle kazen van een bepaalde produktiepartij. De omschrijving van een populatie moet zodanig zijn dat duidelijk is wat de elementen zijn (hier slachtdieren en kazen)en welke elementen wel, respectievelijk niet tot de populatie behoren.

Eigenschappen, variabelen

Meestal is men geinteresseerd in eigenschappen van elementen van populaties. Een voorbeeld van zo'n eigenschap is het aantal liter melk dat een koe in haar 1^e lactatie-periode geeft, het vetgehalte van kaas of de aanwezigheid van diergeneesmiddelen in vlees. Omdat zo'n eigenschap meestal van element tot element varieert wordt vaak de term variabele gebruikt in plaats van eigenschap of kenmerk.

Variabelen kan men onderscheiden in verschillende typen. Allereerst onderscheiden we kwalitatieve en kwantitatieve variabelen. Kwantitatieve variabelen (engels: variates) zijn eigenschappen die in een getal zijn uit te drukken. Bijvoorbeeld het vetgehalte in een kaas. Kwalitatieve variabelen zijn eigenschappen die niet zinvol in een getal zijn uit te drukken, maar ze geven een groepsindeling weer. (Voorbeelden: het ras van een koe, geslacht van een mens, mate van aandoening uitgedrukt in bv. 0, +, ++, +++). De waarden die een kwalitatieve variabele kan aannemen worden niveaus genoemd (Engels: levels).
Kwalitatieve variabelen kan men onderverdelen in twee typen: ordinaal en nominaal. Als de niveaus geordend kunnen worden dan spreken we van een

ordinale variabele (bijvoorbeeld goed, matig, slecht; laag, middel, hoog), anders van een nominale variabele (bijvoorbeeld ras; man, vrouw)

Kwantitatieve variabelen kunnen nog onderscheiden worden in continue en discrete variabelen. Bij een continue variabele zijn voor elk tweetal mogelijke waarden ook alle tussenliggende getallen mogelijk (bijvoorbeeld ook alle gebroken getallen). Voorbeeld: lengte, gewicht, etc. Discrete variabelen kunnen slechts bepaalde afzonderlijke waarden aannemen. Dit is onder meer het geval als slechts gehele getallen mogelijk zijn, bijvoorbeeld: het aantal bladeren van een tomaatplant.

Frequentieverdelingen

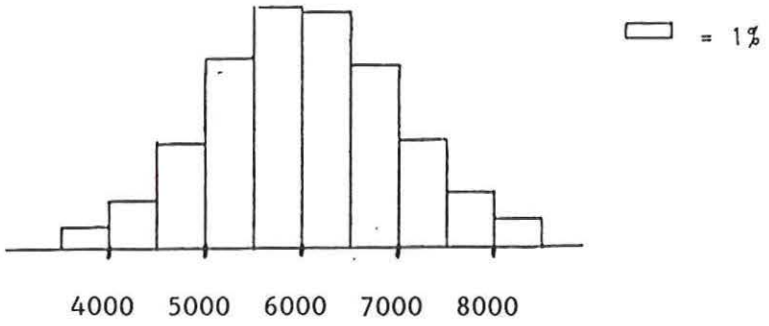
Om te beschrijven hoe een eigenschap varieert in de populatie gebruiken we een frequentieverdeling. Deze verdeling geeft voor elke mogelijke waarde van de eigenschap aan hoe groot de fractie in de populatie is met deze waarde. We zullen dit eerst uitwerken voor een continue variabele.

In het voorbeeld van de melkgift van koeien tijdens de 1^e lactatie-periode kunnen we de verdeling weergeven door een tabel of een histogram. Hiertoe delen we de waarden van de melkgift in in klassen (bijv. <4000, 4000-4500, 4500-5000, etc.). Vervolgens berekenen we voor welke fractie van de dieren de melkgift in die klassen valt en geven die fracties weer zoals in tabel 2.1. Een snelle visuele indruk krijgt men door een histogram te maken als in figuur 2.1.

Tabel 1: Melkgift tijdens de 1^e lactatie van alle koeien van het MRY-ras in Nederland tussen 1980 en 1990 geboren (gefingeerde waarnemingen).

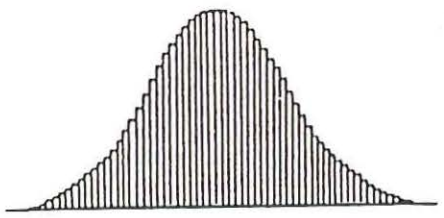
melkgift	<4000	4000-4500	4500-5000	4000-5500	4500-6000	6000-6500	6500-7000	7000-7500	7500-8000	>8000
aantal (x 1000)	84	168	368	608	772	760	596	364	176	104
fractie (in %)	2.1	4.2	9.2	15.2	19.3	19.0	14.9	9.1	4.4	2.6

Figuur 2.1: Histogram van melkgiftgegevens uit tabel 1.

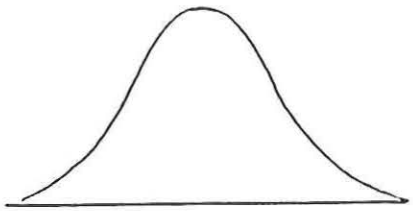


In het histogram zien we dat veel melkgiften in de buurt van 6000 liggen en dat bijna alle waarden (ongeveer 95%) liggen tussen 4000 en 8000.
Als men de breedte van de klassen kleiner maakt, bijvoorbeeld 50, dan verkrijgt men een grafiek als in figuur 2.2. Als men de klassebreedte nog kleiner laat worden en tot nul laat naderen, dan krijgen we bij een grote populatie een grafiek die niet meer afhangt van de klasse-indeling (figuur 2.3).

Figuur 2.2:
Histogram met kleinere klasseindeling

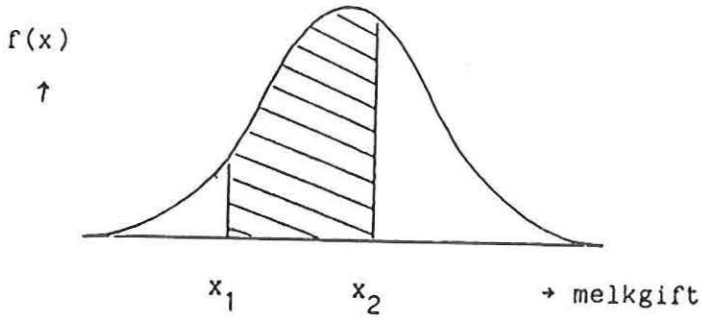


Figuur 2.3:
Relatieve frequentieverdeling



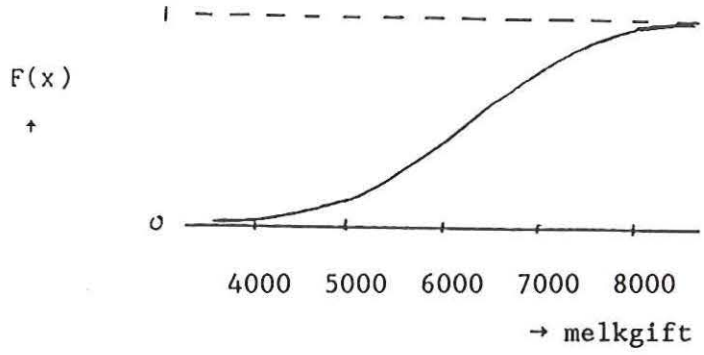
De functie die in fig. 2.3 is weergegeven noemen we de relatieve frequentieverdeling van de melkgift. Als we die functie noteren als f, dan geldt voor die functie (zie figuur 2.4): voor elk tweetal getallen x_1 en x_2 is de fractie van de populatie met melkgift tussen x_1 en x_2 gelijk aan de oppervlakte van het gebied tussen x_1 en x_2 dat onder de grafiek van f ligt.

Figuur 2.4: Fractie dieren met melkgift tussen x_1 en x_2 is gelijk aan oppervlak van gearceerd gebied.



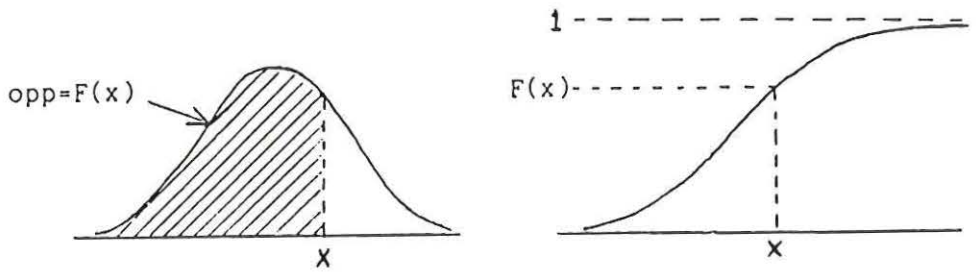
Een andere manier om de verdeling van de melkgiften te beschrijven is de cumulatieve frequentieverdeling. Dit is een functie F, die voor elk getal x aangeeft welke fractie van de koeien een melkgift heeft van ten hoogste x. F is dus een niet-dalende functie.

Figuur 2.5: Cumulatieve frequentieverdeling.



Het verband tussen de relatieve en cumulatieve frequentieverdeling volgt al uit hun definitie en is weergegeven in fig. 2.6. In wiskundige terminologie: F is de integraal van f (ofwel: $F(x) = \int_{-\infty}^{x} f(s)ds$).

Figuur 2.6: Relatie tussen relatieve en cumulatieve frequentieverdeling.



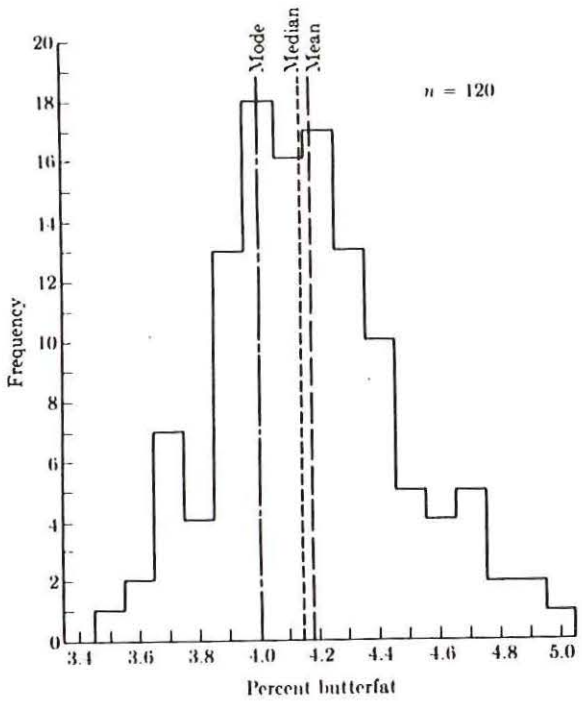
Voor niet-continue variabele kan men de frequentieverdeling in een tabel weergeven. Visualisering is mogelijk via een staafdiagram of eventueel een cirkeldiagram.

Populatiegemiddelde en -variantie

Bij kwantitatieve variabelen is het vaak niet nodig de hele frequentieverdeling op te geven, maar kunnen we volstaan met een beknopte beschrijving d.m.v. een paar kentallen. Een verdeling is al in belangrijke mate gekenschetst door het "centrum" en de "spreiding".

De meest gebruikte maat voor het centrum van een verdeling is het (populatie-)gemiddelde. Dit gemiddelde noteren we vaak als μ (spreek uit: mu) en is gewoon het gemiddelde alle waarden van de eigenschap in de populatie. In formule: $\mu = \Sigma x_i/N$, waarbij N het aantal elementen van de populatie is en x_i de waarde van element i, voor continue verdelingen geldt: $\mu = x.f(x)dx$. Een andere maat voor het centrum is de mediaan. Dit is het getal waarvoor geldt dat de helft van de populatie een grotere waarde heeft dan dit getal en de helft een kleinere. Ofwel: de mediaan is de middelste van alle waarden x_i. Deze maat kan zinvoller zijn als centrummaat dan het gemiddelde als de verdeling erg scheef is, omdat het gemiddelde veel gevoeliger is voor sterk afwijkende waarden. Ten slotte kent men de modus, dit is de waarde die het meeste voorkomt. Zie figuur 2.7.
Bij symmetrische verdelingen zijn het gemiddelde, de mediaan en de modus gelijk.

Figuur 2.7. Modus, mediaan en gemiddelde voor een scheve verdeling



Als maat voor de spreiding ligt het voor de hand zoiets te gebruiken als de "gemiddelde afwijking t.o.v. het gemiddelde". Om te zorgen dat positieve en negatieve afwijkingen elkaar niet compenseren zou men kunnen kijken naar het gemiddelde van de absolute afwijkingen t.o.v. het gemiddelde μ (de absolute afwijkingen $|x_i-\mu|$ zijn de afwijkingen met weglating van de min-tekens). Omdat dit echter mathematisch onaantrekkelijk is hanteert men in de praktijk altijd het gemiddelde van het kwadraat van de afwijkingen t.o.v. μ. Dit heet de (populatie-)<u>variantie</u> en noteren we als σ^2 . In formulevorm:

$$\sigma^2 = \Sigma_i(x_i - \mu)^2 / N$$

(of $\sigma^2 = \int (x-\mu)^2 . f(x)dx$ voor continue verdelingen).

De variantie wordt in tegenstelling met het gemiddelde in andere eenheden uitgedrukt dan de waarnemingen, nl. het kwadraat van de eenheid. Bijvoorbeeld de variantie van melkgift wordt uitgedrukt in liter^2. Om een spreidingsmaat te krijgen in dezelfde eenheid als de waarnemingen trekken we de wortel uit de variantie. Dit noemen we de (populatie)-<u>standaardafwijking</u> en noteren we als σ (Engels: standard deviation, sd).

De standaardafwijking is de gangbare spreidingsmaat. Soms echter wordt relatieve spreidingsmaat gebruikt, nl. de variatiecoefficient, vc, of relatieve standaardafwijking, rsa (coefficient of variation, CV, resp. relative standard deviation, RSD). Deze is gedefinieerd als de standaardafwijking gedeeld door het gemiddelde (vc = σ/μ), maar wordt ook vaak in procenten uitgedrukt. Voorbeeld:

$\mu = 50$ vc = 4/50 = 0.08
$\sigma = 4$ = 8 %

De variatiecoefficient is een bruikbare spreidingsmaat voor variabelen die geen negatieve uitkomsten kunnen hebben en waarvoor in verschillende situaties de standaardafwijking evenredig is met het gemiddelde (bijvoorbeeld bij opbrengsten van gewassen, gehalten van chemische stoffen of de nauwkeurigheid van een meting).

Opmerking: frequentieverdelingen van kwalitatieve verdelingen kunnen niet worden samengevat via gemiddele en variantie.

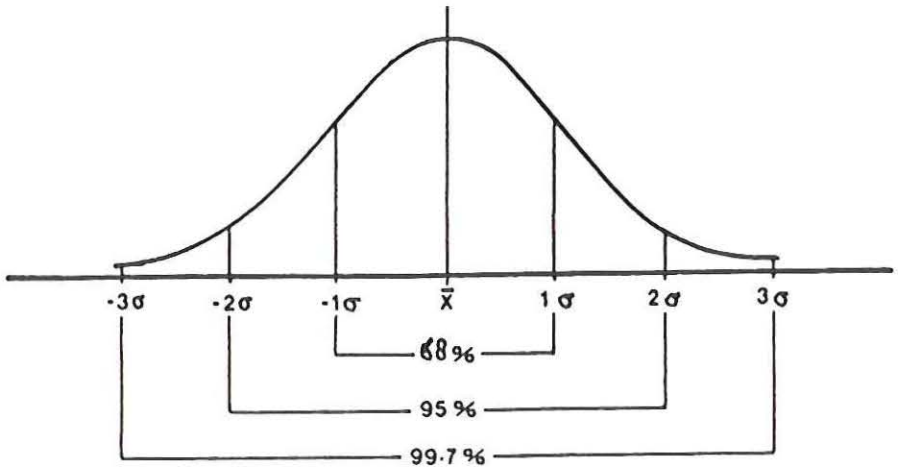
3
NORMALE VERDELING

In de praktijk blijkt dat veel continue variabelen een verdeling hebben van eenzelfde klokvormig type, de zogenaamde *normale verdeling*, zie figuur 3.1. Deze verdeling neemt daarom een belangrijke plaats in in de statistiek. Een variabele heet normaal verdeeld als voor de relatieve frequentieverdeling f(x) geldt

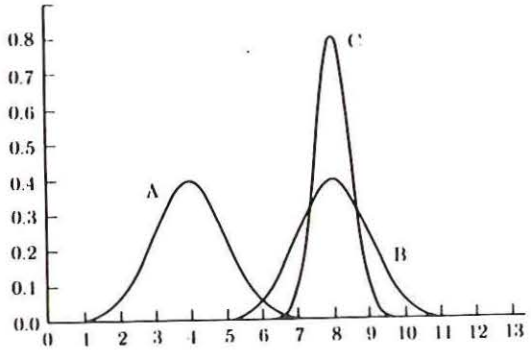
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

waarin μ en σ de *parameters* van de frequentieverdeling zijn.
Uit de formule is in te zien, dat de curve symmetrisch is en dat er een maximum is bij $x = \mu$. Dit is tevens het gemiddelde van de verdeling. Voorts is af te leiden, dat de parameter σ de standaardafwijking van de verdeling is. Daarom zijn voor deze parameters meteen de symbolen μ en σ voor het populatiegemiddelde en de standaardafwijking gebruikt.
De normale verdeling met gemiddelde μ en standaardafwijking σ noteren we afgekort als: $N(\mu,\sigma^2)$

Zoals uit de formule te zien valt wordt een normaal verdeelde variabele geheel gekarakteriseerd door zijn gemiddelde en standaardafwijking. Voor een normaal verdeelde populatie geldt dat (ongeveer) 68% van de populatie ligt tussen $\mu-\sigma$ en $\mu+\sigma$. Verder geldt dat 95% ligt tussen $\mu-2\sigma$ en $\mu+2\sigma$ en tot slot ligt 99.7% tussen $\mu-3\sigma$ en $\mu+3\sigma$, zie figuur 3.2. Figuur 3.3 illustreert hoe de parameters μ en σ de plaats en de vorm van de frequentieverdeling beïnvloeden.



Figuur 3.2. Oppervlakken onder de normale verdelingsfunctie



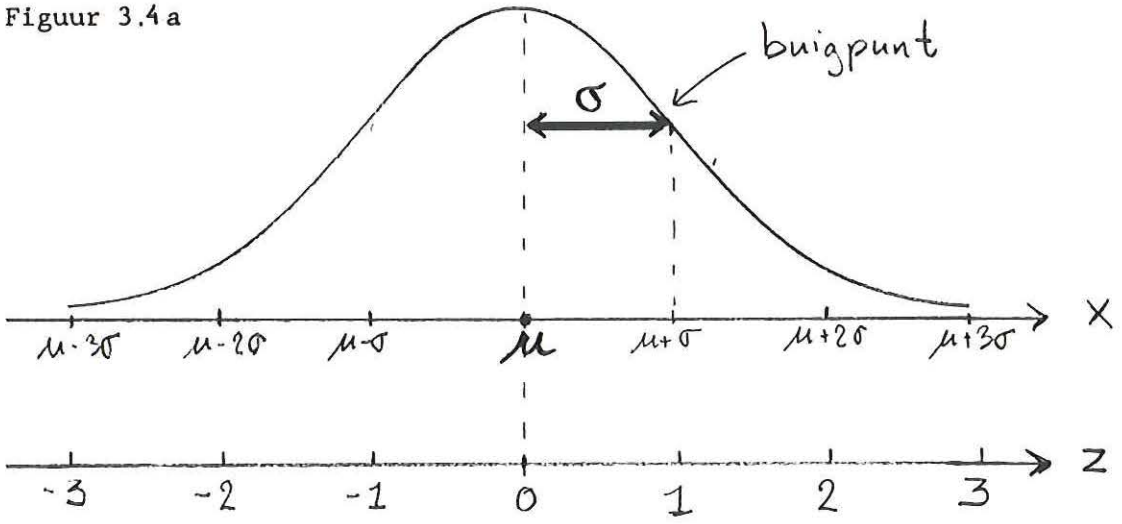
Figuur 3.3. Normale verdelingen. A: $\mu = 4$, $\sigma = 1$; B: $\mu = 8$, $\sigma = 1$; C: $\mu = 8$, $\sigma = 0,5$.

De normale verdeling met gemiddelde 0 en standaardafwijking 1 noemen we de standaardnormale verdeling. Hiervoor zijn tabellen beschikbaar van de cumulatieve verdelingsfunctie (zie tabel II, appendix). Met deze tabel kan men voor elke normale verdeling de populatiefracties berekenen, door die verdeling eerst te standaardiseren.
Dit houdt in dat we de x-variabele van de normale verdeling vervangen door een nieuwe variabele, z, waarvoor geldt

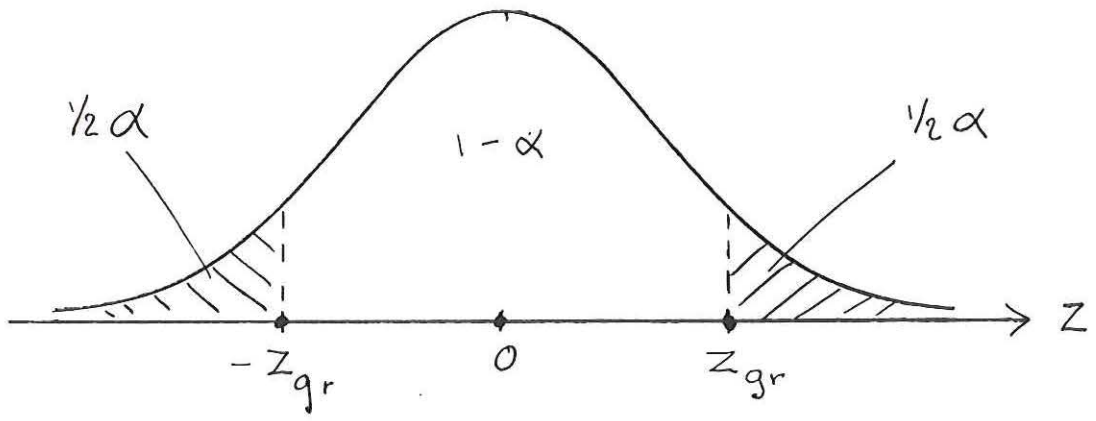
$$z = \frac{x - \mu}{\sigma}$$

Zie figuur 3.4 a. De variabele z is dimensieloos, met verwachting 0 en met de standaardafwijking als eenheid.

Figuur 3.4 a



Figuur 3.4 b



Voor de standaardnormale verdeling zijn de overschrijdingskansen berekend. Men onderscheidt, zie figuur 3.4b, waarin z_{gr} de grenswaarde is:
rechter overschrijdingskans

$P(z > z_{gr})$: het rechter gearceerde gedeelte

linker overschrijdingskans

$P(z < z_{gr})$: het linker gearceerde gedeelte

tweezijdige overschrijdingskans

$\alpha = P(z<-z_{gr}, z>z_{gr})$: beide gearceerde delen

In Tabel 3.1 zijn voor enkele veel gebruikte z_{gr} waarden de kans op niet-overschrijding en overschrijdingskans, α, gegeven. De gegeven. (De rechter- resp. linker-overschrijdingskans zijn ½ α.)

Tabel 3.1 Z-tabel

z_{gr}	1,000	1,650	1,960	2,000	2,330	2,580	3,290
$P(-z_{gr} < z < z_{gr})$	0,683	0,900	0,950	0,954	0,980	0,990	0,999
$\alpha = P\ (z < -z_{gr}, z > z_{gr})$	0,317	0,100	0,050	0,046	0,020	0,010	0,001

Een uitgebreide tabel met rechter overschrijdingskansen wordt gegeven in Tabel II.

Voorbeeld: Stel dat de melkgift tijdens de 1^e lactatie-periode in een populatie van MRY-koeien normaal verdeeld is met gemiddelde 6000 en sa = 1000 en men wil weten welke fractie koeien minder dan 5000 liter per jaar geeft. Dit berekent men als volgt: als we de melkgift als x noteren dan is z = (x-6000)/1000 standaardnormaal verdeeld. De fractie koeien met x kleiner dan 5000 is gelijk aan de fractie koeien met z= (x-6000)/1000 kleiner dan (5000-6000)/10000 = -1. Deze fractie is 16%.

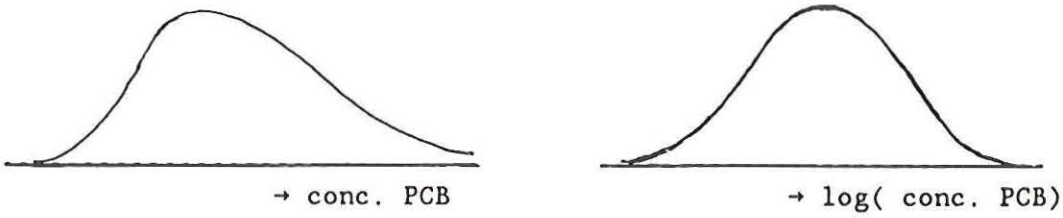
Ander voorbeeld: Uit een groot aantal wegingen van een maatkolf met inhoud hebben we kunnen vaststellen, dat de verdeling van de wegingen normaal verdeeld is met een gemiddelde waarde van μ = 357,525 g en een standaardafwijking σ = 0,015 g. Met behulp van de Z-tabel vinden we nu, dat de kans dat een enkelvoudige weging een uitkomst oplevert tussen 357,495 en 357.555 (dus gekozen z_{gr} = 2) gelijk is aan 95,4 % en dat de kans 4,6 % is dat hij er buiten ligt. Deze overschrijdingskans van ongever 5 % = 2 σ

wordt veel gebruikt.

Lognormale verdeling

Een flink deel van deze cursus zullen we besteden aan normaal verdeelde variabelen. Hoewel niet alle kwantitatieve variabelen normaal verdeeld zijn levert in veel gevallen de normale verdeling een goede benadering. Soms echter kan men een betere beschrijving vinden m.b.v. een verdelingsfunctie van een andere vorm. Een voorbeeld:

Concentraties van chemische stoffen zijn vaak lognormaal verdeeld, bijvoorbeeld het gehalte van PCB in de zeehonden in de Waddenzee. Een lognormale verdeling is een scheve verdeling waarbij afwijkingen naar boven relatief vaker voorkomen dan afwijkingen van dezelfde grootte naar beneden. Een handige uitweg hierbij is de logaritme van de concentratie te beschouwen. Deze variabele is normaal verdeeld.



De omrekening van gemiddelde en sa verloopt als volgt: als een variabele x lognormaal verdeeld is met mediaan m en variatiecoefficient v, dan is log(x) normaal verdeeld met gemiddelde ln(m) en de sa is (ongeveer) gelijk aan v. (met log wordt hier de natuurlijke logaritme bedoeld).

4

STEEKPROEVEN

In het voorgaande hebben we besproken hoe men de populatie kan beschrijven als men deze volledig kent. In de praktijk is dit bijna nooit het geval maar onderzoekt men een gedeelte van die populatie (een zgn. steekproef) en wil men op basis hiervan conclusies trekken over de hele populatie, bijvoorbeeld over het gemiddelde ervan. Deze extrapolatie stelt eisen aan de wijze waarop de steekproef uit die populatie is getrokken. Verder kunnen die uitspraken niet absoluut gesteld worden, maar bezitten deze slechts een bepaalde betrouwbaarheid. Deze betrouwbaarheid willen we kwantificeren.

Voorbeeld: We willen van een bosperceel met douglassparren de gemiddelde hoogte van de bomen vaststellen. Hiertoe kiezen we een tiental bomen waarvan we de hoogte meten. Dit levert de uitkomsten 25.3, 23.8, 28.1, 22.5, 27.4, 24.9, 25.5, 29.1, 26.9, 21.8. Wat kan men dan zeggen over de gemiddelde hoogte van dat bosperceel? Kan men een interval (een zgn. betrouwbaarheidsinterval) opgeven waarvan men redelijk zeker is dat het perceelgemiddelde daar binnenvalt?

Aselecte steekproef

Om de extrapolatie van steekproef naar populatie te kunnen maken geldt als eerste eis dat de steekproef "representatief" is. Dit kan men bereiken door de steekproef aselect te trekken uit de populatie. Hiermee wordt bedoeld dat de trekkingsprocedure zodanig is dat elk element van de populatie een even grote kans heeft om in de steekproef terecht te komen en dat alle elementen onafhankelijk van elkaar gekozen worden.

Bij eindige populaties kan dit bereikt worden door de elementen te nummeren en het gewenste aantal te loten. Hiervoor kan men lotingstabellen gebruiken.

Stochastische grootheid

Als we één willekeurig element (d.w.z. aselect) uit de populatie trekken en hieraan de waarde m.b.t. de te beschouwen variabele vaststellen dan noemen we de uitkomst een toevalsvariabele of stochastische grootheid. Immers het toeval bepaalt welk element getrokken wordt en daarmee wat de uitkomst

wordt. Zo'n stochastische grootheid noteren we vaak als X (vaak wordt een hoofdletter gebruikt om een toevalsgrootheid te kunnen onderscheiden van een getal x). De aselecte trekkingswijze zorgt ervoor dat X voldoet aan een eenvoudig aan te geven kansmechanisme, nl: de kansverdeling van X is gelijk aan de frequentieverdeling in de populatie.

In het voorbeeld van de populatie van koeien beschreven in Hoofdstuk 2 betekent dit dat, wanneer we aselect een koe uit die populatie trekken, de kans dat de melkgift X van deze koe ligt tussen 5000 en 7000 gelijk is aan 0.68 of in verkorte notatie: $P(5000<X<7000) = 0.68$. In de praktijk kennen we de verdeling van de populatie niet, maar via dit kansmechanisme kunnen we uit een aselecte trekking toch iets te weten komen over de populatieverdeling.

Verwachting, variantie

Ook de kansverdeling van een stochastische grootheid X kan men door parameters karakteriseren, die het "centrum" en de "spreiding" aangeven. De verwachtingswaarde van X (vaak afgekort als EX; Engels: expectation) is het gemiddelde van de waarden die X kan aannemen (in feite een gewogen gemiddelde waarbij de kansen de gewichten vormen). Verder kan men spreken van de variantie van X (afgekort var(X)) en van de standaardafwijking van X.

Bij een aselecte trekking van één element uit een populatie met gemiddelde μ en variantie σ^2 geldt voor de uitkomst X van die trekking: $EX = \mu$ en $var(X) = \sigma^2$. (Dit geldt echter niet als er sprake is van meetfouten, zie volgend hoofdstuk.)

5
STEEKPROEFGEMIDDELDEN

Frequentieverdeling van steekproefgemiddelden

Onafhankelijk van de vorm van de frequentieverdeling van de originele populatie van x-waarden, gaat de frequentieverdeling van de gemiddelden x voor herhaalde aselecte steekproeven van omvang n naar een normale verdeling als n toeneemt. Dit belangrijke resultaat uit de statistische theorie heet de *centrale limietstelling*. Zie figuur 5.1.

Deze stelling maakt duidelijk waarom de normale verdeling zo vaak gebruikt

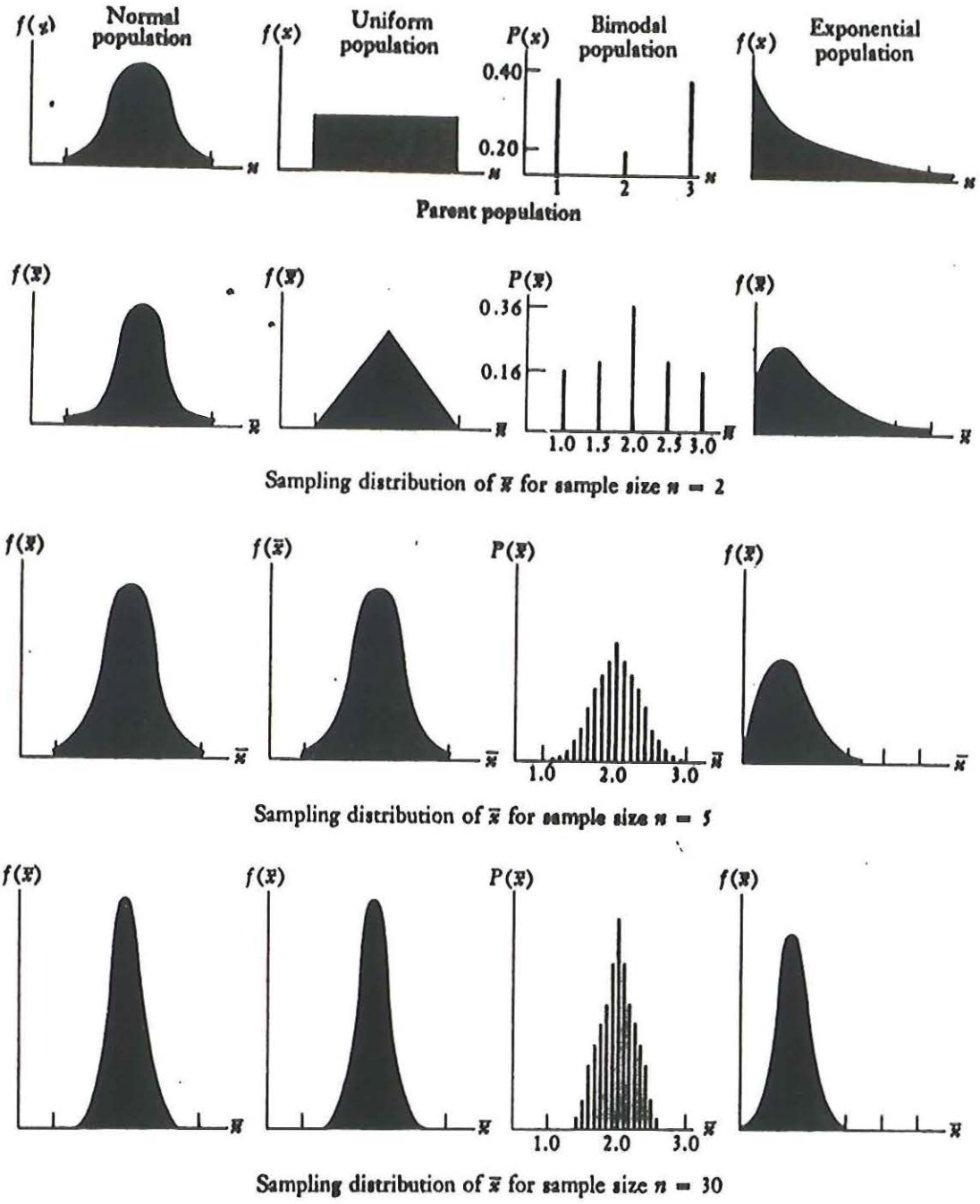


Fig. 5.1 Sampling distribution of $\bar{x}$ for various population distributions when $n = 2$, 5, and 30.

The central limit theorem:

Regardless of the distribution of the parent population (as long as it has a finite mean μ and variance σ^2), the distribution of the means of random samples will approach a normal distribution (with mean μ and variance σ^2/n) as the sample size n goes to infinity.

wordt bij steekproefgemiddelden, zelfs wanneer de oorspronkelijke verdeling niet normaal is. Afgezien van de aselecte trekking vereist de stelling weinig aannamen: het is voldoende dat σ eindig is en de steekproef een aselecte steekproef uit de populatie.

Steekproefgemiddelde en -variantie: schatters voor populatie-parameters
Meestal trekt men niet één maar meerdere elementen aselect uit de populatie. Hiervan berekent men dan het (steekproef-)gemiddelde omdat men verwacht dat dit gemiddelde minder variatie vertoont dan een afzonderlijke waarneming en dus betere informatie oplevert over populatiegemiddelde μ. Men heeft dus n trekkingen X_1, X_2, ..., X_n waarvan men het steekproef-gemiddelde $\bar{X}$ berekent:

$$\bar{X} = \Sigma X_i/n.$$

$\bar{X}$ is ook een stochastische grootheid (immers ook aan het toeval onderhevig) waarvoor men kan bewijzen: $var(\bar{X}) = \sigma^2/n$. De standaardafwijking van het steekproefgemiddelde is dus een factor $\sqrt{n}$ kleiner dan die van de afzonderlijke waarnemingen. Omdat verder geldt $E(\bar{X}) = \mu$, kunnen we $\vec{X}$ nemen als schatter voor het populatiegemiddelde μ; de precisie van die schatter is groter naarmate de steekproefomvang n groter is.

Indien de te meten variabele in de populatie normaal verdeeld is (en dus ook de waarnemingen X_i normaal verdeeld zijn) dan geldt dat X ook normaal verdeeld is. Vanwege de centrale limietstelling geldt dat ook ingeval de eigenschap niet normaal verdeeld is, het steekproef-gemiddelde bij een grote steekproefomvang n toch bij benadering normaal verdeeld is. Zelfs bij nog vrij kleine waarden van n (bv. n = 5) gaat de normale benadering van X al redelijk goed op.

Hetzelfde fenomeen verklaart overigens ook waarom veel eigenschappen bij benadering normaal verdeeld zijn: deze worden vaak door talloze factoren beinvloed en een gerealiseerde uitkomst is een gemiddelde van al die toevallige bijdragen.

Steekproefvariantie
Van een populatie willen we niet alleen het gemiddelde, maar ook de variantie schatten. Op basis van n trekkingen $X_1,\ldots,X_n$ kan men de

populatie-variantie schatten met steekproefvariantie S^2 die gedefinieerd is door:

$$S^2 = \Sigma(X_i - X)^2/(n-1).$$

Ook S^2 hangt van de waarnemingen af en zal in het algemeen niet gelijk zijn aan σ^2 (voor kleine n kan S^2 zelfs flink van σ^2 afwijken). Wel kan men bewijzen dat de schatter zuiver is (in formule: $ES^2 = \sigma^2$), dus S^2 is niet systematisch te hoog of te laag.
De noemer n-1 heet het aantal vrijheidsgraden van S^2 (Engels: degrees of freedom, df). De reden hiervan is dat S berekend wordt uit de afwijkingen van X_i ten opzichte van X. Hiervan kan men er maar n-1 vrij kiezen omdat de som van de afwijkingen nul is.

Precisie van het steekproefgemiddelde

Het opgeven van $\bar{X}$ als een schatting van het populatiegemiddelde μ wint aan betekenis als men ook de precisie van die schatting opgeeft. We zagen al dat de standaardafwijking van $\bar{X}$ gelijk is aan $\sigma/\sqrt{n}$. In de praktijk kennen we σ niet, maar we kunnen hiervoor wel een schatting S invullen (de wortel uit S^2). De (geschatte) onnauwkeurigheid van $\bar{X}$ is dus gelijk aan
$S/\sqrt{n}$. Dit wordt vaak de standaardfout van X genoemd (Engels: standard error (of the mean), afgekort: se of sem). Verwar se niet met sa; se heeft te maken met de precisie van de schatting $\bar{X}$, en sa met de spreiding van de afzonderlijke waarnemingen.

VRAAGSTUKKEN

1.1. Voor het cadmiumgehalte in 15 monsters vlees werden de volgende waarden gevonden ($\mu g/kg \times 10^9$):

4,9 4,6 5,5 9,1 16,3 12,7 6,4 7,1 2,3 3,6 18,0 3,7 7,3 4,4 9,8.

Bereken voor deze waarnemingen het gemiddelde en de standaardafwijking met het volgende rekenschema:

$$n$$

$$\Sigma Y$$

$$\bar{Y} = \Sigma Y/n$$

$$\Sigma Y^2$$

$$\Sigma d^2 = \Sigma Y^2 - (\Sigma Y)^{2/n}$$

$$s^2 = \Sigma d^2/(n-1)$$

1.2. Bereken voor de vleugels uit tabel 1.2 het gemiddelde, de variantie, de standaardafwijking en de variatiecoëfficiënt. Maak ook een histogram van de frequentieverdeling. (De gegevens staan in Data Entry bestand VM.SYS.)

1.3 Trek uit de populatie van vleugels een steekproef van één element (dus één waarde). Doe dit 5 keer, aselect. Simuleer dit laatste door eerst uit Tabel I, vijf tweecijferige getallen te kiezen (bij voorbeeld door te beginnen met regel 11 en dan naar beneden gaande: 80, 78, 05, 22, 61) en dan de elementen met deze nummers uit de tabel 2.2 te trekken.
Bereken het gemiddelde $\bar{\bar{x}}$ en de standaardafwijking s_x.

1.4 Trek nu uit de populatie 5 keer een steekproef van twee elementen. Bepaal voor ieder paar het gemiddelde $\bar{y}_1$, $\bar{y}_2$, ... $\bar{y}_5$ en voor deze gemiddelden het over-all gemiddelde $\bar{\bar{y}}$ en de standaardafwijking van de gemiddelden $s_{\bar{y}}$.

1.5. Trek uit de gegevens voor de melkgiften uit tabel 1.2 met behulp van SPSS 5 keer een aselecte steekproef van 5 elementen. Voor aselecte trekking heeft SPSS het commando SAMPLE ... FROM ..., in dit geval SAMPLE 5 FROM 100.
Bereken voor elke steekproef het gemiddelde $\bar{x}_i$, het gemiddelde $\bar{\bar{x}}$ van deze gemiddelden en de standaardafwijking $s_{\bar{x}}$ van deze gemiddelden.
Doe hetzelfde voor 5 steekproeven van 45 elementen.

Bewaar de resultaten; deze hebben we in de rest van de cursus nodig.

Tabel 1.2. Populatie van vleugellengten en melkgiften

Kolom 1: Volgnummer

Kolom 2: Lengte vleugels van vliegen (mmx 10^{-1})

Kolom 3: Melkgift Friese koeien (kg x 100)

(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
01	36	51	21	42	58	41	45	61	61	47	67	81	49	76
02	37	51	22	42	58	42	45	61	62	47	67	82	49	76
03	38	51	23	42	58	43	45	61	63	47	68	83	49	79
04	38	53	24	43	58	44	45	61	64	47	68	84	49	80
05	39	53	25	43	58	45	45	61	65	47	69	85	50	80
06	39	53	26	43	58	46	45	62	66	47	69	86	50	81
07	40	54	27	43	58	47	45	62	67	47	69	87	50	82
08	40	55	28	43	58	48	45	62	68	47	69	88	50	82
09	40	55	29	43	58	49	45	62	69	47	69	89	50	82
10	40	56	30	43	58	50	45	63	70	48	69	90	50	82
11	41	56	31	43	58	51	46	63	71	48	70	91	51	83
12	41	56	32	44	59	52	46	63	72	48	72	92	51	85
13	41	57	33	44	59	53	46	64	73	48	73	93	51	87
14	41	57	34	44	59	54	46	65	74	48	73	94	51	88
15	41	57	35	44	60	55	46	65	75	48	74	95	52	88
16	41	57	36	44	60	56	46	65	76	48	74	96	52	89
17	42	57	37	44	60	57	46	65	77	48	74	97	53	93
18	42	57	38	44	60	58	46	65	78	49	74	98	53	94
19	42	57	39	44	60	59	46	67	79	49	75	99	54	96
20	42	57	40	44	61	60	46	67	80	49	76	00	55	98

1.6 Van een brievenweger is het weegresultaat normaal verdeeld met een standaardafwijking van 2,0 g ($N(\mu,2)$).

1. Wat is de kans (bij een tariefgrens van 20 g) dat op een brief van 18,0 g teveel porto wordt geplakt?

2. Wat is de kans dat voor deze brief van 18,0 g het weegresultaat tussen 17 en 18 g ligt?

3. Wat is de kans dat op een brief van 21 g te weinig wordt geplakt?

Antw.: 0,1578, 0,2902, 0,3085.

BETROUWBAARHEIDSINTERVALLEN

We hebben de begrippen populatie en steekproef leren kennen.
Een populatie heeft een frequentieverdeling, een μ (centrummaat) en een σ (spreidingsmaat).
Een steekproef heeft een kansverdeling, $\bar{x}$ en s_x.
Een relatief kleine steekproef geeft al veel informatie over de populatie.
Essentiëel is, dat de steekproef *aselect* getrokken is uit de populatie, d.w.z. dat het trekken van een element uit de populatie gebeurt met een kans die voor alle elementen van de populatie dezelfde is; de elementen uit de steekproef zijn onderling *onafhankelijk*.
De eigenschappen van de *kansverdeling van de steekproef* komen dan overeen met de eigenschappen van de *frequentieverdeling van de populatie*.

We hebben verder gezien, dat een populatie die normaal verdeeld is, geheel bepaald wordt door zijn parameters μ en σ. Met behulp van deze parameters konden we de fractie van de populatie berekenen die in een bepaald interval ligt. Dat deden we door de frequentieverdeling te standaardiseren: transformeren naar een z-verdeling, waarbij

$$z = (x-\mu) \;/\; \sigma$$

en dan met behulp van Tabel II de bij de desbetreffende z-waarde de fractie van de frequentieverdeling opzoeken, die rechts van z_{gr} ligt.

Nu nemen we een steekproef van n elementen uit de populatie die normaal verdeeld is, $N(\mu,\sigma^2)$. Voor het gemiddelde $\bar{x}$ van de steekproef geldt dat

$$\frac{\bar{x} - \mu}{\sigma \;/\; \sqrt{n}}$$

standaardnormaal verdeeld is. Dus

$$P\left(-2 < \frac{\bar{x} - \mu}{\sigma \;/\; \sqrt{n}} < 2 \right) = 0{,}95.$$

Omwerken van de ongelijkheid geeft

$$P\left(\bar{x} - \frac{2\,\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{2\,\sigma}{\sqrt{n}} \right) = 0{,}95.$$

Dit is dus het interval, waarbinnen de verwachting μ met een *betrouwbaarheid* van 95 % ligt. Zo'n interval heet het *betrouwbaarheidsinterval voor de meetverwachting* (BI): een interval dat met

een vastgestelde kans de meetverwachting bevat (Ontw. NEN 3114, 6.1), in dit voorbeeld het *95 % betrouwbaarheidsinterval*.
(De meetverwachting is gelijk aan het populatiegemiddelde.)

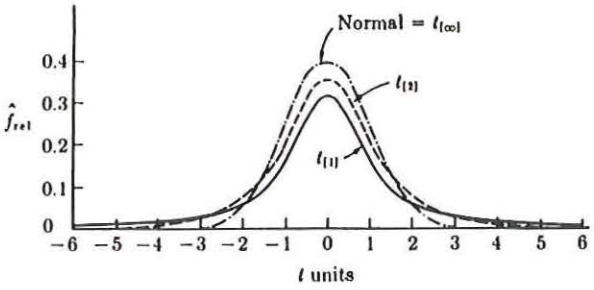
Student verdeling (t-verdeling)

Echter kennen meestal niet de σ van de populatie. We kennen slechts de standaardafwijking, s, van de steekproef en de standaardafwijking van het steekproefgemiddelde $s_{\bar{x}} = s_x / \sqrt{n}$.
Als we nu het betrouwbaarheidsinterval willen berekenen, zouden we daar dan ook s voor mogen gebruiken in plaats van σ? Dus toepassing van de standaardisatie: $z = (x-\mu) / s$?
Het mag niet zonder meer. De reden is dat σ (en ook μ) een *getal* is, dat ligt voor de gegeven populatie vast. Voor steekproeven uit een populatie zijn $\bar{x}$ en s echter geen vaste waarden, maar zijn aan het toeval onderhevig. Elke andere steekproef die we nemen, zal weer een andere s (en $\bar{x}$) opleveren. Daarin zit dus een onzekerheid en dat manifesteert zich daarin, dat we bij 'standaardisatie' van een steekproef niet zo'n mooie standaard-normale verdeling terug krijgen, maar een verdeling die een grotere spreiding vertoont, figuur 6.1. Dat wil zeggen een curve met dikkere 'staarten'. Deze verdeling wordt de t-verdeling of Student-verdeling genoemd; de verdeling is voor het eerst beschreven door W. S. Gosset onder de pseudoniem 'Student'.
Het is verder aan te voelen dat de extra onzekerheid bij een kleine steekproef groter zal zijn dan bij een grote. *De vorm van de t-verdeling is dus mede afhankelijk van het aantal elementen in de steekproef*. De 'staarten' zijn dikker bij een kleine steekproef; naarmate de steekproef groter wordt zal de t-verdeling meer op de normale verdeling gaan lijken, en bij een 'oneindig grote' steekproef daaraan gelijk worden, figuur 6.1.
In plaats van z uit Tabel 3.1 en Tabel II moeten we dus t gebruiken; die gedefiniëerd is als

$$t = (\bar{x} - \mu) / s_{\bar{x}} .$$

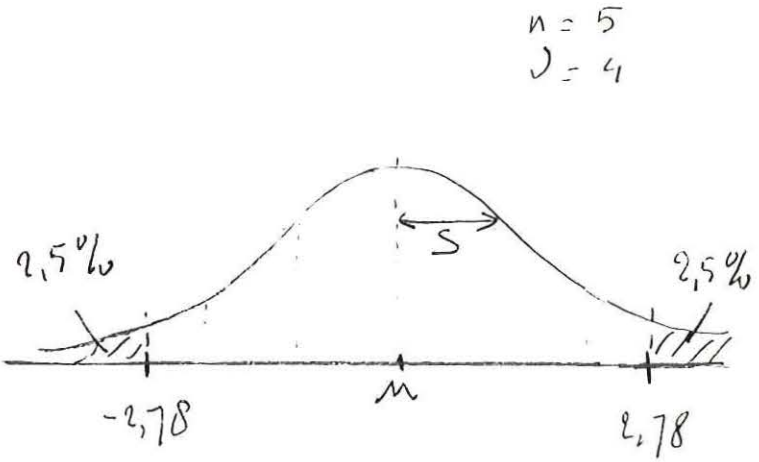
De grootheid t heeft een t-verdeling met n-1 vrijheidsgraden, waarbij n de steekproefomvang is.
Om te zien tussen welke grenzen deze grootheid met een kans P (b.v. 95 %) ligt, hanteren we een tabel van de t-verdeling. Een aantal waarden staan in Tabel 6.1, een uitgebreide tabel in Tabel III, appendix. De t-tabel is ook opgenomen in SPSS en STATCAL. Voor n = 5, dus aantal vrijheidsgraden is 4,

Figuur 6.1



Frequency curves of t distributions for 1 and 2 degrees of freedom compared with the normal distribution.

Figuur 6.2



Tabel 6.1. t-tabel voor 2-zijdige overschrijding

Aantal waarn. n	Aantal vrijh.gr. ν = df= n-1	Grenswaarde t_{ν} , zo dat t met kans P ligt ts. $-t_{\nu}$ en t_{ν} $-t_{\nu} < t < t_{\nu}$ P					
		0,900	0,950	0,980	0,990	0,998	0,999
2	1	6,31	12,71	31,82	63,66	318	637
3	2	2,92	4,30	6,97	9,93	22,3	31,6
4	3	2,35	3,18	4,54	5,84	10,2	12,9
5	4	2,13	2,78	3,75	4,60	7,17	8,61
6	5	2,02	2,57	3,37	4,03	5,89	6,86
7	6	1,94	2,45	3,14	3,71	5,21	5,96
8	7	1,90	2,37	3,00	3,50	4,79	5,41
9	8	1,86	2,31	2,90	3,36	4,50	5,04
10	9	1,83	2,26	2,82	3,25	4,30	4,78
16	10	1,75	2,13	2,60	2,95	3,73	4,07
∞	∞	1,65	1,96	2,33	2,58	3,09	3,29

ligt de grootheid t met een kans van 95 % tussen -2,78 en +2,78, zie figuur 6.2.

Bij een gegeven waarde van n ligt de vorm van de t-verdeling vast en kunnen we weer vragen naar bijvoorbeeld het 95 % betrouwbaarheidsinterval van de verwachting μ van een steekproef.

De procedure is nu gelijk aan de situatie waarin σ bekend is. Alleen staan er in de tabel andere getallen. Er geldt

$$P\left(-2{,}78 < \frac{\bar{x} - \mu}{s_x / \sqrt{5}} < 2{,}78\right) = 0{,}95.$$

Omwerken van deze ongelijkheid geeft

$$P\left\{\bar{x} - \frac{2{,}78\ s_x}{\sqrt{5}} < \mu < \bar{x} + \frac{2{,}78\ s_x}{\sqrt{5}}\right\} = 0{,}95.$$

Voor het betrouwbaarheidsinterval van μ geldt

$$\bar{x} - 2{,}78\ s_x / \sqrt{5} < \mu < \bar{x} + 2{,}78\ s_x / \sqrt{5}.$$

In het algemeen geldt voor het betrouwbaarheidsinterval van μ

$$\bar{x} - t\frac{s_x}{\sqrt{n}} < \mu < \bar{x} + t\frac{s_x}{\sqrt{n}}.$$

(Ontw. NEN 3114)

Voorbeeld:

Voor het in het begin van Hoofdstuk 4 genoemde voorbeeld van de hoogte van 10 douglassparren is het gemiddelde $\bar{x}$ gelijk aan 25.5 en s_x = 2.4. Dus sem $s_{\bar{x}} = s/\sqrt{10} = 0.76$. Een 95%-betrouwbaarheidsinterval voor de gemiddelde hoogte van het hele perceel wordt gegeven door de grenzen 23.8 en 27.2 (immers 25.5 ± 2.26*0.76).

7

TOETSEN VAN HYPOTHESEN

Een melkfabriek heeft een vulmachine voor pakken van 1 kg. Uiteraard zal het vulgewicht van individuele pakken nooit exact 1000 g bedragen maar een spreiding vertonen rond een gemiddeld vulgewicht. Dit gemiddelde vulgewicht μ kan in de loop van de tijd veranderen. Het gemiddelde mag niet te laag zijn (wettelijke eis) en niet te hoog vanwege bedrijfs-economische redenen. Men neemt daartoe elke dag een aselecte steekproef van 10 pakken om te zien of de instelwaarde μ inderdaad nog 1000 g is. Indien nodig kan men deze waarde bijstellen. Op een dag vindt men de volgende gewichten:

982, 1003, 973, 961, 997, 979, 991, 1009, 988, 969.

Op basis van deze 10 waarnemingen moet besloten worden of de machine bijgesteld moet worden. Men wil echter alleen bijstellen als het vrij zeker is dat $\mu \neq 1000$.

Voor de steekproef is

$\bar{x} = 985$

$s = 15,3$

$s_{\bar{x}} = 15,3 / \sqrt{10} = 4,9 = se = sem$

De redenering verloopt nu als volgt. Stel dat het werkelijke gemiddelde vulgewicht nog steeds 1000 is. Dan zou de grootheid t gedefinieerd als $t = (\bar{x} - 1000)/s_{\bar{x}}$ een Studentverdeling moeten hebben met 9 vrijheidsgraden. De in onze steekproef gerealiseerde waarde van t is (985-1000)/4.9 = -3.05. Deze waarde strookt niet goed met de veronderstelde kansverdeling (nl. t ligt met kans 95% tussen -2.26 en 2.26). Dus er zijn sterke aanwijzingen dat de veronderstelling (gemiddeld vulgewicht = 1000) niet houdbaar is. Ofwel: we concluderen op basis van de steekproef dat het gemiddelde vulgewicht significant afwijkt van 1000.

Merk op dat toetsen en betrouwbaarheidsintervallen 2 manieren zijn om min of meer hetzelfde te zeggen: het gemiddelde wijkt significant af van 1000 of het betrouwbaarheidsinterval bevat niet de waarde 1000.

<u>Iets meer over toetsen van hypothesen</u>

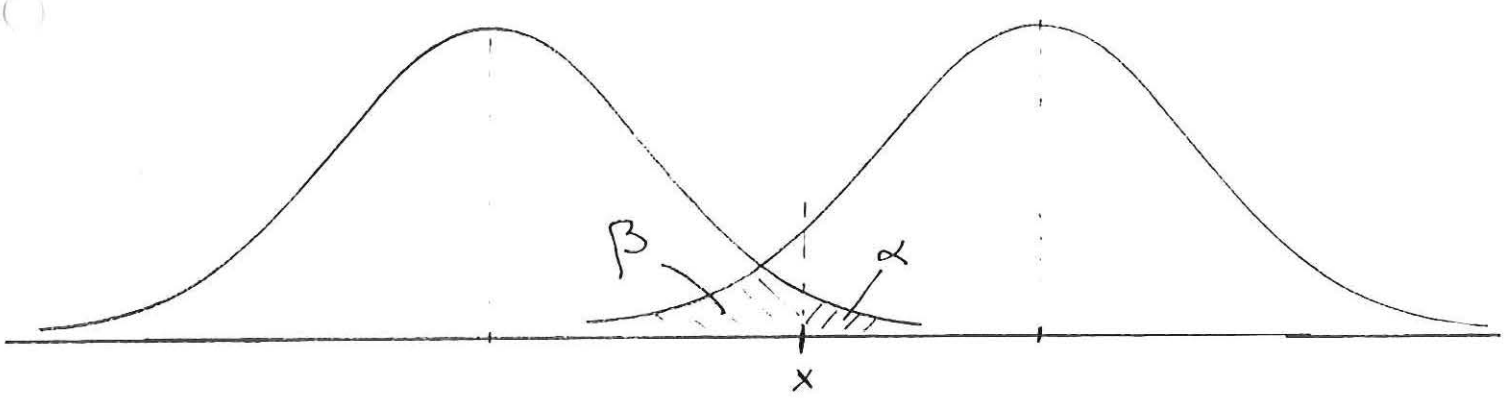
Aan de hand van het bovenstaande voorbeeld zullen we enige terminologie introduceren m.b.t. het toetsen van hypothesen. Er is sprake van twee

hypothesen; de *nulhypothese* $H_0 : \mu = 1000$ en de *alternatieve hypothese* $H_1 : \mu \neq 1000$. We moeten op grond van de waarnemingen beslissen of we H_0 al dan niet verwerpen. Die keuze baseren we op een *toetsingsgrootheid*. Hier is dat de Student grootheid $t = (\bar{X}-1000)/se$. De beslissingsregel luidt: verwerp H_0 als $|t| > 2.26$. Het getal 2.26 heet hier de *kritieke waarde*. Als H_0 verworpen wordt dan noemen we het resultaat *significant*.

De juistheid van de genomen beslissing hangt af van het feit of in werkelijkheid H_0 al dan niet waar is. Dit kan men als volgt schematisch weergeven:

		werkelijke situatie: H_0 waar (μ=1000)	werkelijke situatie: H_1 waar ($\mu\neq$1000)
genomen beslissing:	H_0 niet verwerpen	juist min.kans: $(1-\alpha)$	fout (v. 2e soort) kans: ß
	H_0 verwerpen	fout (v. 1e soort) max. kans: α	juist kans: $(1-ß)$

α = onbetrouwbaarheidsdrempel
$1-ß$ = onderscheidingsvermogen
Zie figuur 7.1



Figuur 7.1.

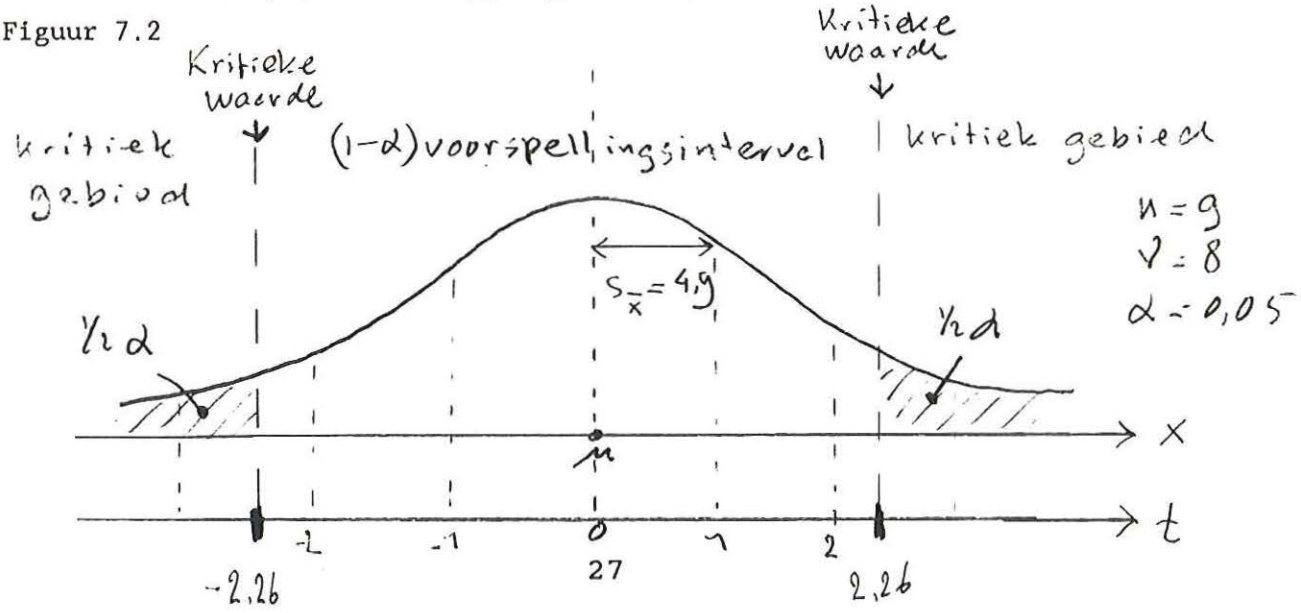
Bij het toetsen van hypothesen wil men allereerst de kans op ten onrechte verwerpen van H_0 klein houden. Dit ten onrechte verwerpen van H_0 noemen we daarom ook wel de *fout van de 1e soort*. De kans op zo'n fout noemen we de *onbetrouwbaarheid* en een voorgeschreven bovengrens voor deze kans de *onbetrouwbaarheidsdrempel* α van de toets. In de praktijk neemt men vaak α = 0.05. Hiermee is gewaarborgd dat men niet lichtvaardig tot een significant verschil besluit. Als men een andere waarde voor α kiest dan verandert ook de kritieke waarde 2.26 in het bovenstaand voorbeeld.

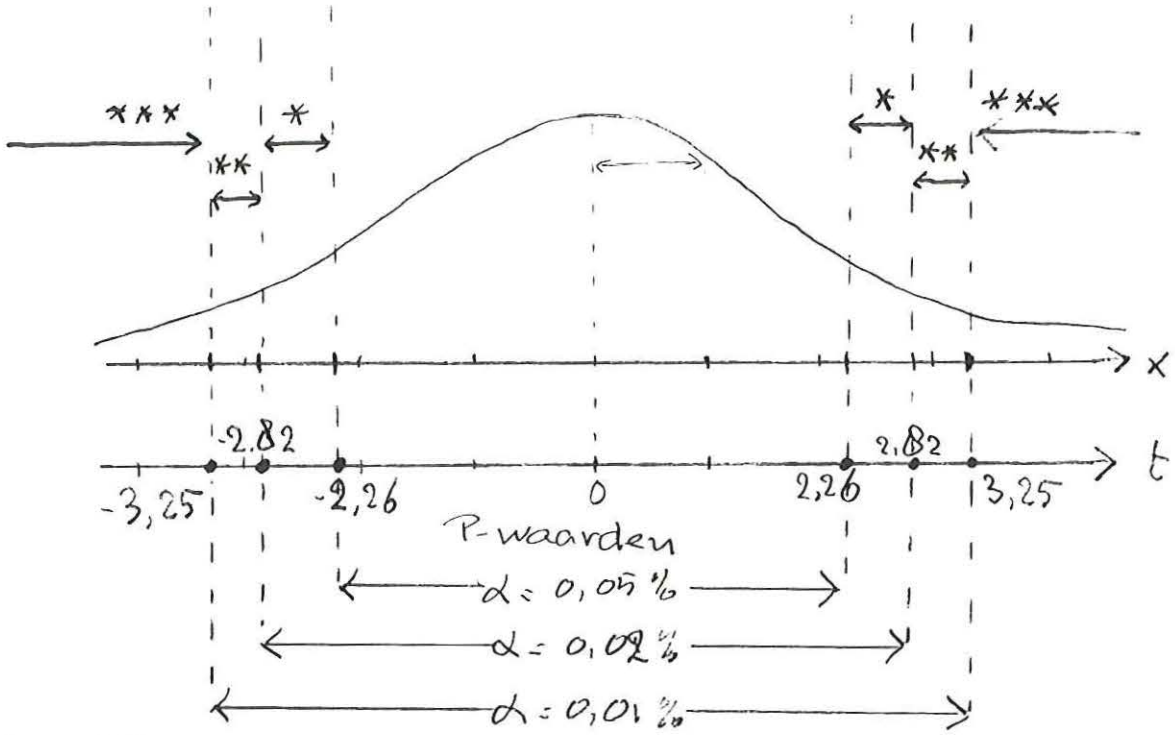
Anders kan het gesteld zijn met de *fout van de 2e soort*: het ten onrechte niet verwerpen van H_0. De kans hierop (soms genoteerd als β) hangt af van de werkelijke waarde van μ. Als het werkelijke vulgewicht μ ver verwijderd is van de nulhypothese dat μ = 1000, dan is deze kans klein, maar als de werkelijke waarde dicht bij de nulhypothese ligt dan is deze kans vrij groot. De toets kan dus waarden van μ die veel van H_0 verschillen beter onderscheiden dan waarden dicht bij H_0. De kans op terecht verwerpen van H_0 wordt ook wel het *onderscheidingsvermogen* (Engels: power) van de gebruikte toets genoemd (dus onderscheidingsvermogen = 1-β); het neemt toe naarmate de werkelijke waarde van μ meer afwijkt van de nulhypothese. Verder geldt dat bij vaste waarde van μ het onderscheidingsvermogen hoger is bij een groter aantal waarnemingen.

Overschrijdingskans

Vaak heeft men behoefte om aan te geven hoe significant de resultaten afwijken van de nulhypothese. In het bovenstaande voorbeeld is de gerealiseerde waarde van de toetsingsgrootheid t veel groter dan de kritieke waarde (3.05 vs 2.26), figuur 7.2.

Figuur 7.2





Figuur 7.3

De mate van significantie kan men aangeven via de *overschrijdingskans*, vaak ook *P-waarde* genoemd. Dit is de kleinste onbetrouwbaarheidsdrempel α waarbij het resultaat nog net significant is. In tabel III (appendix) zien we dat t=-3.05 nog wel significant is bij α=0.02, maar niet meer bij α=0.01. De P-waarde ligt in dit voorbeeld dus tussen deze waarden (P = 0.015). Omdat in de praktijk variabelen meestal slechts bij benadering normaal verdeeld zijn is het meestal niet gewenst de P-waarde tot in vele decimalen op te geven. Men gebruikt daarom ook wel de volgende globale aanduidingen om aan te geven hoe significant een resultaat is:

notatie	omschrijving	P-waarde
~	aanwijzing voor een verschil	0.05 < P < 0.10
*	significant	0.01 < P < 0.05
**	sterk significant	0.001< P < 0.01
***	zeer sterk significant	P < 0.001

Merk op dat er in het toetsen van hypothesen een zekere asymmetrie schuilt: de nulhypothese wordt geacht juist te zijn tenzij de waarnemingen voldoende duidelijk het tegendeel aantonen. Dit heeft gevolgen voor de wijze waarop men conclusies moet beoordelen: als H_0 verworpen wordt, dan mag men vrij zeker zijn dat de conclusie juist is; als echter H_0 niet verworpen wordt,

dan hoeft dat nog niet te betekenen dat H_0 juist is. In de wetenschap hanteert men vaak hetzelfde principe: een theorie wordt aangehouden totdat deze gefalsifieerd is. Ook in het strafrecht geldt een analoge situatie: een verdachte wordt alleen schuldig verklaard als zijn schuld duidelijk bewezen is. Dit houdt in dat vrijspraak niet hoeft te betekenen dat de verdachte onschuldig is, maar dat er ook sprake kan zijn van "gebrek aan bewijs".
Bij inventariserend onderzoek en herkeuringsonderzoek zal men zowel de fout van de eerste soort als de fout van de tweede soort klein willen houden.

Aantal benodigde waarnemingen

Men kan reeds vóór het onderzoek nagaan of het onderscheidingsvermogen voldoende groot is om relevante afwijkingen met voldoende kans te ontdekken. Hiertoe dient men uit te gaan van de grootte van afwijkingen die men relevant vindt en verder moet men een idee hebben hoe groot de spreiding van de waarnemingen zal zijn. Met behulp van Tabel IV (appendix) kan men dan bepalen hoe groot de steekproef moet zijn om relevante afwijkingen met een bepaalde kans als significant aan te merken.

We illustreren dit aan het hierboven besproken voorbeeld van de melkfabriek waar men een steekproef uit de dagproduktie neemt om te zien of het vulgewicht gelijk is aan 1000 g. Uit voorgaande steekproeven is gebleken dat de standaardafwijking van het vulgewicht van individuele pakken uit een dagproduktie gelijk is aan 15.
Hoe groot moet de steekproef zijn om een afwijking van het gemiddelde van 10 g of meer met een kans van minstens 80% te ontdekken?
Men kan de vraag ook anders formuleren: Men wil een betrouwbaarheidsinterval voor μ opstellen en stelt vooraf eisen stelt aan de maximale breedte van het interval; welke steekproefomvang is daarvoor nodig? In dit voorbeeld wil bovengenoemde melkfabriek een 95%-betrouwbaarheidsinterval voor het gemiddelde vulgewicht μ opstellen dat met kans 80% niet breder is dan 20, (neem hier $\mu-\mu_0$ gelijk aan de halve breedte van het interval).
Hiertoe kijken we in Tabel IV en nemen $\mu-\mu_0=10$ en $\sigma=15$ en $\beta=0.20$ en zien dan dat de steekproef moet bestaan uit 21 pakken.

VERGELIJKEN VAN TWEE POPULATIES

Tot nu toe hebben we ons bezig gehouden met uitspraken over één populatie. Vaak is men echter geinteresseerd in het vergelijken van 2 of meer populaties. Bijvoorbeeld geeft het ene tarweras een hogere opbrengst dan een ander ras? Of hebben zeehonden in het westelijk deel van de Waddenzee een hoger cadmiumgehalte dan zeehonden in het oostelijk deel? Verschillen de uitkomsten van één analysemethode van die van een andere?
Vaak is het niet onredelijk om te veronderstellen dat beide populaties normaal verdeeld zijn met gelijke varianties. De populaties zijn dan dus $N(\mu_1,\sigma^2)$ en $N(\mu_2,\sigma^2)$ verdeeld. Onze vraag komt dus neer op: Stel een betrouwbaarheidsinterval op voor $\mu_1-\mu_2$. Of equivalent: we willen toetsen de nulhypothese $H_0 : \mu_1 = \mu_2$.

Als we uit de twee populaties steekproeven trekken ter grootte n_1 en n_2, dan kunnen we hieruit de gemiddelden $\bar{x}_1$ en $\bar{x}_2$ en de varianties s^2_1 en s^2_2 berekenen. Het ligt voor de hand om $\mu_1-\mu_2$ te schatten met $\bar{x}_1-\bar{x}_2$. Deze grootheid is normaal verdeeld met verwachting $\mu_1 - \mu_2$ en variantie $\sigma^2/n_1 + \sigma^2/n_2$. Voor σ^2 hebben we nu twee schatters beschikbaar: s^2_1 en s^2_2. De informatie uit beide schatters kunnen we samenvoegen tot een nieuwe schatter, de zogenaamde gepoolde schatter voor σ^2. Deze noteren we als s^2 en wordt als volgt berekend uit s^2_1 en s^2_2:

$$s^2 = \frac{(n_1-1)s^2_1 + (n_2-1)s^2_2}{(n_1-1) + (n_2-1)}.$$

Deze schatter heeft n_1+n_2-2 vrijheidsgraden. Merk op dat s^2 het gewogen gemiddelde is van s^2_1 en s^2_2 waarbij de aantallen vrijheidsgraden de gewichten vormen.
Met deze gepoolde schatter s^2 kunnen we de variantie van $\bar{x}_1-\bar{x}_2$ schatten: $s^2/n_1 + s^2/n_2$ ofwel $s^2 * (1/n_1 + 1/n_2)$. De standaardfout van $\bar{x}_1-\bar{x}_2$ is gelijk aan de wortel hieruit, dus

$$se(\bar{x}_1-\bar{x}_2) = s.\sqrt{(1/n_1 + 1/n_2)}.$$

Deze se wordt wel sed genoemd: <u>standard error of difference</u> (of means).

Standaardisatie van $\bar{x}_1 - \bar{x}_2$ levert de grootheid

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{sed}$$

die een Student verdeling heeft met n_1+n_2-2 vrijheidsgraden.
Een 95%-betrouwbaarheidsinterval voor $\mu_1-\mu_2$ is dus:

$$\bar{x}_1 - \bar{x}_2 - t_\nu * sed < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + t_\nu * sed$$

waarbij t_ν kan worden opgezocht in Tabel III (t_ν heeft $\nu = n_1+n_2-2$ vrijheidsgraden en is meestal ongeveer 2).

Als we willen toetsen of de gemiddelden van de twee populaties verschillen (H_0: $\mu_1=\mu_2$) dan nemen we als toetsingsgrootheid

$$t = \frac{\bar{x}_1 - \bar{x}_2}{sed}$$

Deze is onder H_0 Student verdeeld met n_1+n_2-2 vrijheidsgraden. Dus als t groter is dan de kritieke waarde t_ν uit Tabel III (of kleiner dan $-t_\nu$) dan verschillen de twee steekproeven significant. Of ook: de twee steekproeven verschillen significant als $\bar{x}_1$ en $\bar{x}_2$ meer verschillen dan t_ν*sed. De laatste uitdrukking heet daarom ook wel het kleinste significante verschil (Engels: least significant difference, afgekort lsd).

Merk op dat ook hier weer geldt dat de nulhypothese $\mu_1=\mu_2$ verworpen wordt dan en slechts dan als 0 niet in het betrouwbaarheidsinterval voor $\mu_1-\mu_2$ ligt.

Voorbeeld 1

Om te onderzoeken of uienras A in de praktijk een andere opbrengst geeft dan ras B, zijn 10 bedrijven geloot die ras A verbouwen en 10 bedrijven geloot die ras B gebruiken. Alle 20 bedrijven zijn ondervraagd naar de opbrengst in 1988 (in ton per ha). De uitkomsten waren als volgt:

ras A: 36, 47, 39, 43, 49, 38, 41, 51, 40, 44
ras B: 45, 47, 34, 39, 31, 38, 41, 37, 43, 40

Dus $\bar{x}_A = 42.8$ en $\bar{x}_B = 39.5$
Uit $s^2_A = 24.4$ en $s^2_B = 23.6$ berekenen we de gepoolde schatter voor de

variantie: $s^2 = \frac{9s^2_A + 9s^2_B}{18} = 24.0$

dus sed = $se(\bar{x}_A - \bar{x}_B) = \sqrt{24.0} . \sqrt{(1/10 + 1/10)} = 2.19$

De kritieke waarde t_{18} is gelijk aan 2.10 (Tabel III)

Omdat $\bar{x}_A - \bar{x}_B$ kleiner is dan lsd = 2.10 * sed = 4.6 concluderen we dat de opbrengst van de 2 rassen in 1988 niet significant verschilt. Het 95%-betrouwbaarheidsinterval voor het gemiddelde verschil is (-1.1, 7.7).

Voorbeeld 2

Om te onderzoeken of het cadmiumgehalte in de lever van zeehonden in de Westelijke Waddenzee systematisch afwijkt van het gehalte in het Oostelijke deel is uit beide delen van de in 1986 dood aangetroffen zeehonden het cadmiumgehalte in de lever bepaald. Allereerst moet men zich afvragen of de steekproef als aselect kan worden opgevat. Immers als vele dieren zijn gestorven als gevolg van een cadmiumvergiftiging dan geven de twee steekproeven geen representatief beeld van de populaties (in de steekproeven zijn dan de dieren met een hoog cadmiumgehalte oververtegenwoordigd). Daarom zijn in de steekproef alleen die zeehonden opgenomen waarvan de doodsoorzaak een duidelijk andere was dan een cadmiumvergiftiging.

De gegevens zijn in ppm:

Westelijke Waddenzee: 96 58 72 205 89 135

Oostelijke Waddenzee: 53 42 38 77 106 66 29 48

Omdat van gehalten aan chemische stoffen bekend is dat de verdeling vaak lognormaal is voeren we de statistische analyse uit op de logaritmen van de gehalten. Verder geldt dat de spreiding van de gehalten niet constant is maar evenredig met het gemiddelde (variatiecoefficient is constant). Dit houdt in dat voor de loggehalten voldaan is aan de veronderstelling van gelijke varianties. De gemiddelden zijn op logschaal $\bar{x}_W$ =4.60 en $\bar{x}_O$ =3.97. Uit s^2_W=0.205 en s^2_O=0.172 volgt de gepoolde schatter s^2=0.186, dus

sed = $\sqrt{0.186} . \sqrt{(1/6 + 1/8)} = 0.233$.

Omdat $\bar{x}_W - \bar{x}_O$ groter is dan lsd = 2.18 * 0.233 = 0.51 concluderen we dat in het Westelijk deel het cadmiumgehalte significant hoger is dan in het Oostelijk deel. Het betrouwbaarheidsinterval voor het verschil van de gemiddelden is op de logschaal 0.63 ± 0.51 ofwel (0.12, 1.14). Omdat verschillen op logschaal overeenkomen met verhoudingen op de

oorspronkelijke schaal concluderen we: het mediane cadmiumgehalte is in het Westelijk deel een factor $e^{0.63} = 1.88$ hoger dan in het Oostelijk deel. Een 95%-betrouwbaarheidsinterval voor deze factor is (1.13, 3.13).

Aantal benodigde waarnemingen

Indien men de gelijkheid van de gemiddelden μ_1 en μ_2 van twee populaties wil toetsen dan geldt ook hier weer dat het onderscheidingsvermogen van de toets groter is naarmate de gemiddelden meer verschillen. Verder hangt dit onderscheidingsvermogen af van σ^2 en n_1 en n_2. Dit betekent dat men de benodigde steekproefgrootte kan bepalen om relevante verschillen met een bepaalde kans te ontdekken als men een schatting heeft voor de binnen-populatie-variantie . Hiervoor kan men Tabel V gebruiken als men n_1 gelijk neemt aan n_2.

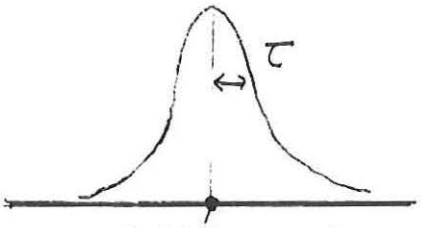
Voorbeeld: Hoeveel bedrijven hadden we in voorbeeld 1 voor beide rassen moeten nemen om een verschil in gemiddelde opbrengst ter grootte 5 met kans 80% te kunnen ontdekken? (Veronderstel hierbij dat 24 een redelijk goede schatter is voor de binnen-populatie-variantie). Neem in Tabel V (appendix) $(\mu_x-\mu)/\sigma$ gelijk aan $5/\sqrt{24} = 1.02$ dan vinden we $n = 17$. Dus bij steekproefomvang $n_1 = n_2 = 17$ zouden verschillen van 5 of groter met een kans van minstens 80% significant geweest zijn.

9

INVLOED VAN MEETFOUTEN

Tot nu toe is gesproken over een kenmerk dat in de populatie een $N(\mu,\sigma^2)$ verdeling bezit. Uit deze populatie trekt men een aselecte steekproefvan n elementen. Van deze elementen wordt de waarde van het kenmerk vastgesteld. Dit levert uitkomsten $x_1,\ldots,x_n$. Deze uitkomsten zijn stochastisch van aard omdat door het toeval bepaald wordt welke elementen in de steekproef terecht komen. Als de meting (d.w.z. de vaststelling van de waarde van het kenmerk) foutloos geschiedt dan is de kansverdeling van de grootheid x_i gelijk aan de populatieverdeling. Dus x_i is $N(\mu,\sigma^2)$ verdeeld.

Als echter de meting niet foutloos geschiedt (hetgeen in de praktijk meestal het geval is) dan heeft x_i een andere kansverdeling. In het eenvoudigste geval levert de meetfout soms een positieve afwijking en soms een negatieve afwijking, beide met even grote kans. Dus als we de meting aan het zelfde element vaak zouden herhalen dan zou dit de volgende resultaten leveren.



We spreken dan van een toevalige meetfout. De variantie van deze meetfout noemen we τ^2, zodat de meetfout (evt. bij benadering) $N(0,\tau^2)$ verdeeld is. De waargenomen x_i is dan $N(\mu,\sigma^2+\tau^2)$ verdeeld. De variantie van x_i is de som van de populatievariantie en de meetfoutvariantie.

Stel dat de meting behalve toevallige fluctuatie ook nog een systematische afwijking vertoont, d.w.z. de meting vertoont gemiddeld genomen een afwijking α t.o.v. de werkelijke waarde. Dan is de grootheid x_i $N(\mu+\alpha, \sigma^2+\tau^2)$ verdeeld.

Hoe zit het met de schatting van het populatiegemiddelde μ resp. de populatievariantie σ^2 als we de schatting ($\bar{x}$ resp s^2) baseren op een

steekproef waarvan we de waarde van het te onderzoeken kenmerk niet foutloos kunnen bepalen? Wat valt er dan te zeggen over de precisie van de schatter $\bar{x}$ en wat de invloed op de betrouwbaarheidsintervalllen en toetsen?

We zullen dit eerst bespreken voor de situatie waarin alleen sprake is van een toevallige fout en de systematische fout afwezig is (dus $\alpha=0$). In dit geval is $\bar{x}$ nog steeds een zuivere schatter van μ, maar s^2 levert een systematisch te hoge schatting voor σ^2. (Er geldt dan immers $Es^2= \sigma^2+\tau^2$). Als echter de meetfout vrij klein is t.o.v. de standaardafwijking van de populatie dan is dit effect nauwelijks merkbaar. Immmers als bijvoorbeeld $\tau= 0.1*\sigma$ dan geldt dat $Es= 1.01*\sigma^2$. Ook de precisie van de schatter $\bar{x}$ heeft in zo'n situatie nauwelijks te lijden onder de meetfout Immers de precisie van $\bar{x}$ wordt dan $\sqrt{((\sigma^2+\tau^2)/n)}$ in plaats van $\sigma/\sqrt{n}$.
De betrouwbaarheidsintervallen resp toetsen blijven geldig als er alleen sprake is van een toevallige meetfout. Als τ vrij klein t.o.v. σ dan zal het betrouwbaarheidsinterval nauwelijks breder worden en ook het onderscheidingsvermogen van de toetsen zal nauwelijks verminderen t.g.v. de meetfout.
Kortom: toevallige meetfouten, mits klein t.o.v. de populatiespreiding, veroorzaken geen problemen.

Als echter de meetfout systematisch is, dan is $\bar{x}$ niet meer een zuivere schatter van μ. Immers dan geldt $E\bar{x} = \mu+\alpha$. De onnauwkeurigheid van $\bar{x}$ is dan gedefinieerd als $E(\bar{x} -\mu)^2$. Deze is gelijk aan $\sqrt{((\sigma^2+\tau^2)/n + \alpha^2)}$. De betrouwbaarheidsintervallen en toetsen voor μ zijn dan niet meer geldig omdat de werkelijke onbetrouwbaarheid niet meer overeenkomt met de opgegeven onbetrouwbaarheid.

Opmerking 1: Als de grootte van de systematische fout α bekend is dan kan men hiervoor corrigeren door x_i met α te verminderen.
Opmerking 2: Een (onbekende) systematische meetfout heeft geen gevolgen voor het vergelijken van 2 populaties (als de grootte van de meetfout niet afhangt van het niveau ervan). Immers toetsen resp. betrouwbaarheidsintervallen voor $\mu_1-\mu_2$ blijven geldig omdat in de grootheid $\bar{x}_1-\bar{x}_2$ de systematische fouten wegvallen.

SAMENVATTING

POPULATIE

Kengetallen:

centrummaat	μ	gemiddelde
		mediaan
		modus
spreidingsmaat	σ^2	variantie
	σ	standaardafwijking
	vc = σ/μ	variatiecoëfficiënt = rel. standaardafwijking

Voor normale verdeling $N(\mu,\sigma^2)$:

standaard-normaal N(0,1):

z-as $\quad z = (x-\mu)/\sigma$

97,5 percentiel: $\mu+2\sigma$ $\quad P(x < \mu+2\sigma) = 97{,}5\ \%$

Ook geldt voor 95 % : $\quad \mu-2\sigma < x < x+2\sigma$

STEEKPROEF

Kengetallen:

centrummaat	$\bar{x}$	gemiddelde
spreidingsmaat	s^2	variantie
	s	standaardafwijking

Voor een steekproef van n waarnemingen:

$x_1, x_2, x_3, x_4, .., x_n$

Gemiddelde $\quad \bar{x} = \Sigma x / n$

s.a. = standaardafwijking voor één waarneming x

$$s_x = \sqrt{\{ \Sigma(x_i - \bar{x})^2 / n-1\}}$$

s.e. = standaardafwijking van het gemiddelde $\bar{x}$

$$s_{\bar{x}} = s_x / \sqrt{n}$$

Betrouwbaarheid:

n waarnemingen:

t-toets $\quad$ t-as $\quad t = (\bar{x} - \mu) / s_{\bar{x}}$

$$= (\bar{x} - \mu) \sqrt{n} / s_x$$

95 % BI: $\quad \bar{x} - 2\, s_x/\sqrt{n} < \mu < \bar{x} + 2\, s_x/\sqrt{n}$

VRAAGSTUKKEN

2.1 Een analysemethode geeft meetresultaten die normaal verdeeld zijn met een standaardafwijking σ van 4 ppm. Er worden vier analysen uitgevoerd met resultaten 13, 16, 17 en 14 ppm. Geef een schatting van het 99 % betrouwbaarheidsinterval voor μ.
Antwoord: $9,84 < \mu < 20,16$.

2.2 Voor de bepaling van het eiwitgehalte in afvalwater worden 10 waarnemingen gedaan. Gevonden wordt (in gram per 10 liter):
36,5 45,2 40,5 42,0 37,7 39,4 41,6 38,8 39,0 43,3.
Bereken: het gemiddelde,
de standaardafwijking van een individuele waarneming,
de standaardafwijking van het gemiddelde.
Geef het 95 % betrouwbaarheidsinterval voor de gezochte waarde.
Antwoord: $38,5 < \mu < 42,3$ (α = 5%; n=10).

2.3 In monsters oppervlaktewater wordt regelmatig het chloorgehalte gecontroleerd. Per keer worden 16 monsters genomen. De standaardafwijking hangt mede af van de aanwezigheid van andere verontreinigingen, daarom wordt elke keer de standaardafwijking opnieuw uit de steekproef berekend. Men vindt op een gegeven moment voor de verdeling van de steekproef een gemiddelde waarde van 1,80 mg/l en een standaardafwijking s_x van 0,40 mg/l. Geef het 95 % betrouwbaarheidsinterval voor het chloridegehalte.

2.4 Men wil van twee appelrassen, Schone van Boskoop en Cox Orange Peppin, wten of het gehalte aan kalium, calcium en magnesium verschillend is. Van elk soort worden 10 appels onderzocht. Gevonden wordt: (zie tabel)

Bereken voor het verschil in kaliumgehalte het 95 % betrouwbaarheidsinterval. Is het gevonden verschil met deze betrouwbaarheid significant?
Bereken de overschrijdingskans voor het gevonden verschil.

Beantwoord deze vragen ook voor calcium en magnesium.

Tabel

Schone van Boskoop			Cox Orange Peppin		
K (%)	Ca(mg/kg)	Mg(mg/kg)	K(%)	Ca(mg/kg)	Mg(mg/kg
0,159	30,9	55,3	0,132	28,3	50,8
0,132	35,7	49,8	0,159	23,6	59,1
0,145	28,6	47,7	0.151	24,8	52,2
0,096	36,5	39,6	0,131	22,7	50,8
0,138	22,6	49,3	0,131	21,8	51,7
0,089	23,0	35,7	0,132	23,1	47,2
0,111	24,9	41,7	0.110	22,8	48,7
0,101	30,4	38,6	0,138	23,6	50,1
0,128	35,6	47,5	0,147	27,2	53,5
0,114	31,0	42,6	0,127	25,3	50,9

10

NAUWKEURIGHEID VAN METINGEN

In de voorgaande hoofdstukken is in algemene zin gesproken over de statistische begrippen populatie en steekproef. Een steekproef wordt uit een populatie genomen om informatie over die populatie te verkrijgen. Binnen het RIKILT komt steekproefonderzoek ook wel voor. Bijvoorbeeld een steekproef uit een maiskuil, verder bij inventariserend, sensorisch of toxicologisch onderzoek. Maar in het algemeen spreken wij niet over populaties en steekproeven: er komt een monster binnen en dat moet worden geanalyseerd. Hoe kunnen we de geleerde berippen toch in deze omgeving toepassen?

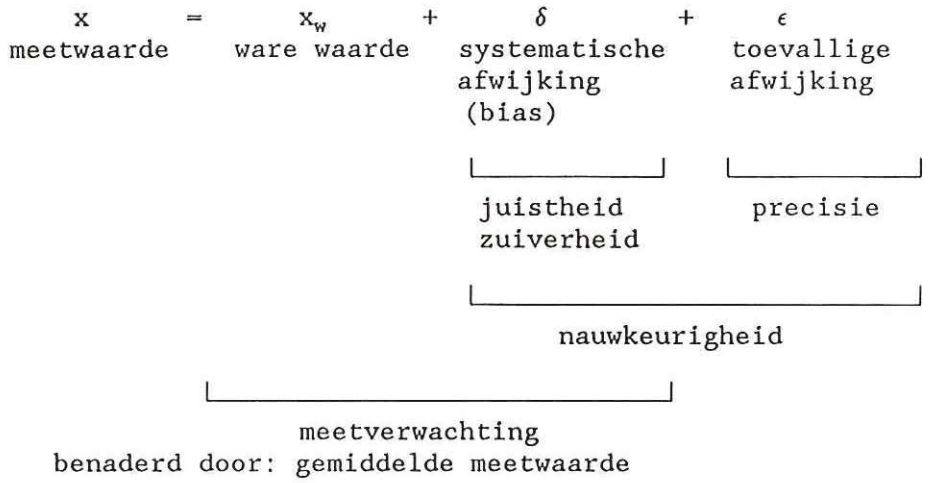
De resultaten van de analysen die op een monster worden uitgevoerd, bijvoorbeeld in duplo, kunnen worden gezien als een steekproef. De populatie is dan de resultaten van alle analysen die in dit monster zouden kunnen worden uitgevoerd.

Daarom is het belangrijk dat van een analysemethode de standaardafwijking bekend is: dan is de standaardafwijking van 'alle analysen in het monster' bekend, en kunnen we uit de duplometingen (de steekproef) conclusies trekken over 'het gehalte' (van de populatie) en de betrouwbaarheid daarvan.

In Hoofdstuk 9 zagen we, dat we ons niet druk hoeven te maken over toevallige meetfouten, als die maar klein zijn t.o.v. de spreiding in de populatie. Deze redenering gaat voor ons niet meer op: bij metingen zonder afwijkingen zouden alle elementen in de populatie exact aan elkaar gelijk zijn en de 'ware waarde' in het monster weergeven. Wij zullen ons dus wel degelijk bezig moeten houden met afwijkingen. (De neutrale term 'afwijking' heeft de voorkeur boven 'fout', het is een statistisch begrip, dat geen waardeoordeel inhoudt. Vergelijk in het Engels 'error' en 'mistake'.)

De Nederlandse norm NEN 3114 handelt over Nauwkeurigheid van meten (Ontwerp 2e druk, 1990) en NEN-ISO 5725 over Precisie van analysemethoden. De hierna te bespreken begrippen zijn aan deze normen ontleend en kunnen als volgt schematisch worden weergegeven.

Statistisch model



Nauwkeurigheid

Een monster dat onderzocht wordt, heeft een *'ware waarde'*, x_w voor de gezochte analyt. Zeg maar het gehalte dat 'echt' in het monster aanwezig is. De bij de analyse verkregen waarde, de *meetwaarde*, of het *waarnemingsresultaat*, x, zal hier van afwijken.

Als deze afwijking gering is, heet de uitkomst *nauwkeurig*: *nauwkeurigheid (accuracy)* is de mate waarin een met een bepaalde analysemethode verkregen waarnemingsresultaat de ware waarde benadert. Vaak spreekt men liever van het omgekeerde, dus de *onnauwkeurigheid*: de mate waarin het waarnemingsresultaat afwijkt van de ware waarde.

De afwijkingen die de onnauwkeurigheid veroorzaken kan men onderscheiden in *toevallige afwijkingen*, ϵ en *systematische afwijkingen*, δ.

Toevallige afwijkingen: Wanneer we een analyse een aantal malen herhalen zullen de waarnemingsresultaten onderling verschillen op een wijze die aan het toeval onderhevig lijkt te zijn en (dus) normaal verdeeld zijn. Het gemiddelde waarnemingsresultaat, $\bar{x}$, zal bij een toenemend aantal waarnemingen tot één waarde naderen: de *meetverwachting*, μ. Onder de toevallige afwijking van één meetresultaat verstaat men nu het verschil tussen dat meetresultaat en de meetverwachting. Een maat hiervoor is de *precisie*, dat is de mate van overeenstemming tussen waarnemingsresultaten van herhalingen.

Systematische afwijkingen: Ook wanneer een methode heel precies is, hoeft het waarnemingsresultaat niet gelijk te zijn aan de ware waarde. De schaalindeling of het nulpunt van een ijkcurve kan niet correct zijn, gebruikte maatapparatuur of ijkoplossingen kunnen afwijkingen vertonen,

temperatuur kan verschillen. Ook kan een bepaalde methode systematisch te lage of te hoge uitkomsten geven. Het is een bekend verschijnsel dat bij niveaucontroles één laboratorium steeds hogere of lagere waarden vindt dan andere laboratoria. In deze gevallen spreken we van een *systematische afwijking*. Soms is de oorzaak bekend, soms ook niet (*bekende*, resp. *onbekende systematische afwijking*). Voor een bekende systematische afwijking kan worden gecorrigeerd.
De systematische afwijking wordt gedefiniëerd als het verschil tussen de meetverwachting en de ware waarde. Bij een *zuivere meetmethode* is de systematische afwijking gelijk aan nul.
Door ISO/DP 5725/1 wordt het begrip *trueness* gedefiniëerd als de mate van overeenstemming tussen de gemiddelde waarde verkregen uit een reeks waarnemingen en de ware waarde; in ORA verband vertaald met *juistheid*. (Chemici verstaan onder zuiver iets anders dan statistici.) In de literatuur wordt ook wel het begrip *nauwkeurigheid van het gemiddelde (accuracy of the mean)* gebruikt, of, verwarrenderwijs, zonder meer *nauwkeurigheid (accuracy)*. Let op: in de meetverwachting zit de systematische afwijking dus nog in.

Totale afwijking: Dit is het verschil tussen een meetwaarde en de ware waarde. De totale afwijking is gelijk aan de som van de de toevallige (ϵ) en de systematische afwijking (δ); in formule

$$x - x_w = \delta + \epsilon = (\mu - x_w) + (x - \mu).$$

Als, zoals gewoonlijk het geval is, de ware waarde μ benaderd wordt door de gevonden gemiddelde waarde $\bar{x}$, wordt de sysematische afwijking benaderd door $(\bar{x} - x_w)$ en de toevallige afwijking door $(x - \bar{x})$.

Toevallige afwijking, precisie

De gebruikelijke maat voor de precisie is de standaardafwijking. Deze geldt dan voor een bepaalde, exact omschreven methode, toegepast op identiek monstermateriaal. Van belang is dat men de meetomstandigheden omschrijft, want de standaardafwijking kan daar sterk van afhangen. Men kan hierbij twee uitersten onderscheiden: alles zo veel mogelijk hetzelfde houden: zelfde laboratorium, zelfde analist, zelfde apparatuur, zelfde ijkoplossingen en -instellingen, analysen gelijktijdig of zo kort mogelijk na elkaar uitvoeren, enz. Dit heten *herhaalbaarheids-omstandigheden*

(repeatability conditions). De gevonden standaardafwijking heet *herhaalbaarheids-standaardafwijking (repeatability standard deviation)*. Onder *herhaalbaarheid* verstaat NEN-ISO 5725 de mate van overeenstemming tussen onderling onafhankelijke analyseresultaten, verkregen onder herhaalbaarheids-omstandigheden.
Het andere uiterste is, dat *alle* genoemde omstandigheden verschillen: verschillende laboratoria met verschillende analisten en verschillende apparatuur. Dit zijn *reproduceerbaarheidsomstandigheden (reproducibility conditions)*. Hiervoor geldt de *reproduceerbaarheids-standaardafwijking (reproducibility standard deviation)*.
En *reproducibiliteit (reproducibility)* is de mate van overeenstemming tussen analyseresultaten verkregen onder reproduceerbaarheids-omstandigheden.

NEN-ISO 5725 gebruikt het volgende statistische model

$$y = m + B + e$$

waarin

y is één enkel analyseresultaat
m is het algemeen gemiddelde
B is de tussen-laboratorium variantie
e is de toevallige afwijking in iedere analyse.

Let op dat in dit model m niet de ware waarde x_w is, maar het niveau van de te onderzoeken eigenschap (b.v. het gehalte) volgens de desbetreffende analysemethode aangeeft. Het verschil $m - x_w$ is de systematische afwijking (bias) van de analysemethode.
De term e representeert een toevallige afwijking die in iedere enkelvoudige analyse optreedt. Voor één laborarotium wordt de variantie van e de *binnen-laboratorium variantie (within-laboratory variable)* genoemd,

$$var(e) = \sigma^2_w.$$

Het is aannemelijk dat σ^2_w voor verschillende laboratoria verschillende waarden zal hebben, maar dat voor een goed gestandaardiseerde methode deze onderlinge verschillen klein zijn, zodat een algemeen geldende, gemiddelde waarde van de binnen-laboratorium variantie voor alle laboratoria die de analysemethode gebruiken kan worden aangenomen. Deze waarde heet de *herhaalbaarheids-variantie*,

$$\overline{var}\ (e) = \sigma^2_r.$$

De *herhaalbaarheidsstandaardafwijking*, σ_r is

$$\sigma_r = \sqrt{\overline{var}}\ (e).$$

De term B beschouwt men voor analysen onder herhaalbaarheidsomstandigheden als constant, maar voor analysen onder reproduceerbaarheidsomstandigheden als een toevallige afwijking. In dit laatste geval heet de variantie van deze parameter de *tussen-laboratorium variantie (between-laboratory variance)*

$$var\ (B) = \sigma^2_L.$$

De *reproduceerbaarheidsstandaardafwijking*, σ_R, is

$$\sigma_R = \sqrt{(\sigma_L^2 + \sigma_r^2)}.$$

Herhaalbaarheid en reproduceerbaarheid

We kunnen ons nu afvragen hoeveel twee enkelvoudige analyseresultaten met 95% betrouwbaarheid mogen verschillen. Daarvoor heeft NEN-ISO 5725 gedefiniëerd:

herhaalbaarheidswaarde (repeatability value), r : De waarde waaronder het absolute verschil tussen twee analyseresultaten verkregen oder herhaalbaarheidsomstandigheden verwacht wordt te liggen met een waarschijnlijkheid van 95 %. Wanneer geen verwarring kan ontstaan over het feit dat de waarde en niet het begrip wordt bedoeld, kort men dit af tot *de herhaalbaarheid*.

Reproducibiliteitswaarde (reproducibility value), R: De waarde waaronder het absolute verschil tussen twee analyseresultaten verkregen oder reproducibiliteitsomstandigheden verwacht wordt te liggen met een waarschijnlijkheid van 95 %. Wanneer geen verwarring kan ontstaan over het feit dat de waarde en niet het begrip wordt bedoeld, kort men dit af tot *de reproducibiliteit*.

Wanneer het aantal analyseresultaten niet te klein is, kan de verdeling daarvan als normaal worden beschouwd en is het 95 % betrouwbaarheidsinterval dus $\pm\ 2\sigma$. Verder is voor duplobepalingen de standaardafwijking $\sqrt{2}$ kleiner dan voor enkelvoudige waarnemingen. Onder deze omstandigheden geldt dan

herhaalbaarheid: $r = 2\sqrt{2}\ s_r \approx 2{,}8\ s_r$

reproduceerbaarheid: $R = 2\sqrt{2}\ s_R \approx 2{,}8\ s_R$.

DOORWERKEN VAN AFWIJKINGEN

Tot nu toe hebben we ons beziggehouden met meetresultaten die afhangen van één grootheid. In de praktijk is het uiteindelijke analyseresultaat, M, afhankelijk van een aantal gemeten grootheden, A, B, C,..:

$$M = f(A, B, C, \ldots).$$

Of, er is een transformatie toegepast, zodat het eindresultaat wel van één gemeten grootheid afhangt, maar niet lineair:

$$M = f(A),$$

bijvoorbeeld $M = \log A$, $M = A^2$, $M = e^A$.

Hoe hangt in zo'n geval de nauwkeurigheid van M af van de afzonderlijk gemeten grootheden A, B, enz.?

Per definitie is

$$s_M^2 = \Sigma (M - \bar{M})^2 / n_M\text{-}1 = \Sigma\, dM^2 / n_M\text{-}1.$$

$$s_A^2 = \Sigma (A - \bar{A})^2 / n_A\text{-}1 = \Sigma\, dA / n_A\text{-}1,$$

enzovoorts voor B, C, .. .

We nemen aan dat we de standaardafwijkingen van de afzonderlijke metingen s_A, s_B, ... kennen en de vraag is nu, hoe we hieruit de standaardafwijking voor M kunnen berekenen. Het verband tussen kleine fluctuaties dM in M en kleine fluctuaties dA, dB, enz. is mathematisch te benaderen door M partiëel te differentiëren:

$$dM = \left(\frac{\partial M}{\partial A}\right). dA + \left(\frac{\partial M}{\partial B}\right). dB + \ldots$$

Kwadrateren van deze uitdrukking geeft

$$dM^2 = \left(\frac{\partial M}{\partial A}\right)^2. dA^2 + \left(\frac{\partial M}{\partial B}\right)^2. dB^2 + ..+ 2\left(\frac{\partial M}{\partial A}\right)\left(\frac{\partial M}{\partial B}\right) dA.dB + \ldots$$

Wanneer de waarnemingen A, B, .. onderling onafhankelijk zijn, dan zijn de dubbelprodukten gemiddeld nul en gaat deze vergelijking, wanneer we middelen over een reeks waarnemingen, over in

$$s_M^2 = \left(\frac{\partial M}{\partial A}\right)^2. s_A^2 + \left(\frac{\partial M}{\partial B}\right)^2. s_B^2 + \ldots$$

Dit is de algemene formule om de variantie van een analyseresultaat uit de varianties van de afzonderlijke bepalingen te berekenen.

Voor een aantal eenvoudige gevallen is deze formule aanzienlijk te vereenvoudigen, namelijk wanneer M uitsluitend de som of het verschil, of het produkt of quotiënt van de afzonderlijke bepalingen is:

Som en verschil

$M = A + B + C + \ldots$ $\qquad s_M^2 = s_A^2 + s_B^2 + s_C^2 + \ldots$

$M = A - B \ldots$

Produkt en quotiënt

$M = A * B * C * \ldots$

$M = A * B / C$

$M = A / B$ $\qquad \left(\frac{s_M}{M}\right)^2 = \left(\frac{s_A}{A}\right)^2 + \left(\frac{s_B}{B}\right)^2 + \ldots$

ofwel: $VC_M = VC_A + VC_B + \ldots$

Logtransformatie

$M = \ln A$ $\qquad s_M^2 = \left(\frac{s_A}{A}\right)^2 = VC_A^2$ ofwel $s_M = VC_A$

Afleiding van variantie voor log-transformatie:

M = ln A, dan is volgens de algemene differentiaalformule

$$s^2_{\ln A} = \left(\frac{\delta \ln A}{\delta A}\right)^2 s_A^2 = \left(\frac{1}{A}\right)^2 . s_A^2 = \left(\frac{s_A}{A}\right)^2$$

De variantie van log A is dus gelijk aan het kwadraat van de variatiecoëfficiënt van A, ofwel de standaardafwijking van log A is gelijk aan de variatiecoëfficiënt van A, zoals reeds in Hoofdstuk 3 is vermeld.

Ook het feit dat de standaardafwijking van n waarnemingen gelijk is aan $1/\sqrt{n}$ van de standaardafwijking van een enkele waarneming volgt uit bovenstaande afleidingen. Voor de berekening van het gemiddelde geldt de somformule:

$$\bar{x} = (x_1 + x_2 + x_3 + \ldots + x_n) / n \text{ , dus}$$

$$s_{\bar{x}}^2 = (1/n)^2 . (s_{x_1}^2 + s_{x_2}^2 + \ldots s_{x_n}^2)$$

Men mag aannemen dat deze waarnemingen onafhankelijk zijn. Ze komen uit dezelfde verdeling en hebben dus dezelfde standaardafwijking s_x , dus

$$s_{\bar{x}}^2 = (1/n)^2 . (n . s_x^2) = s_{\bar{x}}^2 / n$$

en is de standaardafwijking van het gemiddelde

$$s_{\bar{x}} = s_{\bar{x}} / \sqrt{n}.$$

Voorbeeld

Van een blok zijn de afmetingen bepaald:

Lengte	x = 3,0	s_x = 0,1
Breedte	y = 2,5	s_y = 0,1
Hoogte	z = 1,2	s_z = 0,1

Gevraagd wordt de inhoud V en de standaardafwijking van V.

Het partiëel differentiëren kan men als volgt eenvoudig uitvoeren.

V = x . y . z.

	waarde	s	x+dx	y+dy	z+dz
Lengte	3,0	0,1	$x+s_x$= 3,1	3	3
Breedte	2,5	0,1	2,5	$y+s_y$= 2,6	2,5
Hoogte	1,2	0,1	1,2	1,2	$z+s_z$= 1,3
	x.y.z=		(x+dx).y.z=	x.(y+dy).z=	x.y.(z+dz)=
Volume	9,0		9,3	9,36	9,75
Vol.toename			$\frac{\partial V}{\partial x}s_x$ = 0,3	$\frac{\partial V}{\partial y}s_y$ =0,36	$\frac{\partial V}{\partial z}s_z$ =0,75

Variantie V $s^2_V = 0,3^2 + 0,36^2 + 0,75^2 = 0,7821$

Standaardafwijking V $s_V = \sqrt{0,7821} = 0,88$.

Wanneer men daarmee vertrouwd is, kan men dit soort berekeningen heel eenvoudig in een spread sheet uitvoeren.

AFRONDEN

Bij het opgeven van een analyseresultaat rijst de vraag hoe nauwkeurig dit resultaat moet worden opgegeven. Moderne rekenmachines geven zeer veel decimalen die voor de nauwkeurigheid niet meer relevant zijn. Bij handberekeningen bestond nogal eens de neiging de getallen tijdens de berekeningen eenvoudig te houden en maar flink af te ronden, waardoor onjuiste einduitkomsten konden ontstaan. Men moet dus niet te weinig maar ook niet te veel afronden.

Voor afronden worden daarom de volgende regels gehanteerd.

1. Het laatste cijfer moet nog juist significant zijn. Als maat hiervoor wordt door NEN 1047 de standaardafwijking of de spreidingsbreedte gebruikt.
2. De meetwaarde ligt in het interval van ½ eenheid beneden tot ½ eenheid boven dit laatste cijfer, dus

   7.80 betekent 7,795 - 7,805

   7,8 " 7,75 - 7,85.
3. Cijfers <5 worden naar beneden, >5 naar boven afgerond;
   5 zelf wordt naar het dichtstbijzijnde even getal afgerond,

7,55	7,65	7,75	7,85	7,95	afgerond op één decimaal wordt
7,6	7,6	7,8	7,8	8,0.	
4. Indien meer dan 1 decimaal komt te vervallen dient het afronden in één stap te geschieden. Dus:
   bij afronden op 0,1: 7,355 wordt 7,3

   niet: 7,355 wordt 7,35 wordt 7,4.

<u>Berekening van het afrondingsinterval volgens NEN 1047</u>

Deze berekening geschiedt in drie stappen.

1. Eerst wordt een *bovengrens*, b, berekend:

   $b = \frac{1}{2}\sigma$.

   Voorbeeld: $\sigma = 0,52$, dan is $b = 0,26$.
2. Hieruit wordt een *afrondingsinterval*, a, afgeleid. Kies a gelijk aan de grootste decimale eenheid (...;10; 1; 0,1; 0,01; ...) die b niet te boven gaat.

   Voorbeeld: b = 0,26, dan is a = 0,1.
3. Rond af op het dichtstbijzijnde veelvoud van het afrondingsinterval.
   In het voorbeeld dus afronden op tienden: 0,37 wordt 0,4.

Voorbeeld: een waarnemingsresultaat is x = 57,483 met een standaardafwijking s_x = 0,842. Gevraagd wordt x, s en het 95% betrouwbaarheidsinterval.

s = 0,842, dus b = 0,421, dus a = 0.1; dus afronden op 0.1.

x = 57,483 wordt 57,5.

s = 0,842 wordt 0,4.

Het betrouwbaarheidsinterval x ± 2 s ligt tussen

57,483 - 2*0,842 = 55,719 wordt 55,7

en 57,483 + 2*0,842 = 59,247 wordt 59,2.

NEN 5725 geeft ook voorschriften voor het geval dat de standaardafwijking of een redelijke schatting daarvan onbekend is. In dat geval werkt men met de *spreidingsbreedte*.

Stel er zijn k reeksen met elk n waarnemingen (bijvoorbeeld k duplobepalingen, n is dan 2).

1. Bereken voor iedere reeks de spreidingsbreedte w_i (i = 1,...,k), dat is het verschil tussen de hoogste en de laagste waarneming.
2. Bereken de gemiddelde spreidingsbreedte, $\bar{w}$: $\bar{w} = \Sigma w_i / k$.
3. Bereken b uit

$$b = \frac{\bar{w}}{2\sqrt{n}} .$$

Voorbeeld:

Gegeven de waarnemingsreeks (k=1,n=5)

8,72 7,55 8,01 8,25 8,83

De spreidingsbreedte bedraagt w = 8,83 - 7,55 = 1,28.

De bovengrens van het afrondingsinterval is b = 1,28 / (2 √5) = 0,29, zodat a = 0.1. De waarnemingen worden afgerond:

8,7 7,6 8,0 8,2 8,8.

HET VERGELIJKEN VAN MEER DAN TWEE POPULATIES

Indien men meer dan 2 populaties wil vergelijken op basis van steekproeven dan kan men voor elk tweetal steekproeven apart een Student-toets uitvoeren zoals in Hoofdstuk 8. Hierbij is het handig om de varianties van alle steekproeven te gebruiken voor de gepoolde binnen-steekproefvariantie s.

Stel dat men uit k populaties aselecte steekproeven getrokken heeft ter grootte n_1, n_2, ..., n_k . We veronderstellen dat de populaties $N(\mu_1,\sigma^2)$, $N(\mu_2,\sigma^2)$, ... resp. $N(\mu_k,\sigma^2)$ verdeeld zijn, d.w.z. de varianties zijn gelijk maar de gemiddelden kunnen verschillen. Het gemiddelde $\bar{X}_i$ van de i^e steekproef is een schatter voor μ_i. Voor het schatten van σ^2 'poolen' we de varianties S^2_1, S^2_2, ..., S^2_k:

$$S^2 = \frac{(n_1-1)*S^2_1 + (n_2-1)*S^2_2 + ... + (n_k-1)*S^2_k}{\Sigma n_i - k}$$

S^2 heeft Σn_i - k vrijheidsgraden.

Voor het vergelijken van μ_i en $\mu_{i'}$ kijken we naar het verschil $\bar{X}_i-\bar{X}_{i'}$, waarvoor we als standaardfout hanteren:

$$sed = se(\bar{X}_i-\bar{X}_{i'}) = S* \sqrt{(1/n_i + 1/n_{i'})}.$$

Dus de populaties i en i' verschillen significant bij onbetrouwbaarheid α=0.05 als

$$|\bar{X}_i-\bar{X}_{i'}| > lsd = t_\nu * sed ,$$

waarbij t_ν kan worden opgezocht in tabel III bij α= 0.05 en ν= Σn_i -k vrijheidsgraden. Merk op dat sed (en dus ook lsd) niet voor elk paar (i,i') gelijk is als de steekproefgrootten verschillen.

Voorbeeld

Stel dat in voorbeeld 1 uit Hoofdstuk 8 voor drie uienrassen A, B en C bij aselect getrokken bedrijven de opbrengst per ha in 1988 is geregistreerd. De opbrengsten zijn:

A 36, 47, 39, 43, 49, 38, 41, 51, 40, 44

B 45, 47, 34, 39, 31, 38, 41, 37, 43, 40

C 41, 47, 49, 54, 44, 45

De gemiddelden zijn $\bar{X}_A$ = 42.8, $\bar{X}_B$ = 39.5, $\bar{X}_C$ = 46.7 en de varianties:
S^2_A = 24.4, S^2_B = 23.6, S^2_C = 20.3.
Dus S^2 = 23.2 met 23 vrijheidsgraden.

Voor de paarsgewijze vergelijking van de 3 rassen vinden we

vergelijking	verschil	sed	lsd
A vs B	3.3	2.15	4.45
A vs C	-3.9	2.49	5.15
B vs C	-7.2	2.49	5.15

Hieruit volgt dat de opbrengst van rassen B en C significant van elkaar verschillen en die van de overige paren niet.

F-toets

Voor het vergelijken van meer dan 2 populaties wordt deze paarsgewijze vergelijking vaak vooraf gegaan door een zgn. F-toets. Dit is een globale toets die een uitspraak geeft over de vraag of er verschillen zijn tussen de gemiddelden van de k populaties. De F-toets toetst dus de nulhypothese H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$. Deze toets is gebaseerd op het splitsen van de totale variantie van de waarnemingen in variantie tussen steekproeven en variatie binnen steekproeven. Als we de waarnemingen schrijven als:

steekproef 1 : X_{11} , X_{12} , ... , X_{1n_1}

steekproef 2 : X_{21} , X_{22} , ... , X_{2n_2}

.....

steekproef k : X_{k1} , X_{k2} , ... , X_{kn_k}

en verder

$$n = \Sigma n_i, \qquad \bar{X}_{i.} = \frac{1}{n_i} \Sigma_j X_{ij} \qquad \text{en} \qquad \bar{X}_{..} = \frac{1}{n} \Sigma_i \Sigma_j X_{ij}$$

dan kan men de totale kwadraatsom van afwijkingen t.o.v. $\bar{X}_{..}$ (engels: Sum of Squares, SS) als volgt opsplitsen:

$$\Sigma_i \Sigma_j (X_{ij} - \bar{X}_{..})^2 = \Sigma_i \Sigma_j (\bar{X}_{i.} - \bar{X}_{..})^2 + \Sigma_i \Sigma_j (X_{ij} - \bar{X}_{i.})^2$$

ofwel:

$$SS_{totaal} = SS_{tussen} + SS_{binnen} .$$

De aantallen vrijheidsgraden behorende bij deze drie kwadraatsommen zijn: n-1, k-1, en n-k. Als we de kwadraatsommen delen door hun eigen aantal

vrijheidsgraden dan vinden we de gemiddelde kwadraten (Engels: Mean Squares, MS). Voor MS_{binnen} geldt dat deze gelijk is aan de hiervoor gedefinieerde gepoolde schatter S^2. Dus $E(MS_{binnen}) = \sigma^2$. Voor MS_{tussen} geldt dat diens verwachtingswaarde alleen onder H_0 gelijk is aan σ^2. Als echter de μ's sterk verschillen dan is $E(MS_{tussen})$ veel groter dan σ^2.
Als toetsingsgrootheid voor H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ hanteren we dus:

$$F = \frac{MS_{tussen}}{MS_{binnen}} .$$

Onder H_0 is F ongeveer gelijk aan 1. We verwerpen H_0 voor grote waarden van deze grootheid. F heeft onder H_0 een F-verdeling met k-1 en n-k vrijheidsgraden. De kritieke waarden kan men vinden in tabel VI (appendix). Omdat de toets bestaat uit het vergelijken van de variantie tussen steekproeven met de variantie binnen de steekproeven wordt deze toets vaak aangeduid met variantie-analyse (Engels: analysis of variance of ANOVA). Bovenstaande berekeningen kan men overzichtelijk samenvatten in de volgende ANOVA-tabel:

Variatiebron	aantal vr.gr.	kwadraat-sommen	gemiddelde kwadraten	F
tussen steek-proeven	k-1	SS_{tus}	$MS_{tus} = \frac{SS_{tus}}{k-1}$	$F = \frac{MS_{tus}}{MS_{bin}}$
binnen steek-proeven	n-k	SS_{bin}	$MS_{bin} = \frac{SS_{bin}}{k-1}$	
totaal	n-1	SS_{tot}		

In het voorbeeld met de 3 uienrassen ziet de ANOVA-tabel er als volgt uit:

variatiebron	df	SS	MS	F	P
tussen rassen	2	194.9	97.5	4.20	0.028
binnen rassen	23	533.4	23.2		
totaal	25	728.3			

De hier gerealiseerde F-waarde is 4.20 hetgeen groter is dan de kritieke waarde 3.42 (zie tabel VI). De conclusie is dus: er is een significant verschil tussen de opbrengsten van de 3 rassen.
Meestal is met deze F-toets niet het laatste woord gesproken vanwege het globale karakter ervan. Immers, als de 3 rassen verschillen, wil men weten welke rassen verschillen en welke niet. Een F-toets zal dus meestal gevolgd worden door paarsgewijze Student-toetsen zoals die in Hoofdstuk 8 zijn besproken.

3.1 Gegeven een aantal duplobepalingen

Serie	1	2	3	4	5	6
x_1	2,127	1,684	2,568	1,783	2,204	1,612
x_2	2,359	1,735	2,854	1,491	2,358	1,552

Rond deze af volgens NEN 1047.

3.2 Bereken voor het blok uit het voorbeeld bij 'doorwerken van afwijkingen' het oppervlak en de standaardafwijking daarvan.

3.3 Voor de controle op toevoeging van water aan vleesprodukten wordt de bepaling van het Federgetal gebruikt. Hiertoe wordt in het produkt het vocht-, vet-, as- en zetmeelgehalte afzonderlijk bepaald; het Federgetal is gedefiniëerd als

$$F = \frac{W}{100 - W - V - A - Z} ,$$

waarin W = vochtgehalte, V = vetgehalte, A = asgehalt, z = zetmeelgehalte.
Bij de analyse van 3 monsters werden de volgende resultaten verkregen.

	Monster 1			Monster 2		Monster 3		
Vocht	55,10	55,05	55,48	66,16	65,88	53,80	55,90	61,70
Vet	23,10	22,89	22,71	11,35	10,91	18,10	20,60	16,30
As	3,09	3,07		2,75	2,76	3,05	3,08	2,72
Zetmeel	5,58	5,62		4,05	4,12	4,90	4,50	3,80

Bereken voor deze drie monsters het Federgetal en de standaardafwijking daarvan.
De eis is dat het Federgetal niet groter dan 4 mag zijn. Toets of deze monsters met 95 % betrouwbaarheid aan deze eis voldoen.
Hoe groot is de kans dat deze monsters niet aan de eis voldoen?

14

INTERLABORATORIUMONDERZOEK IN DE PRAKTIJK

In dit deel wordt nader ingegaan op de stadia die doorlopen worden bij de voorbereiding, uitvoering en interpretatie van een zogenaamd Gemeenschappelijk Onderzoek (Engels: Collaborative Study). Een dergelijk onderzoek opgezet om bepaalde prestatiekenmerken van een omschreven methode te meten en te beschrijven.
Om de uitkomsten van zo'n onderzoek aanvaardbaar te doen zijn, dienen de omstandigheden waaronder het wordt uitgevoerd te beantwoorden aan in brede kring geaccepteerde condities.
Deze condities hebben betrekking op het aantal deelnemende laboratoria, de aantallen en aard van de onderzochte monsters, de toetsingscriteria om de geldigheid van verkregen uitslagen te testen en op de wijze waarop de reikwijdte daarvan wordt vastgesteld. Op deze punten en op het vlak van de bijbehorende statistische verwerking van de testresultaten geldt de internationale norm ISO 5725. Aangezien deze norm door het NNI is overgenomen, kan ook gesproken worden van NEN/ISO 5725.

1 Planning van het onderzoek

1.1 Verkenning van de methode

Mede afhankelijk van het uiteindelijke doel waarvoor de te onderzoeken methode is ontworpen, wordt allereerst vastgesteld, welke parameters verbonden aan de methode aan het eigenlijke Gemeenschappelijke Onderzoek zullen worden onderworpen. Het is mogelijk slechts een parameter te onderzoeken, maar meerdere simultaan zijn ook heel goed mogelijk.
Het ligt voor de hand dat het vaak om de bepaling van een gehalte van een bepaalde stof of verbinding in een bepaalde matrix zal gaan. Maar ook het meten van smeltpunten van legeringen, pH, (chemisch) zuurstofverbruik ((C)ZV) in oppervlaktewater om enkele andere minder voor de hand liggende analysen te noemen kunnen aan Gemeenschappelijk Onderzoek worden onderworpen.

Voorafgaande aan de uitvoering van het eigenlijke Gemeenschappelijke

Onderzoek dient de te onderzoeken methode een bepaalde mate van toetsing te ondergaan.
Dit wordt gedaan om te voorkomen dat veel moeite wordt gedaan om potentieel deelnemende laboratoria te interesseren, monsters te bereiden en schema's voor verwerking van gegevens op te zetten om dan tijdens het onderzoek te merken dat de methode te veel beperkingen kent, te veel afhangt ven persoonlijke vaardigheden of domweg te veel beïnvloedbaar is door toevalligheden.
De manier waarop een dergelijke toetsing wordt uitgevoerd hangt natuurlijk af van de aard van de methode en de te onderzoeken parameter(s).
Onder de aard van de methode worden de beginselen verstaan waarvan bij de uitvoering gebruik wordt gemaakt: fysisch-chemische als adsorptie/desorptie (o.m. bij chromatografie), oplosbaarheid (o.m. bij extracties), reactiviteit (o.m. bij bereiding derivaten).
Ook kan gedacht worden aan biologische beginselen zoals deze worden toegepast bij microbiologische analysemethoden (bepaling besmettingspatronen en -graden).
Evenzo bepalen de te onderzoeken parameters en de wijzen waarop deze gemeten worden welke stappen in het vooronderzoek onder de loep moeten worden genomen.

Nemen we als voorbeeld een methode, gericht op de bepaling van het gehalte van een bepaalde verbinding in een bepaalde matrix. In dit voorbeeld kan de methode achtereenvolgens de volgende stappen omvatten: extractie - derivatisering tbv gaschromatografische scheiding - gaschromatografische scheiding - detectie.

Enkele van de vragen die gesteld moeten worden zijn:
- is het beschreven extractiemiddel inderdaad optimaal?
- moet voor het extractiemiddel een bepaalde graad van zuiverheid worden voorgeschreven? Zo ja, is die kwaliteit goed verkrijgbaar?
- zijn extractievolumina, -tijden en -temperaturen wellicht kritisch?
- is het derivatiseringsreagens kritisch (verkrijgbaarheid, houdbaarheid), zijn de derivaten kritisch (houdbaarheid)?
- welk oplosmiddel moet worden gekozen tbv de injectie in de gaschromatograaf?
- is het gekozen type kolom voor de gaschromatograaf optimaal? Zijn er

wellicht onder bepaalde omstandigheden verontreinigingen in het extract aanwezig waardoor de kolom onbruikbaar wordt?
- zijn de gekozen temperaturen (injectie, eventueel gekozen oventemperatuurtraject, detectie) mogelijk kritisch?
- is de gekozen methode voor de detectie kritisch? Kunnen eventueel aanwezige verontreinigingen storen?

De antwoorden op deze vragen kunnen gevonden worden door de te testen methode verscheidene malen te herhalen en telkens op de geschetste punten kleine wijzigingen aan te brengen. Sterke en zwakke punten dienen dan naar voren te komen.
Een aspect wat ook aandacht verdient is het verkrijgen van voorinformatie over de mate van precisie die met de methode (vermoedelijk) kan worden behaald. Omdat het voorbereidend onderzoek binnen hetzelfde laboratorium wordt uitgevoerd betreft dit uiteraard alleen de herhaalbaarheid. Inzicht hierin wordt verkregen door, zonder daarin wijzigingen aan te brengen, de methode meerdere malen toe te passen op dezelfde monsterset.

Methoden die op dit punt slecht scoren verdienen in principe niet aan een Gemeenschappelijk Onderzoek onderworpen te worden tenzij er geen betere bestaan.

Vraag: wat is het doel van een Gemeenschappelijk Onderzoek?

1.2 Keuze werkwijze

Tenslotte moet een schatting worden gemaakt van het traject waarover de te onderzoeken parameter(s) van de methode tijdens het Gemeenschappelijk Onderzoek zal/zullen worden onderzocht.
Is het traject groot, dan dient daarmee bij het bereiden van de analyse-monsters rekening te worden gehouden in die zin, dat het van invloed is op het aantal niveau's ("levels") waarop het onderzoek zal worden uitgevoerd. Dit aantal niveau's moet dus in overeenstemming zijn met het beoogde traject; daarbij moet wel worden aangetekend dat het wellicht aantrekkelijk (en wenselijk!) is een mooi en groot experiment te ontwerpen, doch dat vervolgens kan blijken dat het vanwege de kosten niet uitvoerbaar is.

Als het traject groot is, dan zal bovendien kunnen blijken dat herhaalbaarheid r en reproduceerbaarheid R afhankelijkheid van het niveau m vertonen. Als vuistregel mag worden gebruikt dat in dergelijke gevallen ISO 5725 de toepassing van minstens zes niveau's aanbeveelt.

Bij een toets met q niveau's en n replica's zullen dus qn analysen worden uitgevoerd.
Bij de uitvoering van deze analysen moeten de deelnemende laboratoria zorgen dat aan een aantal voorwaarden wordt voldaan:

- de methode moet nauwkeurig volgens het afgesproken voorschrift worden uitgevoerd;
- alle qn analysen moeten door een en dezelfde analist worden uitgevoerd;
- bij de uitvoering moet gebruik worden gemaakt van dezelfde apparatuur;
- iedere groep van n analysen moet onder herhaalbaarheidscondities worden uitgevoerd, hetgeen inhoudt: door dezelfde analist, binnen een aaneengesloten tijdsbestek, zonder tussenliggende calibratie of instelling van apparatuur tenzij deze een integraal deel uitmaakt van de methode;
- het is niet noodzakelijk dat de q groepen van n tests strikt binnen een tijdsinterval worden uitgevoerd, dergelijke groepen mogen verspreid over enkele dagen worden afgewerkt;
- als de situatie zich voordoet dat een analist niet in staat is het gehele onderzoek af te ronden, mag een tweede die taak overnemen. Hierbij geldt als voorwaarde dat deze verandering niet mag plaats vinden binnen een groep van n analysen op een bepaald niveau j, doch alleen tussen de niveau's;
- alle n replica's moeten onder herhaalbaarheidscondities worden uitgevoerd. Dit houdt in dat de analysen moeten worden uitgevoerd als vonden zij plaats in verschillende materialen.

Split Level

De norm ISO 5725 geeft ter overweging om in bijzondere gevallen gebruik te maken van een zogenaamd "split level" experiment. In de daar aangehaalde regels voor het onderzoek is vermeld dat "de analysen moeten worden

uitgevoerd als vonden zij plaats in verschillende materialen".
De analist die het onderzoek uitvoert, moet dan ook dienovereenkomstig worden geinstrueerd. Het is denkbaar dat de leiding van het experiment verwacht dat desondanks ongewenste beinvloeding door voorwetenschap kan optreden. In dergelijke gevallen, en ook als het aantal replica's dat zal worden onderzocht twee is (zie par. 1.4), kan men besluiten van een "split level" experiment gebruik maken.
De opzet van een dergelijke experiment houdt in dat er voor de p deelnemende laboratoria twee series monsters a en b worden bereid waarbij de te toetsen parameter(s) tussen de series licht verschillen. Alle monsters a zijn identiek, evenals alle monsters b. Ieder laboratorium ontvangt een monster a en een monster b. De statistische verwerking van een dergelijk experiment wijkt af van de standaardmethode.
Om deze reden en omdat de te publiceren herziening van de norm ISO 5725 geen rekening meer houdt met deze wijze van werken, wordt bij de nadere toelichting van de statistische verwerking van de resultaten in hoofdstuk 2 niet nader op het "split level" systeem ingegaan.

Vraag: wat zou het gevolg van "voorwetenschap" bij een analist zijn?

1.3 Deelnemende laboratoria

Aan de deelnemende laboratoria moeten eisen worden gesteld.
Om te beginnen moet het aantal voldoende groot zijn, wil de informatie die wordt verkregen ook voldoende betrouwbaar zijn. Het gewenste aantal vertoont ook relatie met het aantal toe te passen toetsniveau's. Op dit punt geeft ISO 5725 enkele criteria:

- er dienen minstens acht deelnemers te zijn;
- als er slechts op één niveau wordt onderzocht, dient het aantal deelnemers minstens vijftien te zijn;
- deelnemende laboratoria moeten competent zijn op het vlak van de te testen methode;
- vanuit een statistisch standpunt gezien, moeten de deelnemende laboratoria zo veel mogelijk willekeurig uit het aantal potentiele kandidaten worden gekozen.

Vraag: welke gevolgen treden op als er slechts enkele deelnemers zijn?

1.4 Monster(voor)bereiding en verzending

Het materiaal dat wordt gebruikt voor de monsters die aan de deelnemende laboratoria worden verzonden, dient zo homogeen als mogelijk te zijn. Indien meerdere niveau's in het experiment zijn voorzien, dan geldt dit uiteraard voor elk niveau apart. Zonodig zorgt het organiserende laboratorium door middel van een (beperkte) monstervoorbehandeling ervoor dat aan deze voorwaarde wordt voldaan.
Tijdens de voorbereidende fase is reeds beoordeeld, of er aanleiding is bijzondere voorzorgen te nemen ten behoeve van de verpakking en de verzending van de monsters. Tijdens het transport mogen er uiteraard geen significante veranderingen in de monsters optreden die de uitslag van de toetsen zouden kunnen beinvloeden. Ook voorzorgen die het ontvangende laboratorium eventueel moet nemen na ontvangst of na opening van de verpakking van de monsters, zijn vanzelfsprekend onderzocht, vastgesteld en aan de deelnemers van het Gemeenschappelijk Onderzoek voorgeschreven.
Hierbij kan gedacht worden aan bewaringstemperatuur en het vermijden van lichtinvloeden, van verdampingsverschijnselen etc.
Ieder verzonden monster moet voor het ontvangende laboratorium op eenduidige wijze zijn benoemd of gecodeerd.

1.5 Rapportage

Naast de uitvoerige beschrijving van de methode krijgen de deelnemers in de toelichting op het programma aanwijzingen van de organisatoren hoe de verwachte rapportage er uit moet zien. Het kan daarbij aanbevelenswaardig zijn voorgedrukte formulieren mee te sturen, waarop de monsternummers of -codes staan voorgedrukt.
Ook zal worden afgesproken in welke eenheden de rapportage plaats zal vinden en de daarbij te betrachten nauwkeurigheid ("cijfers achter de komma". Ten onrechte wordt vaak gesproken van "het aantal decimalen"; met deze uitdrukking wordt het absolute aantal cijfers bedoeld ongeacht de plaats waar de eventuele komma staat!).

Tenslotte wordt ten behoeve van de rapportage afgesproken binnen welk tijdsbestek na ontvangst van de monsters de analysen moeten worden uitgevoerd en gerapporteerd en wat de rapportage dient te omvatten: ruwe data, chromatogrammen etc..

Het resultaat van alle bovengenoemde voorbereidingen moet uiteindelijk zijn: een zo goed mogelijk in zijn facetten uitgezochte methode, op zodanige wijze volledig en helder beschreven en vergezeld van dusdanig gedefinieerde monsters dat deze door een deskundig laboratorium zonder al te veel problemen kan worden nagewerkt.

2 Statistische verwerking

Inleiding

Het doel van de statistische verwerking van de verkregen informatie is een voor eenieder herkenbaar en betrouwbaar beeld te krijgen van de herhaalbaarheid en de reproduceerbaarheid van analyse-uitslagen die met de onderzochte methode kunnen worden verkregen.
Een dergelijke verwerking kan op meerdere manieren worden uitgevoerd. De norm ISO 5725 geeft voor de statistische interpretatie van met Gemeenschappelijk Onderzoek verkregen gegevens uitgebreide aanwijzingen. Zoals reeds in hoofdstuk 10 werd besproken, is de norm gebaseerd op het volgende statistische model:

$$y = m + B + e$$

waarin:

y = uitslag van een enkel analyseresultaat

m = algemeen gemiddelde, geldend voor een dergelijke bepaling, uitgevoerd met de te toetsen methode

B = de tussen-laboratorium variatie

e = de toevallige fout die in elke bepaling kan optreden

De factor B kan geacht worden constant te zijn indien analysen onder herhaalbaarheidscondities worden uitgevoerd. Onder reproduceerbaarheids-

condities daarintegen, gedraagt deze factor zich als een grootheid met willekeurige waarde.

De factor e representeert de willekeurige fout die in elke analyse onder alle omstandigheden voorkomt.

Het is het doel van de statistische analyse om het gedrag van e en B te meten. Samen bepalen deze factoren de precisie waarmee een analyse kan worden uitgevoerd.
De herhaalbaarheid van een methode wordt gebaseerd op het gedrag van e, de reproduceerbaarheid op het gedrag van e en B samen.

Alvorens de gegevens uit het Gemeenschappelijk Onderzoek in mathematische modellen kunnen worden ingevoerd om de grootheden herhaalbaarheid en reproduceerbaarheid te berekenen, moeten de gegevens onderzocht worden, want niet alle ingeleverde cijfers zijn zomaar voor dat doel bruikbaar. Een aantal redenen kan worden gevonden om gerapporteerde waarden niet voor de berekening van de eindresultaten te gebruiken. Dat kan in de volgende situaties mogelijk zijn:

- de uitvoerende laboratoria kunnen kennelijke fouten hebben gemaakt, bijvoorbeeld door zich bij berekeningen factoren 2, 5, 10 etc. te vergissen;
- de laboratoria kunnen monsters kennelijk hebben verwisseld;
- een laboratorium heeft zich niet aan het analysevoorschrift gehouden, de monsters verkeerd behandeld etc.;
- zonder dat herkenbare, kennelijke vergissingen zijn begaan, lopen door een laboratorium gerapporteerde individuele waarden soms toch ten opzichte van de overige deelnemers meer uiteen.
  Acceptatie van dergelijke waarden kan de te berekenen herhaalbaarheid ontoelaatbaar beinvloeden;
- een laboratorium kan ten opzichte van de overige deelnemers gemiddeld op een afwijkend niveau liggen.
  Acceptatie van dergelijke waarden kan de te berekenden reproduceer-baarheid ontoelaatbaar beinvloeden.

Door de verkregen waarnemingen te vergelijken, kunnen eerst de kennelijke

fouten worden opgespoord. In veel gevallen zal het mogelijk zijn door navraag de foutenbron op te sporen en de waarnemingen te corrigeren; waar dat niet ondubbelzinnig mogelijk is, moeten evident foute waarnemingen onherroepelijk uit het gegevensbestand worden verwijderd (niet bij de berekeningen worden gebruikt).
Daartoe worden toetsen toegepast om extreme resultaten op te sporen. Op grond van vergelijking met de bijbehorende toetscontrolewaarden wordt vastgesteld of die extremen statistisch gezien wel of niet tot de populatie behoren (bij de berekeningen mogen worden gebruikt).

Tenslotte worden door middel van variantie-analyse de herhaalbaarheid, de reproduceerbaarheid en de eventuele niveau-afhankelijkheid van beide berekend.

2.1 Verwijderen van extreme resultaten

Extremen in de gerapporteerde analysewaarden kunnen meerdere oorzaken hebben. Naast kennelijke fouten, die vaak gemakkelijk te herkennen zijn, kunnen de oorzaken grofweg in twee categorien worden onderverdeeld: systematische fouten en toevallige fouten. Beide kunnen ook gelijktijdig voorkomen.
In het eerste geval zal het gemiddelde resultaat van het betreffende laboratorium sterk afwijken van dat van de overige; in het tweede geval zal de berekende standaardafwijking in de resultaten verhoudingsgewijs groter zijn.

Opgave: *noem enkele oorzaken voor systematische fouten;*
waar ligt de voornaamste oorzaak van toevallige fouten?

Meetwaarden die aanleiding zijn tot te grote toevallige of systematische afwijkingen worden opgespoord met resp. de Cochran- en de Dixontoets.
Daartoe moet een aantal berekeningen op het gegevensbestand worden losgelaten. Deze berekeningen worden voor ieder niveau separaat uitgevoerd; dit betekent dat de resultaten van laboratoria voor sommige niveau's wel, en voor andere niveau's niet acceptabel zouden kunnen worden bevonden.

De formules die in het navolgende worden genoemd kunnen gevonden worden in de Appendix 3. De Tabellen VII en VIII bevatten toetswaarden voor resp. de Cochran- en de Dixontoets.

De volgende indices worden gebruikt:

i = aanduiding van een bepaald laboratorium
p = aantal laboratoria dat aan het Gezamenlijk Onderzoek deelneemt
j = aanduiding van het niveau dat wordt geanalyseerd
n = aantal replica's dat op ieder niveau wordt geanalyseerd
k = aanduiding van een bepaalde replica in de set van n replica's
H = totaal aantal waarden dat in de Dixontoets wordt getest
C = Cochran-toetswaarde
Q = Dixon-toetswaarde
s = standaardafwijking (in de gebruikte formules de zogenaamde "geschatte standaardafwijking")
r = herhaalbaarheid
R = reproduceerbaarheid
L = tussen-laboratorium

2.1.1 De Cochrantoets

Om significante afwijkingen voortkomende uit te grote toevallige fouten op te sporen, wordt de zogenaamde eenzijdige Cochrantoets op de berekende standaardafwijkingen in de resulaten van de deelnemende laboratoria toegepast. Een situatie met grote toevallige fouten manifesteert zich door een grote spreiding tussen de resultaten, verkregen met de replica's. Een grote spreiding betekent een grote waarde voor de standaardafwijking in de resultaten van het bewuste laboratorium.

Is de toets bevestigend, dan wordt geconcludeerd dat de resultaten van dat laboratorium de homogeniteit in de uitslagen van de het onderzoek als zodanig verstoren en worden ze uit het gegevensbestand verwijderd.
De te doorlopen stappen komen - voor ieder niveau - op het volgende neer:

stap 1 Bereken voor ieder laboratorium de standaardafwijking in de

gerapporteerde resultaten van de n replica's. Daarvoor wordt formule 1 of formule 2 gebruikt.

stap 2 Bereken de Cochran-waarde C volgens formule 3. Hierbij is de teller gelijk aan het kwadraat van de grootste gevonden individuele waarde voor de voor ieder laboratorium berekende standaardafwijking (extreem); de noemer is de som van de kwadraten van de berekende standaardafwijkingen van de p laboratoria samen. Het nummer van het laboratorium waar de waarden, die de grootste berekende standaardafwijking opleverden werd gevonden, wordt genoteerd

stap 3 Vergelijk vervolgens de gevonden waarde C met de zogenaamde kritische waarden die voor de Cochrantoets gelden. Deze zijn in appendix 2 gerubriceerd voor het 95% resp. 99% betrouwbaarheidsniveau (= foutkans resp. 5% en 1%), voor waarden van p = 2 - 40 en n = 2 - 6.

stap 4 De volgende waarnemingen kunnen worden gedaan en daarop gebaseerde conclusies worden getrokken:

- De gevonden waarde voor C < kritische waarde 5%.
  De uitslagen van het laboratorium met het gevonden extreem behoren statistisch gezien wel tot de populatie.

- kritische waarde 5% < C < kritische waarde 1%.
  Dit wil zeggen dat de berekende toetswaarde tussen het 95% en het 99% betrouwbaarheidsniveau ligt. De kans dat de waarden van het bewuste laboratorium als onbetrouwbaar (= niet tot de populatie behorend) moeten worden beschouwd is statistisch gezien "aannemelijk". In statistische termen wordt een dergelijk laboratorium een "straggler" genoemd. De consequenties zijn dat dergelijke waarnemingen nauwkeurig op een mogelijke verklaring moeten worden onderzocht.

  Is die er (denk aan foutieve berekeningen, verwisselingen) dan kunnen eventueel correcties worden aangebracht. Zo niet,

dan kunnen de waarnemingen alsnog worden verworpen (zie par. 2.1.2 en 2.1.3).

- C > kritische waarde 1%.
  De meetwaarden van het laboratorium waarvoor de testwaarde C werd berekend, moeten statistisch als bewezen onbetrouwbaar en niet behorend tot de populatie worden beschouwd (een "outlier"). De waarden (en daarmee de uitslagen van het laboratorium dat ze aanleverde) moeten uit het gegevensbestand worden verwijderd.

stap 5 Indien meetwaarden moesten worden verwijderd wordt de toets vanaf stap 2 herhaald totdat er geen statistisch significante extremen meer worden gevonden.

2.1.2 De Dixontoets

De Dixontoets wordt als volgt toegepast:

- Als de Cochrantoets een niet te verklaren "straggler" heeft aangewezen.
  De toets wordt toegepast op de individuele gemeten waarden voor de replica's van dat laboratorium, mits n >= 3.
  Op deze wijze toegepast kan met de toets inzicht worden verkregen of een (of meer) replica-uitslag(en) in het bijzonder aangewezen kan (kunnen) worden voor een met de Cochrantoets gesignaleerde extreme spreiding;

- Op de gemiddelden van de waarden van de replica's van ieder laboratorium, mits p >= 3.
  Op deze wijze toegepast kan met de toets inzicht worden verkregen of er laboratoria zijn met een te grote systematische afwijking ten opzichte van de andere laboratoria.

De Dixontoets toetst de hoogste of de laagste waarde uit de aan de toets onderworpen set waarden. De keuze tussen deze twee is gebaseerd op de respectievelijke mate van afwijking ten opzichte van de overige waarden.

Uitvoering

De te toetsen waarden worden in oplopende grootte gerangschikt.
Gebruikt worden: de waarden van de replica's van een laboratorium dat in de Cochrantoets als "straggler" is aangemerkt of, indien de Cochrantoets geen afwijkingen (meer) rapporteert, de gemiddelde waarden van de replica's van alle laboratoria.

Indien het aantal te toetsen waarden 3-7 bedraagt, worden de uitkomsten Q van formule 4a en 4b berekend. De hoogste van de twee levert de te toetsen waarde op. Evenzo worden bij te toetsen aantallen van 8-12 de formules 5a en 5b gebruikt en bij 13 of meer de formules 6a en 6b. De te toetsen waarde is telkens de hoogste van de twee uitkomsten.

De formules 5 en 6 zullen, gezien het grote aantal te toetsen waarden, vooral worden benut als gemiddelden worden getoetst.

De volgende waarnemingen kunnen worden gedaan en daarop gebaseerde conclusies worden getrokken:

- De gevonden waarde voor Q < de kritische waarde 5%.
  De getoetste replica-uitslag of het gemiddelde van de set replica-uitslagen van het laboratorium behoort statistisch gezien wel tot de populatie.

- kritische waarde 5% < Q < kritische waarde 1%.
  Dit wil zeggen dat de getoetste waarde tussen het 95% en het 99% betrouwbaarheidsniveau ligt. De kans dat die waarde als onbetrouwbaar (niet tot de populatie behoort) is "aannemelijk".

  In statistische termen wordt een dergelijke waarde evenals bij de Cochrantoets een "straggler" genoemd. De consequentie is dat dergelijke waarnemingen nauwkeurig op mogelijke verklaringen van de afwijking moeten worden onderzocht.

  Zijn die er (denk aan foutieve berekeningen, verwisselingen) dan kunnen eventueel correcties worden aangebracht. Zo niet, dan kunnen de

waarnemingen alsnog worden verworpen. (zie ook par. 2.1.3)

- Q > kritische waarde 1%.
  De getoetste waarde waarvoor de testwaarde Q werd berekend, moet statistisch als bewezen onbetrouwbaar en niet behorend tot de populatie worden beschouwd. De getoetste replicawaarde c.q. alle replicawaarden van het getoetste laboratorium moet(en) uit het gegevensbestand worden verwijderd ("outlier(s)").

Indien meetwaarden moesten worden verwijderd wordt de Dixontoets herhaald totdat er geen statistisch significante extremen meer worden gevonden.

Vraag: *waarom moeten de Cochran- en Dixontoets worden herhaald als er resultaten uit het gegevensbestand zijn verwijderd?*

2.1.3 Opmerkingen bij de uitvoering van de toetsen

Er dient zorgvuldig te worden beoordeeld of meetwaarden, die met de Cochran of Dixontoets als verdacht ("straggler") of verwerpbaar ("outlier") worden aangemerkt, ook inderdaad moeten worden verworpen. Vooral indien zoals is aangeven de toetsen worden herhaald en er voortdurend nieuwe aanwijzingen voor afwijkende waarden worden gevonden.

Als vuistregel geldt dat, indien waarnemingen moeten worden verwijderd, dit niet mag leiden tot een reductie in aantal meetellende laboratoria (per niveau!) van minder dan 80% van het oorspronkelijk aan het Gemeenschappelijk Onderzoek deelnemende aantal.
Als zich de situatie voordoet dat dit op grond van de toetsen wel het geval zou zijn, dan is het verstandig in eerste instantie toch alle gerapporteerde waarden in de berekeningen voor de herhaalbaarheid en de reproduceerbaarheid mee te nemen. In tweede instantie moet vervolgens nauwkeurig onderzocht worden of de waargenomen discrepanties uit methodologische problemen kunnen worden verklaard en of mogelijk de methode zelf moet worden herzien.

2.1.4 Afrondingen

Sinds zakrekenmachines en computers hun intrede hebben gedaan, lijken uitkomsten van berekeningen met steeds grotere cijferreeksen te kunnen worden gepresenteerd. Uiteraard is dit maar schijn. De nauwkeurigheid van de uitkomst van een bewerking op een of meer getallen hangt af van de nauwkeurigheid van die getallen zelf.

Bij de presentatie van de uitslag van de berekening van gemiddelden mag een decimale plaats meer worden gebruikt dan de getallen die daarvoor als basis dienen; dit geldt ook voor berekende standaardafwijkingen.

Voor de berekening van deze standaardafwijkingen zijn in Appendix 3 twee formules gegeven: formule 1 en 2. Het verschil tussen beide is dat ten behoeve van het verminderen van de hoeveelheid rekenwerk in formule 1 gebruik wordt gemaakt van een tevoren berekend gemiddelde $\bar{y}_i$, waarbij, gekoppeld aaan het berekenen van het verschil tussen y_{ik} en $\bar{y}_i$, bij kleine verschillen soms aanmerkelijke afrondingsfouten kunnen ontstaan.

Formule 2 gaat uit van de feitelijke definitie en kent dit probleem niet, doch is duidelijk bewerkelijker bij het gebruik wanneer geen rekenapparatuur ter beschikking staat.

2.1.5 Ontbrekende waarden

In principe mogen de Cochran- en de Dixontoetsen alleen worden toegepast als voor alle te toetsen laboratoria (binnen een onderzocht niveau) het aantal replica's n gelijk is. Het is evenwel toegestaan om de standaardafwijkingen die voor dergelijke toetsen moeten worden berekend, te baseren op niet gelijke aantallen n.

Deze ongelijkheid kan ontstaan tengevolge van schrappen van uitslagen (kennelijke fouten), het ontbreken van uitslagen of tengevolge van het schrappen van waarden na toepassing van de Dixontoets.

Het omgekeerde kan zich voordoen indien een laboratorium meer dan het

afgesproken aantal waarden voor replica's indient. In dat geval beveelt ISO 5725 aan ook deze waarden te benutten, daar zij bijdragen tot de vergroting van de betrouwbaarheid van de uiteindelijke conclusies.

2.2 Berekenen van het gemiddelde, de herhaalbaarheid en de reproduceerbaarheid

Per onderzocht niveau worden het gemiddelde en respectievelijk de herhaalbaarheid en de reproduceerbaarheid bij dat gemiddelde berekend.

Vraag: wat is waar:
a de waarde voor de herhaalbaarheid kan groter zijn dan die voor de reproduceerbaarheid
b de waarde voor de reproduceerbaarheid is nooit kleiner dan die voor de herhaalbaarheid

Voor de berekeningen wordt voor ieder niveau j het volgende schema gebruikt (zie ook Appendix 3):

---

aantal laboratoria: p $p = \ldots$

aantal replica's : n $n = \ldots$

$T1 = \Sigma\bar{y}_i$ $T_1 = \ldots$

$T2 = \Sigma\bar{y}_i^2$ $T_2 = \ldots$

$T3 = \Sigma s_i^2$ formule 1 of 2

$$s_r^2 = \frac{T_3}{p}$$ formule 7

$$s_L^2 = \frac{p * T_2 - T_1^2}{p * (p - 1)} - \frac{s_r^2}{n}$$ formule 8

$s_R^2 = s_L^2 + s_r^2$ $s_R^2 = \ldots$

$$m = \frac{T_1}{p}$$ $m = \ldots$

$r = 2{,}8 \sqrt{s_r^2}$ $r = \ldots$

$R = 2{,}8 \sqrt{s_R^2}$ $R = \ldots$

2.3 Beoordelen niveau-afhankelijkheid van de herhaalbaarheid en de reproduceerbaarheid

Als in een te toetsen methode een bepaalde parameter over een behoorlijk traject kan worden bepaald, dan zullen in een Gemeenschappelijk Onderzoek monsters worden onderzocht op verschillende niveau's. Het is niet onwaarschijnlijk dat blijkt dat de te berekenen herhaalbaarheid en reproduceerbaarheid niet op elk onderzocht parameterniveau van gelijke orde zijn. In dergelijke gevallen moet onderzocht worden of er een relatie is tussen het niveau waarop de parameter wordt bepaald en de bijbehorende herhaalbaarheid en reproduceerbaarheid. Kan een dergelijke relatie worden vastgesteld, dan kan deze niveaurelatie in de praktijk worden toegepast bij de beoordeling of bepalingen van replica's binnen de geldende herhaalbaarheid zijn gevonden. Evenzo kan de niveauafhankelijkheid van de reproduceerbaarheid worden benut.

Verscheidene relatiemodellen kunnen in formulevorm worden beschreven. De norm ISO 5725 gebruikt er drie:

type I, een rechte lijn door de oorsprong
r of $R = b * m$

type II, een rechte lijn met intercept
r of $R = a + b * m$

type III, een exponentiele relatie
$\log (r)$ of $\log (R) = c + d \log (m)$
equivalent aan:

r (of R) $= 10^c * m^d$

waarin bij alle vergelijkingen m de waarde is van het niveau van de analyse-uitslag

Door voor ieder der drie modellen de factoren b, a, c en d uit te rekenen, kunnen vervolgens grafisch de resulterende krommen r of R tegen de waarde van het niveau m worden uitgezet.

In het algemeen geldt dat indien de modellen op basis van de berekende factoren min of meer horizontaal blijken te lopen, er van afhankelijkheid geen sprake is.
Vertonen de modellen enige helling en blijken de werkelijk waargenomen waarden min of meer te corresponderen, dan wordt beoordeeld welke der drie modellen de beste is.
In het algemeen is het gemak waarmee dat mogelijk is afhankelijk van het aantal niveau's waarop de modellen zijn gebaseerd: vier is daarvoor eigenlijk wel een minimum. (Dit is tevens een reden waarom ISO 5725 in dergelijke gevallen zes te toetsen niveau's adviseert!).

Opgave: bedenk een tweede reden waarom niveau-afhankelijkheid bij meer dan twee niveau's moet worden onderzocht.

2.3.1 Berekeningswijze

Vanuit statistisch standpunt gezien, is het ontwikkelen van een rechte lijn (type I en type II) enigzins gecompliceerd omdat zowel r (of R) als m tijdens het experiment worden bepaald. Er moet dus een relatie worden bepaald tussen twee variabelen die beide onderhevig zijn aan fouten.
De fout in m telt minder en kan daarom worden genegeerd, omdat in het algemeen de steilheid van de lijnen (factor b) niet groot is.

De invloed van de onzekerheid in r (of R) is echter van dien aard, dat een correctie moet worden aangebracht in de vorm van een weging. Weging houdt in dat naarmate de waarde van r of R groter is, de invloed daarvan bij de berekening van de factoren voor de formule minder wordt meegeteld.

De relatie type III is door zijn exponentiele aard niet onderhevig aan de hierboven geschetste invloeden. Hierbij kan weging dus achterwege blijven.

2.3.1.1 Rechte lijn door oorsprong

Bij algebraische uitwerking van de formules voor de bepaling van de factor b blijkt in dit geval de weegfactor te kunnen worden geelimineerd. Er resulteert een vereenvoudigde formule (formule 11).

2.3.1.2 Rechte lijn met intercept

In dit geval moet bij de berekening van de factoren b en a wel van een weegfactor gebruik worden gemaakt. Van deze weegfactor worden samengestelde factoren afgeleid, die in eerste instantie worden gebaseerd op de in het Gemeenschappelijk Onderzoek gevonden waarden voor r resp. R en m.
Er resulteert een eerste benadering voor de vergelijking:

r of R = a + b * m.

Ter verhoging van de nauwkeurigheid waarmee b en a worden bepaald, wordt de berekening nog eenmaal herhaald, waarbij ditmaal in de formules voor de samengestelde berekeningsfactoren waarden voor r en R worden ingevoerd die worden verkregen door de voor de niveau's berekende waarden van m te substitueren in de vergelijking:
r of R = a + b * m in eerste benadering.

Er wordt gebruik gemaakt van de weegfactor $W = 1/r_j^2$ resp. $1/R_j^2$, waarbij j het niveau is waarvoor r resp R werden bepaald.

Uit deze weegfactor worden vijf samengestelde berekeningsfactoren afgeleid:

$F1 = \Sigma W_j$

$F2 = \Sigma m_j * W_j$

$F3 = \Sigma m2_j * W_j$

$F4 = \Sigma r_j * W_j$

$F5 = \Sigma m_j * r_j * W_j$

(bovenstaande berekeningsfactoren F1 t/m F5 worden uiteraard ook op dezelfde wijze berekend voor R in plaats van r).

Uit deze samengestelde berekeningsfactoren F1 t/m F5 worden de resp. formules voor de factoren b en a afgeleid (formules 12 en 13).

Nadat b en a berekend zijn, worden de respectievelijke waarden voor de niveau's m in de vergelijking:
r of R = a + b * m ingevuld en worden r1 resp R1 voor ieder niveau van m berekend.
Deze waarden r1 en R1 worden weer gebruikt om opnieuw W en F1 t/m F5 te bepalen, waarna de formules 12 en 13 opnieuw worden toegepast.
Dit proces wordt iteratie genoemd. We zouden de iteratie nog kunnen voortzetten, maar de praktijk leert dat in het algemeen geen belangrijke verbetering in de nauwkeurigheid
van de factoren b en a meer wordt bereikt.

2.3.1.3 Exponentiele relatie

Zoals eerder vermeld behoeft in dit geval geen weging te worden toegepast.
Voor de berekening van de factoren c en d worden vier samengestelde berekeningsfactoren gebruikt:

$G1 = \Sigma \log (m_j)$

$G2 = \Sigma (\log (m_j))^2$

$G3 = \Sigma \log (r_j)$

$G4 = \Sigma (\log (m_j)) * (\log (r_j))$

Voor de berekening van c geldt formule 14, voor de berekening van d geldt formule 15.

(bovenstaande berekeningsfactoren G1 t/m G4 worden uiteraard ook op dezelfde wijze berekend voor R in plaats van r)

2.3.2 Voorbeeld van de berekening van de afhankelijkheid van r van het niveau m

De gegevens in dit voorbeeld zijn fictieve gegevens, die verkregen zouden kunnen zijn uit een Gemeenschappelijk Onderzoek.

niveau	1	2	3	4	5
m_j	3,94	8,28	14,18	15,59	20,41
r_j	0,258	0,501	0,355	0,943	1,102

model type I:

niveau	1	2	3	4	5
r_j / m_j	0,065	0,060	0,025	0,060	0,054

Na invulling van deze waarden in formule 11 uit Appendix 3 kan berekend worden:

$$b = \frac{0,265}{5} = 0,053$$

waaruit volgt:

$$r = 0,053 * m$$

Model type II

niveau	1	2	3	4	5
W_{0j}	15	4,0	7,9	1,1	0,82

Na invulling van deze waarden in de formules voor F1 t/m F5 en na gebruikmaking van de formules 12 en 13 uit appendix 3 kan berekend worden:

$$r = 0,161 + 0,025 * m$$

Invullen van de waarden van m in deze formule voor r in eerste aanleg geeft:

niveau	1	2	3	4	5
r_{1j}	0,260	0,369	0,517	0,552	0,673
W_{1j}	15	7,3	3,7	3,3	2,2

Na invulling van deze waarden in de formules voor F1 t/m F5en na gebruikmaking van de formules 12 en 13 uit Appendix 3 kan berekend worden:

$r = 0,085 + 0,043 * m$

Een 2e iteratie zou geven:

niveau	1	2	3	4	5
r_{2j}	0,257	0,446	0,703	0,765	0,975
W_{2j}	15	5,0	2,0	1,7	1,0

Na invulling van deze waarden in de formules voor F1 t/m F5en na gebruikmaking van de formules 12 en 13 uit Appendix 3 kan berekend worden:

$r = 0,090 + 0,043 * m$

Overeenkomstig eerder gesteld blijkt deze tweede iteratie overbodig.

Model type III

niveau	1	2	3	4	5
log (m_j)	+0,595	+0.918	+1,152	+1,193	+1,350
log (r_j)	-0,588	-0,300	-0,450	-0,025	+0,042

Na invulling van deze waarden in de formules voor G1 t/m G4en na gebruikmaking van de formules 14 en 15 uit Appendix 3 kan berekend worden:

$$\log (r) = -1,057 + 0,767 * \log (m)$$

Deze logarithmische formule voor r kan herschreven worden als:

$$r = 10^{-1,057} * m^{0,77} = 0,088 * m^{0,77}.$$

Bij substitutie van de oorspronkelijke waarden van m levert deze formule de volgende waarden voor r_j op:

niveau	1	2	3	4	5
r_j	0,253	0,448	0,678	0,729	0,898

Bekijken we de in het Gemeenschappelijk Onderzoek gevonden originele waarden van r voor de vijf niveau's m_j, dan valt makkelijk te concluderen dat bij toenemende m, ook een toenemende r hoort en dat er dus sprake is van niveau-afhankelijkheid.

Een en ander wordt nader verduidelijkt in figuur 1. Hierin zijn de drie typen modellen voor r als functie van m uitgezet. De origineel waargenomen waarden voor r zijn in de figuur met een plusje (+) aangegeven.

Figuur 1 Relatie r versus m type I, II en III

Na beoordeling van de mate waarin de origineel waargenomen waarden voor rj bij de vijf gemeten niveau's van m passen op de berekende krommen, kan geconcludeerd worden dat de lijn door de oorsprong (type I) het best past.

Een geheel vergelijkbare berekening wordt gevolgd bij de beoordeling van de afhankelijkheid van R van m.

3 Uitgewerkt voorbeeld van een Gemeenschappelijk Onderzoek

In dit voorbeeld wordt een methode ten tonele gevoerd, ontworpen voor de bepaling van het element "TEST" in monsters met de matrix "cursus". De methode berust op een eenvoudige extractie van de monsters, gevolgd door indampingvan het extract, waarna het residu bestaande uit een zuivere hoeveelheid "TEST" wordt gewogen. Het bereik van de bepalingis gering, vandaar dat alle monsters op een niveau worden onderzocht.

3.1 Uitgangsgegevens

Er nemen aan het onderzoek vier laboratoria deel, die ieder drie monsters "cursus" te analyseren krijgen. Alle monsters zijn volkomen homogeen en identiek.
Opmerking: ter wille van de eenvoud van het voorbeeld is hierbij voorbij gegaan aan de aanbeveling van ISO 5725 om tenminste vijftien laboratoria in te schakelen.

De deelnemende laboratoria rapporteren de volgende resultaten (in grammen):

lab. no.1	11	15	17
lab. no.2	12	19	17
lab. no.3	10	8	11
lab. no.4	8	7	12

3.2 Uitwerking

Volgens het schema van par. 2.4 kunnen de volgende kenwaarden worden berekend:

$p = 4$
$n = 3$
$T1 = 49,0$
$T2 = 635,9$
$T3 = 31,7$
$s^2_r = 7,9$
$s^2_L = 9,2$
$s^2_R = 17,2$
$m = 12,2$
$r = 7,9$
$R = 11,6$

Individuele waarden voor de laboratoria zijn:

lab nr	waarden			gemiddelde	standaard-afwijking	n
1	11	15	17	14,3	3,1	3
2	12	19	17	16,0	3,6	3
3	10	8	11	9,7	1,5	3
4	8	7	12	9,0	2,6	3

Worden de in de bovenstaande tabel weergegeven afgeronde waarden voor de per laboratorium berekende standaardafwijkingen gebruikt, dan luidt de berekening van de waarde C voor de Cochrantoets:

$$(3,6)^2 / \{(3,1)^2 + (3,6)^2 + (1,5)^2 + (2,6)^2\} = 0,38$$

Deze waarde is gekoppeld aan laboratorium no.2.
In Tabel VII staan voor p = 4 en n = 3 de volgende kritische toetswaarden voor de Cochrantoets getabelleerd:

5% : 0,768
1% : 0,864

De berekende toetswaarde C is ruimschoots kleiner dan de kritische toetswaarde voor 5% onzekerheid. Dit betekent dat alle gerapporteerde waarden als homogeen kunnen worden beschouwd, er zijn geen "outliers" of "stragglers".

Aangezien er geen "stragglers" zijn gevonden met de Cochran-toets, behoeven er geen replicasets met de Dixontoets worden geanalyseerd. Wel kan de Dixontoets worden toegepast op de berekende gemiddelden om te zien of een der laboratoria een te grote systematische afwijking vertoont ten opzichte van de andere.

Het aantal te toetsen waarden bedraagt vier, daarom worden formules 4a en 4b uit de Appendix 3 gebruikt.
De toets levert het navolgende beeld op.

lab. no.	4	3	1	2
gemiddelde	9,0	9,7	14,3	16,0

formule 4a: (9,7 - 9,0) / (16,0 - 9,0) = 0,10
toetsen van laagste lab.: no. 4

formule 4b: (16,0 - 14,3) / (16,0 - 9,0) = 0,24
toetsen van hoogste lab.: no. 2

Het gemiddelde, bepaald door laboratorium no. 2 vertoont verhoudingsgewijs het grootste verschil met de overige laboratoria en wordt dus met de Dixontoets beoordeeld.

Uit de Tabel VII kunnen voor H = 4 de volgende kritische waarden voor de Dixontoets worden afgelezen:

5% : 0,829
1% : 0,926

De berekende toetswaarde Q = 0,24 is ruimschoots kleiner dan de kritische waarde voor 5% onzekerheid. Dit betekent dat laboratorium no. 2 statistisch gezien niet systematisch afwijkt van de drie andere.

3.3 Eindrapportage

De eindrapportage van het Gemeenschappelijk Onderzoek luidt: "met de methode 'analyse van TEST in de matrix Cursus' kan het element TEST met de volgende precisie worden bepaald:

m = 12,2
r = 7,9
R = 11,6".

In praktisch opzicht betekent de uitslag van dit Gemeenschappelijk Onderzoek dat, wanneer een laboratorium het element "TEST" op het niveau van ca 10 g in duplo bepaalt in de matrix "cursus" en de bepaling onder herhaalbaar-heidscondities uitvoert, het verschil tussen beide bepalingen niet groter mag zijn dan 7,9 gram. Is dat wel het geval, dan is de analyse kennelijk niet juist uitgevoerd.
Wordt dezelfde analyse in hetzelfde monster door een ander laboratorium met gebruikmaking van dezelfde methode uitgevoerd (bijvoorbeeld in het geval van arbitrage), dan mag het verschil tussen wat het ene laboratorium vindt en wat het andere vindt maximaal 11,6 gram bedragen. Is het verschil groter, dan bestaat het vermoeden dat een van hen of beide de bepaling onjuist hebben uitgevoerd.

3.4 Beoordeling van de resultaten

Als we naar de kwaliteit van de resultaten van dit voorbeeld van een Gemeenschappelijk Onderzoek met een bepalingsmethode van het element "TEST" in "cursus" kijken, dan valt op dat de gemeten herhaalbaarheid en reproduceerbaarheid van de analysemethode op het getoetste niveau bepaald niet gunstigzijn. Voor een belangrijk deel is dit zonder meer het gevolg van het nog al uiteenliggen van de zogenaamd gerapporteerde analyseresultaten; de methode is dus niet erg precies. Vooreen ander deel is dit het gevolg van de beperkte opzet die in het gekozen voorbeeld voor het Gemeenschappelijk Onderzoek is gevolgd: slechts vier deelnemende laboratoria die ieder slechts drie replica's analyseren.
In dit voorbeeld is dit gedaan om de uit te voeren berekeningen eenvoudig te houden; de prijs die daarvoor moet worden betaald is dat onder

dergelijke omstandigheden de Cochran- en de Dixontoets alleen zeer afwijkende resultaten kunnen aantonen. Met andere woorden: relatief grote afwijkingen zijn dan statistisch nog steeds toelaatbaar; daardoor blijft een resultaat als voor replica 2 van lab. 4(7) naast dat voor replica 2 van lab. 2 (19) geldig.
Dit voorbeeld maakt dus duidelijk waarom een voldoende groot aantal resultaten moeten worden verwerkt wil een Gemeenschappelijk Onderzoek succesvol kunnen worden afgesloten. Het maakt ook duidelijk waarom een methode aan kritisch onderzoek moet worden onderworpen wanneer te veel gegevens op statistische gronden uit het gegevensbestand moeten worden verwijderd en slechts weinig geldige waarden overblijven. Er bestaat dan immers een grote kans dat de berekende resultaten slechts in geringe mate representatief zijn.

4 Samenvatting

In deze les hebben we kennis gemaakt met de techniek waarmee een Gemeenschappelijk Onderzoek wordt opgezet, uitgevoerd engeevalueerd. De opzet, uitvoering en evaluatie van het Gemeenschappelijk Onderzoek is vastgelegd in de internationale norm ISO 5725.
Een Gemeenschappelijk Onderzoek heeft ten doel de kwaliteit van met name genoemde parameters van een nauwkeurig omschreven analysemethode te meten. Die kwaliteit komt tot uitdrukking in de begrippen herhaalbaarheid en reproduvceerbaarheid, verbonden aan de onderzochteparameter(s).
De norm geeft bovendien aanwijzingen hoe vastgesteld kan worden of deze kwaliteitsparameters afhankelijk zijn van het niveau waarop de parameter(s) wordt(en) bepaald.

Voordat een Gemeenschappelijk Onderzoek van start kan gaan, moet tevoren door de organisatoren orienterend onderzoek zijn gedaan naar de mate waarin de te onderzoeken analysemethode geschikt is om in een dergelijk onderzoek te worden getoetst. Aan deelnemers worden voorwaarden gesteld ten aanzien van de competentie om het beoogde onderzoek uit te voeren.

5 Lijst van begrippen en definities

Populatie	Verzameling individuele grootheden waarvan eigenschappen kunnen worden bepaald.
Fout	Statistische term voor een afwijking in een analyseresultaat ten opzichte een theoretische waarde.
Systematische fout	Een fout, gerelateerd aan een bepaalde methode/analysetechniek.
Toevallige fout	Een fout, gerelateerd aan mens en omgeving.
Herhaalbaarheid	De variatie in respons van een gestandaardiseerd monster, kort na elkaar meermalen geanalyseerd onder gelijke omstandigheden (zelfde lab, zelfde analist, een (1) serie).
Reproduceerbaarheid	De variatie in respons van een gestandaardiseerd monster, meermalen geanalyseerd onder variabele omstandigheden (verschillende laboratoria, verschillende analisten, verschillende series).
Herhaalbaarheidscondities	Uitvoering van een set analyses in hetzelfde monster door dezelfde analist, met hetzelfde apparaat, met gebruikmaking van dezelfde set chemicalien, kort na elkaar uitgevoerd.
Prestatiekenmerken	Kenmerken van analysemethoden waarmee de bruikbaarheid daarvan wordt gemeten. Onder P. worden o.m. begrepen: herhaalbaarheid, aantoonbaarheidsgrens, juistheid.

Level	Het niveau van een te bepalen analyt waarop een prestatiekenmerk van een methode wordt vastgesteld (doorgaans: gehaltemeting - herhaalbaarheid) Uniform L. Het onderzoek wordt verricht met sets monsters die binnen de set een gelijk niveau bezitten. Split L. Het onderzoek wordt verricht met sets monsters waarvan het niveau binnen de set verschilt.
Traject	Het gebied tussen het hoogste en het laagste niveau waarvoor een prestatiekenmerk van een methode wordt vastgesteld (doorgaans: gehaltemeting - herhaalbaarheid)
Outlier	Een individuele waarneming in een verzameling waarnemingen waarvan door middel van statistische toetsing is vastgesteld dat deze niet tot de groep behoort
Straggler	Een individuele waarneming in een verzameling waarnemingen waarvan door middel van statistische toetsing is vastgesteld dat het twijfelachtig is of deze wel tot de groep behoort
NNI	Nederlands Normalisatie Instituut
ISO	International Organisation for Standardization

6 Antwoorden tussentekstvragen

par. 1.1

Het doel van een Gemeenschappelijk Onderzoek is het vaststellen van de prestatiekenmerken, verbonden aan een bepaalde analysemethode. Deze methode moet nauwkeurig zijn beschreven. Bij toepassing van de norm ISO 5725 kan uit de resultaten de precisie (herhaalbaarheid enreproduceerbaarheid) van een of meer parameters bepaald worden.

par. 1.2

Voorwetenschap betekent dat een analist een min of meer reële kennis heeft van wat de uitslag van een komende analyse kan zijn. Dit kan er toe leiden dat hij min of meer naar deze uitslag toe werkt. Het gevolg hiervan is dat de invloed van toevallige fouten minder groot lijkt dan deze anders zou zijn geweest.

par. 1.3

Indien er slechts enkele deelnemers zijn, is de kans dat de gerapporteerde resultaten niet homogeen verdeeld zijn, nogal groot. Omdat daarnaast op basis van weinig uitslagen statistische toetsing nauwelijks mogelijk is, is de kans groot dat de eindresultaten van het Gemeenschappelijk Onderzoek vertekend worden door een of twee deelnemers die "rare" resulaten rapporteren.

par. 2.1

Oorzaken van systematische fouten kunnen zijn:

- verkeerd gecalibreerde apparatuur
- verkeerd gewaardeerde controle/ijkmonsters
- een op onjuiste principes gebaseerde analysemethode

De voornaamste bron van toevallige fouten is het handelen van de analist zelf, m.a.w. de menselijke factor.

par. 2.1.2

De Cochran- en de Dixontoets moeten herhaald worden nadat op grond van de uitgevoerde toets meetwaarden uit het gegevensbestand zijn verwijderd. Deze herhaling is noodzakelijk omdat de verdeling van de eigenschap die getoetst wordt (Cochrantoets: variantie; Dixontoets: uiterste waarden) door deze

verwijdering gewijzigd wordt.
Hierdoor kunnen nieuwe extremen statistisch significant gaan afwijken. Herhaalde toepassing van de Cochran- resp. de Dixontoets moet deze nieuwe extremen toetsen.

par. 2.4
a - onjuist
b - juist
De reproduceerbaarheid omvat de herhaalbaarheid (binnen een laboratorium) plus de variatie die het gevolg is van de verschillen tussen de verschillende laboratoria, veroorzaakt door: andere mensen, andere hulpmaterialen, andere apparatuur. Deze toegevoegde component is alleen in uitzonderingsgevallen klein en nooit nul.

par 2.5
Als slechts bij twee niveau's onderzocht zou worden, dan zijn er bij relatiemodel I slechts drie meetpunten (de oorsprong en de twee gemeten niveau's) en bij de modellen IIen III slechts twee.
De modellen II en III leveren een lijn die altijd door de punten van beide niveau's gaat. De mate van zekerheid dat de berekende relatie inderdaad representatief is voor de onderzochte situatie, wordt daardoor twijfelachtig.

15

ACCEPTEREN OF VERWERPEN

Voor de (kwantitatieve) bepaling van een analyt in een monster wordt op het RIKILT de analysemethode meestal ten minste in duplo uitgevoerd onder herhaalbaarheidsomstandigheden.
Duplowaarnemingen worden uitgevoerd
- om de precisie van het resultaat te vergroten: de standaardafwijking van duplowaarnemingen is $\sqrt{2}$ kleiner dan die van een enkele waarneming;
- om te zien of het analyseresultaat 'geloofwaardig' is: 'kloppende duplo's.

Als van methode niets over spreidingen bekend zou zijn, is uit duplo's ook niet te zeggen of ze 'kloppen'. Het is daarom belangrijk dat er wel wat bekend is over 'de standaardafwijking van de methode', dat wil zeggen, dat bekend is wat de standaardafwijking σ_x van de bepaling is, als deze goed wordt uitgevoerd. In dat geval zijn de duplowaarden op te vatten als een steekproef uit de populatie van 'alle mogelijke analyseresultaten die voor dit monster zouden kunnen worden verkregen'. Voor de variantie van de absolute waarde van het verschil van de duplowaarden geldt dan (toepassing van formule voor verschil op blz.45)

$$\text{var}\ (x_1 - x_2) = 2\ \sigma_x^2 .$$

Dus

$$P\ (\ |x_1 - x_2|\ < 2.\ \sqrt{2}.\ \sigma_x\ ) = 0{,}95.$$

Hoe komen we aan deze standaardafwijking?
a. Voor de methode zijn r en R bepaald in een gemeenschappelijk onderzoek. Volgens de moderne opvattingen van een analysevoorschrift behoort het voorschrift een paragraaf te bevatten met een opgave van de precisie van de methode, uitgedrukt in r en R. Dit is ook een eis van GLP.
Per definitie geldt

Herhaalbaarheid: $P\ (\ |x_1 - x_2| \leq r\ ) = 0{,}95.$
Reproduceerbaarheid: $P\ (\ |x_1 - x_2| \leq R\ ) = 0{,}95.$

Waarschuwing: de in een gemeenschappelijk onderzoek verkregen waarden zijn een momentopname van de bereikte precisie, zoals die toen, ten tijde van

het ringonderzoek door de laboratoria die daar aan nee deden, gold. Binnen het eigen laboratorium, op andere tijden en andere omstandigheden, kunnen andere waarden voor r worden verkregen.

b. r zelf te bepalen in eigen instituut uit grote series duplowaarden die in de loop van de tijd zijn verkregen.

Hiervoor geldt de formule

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \sum_{j=1}^{2} (x_{ij} - \bar{x}_{i.})^2$$

Dit is in wezen de formule voor de gepoolde variantie voor het vergelijken van meer dan twee populaties (blz. 49). Hier worden n variantieschattingen uit de n duplobepalingen, elk met één vrijheidsgraad gepoold tot één schatting van de binnen-duplo's variantie.

Verwerpen van resultaten.

Wanneer van een methode de herhaalbaarheid bekend is, kunnen we op grond daarvan een oordeel vellen over de kwaliteit van duplowaarnemingen. In de 'Draft proposal' ISO/DP 5725-6: worden verschillende schema's voor verwerpen van duplo's en analyseresultaten in n voud gegeven, mede afhankelijk van de omstandigheid of het een goedkope of een dure methode betreft. In IUPAC 'Draft harmonized protocols for the adoption and the presentation of anlytical methods' wordt één eenvoudig schema, zie figuur 15.1 gehanteerd.

In EEG verband zijn eisen gesteld aan de maximaal toelaatbare variatiecoëfficiënt, in afhankelijkheid van het gehalte, zie Appendix 1.

Bij het uitvoeren van een analysemethode moet van te voren vast liggen, hoe er gehandeld moet worden als duplo's (of muliplo's) niet kloppen.

16

RAPPORTEREN VAN ANALYSERESULTATEN

Wanneer een analyse wordt uitgevoerd moet vooraf duidelijk zijn

- wat geanalyseerd moet worden
- wat en hoe gerapporteerd wordt.

Dit is geen kwestie van statistiek, maar van afspraken!

Wat geanalyseerd moet worden

Er moet vooraf gedefinieerd zijn:

- laboratoriummonster
- analysemonster
- analyseportie.

De definities voor deze begrippen staan in Appendix 1.

Met de opdrachtgever moet van te voren afgesproken zijn hoe met de uitkomsten moet worden gehandeld. Ook dit is geen statistiek maar een zaak van afspraken.

17

PROEFOPZETTEN EN VARIANTIEANALYSE

Variantieanalyse is een statistische techniek die nuttig is gebleken in het onderzoek naar verschillen. Het gáát echter om het onderzoek dat men wil uitvoeren, dit moet centraal staan en variantieanalyse is een hulpmiddel, dat, mits juist toegepast, ons kan helpen de resultaten van het onderzoek te interpreteren.
De computer heeft hierbij voor een geweldige verandering voor de onderzoeker gezorgd. Vroeger moest men zich door een rijstebrij van mathematische formules heenwerken om zijn doel te bereiken. Tegenwoordig hoeft men dat allemaal niet meer te weten, dat doet het computerprogramma wel en kan men zich zelf geheel op het onderzoek concentreren. Moderne zogenaamde expert-systemen kunnen voor een opgegeven doel 'zelf bedenken', op welke wijze dat doel verwezenlijkt moet worden, welke berekeningen daarvoor moeten worden uitgevoerd en komen dan met een resultaat, zonder dat de gebruiker precies weet en hoeft te weten, hoe dat tot stand is gekomen.

Centraal staat dus het onderzoek. We kunnen hierbij onderscheiden: observationeel onderzoek en experimenteel onderzoek.
Bij *observationeel onderzoek* willen we iets weten over een bestaande populatie. Daaruit wordt een steekproef getrokken. Deze wordt onderzocht. De resultaten uit de steekproef worden geëxtrapoleerd naar de populatie. In de vorige hoofdstukken hebben we de statistische begrippen toegelicht aan de hand van observationeel onderzoek. Zo werd een vraag als "hebben boeren die uienras A verbouwen gemiddeld een hogere opbrengst dan boeren die uienras B verbouwen?" beantwoord op basis van een aselekte steekproef uit beide populaties.
Bij *experimenteel onderzoek* is er geen sprake van een populatie of steekproef, maar zet men zelf een experiment op, om een bepaald resultaat te verkrijgen. In het bovengenoemde voorbeeld kan men, in plaats van zich te baseren op een steekproefsgewijze inventarisatie het volgende experiment uitvoeren: men deelt een stuk grond op in 10 veldjes, hieruit loot men 5 veldjes die beplant worden met uienras A en de overige 5 worden beplant met ras B. Van elk veld bepaalt men de opbrengst en men vergelijkt de gemiddelde opbrengst van ras A en B.
In het algemeen komt experimenteel onderzoek erop neer dat men een aantal

experimentele eenheden neemt (hier de 10 veldjes) en deze de te vergelijken behandelingen toedient (hier beplanting met ras A resp. ras B). Indien systematische verschillen optreden tussen de behandelingen dan kan men die toeschrijven aan de zelf aangebrachte behandelingen. M.a.w. de behandelingen kunnen dan als oorzaak van de gevonden verschillen worden beschouwd. Dit oorzakelijk verband kan niet geclaimd worden bij observationeel onderzoek. Bijvoorbeeld het feit dat in het voorbeeld ras A een hogere opbrengst gaf dan ras B kan een andere oorzaak hebben, zoals bijvoorbeeld: de boeren die ras A verbouwen zijn alle gevestigd op kleigrond en de boeren met ras B alle op zandgrond!

Observationeel onderzoek kan gebruikt worden om relaties op te sporen en kan daarom in de beginfase van het onderzoek (exploratieve fase) zeer nuttig zijn. De gevonden relaties mogen echter niet zonder meer als causaal geïnterpreteerd worden. Voor dat laatste is experimenteel onderzoek wel geschikt.

In bijvoorbeeld sociale, psychologische, medische, biologische en ecologische wetenschappen is het soms niet eenvoudig experimentele methoden toe te passen. De bewijsvoering m.b.t. de causaliteit van effecten verloopt dan vaak moeizaam. Denk bijvoorbeeld aan de discussies rond melkvet en cholesterol in relatie tot hart- en vaatziekten. In de landbouwwetenschap bestudeert men echter vaak juist het effect van menselijk ingrijpen. Bijvoorbeeld: wat is het effect van stikstofgift op de opbrengst van een bepaald gewas, wat is het effect van verschillende bewerkings- en bemestingsmethoden op het bodemleven of wat is het effect van een bepaalde enzymbehandeling op de kwaliteit en bewaarbaarheid van een bepaald voedselprodukt. In de RIKILT omgeving: geeft een nieuwe (of: een eenvoudige) analysemethode dezelfde resultaten als een bestaande (of: referentie-) methode, niveaucontrole tussen laboratoria, onderzoek naar de precisie van een analysemethode.

Observationeel onderzoek binnen het RIKILT is bijvoorbeeld inventariserend onderzoek naar bepaalde contaminanten in plantaardige of dierlijke produkten. In het milieuonderzoek wordt veel observationeel onderzoek verricht.

Experimenteel onderzoek

Bij experimenteel onderzoek bestudeert men het effect van een of meer behandelingsfactoren (*instelvariabelen*) op een of meer *responsvariabelen*. In het experiment gebruikt men een aantal experimentele eenheden, men

verloot de behandelingen over deze eenheden, voert de proef uit en meet aan elke eenheid de responsvariabele. Naast de keuze van de behandelingsfactoren en de instellingen daarvan en de keuze van de experimentele eenheden zijn in een experiment 3 zaken belangrijk: herhaling, verloting en blokvorming.

Herhaling is van belang om de natuurlijke variatie tussen de experimentele eenheden te kunnen vaststellen. Als men één patiënt behandeld heeft met medicijn A en deze geneest beter dan één andere patient die medicijn B toegediend kreeg, dan kan men hieruit niet concluderen dat medicijn A beter werkt dan medicijn B. Immers, ook na toediening van hetzelfde medicijn zullen de twee patienten (waarschijnlijk) verschillend reageren. Per behandeling zijn dus meer experimentele eenheden nodig.

Verloting is essentieel om zoveel mogelijk te voorkomen dat de verschillen tussen de behandelingen verstrengeld raken met (onbewuste) subjectieve indelingen van de experimentele eenheden. Bijvoorbeeld bij vergelijking van rassen moeten deze rassen verloot worden over de beschikbare veldjes om zoveel mogelijk te voorkomen dat men onbewust aan een van de rassen de betere veldjes toewijst.

Om rekening te houden met vooraf bekende (of vermoede) verschillen tussen de experimentele eenheden kan men *blokken* vormen. Ook als er sprake is van te verwachten invloeden van andere variatiebronnen of omstandigheden kan men het effect hiervan op het te meten behandelingsverschil uitschakelen d.m.v. blokvorming. Als men twee analysemethoden wil vergelijken, moet men dus niet alle experimenten voor de ene methode achter elkaar in één periode uitvoeren en die voor de andere methode in een andere periode, maar beide perioden verdelen in blokken voor de ene en voor de andere methode.

'Random' en 'fixed' onderzoek

Wanneer men iets onderzoekt, kan men onderscheid maken tussen 'random' en 'fixed' onderzoek.

Bij *'random' onderzoek* zijn de voor een bepaalde instellingsvariabele geselecteerde elementen zelf (bijvoorbeeld welke laboratoria) niet van belang. Ze worden geacht willekeurig te zijn gekozen uit een veel grotere populatie. Een voorbeeld hiervan is een gemeenschappelijk onderzoek naar de

precisie van een analysemethode: een aantal laboratoria onderzoekt een identiek monster en het resultaat van het onderzoek is de vaststelling van een r en een R voor de methode. Maar welke laboratoria aan het onderzoek deelnamen, doet er verder niet toe. Er vindt hier geen toetsing plaats.
Bij *'fixed' onderzoek* gaat het wèl om de individuele elementen van een instellingsvariabele. Een voorbeeld hiervan is de niveaucontrole van laboratoria: het gaat daarbij om het opsporen van significante verschillen tussen een aantal, met name genoemde, laboratoria. Dit doet men door toetsen, bij voorbeeld met de t-toets, eventueel voorafgegaan door de hierna te bespreken F-toets.
Wanneer in een gemeenschappelijk onderzoek in tien laboratoria drie bepaalde analysemethoden door twee analisten worden onderzocht, zijn de instelvariabelen Laboratorium en Analist random, en is Methode fixed.
De eigenlijke variantieanalyse is in beide gevallen gelijk, maar wat men met de gegevens verder doet, verschilt voor random en fixed onderzoek.

Variantieanalyse in observationeel onderzoek

Bij dit onderzoek gaat het dus over populaties, die we met elkaar willen vergelijken. We doen dit door uit de populaties steekproeven te trekken en daarop variantieanalyse toe te passen. Dit leidt tot een uitspraak of er, met een bepaalde zekerheid, significante verschillen bestaan tussen de populaties.

Om variantieanalyse toe te mogen passen, zijn een aantal veronderstellingen over de aard van het cijfermateriaal nodig, en wel:

1. De elementen per steekproef zijn aselect getrokken uit een populatie.
2. De steekproeven zijn onafhankelijk van elkaar uit hun populaties getrokken.
3. De populaties zijn normaal verdeeld.
4. De standaardafwijkingen van de populaties, σ_1, σ_2, ..., σ_i, ..., σ_k, zijn gelijk en kunnen dus worden aangeduid door één symbool σ.
5. De gemiddelden van de populaties, μ_1, μ_2, ..., μ_i, ..., μ_k kunnen verschillend zijn.

We hebben dus k populaties $N(\mu_1,\sigma^2)$, $N(\mu_2,\sigma^2)$, ... , $N(\mu_k,\sigma^2)$ en het doel van de variantieanalyse is na te gaan of μ_1, μ_2, ..., μ_k al dan niet aan elkaar gelijk zijn.

Variantieanalyse in experimenteel onderzoek

Hoewel hier geen sprake is van steekproeven, mag de techniek van de variantieanalyse ook hier toegepast worden. De onderliggende gedachte is, dat het uitgevoerde experiment als een steekproef gedacht kan worden uit een populatie van veel van deze experimenten die uitgevoerd zouden kunnen zijn. Wanneer bijvoorbeeld in een gemeenschappelijk onderzoek voor een analysemethode een r en een R zijn vastgesteld, worden deze op grond van deze redenering geldig geacht voor 'alle' analysen die met de getoetste methode worden uitgevoerd.

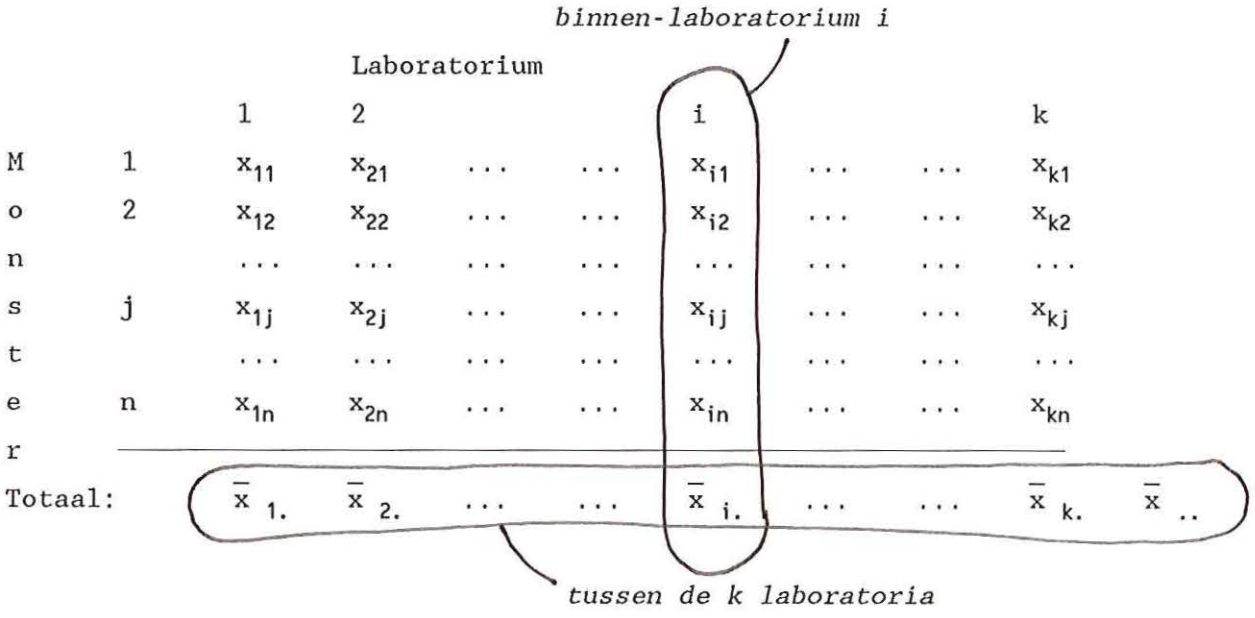
Eenvoudige variantieanalyse

We zullen nu de variantieanalyse bespreken voor een eenvoudig geval: een aantal laboratoria onderzoekt een aantal monsters.

We hebben k laboratoria. Elk laboratorium heeft in identieke monsters n analysen uitgevoerd. En de vraag is nu: bestaan er tussen de analyseresultaten van de laboratoria significante verschillen?

We weten intussen, dat om statistisch een uitspraak te doen of verschillen 'significant' zijn, men kennis moet hebben van de varianties. We moeten dus op zoek naar de variantie σ^2, die, zoals vooronderstelt is, geldt voor alle laboratoria die we beschouwen.

1. Eerst schrijven we de analyseresultaten in een tabel:



2. Wij kunnen nu voor elk laboratorium voor de n analyseresultaten een gemiddelde en een variantie berekenen. Voor het i de laboratorium (in de tabel omlijnd en aangeduid met *binnen-laboratorium i*)

gemiddelde: $\bar{x}_{i.}$ 1

(deze waarden hebben we meteen onderaan in de tabel gezet.)

variantie: $s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{i.})^2 = s^2_{\text{binnen lab i}}$. 2

We weten dat de variantie van een steekproef een schatter is voor de variantie van de populatie waar deze steekproef uit getrokken is. Dus is $s^2_{\text{binnen lab i}}$ een schatter voor σ_i^2. Het aantal vrijheidsgraden van deze variantie bedraagt n-1.

We waren er echter van uit gegaan, dat voor alle populaties dezelfde variantie σ^2 geldt. De variantie van elke steekproef is dus een schatter voor dezelfde σ^2. We mogen dus de varianties binnen alle k steekproeven (d.w.z. de resultaten van de k laboratoria) middelen ('poolen') tot één *binnen-laboratorium* variantie. We hebben k varianties, elk met n-1 vrijheidsgraden; het aantal vrijheidsgraden van de gepoolde variantie bedraagt dus k(n-1). Dus wordt de binnen-laboratorium variantie

$$s^2_{\text{binnen}} = \frac{1}{k} \sum_{i=1}^{k} s_i^2 = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{i.})^2 \qquad 3$$

Deze s^2_{binnen} is een schatter van σ^2, ofwel de verwachtingswaarde van $s^2_{\text{binnen}} = \sigma^2$:

$$E(s^2_{\text{binnen}}) = \sigma^2. \qquad 4$$

E staat hier voor verwachting (expectation).

3. Als we voor elk laboratorium een gemiddelde $\bar{x}_{1.}$, $\bar{x}_{2.}$, ..., $\bar{x}_{i.}$, ..., $\bar{x}_{k.}$ hebben berekend, kunnen we ook voor deze waarden een gemiddelde en een variantie berekenen:

gemiddelde: $\bar{x}_{..}$ 5

variantie: $s^2_{\text{tussen}} = \frac{1}{k-1} \sum_{i=1}^{k} (\bar{x}_{i.} - \bar{x}_{..})^2$. 6

Als er geen verschillen tussen de populaties bestaan, vormen deze gemiddelden en ook het totaalgemiddelde $\bar{x}_{..}$ een schatting van 'het' gemiddelde μ van alle populaties bij elkaar. We hebben geleerd dat de variantie van het gemiddelde van een steekproef van n waarnemingen uit een populatie met variantie σ^2 gelijk is aan σ^2/n.

Dus is s^2_{tussen} een schatter voor σ^2/n, ofwel is

$$n \cdot s^2_{tussen} = \frac{n}{k-1} \sum_{i=1}^{k} (\bar{x}_{i.} - \bar{x}_{..})^2 \qquad 7$$

een schatter voor σ^2. De verwachting van $n \cdot s^2_{tussen}$ is σ^2, in formule:

$$E(n.s^2_{tussen}) = \sigma^2. \qquad 8$$

Echter, dit gaat alleen op als er géén verschillen tussen de steekproeven zijn. Waren die er wel, dan zullen die verschillen ook tot uiting komen in de gemiddelden $\bar{x}_{1.}, \bar{x}_{2.}, \ldots, \bar{x}_{i.}, \ldots, \bar{x}_{k.}$. In dat geval zal $n.s^2_{tussen} >> \sigma^2$ zijn:

$$E(n.s^2_{tussen}) >> \sigma^2 \quad \text{bij significant verschil.} \qquad 9$$

4. Hiermee hebben we een middel in handen om te zien of er significante verschillen bestaan tussen de steekproeven. Deel $n.s^2_{tussen}$ door s^2_{binnen}. Is er geen significant verschil, dan moet er (ongeveer) 1 uitkomen (formule 8/ formule 4). Vinden we een waarde >> 1, dan is er wèl een significant verschil (formule 9/ formule 4).

5. Als er géén verschillen tussen de laboratoria bestaan, kunnen we alle waarnemingsresultaten 'poolen'. Ook voor deze resultaten zijn weer een-gemiddelde en een variantie te berekenen:

gemiddelde: $\bar{x}_{..}$ 10

variantie: $$s^2_{Totaal} = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{..})^2 / (kn-1). \qquad 11$$

In wezen hebben we hiermee een variantieanalyse uitgevoerd. Men is gewoon een variantieanalyse volgens een bepaald vast patroon op te schrijven met gebruikmaking van een eigen nomenclatuur.

De grondgedachte is, dat we het verschil tussen een individuele waarneming, x_{ij} en het totaalgemiddelde, $\bar{x}_{..}$ op kunnen splitsen in een verschil tussen x_{ij} en zijn eigen steekproefgemiddelde, $\bar{x}_{i.}$, en een verschil tussen dit steekproefgemiddelde en het totaalgemiddelde:

$$x_{ij} - \bar{x}_{..} = (\bar{x}_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x}_{..}).$$

Dit geldt voor elke individule waarneming, x_{ij}, en dus ook voor de som van alle individuele waarnemingen

$$\sum_i \sum_j (x_{ij} - \bar{x}_{..}) = \sum_i \sum_j \{(x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x}_{..}).$$

Kwadrateren van deze uitdrukking levert

$$\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = \{ \sum_i \sum_j \{(x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x}_{..})\}^2.$$

Uitwerken van het rechterlid geeft

$$= \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 + \sum_i \sum_j (\bar{x}_{i.} - \bar{x}_{..})^2 + 2 \sum_i \sum_j (x_{ij} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{x}_{..}).$$

Hierin is $\sum_j = n$, en $= 0$, want dit uitgewerkt voor elke kolom is

$$\sum_{j=1}^{n} (x_{ij} - x_{i.}) = \underbrace{x_{i1} + x_{i2} + \ldots + x_{in}}_{= n \,.\, x_{i.}} - n.x_{i.} = 0.$$

De dubbelprodukten zijn dus allemaal = 0 en de uitdrukking wordt

$$\sum_i \sum_j (x_{ij} - x_{..})^2 = \sum_i \sum_j (x_{ij} - x_{i.})^2 + n \sum_i (x_{i.} - x_{..})^2.$$

Men noemt deze uitdrukkingen *kwadraatsommen, KS (sums of squares, SS)* en wel resp.

‖	‖	‖
Totale kwadraatsom	Kwadraatsom binnen	Kwadraatsom tussen
KS_{totaal}	laboratoria KS_{binnen}	laboratoria KS_{tussen}.

In het Engels, wat algemener geformuleerd

Total sum of squares	Sum of squares within	Sum of squares between
SS_{total}	groups SS_{within}	groups $SS_{between}$.

Wanneer we deze kwadraatsommen delen door de bijbehorende vrijheidsgraden, dus resp. door

totaal: k(n-1) binnen: k(n-1) tussen: p-1

krijgen we

$$\frac{\sum_i \sum_j (x_{ij} - \bar{x}_{..})^2}{kn - 1} \qquad \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2}{k\,(n - 1)} \qquad \frac{n \sum_i (\bar{x}_{i.} - \bar{x}_{..})^2}{k - 1}. \qquad 12,13,14$$

Maar dit zijn precies de varianties, die we uitgerekend hebben bij 11, 3 en 7, dus

‖	‖	‖
s^2_{totaal}	s^2_{binnen}	$n.s^2_{tussen}$.

In het taalveld van de variantieanalyse worden 12, 13 en 14 niet varianties genoemd, maar *gemiddelde kwadraten, GK (mean squares, MS)*.

Dus: om te zien of er een invloed van de laboratoria op de resultaten bestaat, delen we de gemiddelde kwadraten 'tussen laboratoria' en 'binnen laboratoria' op elkaar. De uitkomst hiervan wordt F genoemd:

$$F = \frac{GK_{tussen}}{GK_{binnen}} \qquad \frac{MS_{between}}{MS_{within}}.$$

Als $F \approx 1$, dan: geen verschillen tussen de laboratoria.

F >> 1, dan: wel verschillen tussen de laboratoria.

De resultaten van een variantieanalyse wordt als volgt in tabelvorm genoteerd

VARIANTIEANALYSE

Variatiebron (1)	Aant.vr.gr. (2)	Kwadraatsom (3)	Gem.kwadraten (4) = (3)/(2)	F
Tussen lab.	k-1	KS_{tussen}	GK_{tussen}	$F = GK_{tus}/GK_{bin}$
Binnen lab.	k(n-1)	KS_{binnen}	GK_{binnen}	
Totaal	nk-1	KS_{totaal}	GK_{totaal}	

Opmerkingen

1 De som van de uitgesplitste vrijheidsgraden is gelijk aan het aantal vrijheidsgraden van 'totaal'.

2 Evenzo is de som van de uitgesplitste kwadraatsommen gelijk aan KS_{totaal}. Ook bij meer ingewikkelde variantieanalyses gaat een en ander op.

3 Het SPSS commando voor bovenstaande variantieanalyse is
ANOVA gehalte BY laboratorium(1,k) analyse(1,n).

F toets

Met de in de vorige paragraaf berekende waarde voor F zullen we nog niet helemaal gelukkig zijn. We zullen ook hier de betrouwbaarheid willen weten, dat wil zeggen, de kans dat als er een laboratoriumeffect is, een bepaalde waarde van F wordt overschreden.

De algemene definitie voor F voor het vergelijken van de varianties van twee steekproeven, 1 en 2, is

$$F = \frac{s_1^2}{s_2^2}$$

Stel aantal vrijheidsgraden in 1: ν_1
" " " 2: ν_2

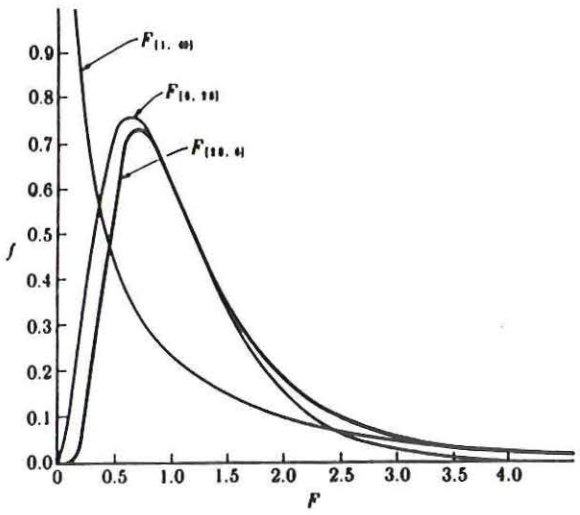
Het zal duidelijk zijn, dat de betrouwbaarheid van F afhangt van de omvang van de steekproeven, m. a. w. van het aantal vrijheidsgraden ν_1 en ν_2.

Voor elke combinatie van ν_1 en ν_2 bestaat er een F verdeling, genoemd naar de statisticus R.A. Fisher, zie figuur 17.1.

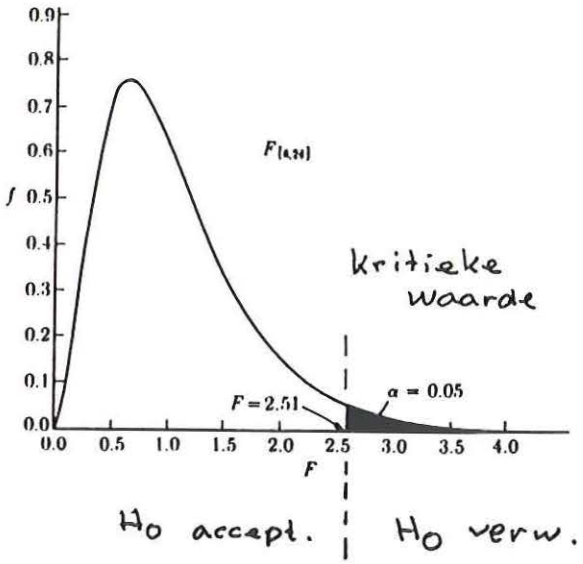
In ons uitgewerkte voorbeeld boven was

$$F = \frac{GK_{tussen}}{GK_{binnen}}$$

Als H_o waar is, dus als $\mu_1 = \mu_2 = \ldots = \mu_k$, dan zullen GK_{tussen} en GK_{binnen} niet exact gelijk zijn aan σ^2, maar fluctuaties vertonen. Dus zal F fluctueren om 1 heen als H_o waar is, en wel volgens de F verdeling voor het



Figuur 17.1 Drie voorbeelden van een F verdeling



Figuur 17.2 Rechtszijdige overschrijdingskans $\alpha = 0{,}05$

aantal vrijheidsgraden van GK_{tussen} en GK_{binnen}. We zullen tot significantie besluiten, en dus H_o verwerpen, als de uitkomst van F niet meer past bij de verdeling van F onder H_o. 'Niet meer past' houdt in, dat de gevonden F-waarde een bepaalde, kritieke waarde van de F verdeling overschrijdt. Deze waarde is op te zoeken in een *tabel met kritische waarden van de F verdeling*, zie Tabel X, Appendix 2.

Als voor s_1^2 en s_2^2 een waarde voor F gevonden wordt > $F_{kritiek}$ voor 95 % betrouwbaarheid, dan kan met een betrouwbaarheid van 95 % gesteld worden, dat $s_1^2 > s_2^2$. Stel:

$s_1^2 = 8 \quad \nu_1 = 24$

$s_2^2 = 2 \quad \nu_2 = 6.$

We zoeken op in de tabel: in de kolom $\nu_{teller} = 24$, rij $\nu_{noemer} = 6$ en vinden voor $\alpha = 0{,}05$: 3,84, dit schrijven we $F_{0,05[24,6]} = 3{,}84$,

voor $\alpha = 0{,}025$: 5,12 $\quad F_{0,025[24,6]} = 5{,}12$

en voor $\alpha = 0{,}01$: 7,31 $\quad F_{0,01[24,6]} = 7{,}31$

De gevonden waarde voor F = 8 / 2 = 4 is groter dan 3,84; de varianties s_1^2 en s_2^2 zijn dus met 95 % betrouwbaarheid van elkaar verschillend, zie figuur 17.2. De waarde is echter < 5,12 en < 7,31, dus met 97,5 en 99% betrouwbaarheid kan er geen verschil worden aangetoond.

De F tabel is ook opgenomen in Statcal, 1 = Tables of statistical distribution; 4 = F-distribution. Er zijn 4 opties:

1. Probability P, entering F. In ons geval levert df1=24, df2=6, F=4 een cumulative probability above F=4 = 0,0455 op.
2. Coordinate F, entering left-sided probability.

3. Coordinate F, entering right-sided probability. In ons geval df1=24, df2=6, cumulative probability exeeding desired F: .05, geeft 3.841.
4. Border values F (95%, 99%, 99.9%).
Uiteraard is een F tabel ook aanwezig in de grote statistische paketten, zoals SPSS en Genstat.

Statistisch model

Voor de waarnemingen van de laboratoria kan men het volgende statistische model opzetten

$$x_{ij} = \mu + \alpha_i + \epsilon_{j(i)}.$$

Hierin is

μ = het gemiddelde van de populatie van alle waarnemingen die men in het analysemonster had kunnen uitvoeren;

α_i = de systematische afwijking van laboratorium i en is $\mu + \alpha_i = \bar{x}_{i.}$

$\epsilon_{j(i)}$ = de toevallige afwijking tussen x_{ij} en $\bar{x}_{i.}$.

We hebben al gezien, dat bij systematische verschillen tussen de laboratoria $n.s^2_{tussen} > \sigma$. Aan de hand van het model kunnen we dit meer kwantitatief onderbouwen.

Als $\bar{\epsilon}_i$ het gemiddelde is van de n $\epsilon_{j(i)}$ van n waarnemingen van laboratorium i dan is

$$\bar{x}_{i.} = \mu + \alpha_i + \bar{\epsilon}_i.$$

De variantie van $\epsilon_{i(j)}$ is s^2_{binnen}, wat een schatter was voor σ, zodat de variantie van $\bar{\epsilon}_i$ een schatter is voor σ/n: $\hat{s}^2(\bar{\epsilon}_i) = \hat{s}^2_{binnen} = \sigma/n$.

Laat de verwachting van de variantie voor de systematische afwijking $\hat{s}^2(\alpha_i) = \sigma^2_\alpha$ zijn.

Als bovendien, door een juiste uitvoering van het experiment, mag worden aangenomen dat de grootte van de $\epsilon_{j(i)}$ onafhankelijk is van de grootte van α_i, dan is volgens de optelregel voor varianties

$$\hat{s}^2(\bar{x}_{i.}) = \hat{s}^2_{tussen} = \hat{s}^2(\alpha_i) + \hat{s}^2(\bar{\epsilon}_i) = \sigma^2_\alpha + \sigma^2/n.$$

Derhalve is $E(GK_{tussen}) = n.\hat{s}^2_{tussen} = \sigma^2 + n.\sigma^2_\alpha$

Algemeen noteert men, ook in Nederlandse literatuur, de verwachting van de gemiddelde kwadraten (expected mean squares) als EMS. Dus in ons geval is

$EMS_{tussen} = \sigma^2 + n.\sigma^2_\alpha$ en

$EMS_{binnen} = \sigma^2$,

en is F in ons voorbeeld

$$\hat{F} = EMS_{tussen} / EMS_{binnen} = (\sigma^2 + n.\sigma^2_\alpha) / \sigma^2.$$

Geldigheid van de aannamen voor de variantieanalyse

De vooronderstellingen om een variantieanalyse te mogen uitvoeren waren: normaliteit van de verdelingen, gelijkheid van de varianties, aselectheid en onafhankelijkheid van de steekproeven.

Normaliteit van de verdelingen

Er wordt aangenomen dat de verdelingen in de populaties, waaruit de steekproeven getrokken zijn (en dus ook de steekproeven), normaal zijn. Dit is meestal moeilijk te controleren, omdat het aantal waarnemingen niet erg groot is. In ettelijke studies is echter aangetoond, dat vrij grote afwijkingen van de normaliteit kunnen optreden zonder dat de geldigheid van de variantieanalyse geschaad wordt.

Gelijkheid van de varianties

De aanname dat de populaties dezelfde varianties hebben is wel te controleren. Daartoe neemt men van de groepsgemiddelden de grootste en de kleinste waarde en past daarop de t-toets toe. Als de varianties niet gelijk zijn, kan men dat soms verbeteren door een transformatie toe te passen. Hier wordt de zin duidelijk van een log-transformatie voor chemische analysen, waarvoor de variantie evenredig toeneemt met het gehalte.

Niet te grote afwijkingen van de aanname heeft ook hier geen grote invloed op de geldigheid van de variantieanalyse.

Aselectheid en onafhankelijkheid

De gevolgen van niet-aselectheid en afhankelijkheid kunnen zeer ernstig zijn. Het is daarom zeer belangrijk dat een proefopzet van te voren goed wordt overwogen en wordt doorgesproken zowel met deskundigen op het terrein van het onderwerp van onderzoek als met statistisch deskundigen!

Wat er gebeuren kan als een variantieanalyse verkeerd wordt toegepast, illustreert het volgende voorbeeld. Bij een proef met suikerbieten werden 3 groenbemesters in 5 niveaus toegepast in 3 herhalingen. De groenbemesters waren geloot over velden binnen blokken en de stikstoftrappen over subveldjes binnen velden. Als men geen rekening houdt met de verlotingen, maar alleen naar de factoren Groenbemesting en Stikstof kijkt, verkrijgt men de (foute) variantieanalyse-tabel:

Variatiebron	df	SS	MS	F	Sign of F
blok	2	1,4	0,72		
groenbem	2	21,2	10,6	7,05	0,003
stikstof	4	79,3	19,8	13,17	<0,001
groenb.stikst	8	11,4	1,4	0,94	0,50
rest	28	42,1	1,5		
totaal	44	155,5			

De juiste variantieanalyse-tabel ziet er anders uit:

Variatiebron	df	SS	MS	F	Sign of F
blok	2	1,4	0,72		
blok.veld					
groenbem	2	21,2	10,6	2,09	0,239
rest	4	20,3	5,1		
blok.veld.subveld					
stikstof	4	79,3	19,8	21,76	<0,001
groenb.stikst	8	11,4	1,4	1,56	0,190
rest	28	42,1	1,5		
totaal	44	155,5			

Hier zien we duidelijk dat een, gezien de proefopzet, foute analyse tot een verkeerde conclusie kan leiden: Bij de eerste analyse lijkt het groenbemester-effect zeer betrouwbaar te zijn (F= 7,05 dus >>1, p=0,003), terwijl we in werkelijkheid deze conclusie helemaal niet kunnen trekken (F= 2,09, p= 0,239)!

Soorten proefopzetten

In het voorgaande hebben we een heel eenvoudige proefopzet besproken: een aantal groepen (laboratoria) met één factor binnen de groepen (de monsters). De variantieanalyse daarvoor is nog redelijk te doorzien. Er zijn veel ingewikkeldere omstandigheden denkbaar. De computer is gewillig, maar het moet voor de onderzoeker duidelijk zijn, wat hij aan het doen is en wat hij wil weten. In het volgende zullen we een aantal mogelijkheden kort aangeven.

Gekruiste classificatie.

De factoren zijn gelijkwaardig. Bijvoorbeeld 3 laboratoria onderzoeken 4 monsters met verschillende gehalten in enkelvoud. Essentiëel is dat de alle niveaus van de éne factor voorkomt bij alle niveaus van de andere: Alle laboratoria onderzoeken alle gehalten. Aan te geven met lab*gehalte.

Dit kan ook met meer dan twee factoren: 3 laboratoria onderzoeken 4 gehalten met 2 methoden, dat is: lab*gehalte*methode. Men heeft dan een driedimensionale gegevenstabel, deze schrijft men tweedimensionaal op:

		Laboratorium 1		2		3	
		Methode 1	2	1	2	1	2
Gehalte	1	x	x	x	x	x	x
	2						
	3						
	4						

of:

				Laboratorium 1	2	3
Gehalte	1	Methode	1	x		
			2	x		
	2		1	x		
			2	x		
	3		1	x		
			2	x		
	4		1	x		
			2	x		

Hiërarchische classificatie (genest)

De ene factor is 'ondergeschikt' aan de andere. Bijvoorbeeld op 3 laboratoria worden 4 gehalten in duplo onderzocht. Hier heeft elk laboratorium z'n eigen duplo's. De duplowaarden zijn ondergeschikt aan de laboratoria. Dit kan men aangeven met lab*gehalte/duplo.

<u>Verder verloop na de variantieanalyse</u>

Als de proefopzet random was zal men de varianties willen weten, in het besproken geval van een aantal laboratoria die een aantal monsters hebben onderzocht de precisie van de methode. We hebben afgeleid dat

$$EMS_{binnen} = \sigma^2_r$$

$$EMS_{tussen} = \sigma^2_r + n.\sigma^2_L.$$

Dus is

$$\hat{\sigma}^2_r = MS_{binnen}$$

$$\hat{\sigma}^2_L = (MS_{tussen} - MS_{binnen}) / n.$$

In het geval van een fixed proefopzet zal men eerst kijken of de F-waarde die de variantieanalyse heeft opgeleverd, uitwijst, dat er significante

verschillen tussen -in ons voorbeeld- de laboratoria bestaan. Is dit het geval, dus is F >> 1, dan zal men willen weten tussen welke laboratoria er significante verschillen bestaan.

We passen de berekening voor het vergelijken van twee populaties (Hoofdstuk 8) toe. Eerst berekenen we de sed. Dan zoeken we t op. Vermenigvuldiging van beide geeft het kleinste significante verschil tussen twee laboratorium-gemiddelden, lsd:

lsd = t . sed.

Stel sed = 4,78; t zal ongeveer 2 zijn, dan zal bij een verschil tussen twee laboratoriumgemiddelden groter dan ongeveer 10 het verschil significant zijn. De laboratoriumverschillen kunnen we eveneens opvragen. Stel, dat dit geeft:

Laboratorium:	A	B	C	D	E	F	G
Lab.gemiddelde:	20,8	24,5	9,5	16,8	15,5	5,7	14,3

dan zien we meteen, dat C significant afwijkt van A en B, en G van A t/m F. Deze methode heet de LSD methode, of ook 'volgens Fisher'. Er zijn ook andere methoden: onder andere de Student-Newman-Keuls (SNK) toets, Tukey toets en Duncan toets. De Groep Landbouwwiskunde geeft de voorkeur aan de LSD toets.

HOOFDSTUK 18

LINEAIRE REGRESSIE-ANALYSE

18.1 Inleiding

Indien we de samenhang tussen 2 (of meer) kwantitatieve variabelen willen onderzoeken, dan kan dat met behulp van regressie-analyse. Men kan in zo'n verband geinteresseerd zijn om twee redenen:

- men wil een hypothetisch model voor een biologisch systeem opstellen resp. valideren,
- men wil de ene variabele voorspellen uit de andere.

In sommige onderzoeken zijn beide motieven aanwezig.

Voorbeelden waarin regressie-analyse wordt gebruikt zijn:

a) Men wil beschrijven hoe de opbrengst van een tarweras afhangt van de mestgift (hoeveelheden N, P, K, Mg en Ca) om bijvoorbeeld een optimale bemesting te kunnen geven.

b) Men wil onderzoeken of de hoeveelheid mycorrhiza op de wortels van douglassparren samenhangt met de hoeveelheid vrij aluminium in de bodem Mycorrhiza is een schimmel die leeft op de wortel en die de opname van mineralen door de boom verbetert. De hypothese is dat onder invloed van zure depositie het in zandgronden in gebonden vorm aanwezige aluminium vrijkomt in de vorm van Al^{3+}-ionen, welke laatste een toxische werking zouden hebben op de mycorrhiza's.

c) De hoeveelheid van een bepaald eiwit in plantcellen kan men indirect bepalen na een kleurreactie m.b.v. een spectrofotometer door bij een bepaalde golflengte de absorptie te meten. Dit gaat als volgt. Eerst bepaalt men voor een aantal monsters zowel het eiwitgehalte als de absorptie en op basis van deze waarnemingen beschrijft men het verband tussen eiwitgehalte en de hoeveelheid absorptie. Daarna wordt voor nieuwe monsters alleen de absorptie gemeten en gebruikt men het gevonden verband om het eiwitgehalte te berekenen. Deze procedure wordt calibratie genoemd. Bij calibratie gaat het altijd om een moeilijk te meten variabele die men wil vervangen door een eenvoudiger (of

goedkoper) te meten variabele (hier eiwitgehalte meten via absorptie). De verzameling van objecten waarvoor beide variabelen bekend zijn noemen we ijkset of calibratieset.

Bij regressie-analyse onderzoeken we hoe goed een responsvariabele voorspeld wordt uit een of meer predictorvariabelen. De responsvariabele wordt ook wel de te verklaren variabele of afhankelijke variabele genoemd en de predictorvariabele heet ook wel eens stimulusvariabele of verklarende of onafhankelijke variabele.
In voorbeeld a) nemen we de opbrengst als responsvariabele en in voorbeeld b) de hoeveelheid mycorrhiza. In c) zou men het eiwitgehalte als responsvariabele kunnen nemen omdat deze voorspeld moet worden uit de absorptie (zie verderop).

Soms wil men een waargenomen relatie interpreteren als een causaal verband. Of dit mogelijk is hangt af van de wijze waarop het onderzoek is uitgevoerd (experimenteel versus observationeel onderzoek). Bij experimenteel onderzoek brengt men zelf de stimulus variabele(n) aan bij vergelijkbare experimentele eenheden, zodat een waargenomen relatie als een causaal verband mag worden geinterpreteerd (bijvoorbeeld als men in het hierboven beschreven voorbeeld a) vergelijkbare veldjes gebruikt met verschillende bemestingsgiften). Bij observationeel (inventariserend) onderzoek echter vormt een gevonden relatie geen sluitend bewijs voor causaliteit. Dit is o.a. het geval als in voorbeeld b) verschillende plekken met douglasbomen worden bemonsterd. Het waarnemen van samenhang tussen het Al^{3+}-gehalte en de hoeveelheid mycorrhiza's sluit niet uit dat de hoeveelheid mycorrhiza's door andere oorzaken bepaald wordt dan door Al^{3+}. Soms kan bij een inventariserend onderzoek een samenhang tussen 2 variabelen naar voren komen omdat beide variabelen samenhangen met een derde variabele. In voorbeeld b) zou dat bodemtype kunnen zijn (of de zuurgraad zelf zonder dat Al^{3+} als toxisch intermediair optreedt!).

Bij calibratie (zie voorbeeld c) is causaliteit minder relevant. Men is slechts op zoek naar een nauwkeurige voorspelling van de moeilijk te meten variabele uit de eenvoudig te meten variabele.

Bij regressie-analyse hanteert men een model dat beschrijft hoe de responsvariabele Y samenhangt met de predictorvariabele x. Hierbij geeft

men bij elke vaste waarde van x de verdeling van Y op en daarmee dus ook de verwachtingswaarde EY en de variantie var(Y).

Het meest eenvoudige regressiemodel is dat waarbij EY op een lineaire wijze van x afhangt en waarbij Y normaal verdeeld is met variantie σ^2 die niet van x afhangt.
Dus

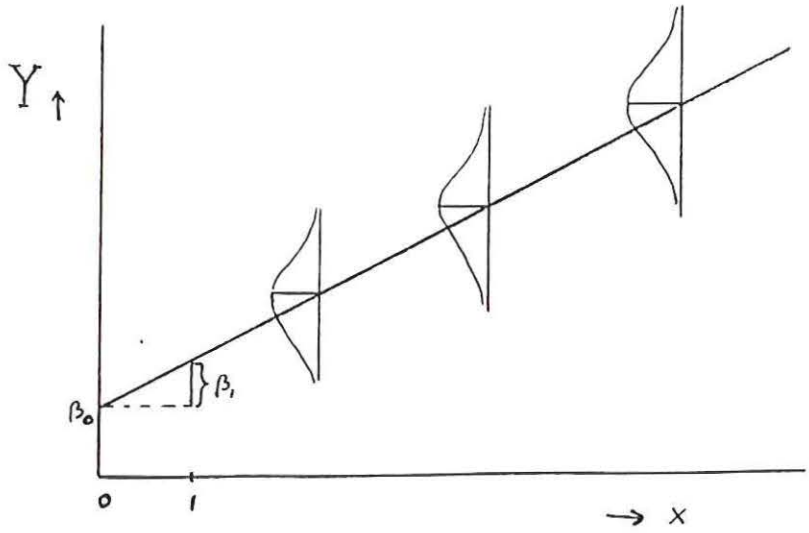
$$EY = \beta_0 + \beta_1 . x \qquad \text{en Y normaal verdeeld met variantie } \sigma^2.$$

Men kan het model ook anders schrijven:

$$Y = \beta_0 + \beta_1 x + E$$

waarin het E normaal verdeeld is met verwachting 0 en variantie σ^2.

De rechte lijn $EY = \beta_0 + \beta_1 x$ noemen we de regressielijn van Y op x. Van deze lijn is β_0 het intercept en β_1 de richtingscoefficient (ofwel helling). De term E noemen we de niet door x verklaarde variatie van Y of ook wel residuele variatie of de ruisterm van het model. We hebben het model dus opgedeeld in een systematisch deel EY en een stochastisch deel E. Het systematische deel gebruiken we voor de voorspelling van Y uit x; de standaarddeviatie σ van het stochastische deel geeft de onnauwkeurigheid van die voorspelling aan (geeft aan hoeveel Y kan variëren bij vaste x).



De vraag is nu hoe men op basis van waarnemingen de regressiecoefficienten (β_0 en β_1) en de restvariantie σ^2 kan schatten opdat de voorspelformule kan

worden opgesteld en de nauwkeurigheid ervan kan worden aangeven. Bij de beantwoording van deze vraag wordt ervan uitgegaan dat de variabelen Y en x gemeten zijn aan onafhankelijke objecten (experimentele eenheden).

In de latere paragrafen zullen we allerlei generalisaties bespreken van dit simpele regressiemodel. Het kan bijvoorbeeld voorkomen dat een rechte lijn geen adequate beschrijving levert. In dat geval kan men in het systematische deel van het model kiezen voor een kwadratische curve, polynomen van hogere graad, een exponentiële curve of talloze andere functies. Ook kan het voorkomen dat we de responsvariabele Y willen verklaren (voorspellen) uit meerdere predictor variabelen. (multipele regressieanalyse). Hierbij speelt ook selectie van variabelen vaak een belangrijke rol. Een combinatie van beide generalisaties vindt men bij responsieoppervlakken (voorbeeld: welke hoeveelheden N, P, K, etc. geven een optimale opbrengst?). Hierbij spelen meerdere predictor variabelen een rol en het verband is niet-lineair. Tenslotte kan men kwalitatieve variabelen als predictor opnemen in regressiemodellen.

Een andere generalisatie betreft de verdeling van de responsvariabele bij vaste x. Als de variantie niet constant is kan men dit nog verhelpen via een transformatie of men kan de waarnemingen gewichten geven. Als de verdeling van de responsvariabele niet normaal is dan kan men werken met de zgn. gegeneraliseerde lineaire regressiemodellen. Hierbij kan men de verdeling van Y opgeven en men kan los daarvan de relatie tussen EY en de x-variabele(n) specificeren. Voorbeelden hiervan zijn loglineaire modellen voor tellingsgegevens (als deze gegevens poisson verdeeld zijn) en logistieke regressie voor percentagegegevens (waarbij de percentages gebaseerd zijn op binomiaal verdeelde aantallen).

18.2 Enkelvoudige lineaire regressie

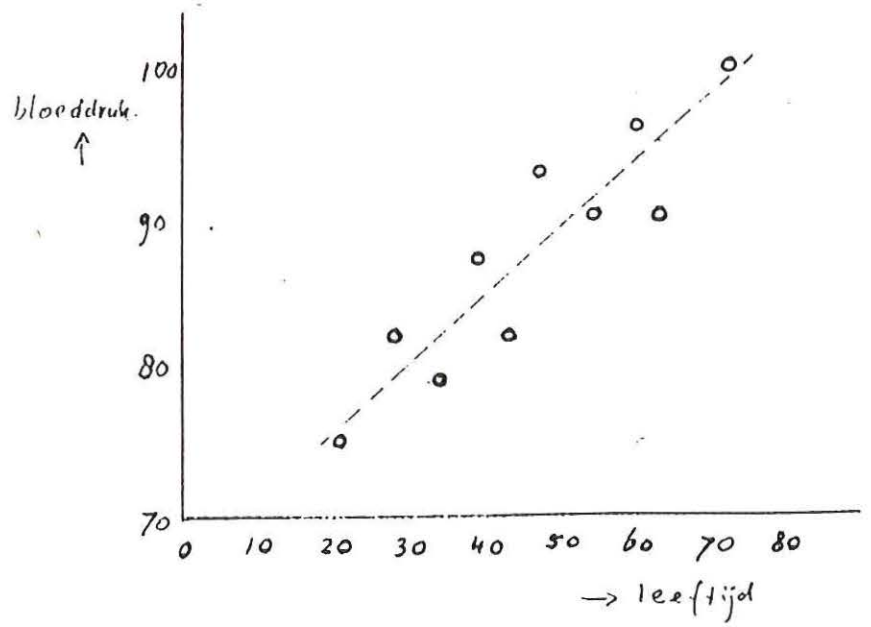
We bespreken eerst de meest eenvoudige vorm van regressie-analyse aan de hand van het volgende voorbeeld waarin de samenhang van de bloeddruk met de leeftijd wordt onderzocht voor vrouwen. Hierbij is voor een 10-tal vrouwen de leeftijd en de (diastolische) bloeddruk bepaald. De gegevens zijn

leeftijd	21	28	34	39	43	47	54	60	63	72
bloeddruk	75	82	79	87	82	93	90	96	90	100

Als we in een grafiek de bloeddruk tegen leeftijd uitzetten dan lijkt het verband tussen bloeddruk (Y) en leeftijd (x) lineair zodat het zinvol lijkt de samenhang te beschrijven met het volgende model $EY = \beta_0 + \beta_1 x$ waarbij we Y normaal verdeeld veronderstellen met variantie σ^2. Op basis van de waarnemingen willen we schattingen vinden voor de regressie-coefficienten β_0 en β_1 en voor de restvariantie σ^2. De schattingen voor β_0 en β_1 noteren we als b_0 en b_1 en de schatting voor σ^2 noteren we als S^2. Op deze manier onderscheiden we de werkelijke (doch onbekende) modelparameters β_0, β_1 en σ^2 van hun schattingen b_1, b_2 en S^2. De schattingen hangen van de waarnemingen af en zijn dus stochastisch. Men kan daarom spreken over hun nauwkeurigheid. Er zijn verschillende methoden die schattingen voor de onbekende modelparameters kunnen leveren, maar als Y normaal verdeeld is met constante variantie dan worden de meest nauwkeurige schattingen geleverd door de kleinste kwadraten methode. Deze methode kiest de schattingen b_0 en b_1 zodanig dat de kwadratensom van afwijkingen van de waarnemingen t.o.v. de geschatte lijn $b_0 + b_1 x$ minimaal is. (In de figuur is dit de som van kwadraten van de vertikale afstanden). Ofwel in formulevorm: kies b_0 en b_1 zodanig dat de kwadratensom

$$\Sigma_i (Y_i - (b_0 + b_1 x_i))^2$$

minimaal is, waarbij (Y_i, x_i) de waarneming voor de i vrouw voorstelt (i = 1,...,10).



Men kan bewijzen (door de kwadratensom te minimaliseren naar b_0 en b_1) dat dit de volgende schattingen oplevert:

$$b_1 = \frac{\Sigma_i (x_i - \bar{x})(Y_i - \bar{Y})}{\Sigma_i (x_i - \bar{x})^2}$$

$$b_0 = \bar{Y} - b_1 * \bar{x}$$

De restvariantie σ^2 schatten we met:

$$S^2 = \frac{1}{n-2} \Sigma_i (Y_i - (b_0 + b_1 x_i))^2$$

De schatter S^2 heeft n-2 vrijheidsgraden omdat er 2 vrijheidsgraden gebruikt worden voor de schatting van $ß_0$ en $ß_1$.

S^2 is de niet-verklaarde variantie van de responsvariabele. Als we deze vergelijken met de totale variantie van Y gedefiniëerd door

$$S^2_{tot} = \frac{1}{n-1} \Sigma_i (Y_i - \bar{Y})^2,$$

dan zou men

$$R^2_{adj} = \frac{S^2_{tot} - S^2}{S^2_{tot}} * 100\%$$

kunnen definiëren als het percentage van de variantie van Y dat door de regressie verklaard wordt.

In het bovenstaande voorbeeld m.b.t. de bloeddruk van vrouwen vinden we op basis van de 10 waarnemingen de volgende schattingen $b_0 = 67.1$, $b_1 = 0.44$, $S^2 = 12.6$. Dit levert de volgende voorspelformule op: voorspelde bloeddruk = 67.1 + 0.44 * leeftijd. De individuele spreiding van de bloeddruk van vrouwen bij een vaste leeftijd wordt gegeven door de residuele standaard-

deviatie: sd = $\sqrt{12.6}$ = 3.5. Omdat S^2_{tot} = 62.3, is het percentage door de regressie verklaarde variantie gelijk aan

$$\frac{62.3 - 12.6}{62.3} * 100\% = 80\%$$

We hebben m.b.v. dit model de waargenomen bloeddruk zo goed mogelijk opgesplitst in de voorspelde waarde (engels: fitted value) en de individuele afwijking t.o.v. de voorspelling (engels: residual).

waargenomen bloeddruk		voorspelde bloeddruk		residuele afwijking
75		76.4		-1.4
82		79.4		2.6
79		82.1		-3.1
87		84.3		2.7
82	=	86.0	+	-4.0
93		87.8		5.2
90		90.9		-0.9
96		93.5		2.5
90		94.8		-4.8
100		98.8		1.2

Om te zien of het gebruikte regressiemodel voorspellende waarde heeft gebruiken we een F-toets die we afleiden door de kwadraatsom van Y als volgt op te splitsen

$$SS_{tot} = SS_{regr} + SS_{res}$$

Dit zijn precies de kwadraatsommen voor de 3 bovenstaande kolomvectoren. Als we deze 3 kwadraatsommen delen door hun eigen aantal vrijheidsgraden dan vinden we de gemiddelde kwadraatsommen MS:

$MS_{tot} = SS_{tot} / (n-1)$ $(= S^2_{tot})$
$MS_{regr} = SS_{regr} / 1$
$MS_{res} = SSres / (n-2)$ $(=S^2)$

Onder de nulhypothese dat er geen (rechtlijnige) samenhang bestaat tussen bloeddruk en leeftijd (ofwel H_0: $ß_1 = 0$) geldt dat MS_{regr} en MS_{res} beide ongeveer gelijk zijn aan σ^2. Als er echter wel lineaire samenhang is dan

zal MS_{regr} veel groter zijn dan MS_{res}. Als toetsingsgrootheid nemen we daarom de grootheid

$$F = \frac{MS_{regr}}{MS_{res}}$$

Deze heeft onder H_0 een F-verdeling met 1 en n-2 vrijheidsgraden. We noemen de samenhang significant als F groter is dan de in tabel A3 gegeven kritieke waarde. Bij regressie kan men dus een variantie-analyse uitvoeren om te zien of het regressiemodel wel of geen voorspellende waarde heeft.

Een andere methode om te toetsen of er samenhang bestaat is het toetsen van H_0: $ß_1 = 0$). De toetsgrootheid hiervoor vindt men door b_1 te delen door zijn eigen standaardfout. Dit levert een grootheid die onder H_0 t_{n-2} verdeeld is. Dus men concludeert dat de samenhang significant is als $|b_1/se(b_1)|$ groter is dan een kritieke waarde t (t wordt gevonden in een Student-tabel bij n-2 vrijheids-graden en zal meestal in de buurt van 2 liggen). Voor de berekening van de standaardfout van b_1 hebben we (als we geen computer-programma gebruiken) nodig dat

$$var(b_1) = \frac{\sigma^2}{\Sigma_i (x_i - \bar{x})^2}$$

Door in deze formule σ^2 te vervangen door S^2 vinden we na worteltrekken $se(b_1)$.
Voor de eenvoudige regressie met één predictor-variabele zijn de t-toets en de hierboven besproken F-toets equivalent: de F-grootheid is het kwadraat van de t-grootheid. Een 95%-betrouwbaarheidsinterval voor de helling $ß_1$ is gelijk aan $b_1 \pm t*se(b_1)$.

Voor de bovenstaande waarnemingen geldt dat $MS_{regr} = 460$ en $MS_{res} = 12.6$ dus $F = 460/12.6 = 36.5$. F is groter dan de kritieke waarde 5.32, dus concluderen we dat er samenhang bestaat tussen bloeddruk en leeftijd. Evenzo vinden we $se(b_1) = 0.073$, dus $b_1/se(b_1) = 0.44/0.073 = 6.0$, hetgeen op een significante samenhang duidt.

Opmerking:
Vroeger werd vaak de correlatiecoefficient R gebruikt om de mate van samenhang tussen de y- en x-variabele weer te geven. Omdat deze grootheid in de regressie situatie geen informatie toevoegt aan de beschrijving, zullen we het begrip correlatie onbesproken laten.
Voor het kwadraat van de correlatiecoefficient geldt dat

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{residual}}{SS_{tot}}$$

Deze R^2 wordt vaak de determinatiecoefficient genoemd.
Wij hanteren R^2_{adj} in plaats van R^2 omdat deze rekening houdt met het aantal x-variabelen dat in het regressiemodellen is opgenomen.

Gebruik van regressie voor voorspelling
De berekende regressielijn $EY = b_0 + b_1x$ $(= \bar{Y} + b_1(x-\bar{x}))$ wordt vaak gebruikt om voor nieuwe objecten de respons Y te voorspellen uit de predictorvariabele x. Een significante samenhang tussen Y en x hoeft nog niet in te houden dat die voorspelling de gewenste precisie bezit. We zullen nu die precisie bepalen. Hiervoor moeten we met twee bronnen van onnauwkeurigheden rekening houden. Ten eerste bezit de regressielijn een bepaalde onnauwkeurigheid (de schattingen b_0 en b_1 zijn stochastisch omdat ze gebaseerd zijn op een beperkt aantal waarnemingen). Ten tweede zal een nieuw object een variantie rond de regressielijn vertonen. Voor de onnauwkeurigheid van de voorspelling van een nieuw object waarvoor $x = x_0$ geldt:

$$\text{var(voorspelling)} = (\text{se}(b_0 + b_1x))^2 + \sigma^2$$

Voor de onnauwkeurigheid van de regressielijn $b_0 + b_1x$ in $x = x_0$ geldt:

$$\text{se}(b_0+b_1x) = \sqrt{(S^2 \cdot [\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}])} \qquad (18.1)$$

(omdat $\text{var}(\bar{Y} + b_1(x_0 - \bar{x})) = \sigma^2/n + (x_0 - \bar{x})^2.\text{var } b_1$).

De onnauwkeurigheid van de voorspelling van Y voor een nieuw object met

$x = x_0$ is dus:

$$\text{se(voorspelling)} = \sqrt{(S^2 \cdot [\, 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\,])} \qquad (18.2)$$

(Merk op: x_i en $\bar{x}$ slaan op de waarnemingen waarop de regressielijn gebaseerd is en x_0 is de waarde van een nieuw object.)

Een 95%-betrouwbaarheidsinterval voor de voorspelling is dan gegeven door
$b_0 + b_1 x \pm t_{n-2} * \text{se(voorspelling)}$.

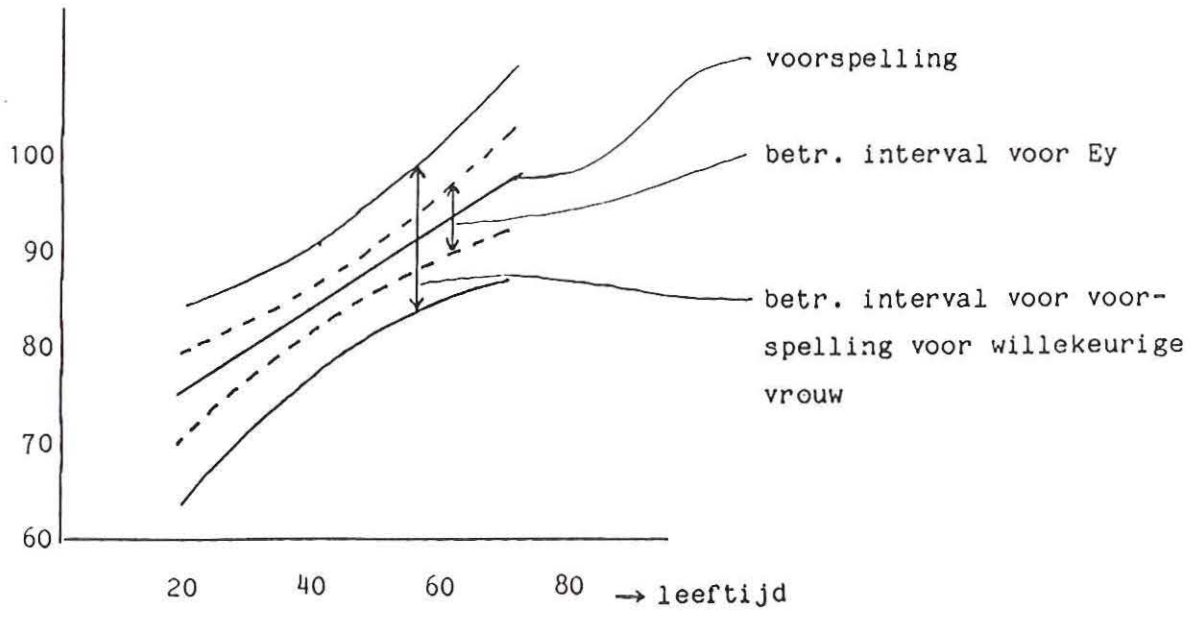
Voor de hierboven beschreven gegevens m.b.t. bloeddruk vinden we:
$\bar{x} = 46.1$ en $\Sigma(x_i - \bar{x})^2 = 2377$. Voor vrouwen van 30 jaar is de verwachte bloeddruk gelijk aan $67.1 + 0.44 * 30 = 80.3$ met

$se = \sqrt{(\, 12.6 * (\, 1/10 + (30-46.1)^2/2377)\,)} = 1.6$

Dus is het 95%-betrouwbaarheidsinterval voor de gemiddelde bloeddruk van vrouwen van 30 gelijk aan $80.3 \pm 2.31 * 1.6$ ofwel (76.6, 84.0). Voor de voorspelde bloeddruk van een willekeurige vrouw van 30 jaar geldt echter

$se = \sqrt{(\, 12.6 * (\, 1 + 1/10 + (30 - 46.1)^2/2377)\,)} = 3.9$

zodat een 95%-betrouwbaarheidsinterval voor de bloeddruk van een willekeurige vrouw van 30 gelijk is aan $80.3 \pm 2.31 * 3.9$ ofwel (71.3, 89.3).
In de volgende figuur zijn voor elke leeftijd tussen 20 en 70 jaar de 2 betrouwbaarheids-intervallen weergegeven.



Aan de formules (18.1) en (18.2) kan men zien dat de voorspelling het nauwkeurigst is als x_0 gelijk is aan $\bar{x}$. De onnauwkeurigheid neemt toe naarmate x_0 verder verwijderd ligt van $\bar{x}$, vooral buiten het bereik van de voor de 0 regressie gebruikte waarden x_i wordt dit merkbaar (echter: zulke extrapolaties zijn altijd gevaarlijk omdat men de lineariteit buiten het bereik niet gecontroleerd heeft).

Verder zien we aan formules (18.1) en (18.2) dat men de onnauwkeurigheid van de regressielijn tot nul kan laten reduceren door zeer veel waarnemingen te verrichten (d.w.z. n). Echter de se van de voorspelling wordt nooit kleiner dan σ. Voor herhaald gebruik van een eenmaal vastgestelde relatie voor voorspellingsdoeleinden moet de eis worden gesteld dat de onnauwkeurigheid van de regressielijn relatief klein is t.o.v. σ.

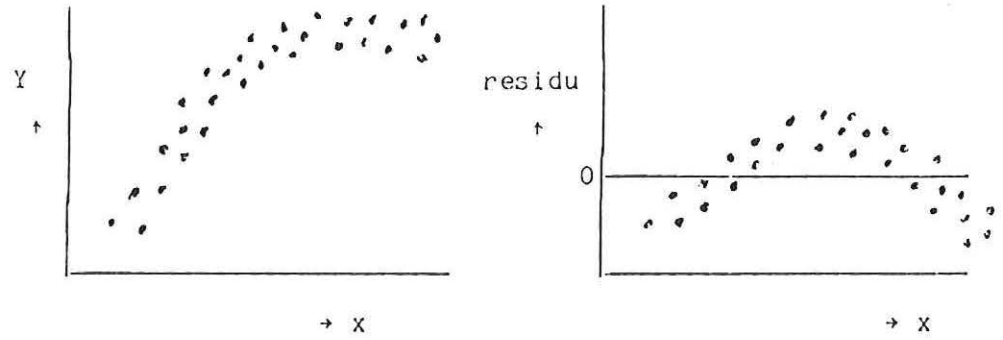
Met betrekking tot het kiezen van een proefopzet (d.w.z. de keuze van de instelwaarden x_i) zien we ook nog het volgende aan de formules (18.1) resp. (18.2). Bij een vast aantal waarnemingen n is de voorspelfout minimaal als $(x_i - \bar{x})^2$ maximaal is, dus als de helft van de waarnemingen uiterst links van het bereik van x ligt en de andere helft uiterst rechts. Op deze wijze kan men echter niet controleren of het verband wel rechtlijnig is. Als men dat wenst te doen zal men dus ook een aantal waarnemingen moeten doen in het midden van het bereik van x.

18.3 Modelcontrole bij regressie-analyse

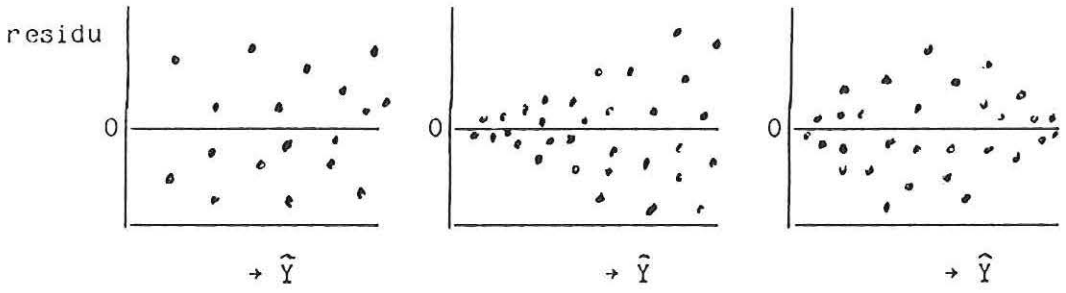
Regressie-analyse wordt vaak toegepast, maar als men de techniek blindelings gebruikt dan hoeft het resultaat niet altijd even zinvol zijn. Het verdient aanbeveling te controleren of aan de modelveronderstellingen voldaan is. Voor modellen met één predictorvariabele x kunnen modelafwijkingen vrij eenvoudig op grafische wijze worden opgespoord. Door echter maten voor de afwijkingen te definiëren zijn generalisaties mogelijk voor de situatie met meer predictorvariabelen.

1. De eerste modelafwijking betreft het systematische deel van het model. De aangepaste rechte lijn vormt geen adequate beschrijving, als het verband kromlijnig is. Dit kan men zien door een grafiek te maken van Y tegen x. Als men toch een rechte lijn aanpast dan geeft ook de grafiek

van de residuen tegen x een aanwijzing voor kromlijnigheid. Er bestaan formele toetsen voor het vaststellen van afwijkingen van het model (lack-of-fit toetsen, zie verderop).



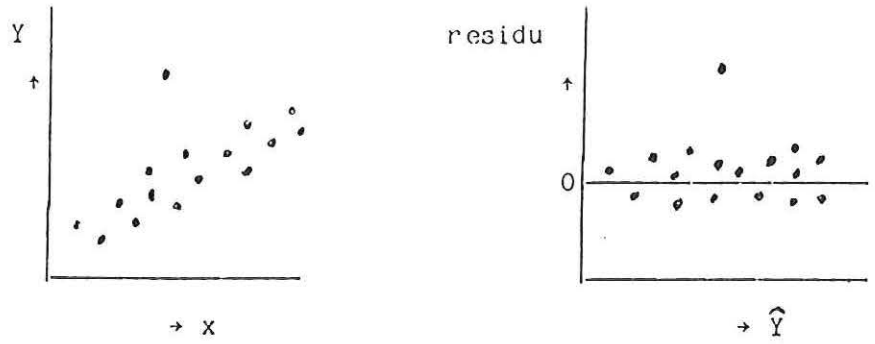
2. Met betrekking tot het stochastische deel van het model is een belangrijke veronderstelling dat de variantie van Y constant is. Dit kan men controleren door de residuen uit te zetten tegen de gefitte waarden $\hat{Y}$.



De linker figuur treedt op als er geen modelafwijkingen zijn. In de middelste figuur is de residuele standaardafwijking evenredig met het niveau van de voorspelling (ofwel de variatiecoefficient is constant). In deze situatie zou men een logtransformatie van Y kunnen overwegen om de variantie te stabiliseren. In de rechter figuur is de variantie het kleinst voor lage en hoge waarden van Y en het grootst voor intermediaire waarden van Y. Dit komt o.a. voor als Y een percentage of fractie voorstelt. In zo'n siutatie is een logistische regressie-analyse aan te bevelen.

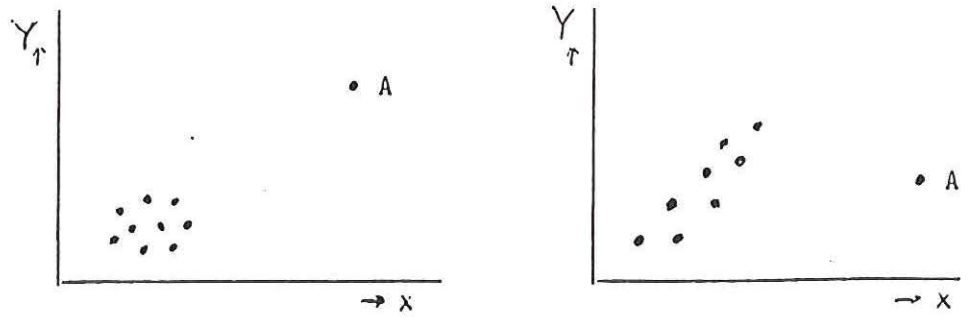
3. De normaliteit van het stochastische deel kan men controleren door een histogram te maken van de residuen. Afwijkingen van normaliteit kan men echter alleen constateren als het aantal waarnemingen groot is. Geringe afwijkingen van normaliteit blijken echter toelaatbaar bij regressie-

analyse. Een van de belangrijkste afwijkingen van normaliteit vormen de zgn. uitbijters (engels: outliers). Dit zijn waarnemingen die niet goed passen bij de rest van de gegevens; ze kunnen het aanpassen van het regressiemodel verstoren omdat ze een belangrijke invloed hebben in de te minimaliseren kwadratensom.



Men kan uitbijters opsporen door de waarnemingen uit te zetten tegen x of door de residuen uit te zetten tegen de gefitte waarden $\hat{Y}$. Men kan ook kijken naar de gestandaardiseerde residuen (worden gedefinieerd aan het eind van deze paragraaf). Deze zouden voor 95% van de waarnemingen tussen -2 en 2 moeten liggen. Als een (of enkele) van de gestandaardiseerde residuen ver buiten dit interval ligt, dan moet men die waarneming bijzondere aandacht geven. Als blijkt (bv. na inspectie van het lab-journaal) dat bij uitvoering van deze waarneming iets is misgegaan, dan kan men deze weglaten. Anders moet men zo'n waarneming handhaven. Zo'n waarneming kan een sleutel zijn tot verder experimenteren (er kan bv. een 2^e predictorvariabele nodig zijn).

4. Een andere vorm van extreme waarnemingen zijn afwijkingen m.b.t. de x-variabele. Voorbeelden hiervan zijn de punten A in de onderstaande figuren.



De x-waarde van de punten A ligt ver buiten het bereik van de overige punten en hebben daarom een grote invloed op de ligging van de regres-

sielijn. Als men A zou weglaten, zou het regressiemodel er heel anders uitzien (in de linker figuur zou het verband zelfs afwezig zijn na weglating van A). Bij een predictorvariabele kan men zulke invloedrijke punten opmerken in een grafiek van Y tegen x. Men kan ze ook opsporen door de zgn. "hefboomwerking" (engels: leverage) te berekenen. Bij één predictorvariabele is de leverage van de waarneming (x_i, Y_i) gedefinieerd door

$$h_i = \frac{1}{n} + (x_i - \bar{x})^2 \; / \; \Sigma_j (x_j - \bar{x})^2.$$

De leverage is gemiddeld gelijk aan 2/n. Punten met een hoge leverage (bv. groter dan 4/n of 5/n) liggen dus aan de buitenkant van het bereik van x. Om te zien of dergelijke punten werkelijk invloed hebben op de regressielijn (resp. op de conclusies) zou men de regressie nogmaals kunnen uitvoeren zonder deze punten. Uiteraard is het raadzaam meer waarnemingen te verzamelen in de omgeving van een punt met een hoge leverage als men vindt dat zo'n punt wel hoort bij het gewenste bereik van x.

5. Controle op onafhankelijkheid van de waarnemingen is soms ook gewenst. Daarvoor zijn methoden beschikbaar, maar daarop wordt hier niet ingegaan.

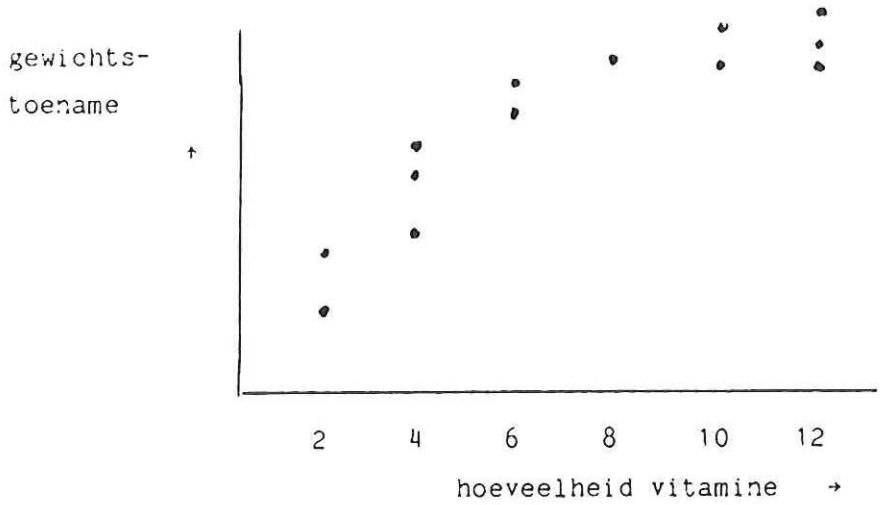
Opmerking:
De in punt 3 genoemde gestandaardiseerde residuen worden gedefinieerd als
$(Y_i - \hat{Y}_i)/S.\sqrt{(1-h_i)}$. Hierbij worden de residuen $Y_i - \hat{Y}_i$ gedeeld door hun eigen standaardfout om hun variantie gelijk te maken. We zullen hier niet formeel bewijzen dat $se(Y_i - \hat{Y}_i) = S.\sqrt{(1-h_i)}$. De factor $\sqrt{(1-h_i)}$ corrigeert hierbij voor het feit dat invloedrijke punten (d.w.z. hoge l_i) de regressielijn sterker naar zich toe trekken dan minder invloedrijke punten).

18.4 Een toets voor lack-of-fit

Een van de in paragraaf 18.3 besproken modelcontroles betreft het

systematische deel van het model: past een rechte lijn bij de waarnemingen? Dit kan men op grafische wijze controleren, maar als men bij de opzet zorgt voor herhalingen bij dezelfde x-waarden dan is het mogelijk formeel te toetsen of de lijn past bij de waarnemingen (een zgn. toets voor lack-of-fit).

Neem bijvoorbeeld een proef waarbij men wil onderzoeken hoe de gewichtstoename van ratten afhangt van de hoeveelheid van een bepaald vitamine dat wordt toegevoegd aan een basisvoer. De waarnemingen zijn weergegeven in de onderstaande figuur.



In dit voorbeeld zijn er 6 x-waarden die voorkomen in groepen van 2, 3, 2, 1, 2 en 3 herhalingen. Voor die groepen kan men een schatting berekenen vande binnen-groepen-variantie (= de in hoofdstuk 13 geïntroduceerde gepoolde variantieschatter). Deze $MS_{\text{binnen herh.}}$ (hier met 7 vrijheidsgraden) is een zuivere schatting van de variantie van Y bij vaste x (de zgn. pure error). De bij de regressie verkregen MS_{residual} bevat behalve deze pure error ook de afwijking van de werkelijke curve t.o.v. de aangepaste lijn (deze afwijking heet lack-of-fit). Voor de toetsing splitsen we SS_{residual} op als volgt

$$SS_{\text{residual}} = SS_{\text{binnen herh.}} + SS_{\text{lack-of-fit}}.$$

Deze kwadraatsommen hebben in bovenstaand voorbeeld, als men een rechte lijn aanpast, resp. 11, 7 en 4 vrijheidsgraden ($SS_{\text{lack-of-fit}}$ berekent men als het verschil van de twee andere kwadraatsommen).

Als de rechte lijn goed past bij de waarnemingen dan zullen $MS_{residual}$, $MS_{binnen\ herh.}$ en $MS_{lack\text{-}of\text{-}fit}$ alle ongeveer gelijk zijn aan σ^2. Afwijkingen van lineariteit komen terecht in $SS_{lack\text{-}of\text{-}fit}$. Dit betekent dat

$$F = \frac{MS_{lack\text{-}of\text{-}fit}}{MS_{binnen\ herh.}}$$

ongeveer 1 is (in feite F verdeeld is) ingeval van lineariteit en veel groter is als de lijn niet past bij de waarnemingen. In dit voorbeeld zullen we de nulhypothese van lineariteit verwerpen als $F > 4.12$. We zullen dan op zoek moeten naar een andere curve om de samenhang te beschrijven.

Merk op dat deze methode alleen gebruikt kan worden als men "echte" herhalingen heeft, d.w.z. waarnemingen die gemeten zijn aan verschillende experimentele eenheden. Duplometingen komen niet in aanmerking want de spreiding ervan is kleiner dan de spreiding van Y van verschillende experimentele eenheden met eenzelfde waarde voor x.

Als men niet over herhalingen beschikt dan bestaan er methoden waarbij men gebruik maakt van de spreiding van naburige punten (naburig m.b.t. de x-waarden).

HOOFDSTUK 19

CALIBRATIE

Een van de voorbeelden waarbij regressie-analyse wordt gebruikt had betrekking op calibratie. Calibratie kan worden gedefinieerd als het vaststellen van een functionele relatie tussen de resultaten verkregen met een meetmethode, meetinstrument of laboratorium en bekende of gemeten referentiewaarden. De verkregen relatie dient voor het herleiden van toekomstige meetresultaten naar de referentieschaal. Bij calibratie kunnen een aantal situaties aan de orde zijn waarvan herkenning van belang is voor een juiste interpretatie van de verkregen resultaten. Deze zullen we achtereenvolgens bespreken.

1. Klassieke calibratie

In de klassieke calibratie wordt een aantal instelpunten van de te meten grootheid gekozen, bijvoorbeeld een aantal kwikgehalten in grondmonsters of een aantal doses penicilline in melk. Vervolgens wordt met de te ijken meetprocedure bij ieder instelpunt het resultaat gemeten. Als een lineaire relatie tussen de metingen Y en de instellingen x aannemelijk is kan men voor deze ijkset een lineaire regressie berekenen volgens het in hfdst. 18 besproken model:

$$EY = \alpha + \beta x$$

$$Y = EY + E \quad \text{met var } E = \sigma^2$$

De eerste uitdrukking specificeert de functionele relatie tussen Y en x en de tweede uitdrukking de meetfoutstructuur. De regressieberekening resulteert ondermeer in schattingen voor α, β en σ^2. Met de aldus gespecificeerde ijklijn kan men voor nieuwe metingen Y de bijbehorende waarde van x vaststellen als volgt:

$$\hat{x} = (Y - \alpha) / \beta$$

Aannemende dat het aantal punten in de ijkset voldoende groot is om de onnauwkeurigheid van de schattingen α en β te kunnen verwaarlozen vindt men:

$$se(\hat{x}) = \sigma/\beta$$

De waarneming Y wordt dus als het ware gecorrigeerd voor de niveau-afwijking α en voor de schaalfactor β tussen de meetschalen van Y en x. De aldus verkregen schatter $\hat{x}$ wordt de klassieke schatter genoemd.

Voor de proeftechnische uitvoering van een dergelijk experiment zijn enkele opmerkingen relevant:

- Aantal en plaats van de instelpunten kan men in principe vrij kiezen. Het is echter verstandig er voor te zorgen dat het interval waarin toekomstige meetwaarden verwacht worden volledig wordt bestreken. Voorts is het uit een oogpunt van efficiëntie gewenst, zeker in gevallen waarin de aard van de relatie voldoende bekend is, het aantal instelpunten zoveel mogelijk te beperken (zie hfdst. 18).
- Het meetresultaat Y is het enkelvoudige resultaat zoals dat volgens het meetvoorschrift wordt verkregen. Als bijvoorbeeld duplometingen worden voorgeschreven, met regels voor herhaling van metingen bij te grote duploverschillen, dienen niet de afzonderlijke metingen te worden gebruikt maar het overeenkomstig het voorschrift vastgestelde eindresultaat (meestal gemiddelde of mediaan). Voor een juiste waardering van de op te geven precisie is het van groot belang dat zowel bij de ijkset als bij de toekomstige metingen dezelfde procedure wordt gehanteerd.
- In principe dienen alle metingen in aselecte volgorde te worden uitgevoerd. Eventueel kan men ingewikkelder proefopzetten met blokindelingen, instrumenten, analisten e.d. inpassen.

2. Onderlinge calibratie van meetmethoden met foutvrije referentiemethode

Soms is het niet mogelijk de instellingen van de te meten variabele te fixeren op een aantal willekeurig te kiezen instelpunten maar is wel een referentiemethode beschikbaar waarmee men de variabele in een aantal monsters "foutvrij" kan vaststellen (bijvoorbeeld Kjeldahl-destructie bij eiwitbepalingen). Voor deze situatie kan hetzelfde model worden gesteld als bij de klassieke calibratie en kan ook de klassieke schatter $\hat{x}$ worden berekend. Een verschil met de klassieke calibratie-situatie is dat de ijkset zou kunnen worden gekozen als een aselecte steekproef uit de populatie van toekomstige meetwaarden. Aannemende dat deze normaal verdeeld zijn is ook de regressie van x op Y een goed gedefinieerde relatie:

$$x = \gamma + \delta Y + F \quad \text{met var } F = \tau^2$$

De parameters γ, δ en τ^2 kan men eveneens met de behandelde regressieberekeningen vaststellen. De voor de hand liggende schatter voor x wordt nu

$$\hat{x} = \gamma + \delta Y$$

Deze schatter, die in het algemeen verschilt van de klassieke schatter, wordt de inverse schatter genoemd. Men kan bewijzen dat de variantie van de inverse schatter, τ^2, bij verwaarlozing van de onnauwkeurigheid in γ en δ,

gelijk is aan $\rho^2\sigma^2/\beta^2$, waarin ρ de correlatie-coëfficiënt tussen x en Y is. Omdat $\rho < 1$ is de precisie van de inverse schatter dan ook beter dan die van de klassieke schatter en deze verdient derhalve de voorkeur. De grotere precisie van deze schatter is in feite te danken aan het gebruiken van de extra informatie uit het gegeven dat de x-waarden aselect getrokken zijn uit de toekomstige populatie van meetwaarden. Men kan bewijzen dat de inverse schatter geschreven kan worden als een gewogen gemiddelde van de klassieke schatter en de uit de extra informatie af te leiden schatter:

$$\hat{X}_{inverse} = \rho^2 . \hat{x}_{klassiek} + (1-\rho^2) . \hat{\mu}_x$$

waarin $\hat{\mu}_x$ het gemiddelde van de x-waarden van de ijkset is (weer uitgaande van een grote ijkset). Aan deze uitdrukking ziet men ook dat als $\rho^2 \sim 1$ dat de beide schatters elkaar niet veel zullen ontlopen. Met name bij grote meetfouten is de keuze wel van belang.

Ook hier zijn enkele proeftechnische opmerkingen relevant:

- De inverse regressie kan strikt genomen alleen berekend worden als de x-waarden in de ijkset werkelijk een aselecte steekproef uit een populatie vormen en de bijbehorende inverse schatter is alleen voor die populatie bruikbaar. Men kan echter ook altijd op indirecte wijze de inverse schatter berekenen, zoals boven aangegeven, met behulp van afzonderlijk verkregen informatie over de klassieke calibratie relatie en over de doelpopulatie. Hiervoor is het nog relevant te weten dat ρ^2 geschreven kan worden als volgt:

  $$\rho^2 = \beta^2\sigma_X^2/(\beta^2\sigma_X^2 + \sigma^2)$$

  waarin σ_X^2 de variantie van de x-waarden in de doelpopulatie is. Deze wijze van werken biedt de mogelijkheid enerzijds gebruik te maken van een optimale keuze van instelpunten voor het vaststellen van de functionele relatie en anderzijds de inverse schatter toe te snijden op eventueel verschillende doelpopulaties.
- Hoewel onder de aangegeven omstandigheden de inverse regressie goed gedefinieerd is, is deze zelf niet op een natuurlijke wijze te interpreteren als een functionele relatie omdat Y in feite een aan toevalsvariatie onderhevige meetuitkomst voorstelt. Als de x-waarden niet aselect getrokken zijn uit de doelpopulatie dan is de inverse regressie niet goed gedefinieerd. Men kan dan op de boven beschreven indirecte wijze te werk gaan. Men kan echter ook - zonder rekening te houden met scrupules omtrent de juistheid van het model - de rekenregels voor de inverse regressie toepassen. Hoewel men dan theoretisch niet meer uit de voeten

kan is uit simulatie-studies meer malen gebleken dat ook in deze situatie de inverse regressie en de daarop gebaseerde inverse schatter betere resultaten oplevert dan het gebruiken van de klassieke schatter. Dit voert ons tot de voorzichtige conclusie dat het in de praktijk altijd raadzaam is de inverse regressie te berekenen en daarop de schatter voor x te baseren. Met andere woorden: in de praktijk zal men de te voorspellen variabele altijd als de afhankelijke variabele in de regressie kunnen nemen.

3. Metingen met de referentiemethode zijn niet foutvrij

Soms is wel een referentiemethode beschikbaar, die in principe zuivere resultaten oplevert, maar die zelf ook relatief belangrijke meetfouten bevatten. Dit zal zeker het geval kunnen zijn als de referentiemethode het uitvoeren van een omslachtige procedure op een kleine hoeveelheid materiaal betreft, terwijl de andere methode veel eenvoudiger uitvoerbaar is en bovendien per meting betrekking heeft op een relatief veel grotere hoeveelheid materiaal.

Voorbeeld: met het microscoop tellen van het aantal zaadcellen in verdund sperma versus een colorimetrische bepaling van de spermadichtheid.

In deze situatie kan men zonder bezwaar de inverse regressie berekenen en deze gebruiken voor het vaststellen van de inverse schatter. De variantie van de punten rond de regressielijn, τ^2, bevat nu echter ook een bijdrage van de meetfouten van de referentiemethode. Deze is derhalve niet rechtstreeks bruikbaar om de onnauwkeurigheid van de eenvoudige snelle methode te karakteriseren. Daarvoor dient men τ^2 eerst te verminderen met de variantie σ_0^2 van de meetfout van de referentiemethode:

$$\tau_*^2 = \tau^2 - \sigma_0^2.$$

Enkele proeftechnische opmerkingen:

- σ_0^2 moet bekend zijn om τ^2 daarvoor te kunnen corrigeren. Als σ_0^2 niet bekend is moet deze worden geschat uit onafhankelijke herhalingen van de metingen met de referentiemethode. Onafhankelijk betekent hier dat niet zonder meer een duplo-bepaling onder herhaalbaarheidscondities moet worden uitgevoerd maar onder nader te definiëren intermediaire reproduceerbaarheidscondities.
- Het vaststellen van de functionele relatie tussen Y en de met meetfout belaste x kan niet met standaard regressietheorie en vraagt het toepassen van methoden die uitgaan boven wat in deze basiscursus aan de orde kan komen; zie echter ook paragraaf 6. Men zal deze functionele relaties

echter zelden nodig hebben, omdat men op de boven aangegeven wijze altijd een schatter $\hat{x}$ met zijn onnauwkeurigheid kan bepalen.
Wel kan het voor een goed begrip van de situatie wenselijk zijn functionele relaties te gebruiken voor de beschrijving daarvan.

4. Soms is er geen referentiemethode

Als een referentiemethode niet direct gerelateerd is aan in principe goed gedefinieerde kenmerken, zoals het aantal zaadcellen per ml sperma of het aantal IE vitamine per gram, maar uitsluitend gebaseerd is op afspraak of gewoonte kan de situatie ontstaan dat men een nieuwe methode meer vertrouwt dan de actuele referentie. Het kan in een dergelijke situatie zeer wel zinvol zijn niet te calibreren op de oude referentie. In feite is er dan geen referentiemethode. Er dient een nieuwe referentie te worden aangewezen. Denk bijvoorbeeld aan nieuwe en oude methoden voor het meten van de longinhoud van mensen: het begrip longinhoud is geen absoluut vaststelbaar kenmerk maar wordt in feite gedefinieerd door de wijze waarop deze wordt gemeten. Calibratie-relaties kunnen in dergelijke situaties worden gebruikt voor het omrekenen van resultaten verkregen met de oude methode naar de nieuwe of omgekeerd. Men gebruikt hiervoor dus de regressie van de nieuwe methode op de oude, respectievelijk de regressie van de oude methode op de nieuwe.

5. Calibratie vraagt niet altijd het berekenen van regressielijnen

Soms is het aannemelijk dat de relatie tussen de te calibreren methode en de referentiemethode kan worden weergegeven als een rechte lijn door de oorsprong, waarvan de helling niet noodzakelijk gelijk is aan 1. Voorts ziet men in zo'n situatie vaak dat de grootte van de meetfout toeneemt met de grootte van de meetuitkomst zodat een logtransformatie op zijn plaats kan zijn. Door deze logtransformatie wordt het regressiemodel $y = \beta x + e$ in feite vervangen door het model

$$\ln y = \ln \beta + \ln x + \epsilon$$

of

$$\ln y - \ln x = \ln \beta + \epsilon$$

waarin ϵ een constante variantie heeft en $\ln \beta$ de te schatten parameter is. Onder dit model vindt men als schatting voor $\ln \beta$ het gewone gemiddelde van de verschillen $(\ln y - \ln x) = \ln (y/x)$.

Voorbeeld:

Onderstaande meetuitkomsten hebben betrekking op de meting van Hg in monsters rivierklei met respectievelijk een standaardmethode (neutronen-

activeringsanalyse = x) en een andere methode (koude dampatomaire absorptie spectrometrie = y). Voor later gebruik zijn in de tabel nog kolommen z, ln z/x en ln z/y toegevoegd waarin z betrekking heeft op een variant van y.

x	y	ln y/x	z	ln z/x	lnz/y
0.96	0.88	- 0.087	0.89	- 0.076	+ 0.0011
0.17	0.15	- 0.125	0.15	- 0.125	0
0.23	0.20	- 0.140	0.20	- 0.140	0
0.08	0.08	- 0.000	0.08	0	0
0.42	0.43	+ 0.024	0.38	- 0.100	- 0.124
0.08	0.07	- 0.134	0.06	- 0.288	- 0.154
0.19	0.16	- 0.172	0.15	- 0.236	- 0.065
0.07	0.05	- 0.336	0.05	- 0.336	0
0.22	0.18	- 0.201	0.18	- 0.201	0
0.61	0.59	- 0.033	0.57	- 0.068	- 0.034
0.15	0.14	- 0.069	0.12	- 0.223	- 0.154
1.23	1.16	- 0.059	1.14	- 0.076	- 0.017

Gemiddelde en standaardafwijking van ln (y/x) zijn respectievelijk - 0.111 en 0.098. Derhalve kan men β schatten als $e^{-0.111} = 0.895$ met een standaardafwijking van $(0.098/\sqrt{12}) \times 100\% = 2.8\%$. β is dus significant lager dan 1 (t-toets met t_{11} = - 0.111/0.028 = - 3.96; P = 0.00).

Met behulp van variantie-analyse kan men nagaan of de relatie tussen de beide methoden verschillend is voor de 4 grondsoorten. Dit blijkt overtuigend het geval te zijn. Zie de F-toets in onderstaande variantie-analyse:

Bron van variatie	df	MS	F	P
Grondsoorten	3	0.1160	8.11	0.00
Rest	19	0.0143		
Totaal	22			

Gemiddelden:	rivierklei	- 0.111	n = 12
	zeeklei	- 0.065	n = 5
	veengrond	- 0.057	n = 4
	dalgrond	+ 0.339	n = 2

Met name dalgrond geeft afwijkende resultaten.

Bij deze analyse kunnen enkele opmerkingen worden gemaakt

- door berekening van regressie van ln y op ln x (of omgekeerd) kan men nagaan of de relatie tussen y en x inderdaad kan worden beschouwd als

een rechte lijn door de oorsprong: de regressiecoëfficiënt van ln x zal dan niet duidelijk van 1 afwijken.

- voor de lagere waarden van x en y hebben afrondingsfouten een relatief grote invloed op de uitkomsten. Het was beter geweest ook daar de resultaten in tenminste 2 significante cijfers weer te geven.
- wanneer het aannemelijk is dat de standaardafwijking van de meetfout in x verwaarloosbaar klein is t.o.v. die in y kan men uiteraard de residuele standaardafwijking beschouwen als maat voor de meetfout van y. Veelal zal echter ook x niet foutvrij worden gemeten. In dat geval is de residuele variantie de som van de varianties van ln x en ln y. Als men de variantie van ln x kent kan men die voor ln y verkrijgen door de variantie van ln y/x te verminderen met de variantie van ln x; modelafwijkingen komen zodoende terecht in de variantie van ln y. Voor het geval dat men de variantie van ln x niet kent bestaan er methoden om de varianties van de meetfouten van de afzonderlijke meetmethoden vast te stellen op grond van enkelvoudige waarnemingen van ln x en ln y aan een aantal monsters. Deze methoden zijn echter zeer gevoelig voor afwijkingen van de juistheid van het model. Zie de volgende paragraaf.

6. Vergelijking van de precisie van een aantal meetmethoden op basis van enkelvoudige waarnemingen

Voor het algemene model voor functionele relaties:

$$EY = \alpha + \beta Ex$$
$$Y = EY + E \qquad \text{met var } E = \sigma_1^2$$
$$X = EX + F \qquad \text{met var } F = \sigma_0^2$$

kan het vaststellen van de functionele relatie tussen de werkelijke waarden EY en EX en van σ_0^2 en σ_1^2 niet worden uitgevoerd met de gewone regressiemethoden. Als men zou weten dat β=1 kan men α en de precisie daarvan bepalen met eenvoudige methoden zoals in paragraaf 5 geïllustreerd voor de logaritmen van de metingen, waarvoor dit model geldt. Er bestaan voor deze situatie ook manieren om de varianties van de meetfouten van de afzonderlijke methoden vast te stellen op basis van de enkelvoudige meetuitkomsten. Dit is in principe aantrekkelijk omdat men dan geen onafhankelijke herhalingen nodig heeft van de metingen aan afzonderlijke monsters, met alle problemen van dien. De methoden zijn beschreven door Grubbs (1948, 1973) en door Russel & Bradley (1958); voor een samenvatting

zie ook Jansen (1977). Helaas zijn de methoden extreem gevoelig voor modelafwijkingen, zodat hun bruikbaarheid in de praktijk niet groot is.

Voor een goed begrip wordt het model als volgt opgeschreven voor 2 meetmethoden die de meetresultaten Y_0 en Y_1 produceren aangaande de te meten grootheid x:

$$Y_0 = x + F_0 \qquad \text{var } F_0 = \sigma_0^2$$
$$Y_1 = \alpha + x + F_1 \qquad \text{var } F_1 = \sigma_1^2$$

Voor het verschil en de som geldt dan

$$Y_1 - Y_0 = \alpha + F_1 - F_0 \qquad \text{var } (F_1 - F_0) = \sigma_0^2 + \sigma_1^2$$
$$Y_1 + Y_0 = \alpha + 2x + F_1 + F_0 \qquad \text{var } (F_1 + F_0) = \sigma_0^2 + \sigma_1^2$$

Omdat het verschil onafhankelijk is van x zal de regressiecoëfficiënt β van de som op het verschil ook geen bijdrage van x bevatten; deze is gelijk aan de volgende functie van de modelparameters

$$\beta = (\sigma_1^2 - \sigma_0^2)/(\sigma_0^2 + \sigma_1^2).$$

Hiermee heeft men dus een middel in handen om enerzijds de hypothese $\sigma_1^2 = \sigma_0^2$ te toetsen op basis van de toets op correlatie en om anderzijds σ_0^2 en σ_1^2 afzonderlijk te schatten op basis van

$$\sigma_0^2 = \text{var } (Y_1 - Y_0) * (1-\beta)/2$$
$$\sigma_1^2 = \text{var } (T_1 - Y_0) * (1+\beta)/2$$

De precisie van deze schatters hangt af van de variabiliteit van x (de produktvariabiliteit) omdat deze deel uitmaakt van de spreiding van de punten rond de regressielijn van de som op het verschil. Deze zal in praktische situaties relatief groot zijn; de precisie zal daarom beperkt zijn en bijgevolg ook het vermogen om kleine verschillen tussen σ_0^2 en σ_1^2 te onderscheiden m.b.v. de toets. Een betere schattingsmethode is beschikbaar voor 3 of meer (p) meetmethoden waarvoor als model geldt

$$Y_{ij} = \alpha_i + x_j + F_{ij} \quad \text{met var } F_{ij} = \sigma_i^2$$

waarin de index j betrekking heeft op de (n) verschillende monsters. Als men de waarnemingen vermindert met rij- en kolomgemiddelden krijgt men de residuen

$$d_{ij} = Y_{ij} - Y_{i.} - Y_{.j} + Y_{..}$$

waarin de constanten α_i en x_j niet meer voorkomen. Men kan afleiden (zie de literatuur) dat als schatter voor σ_i^2 kan worden genomen:

$$\hat{\sigma}_i^2 = [p(p-1)\ \Sigma_j\ d_{ij}^2 - \Sigma_i \Sigma_j\ d_{ij}^2]/[(n-1)(p-1)(p-2)].$$

Voorbeeld

In het boven behandelde voorbeeld met de metingen $Y_1 = \ln y$ en $Y_0 = \ln x$ voor rivierklei verkrijgt men negatieve schattingen voor een van de varianties, zodat hiervoor de methode onbruikbaar is. Oorzaken hiervan kunnen zijn de geringe precisie van schatten voor het geval met 2 meetmethoden, maar ook de relatief grote afrondingsfouten en voorts modelafwijkingen.

Toepassen van de formule voor 3 meetmethoden, gebruik maken van de gegevens over x, y en z, levert op:

$$\sigma_0^2 = 0.0079$$
$$\sigma_1^2 = 0.0018$$
$$\sigma_2^2 = 0.0024$$

Dit resultaat lijkt evenmin erg overtuigend; het voorbeeld lijkt niet erg geschikt voor een illustratie van deze methoden.

Bij deze analysemethode dienen de volgende kanttekeningen te worden gemaakt:

- de methode voor 2 meetmethoden is relatief onnauwkeurig; alle methoden kunnen resulteren in negatieve schattingen voor de varianties.
- de methoden zijn zeer gevoelig voor modelafwijkingen m.b.t. de aanname dat β=1. Kleine afwijkingen kunnen, doordat deze worden vermenigvuldigd met de variantie tussen de monsters, de uitkomsten beïnvloeden met als mogelijk gevolg ongeloofwaardige en/of negatieve schattingen voor de varianties van de meetfouten. Het voorkomen hiervan is derhalve een indicatie dat het model niet juist is.

7. Interferentie-effecten en instabiliteit

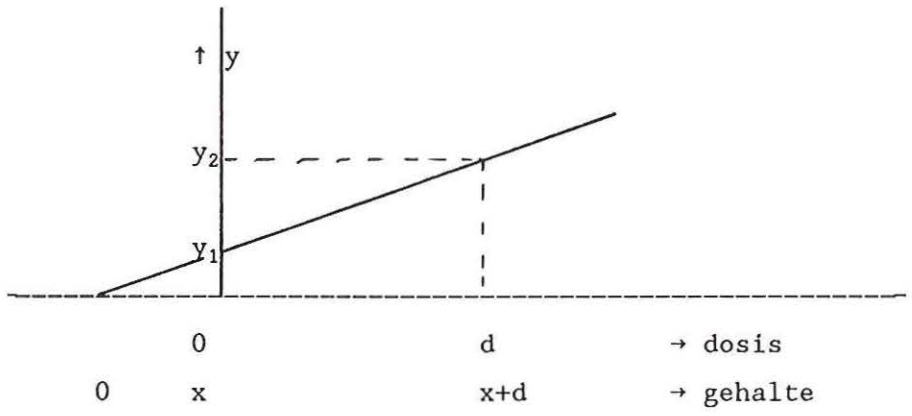
In de cursus is aandacht besteed aan de begrippen herhaalbaarheid en reproduceerbaarheid die beogen op een genormeerde wijze de meetfout van een meetmethode te kwantificeren onder respectievelijk zoveel mogelijk constante en zoveel mogelijk verschillende omstandigheden. Soms zijn deze begrippen niet afdoende om de onnauwkeurigheid van een meetmethode vast te leggen. Stel dat interferentie-effecten (ook wel matrix-effecten genoemd?) optreden als gevolg van de aanwezigheid van wisselende onbekende hoeveelheden storende substanties bij een eenvoudige of snelle meetmethode (zoals

bijvoorbeeld een NIR-meting) maar niet bij de referentiemethode. Dan zal de variantie van de punten rond de regressielijn terecht de bijdrage van deze storende invloeden bevatten omdat deze geacht moeten worden deel uit te maken van de meetfout van de snelle methode. Omdat de vaststelling van de herhaalbaarheid en de reproduceerbaarheid overeenkomstig de normvoorschriften plaats vinden op zoveel mogelijk identiek materiaal zullen de bijdragen van storende substanties aan de meetfout niet in deze maten kunnen terecht komen. Immers de storende substantie is in alle te meten identieke monsters gelijkelijk aanwezig. Het begrip meetfout moet in dergelijke gevallen worden uitgebreid zodat ook interferentie-effecten daaronder vallen.

Soms kan men het effect van interferentie-effecten elimineren door aan de monsters bekende hoeveelheden van de te meten substantie toe te voegen. De regressie van de meetuitkomsten Y op de bekende toevoegingen kan worden gebruikt om de in het monster aanwezige hoeveelheid te bepalen, als de meetuitkomst Y evenredig is met de aanwezige hoeveelheid. Stel bijvoorbeeld y_1 en y_2 zijn metingen bij het onbekende gehalte x en bij x+d respectievelijk, waarvoor dus geldt:

$$y_1 = \beta x + e_1$$
$$y_2 = \beta(x+d) + e_2.$$

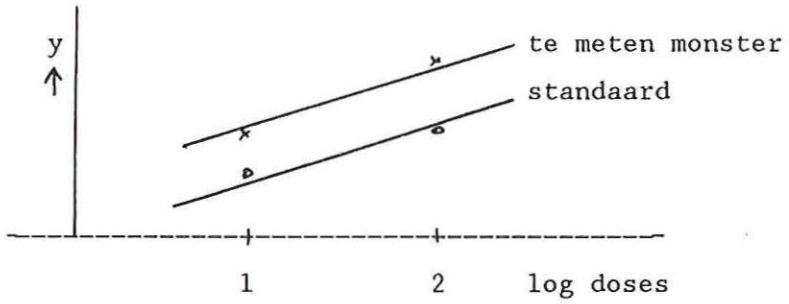
Men kan eenvoudig zien dat $b = (y_2-y_1)/d$ een schatting voor β oplevert. Vervolgens kan men $\hat{x}$ berekenen als $\hat{x} = y_1/b$.



Enkele opmerkingen:

- Ook deze methode is slechts geldig als het model juist is, d.w.z. de relatie tussen y en x moet een rechte lijn zijn door het punt y=0 bij gehalte x+d=0.

- Voor het schatten van de meetfout en/of voor modelcontrole zal het in de praktijk wenselijk kunnen zijn bij meer dan 1 toegevoegde dosis metingen te verrichten.
- Log-transformatie van y en x+d kan wenselijk zijn i.v.m. het niet constant zijn van meetfouten. Het model wordt dan echter gecompliceerder omdat x niet bekend is (het model wordt een niet-lineaire of een gegeneraliseerd lineair model).
- Interferentie-effecten zijn een vorm van instabiliteit, waarvan de oorzaak ligt in de produktvariabiliteit. Het is ook mogelijk dat de meting zelf (y dus) instabiel is doordat deze afhangt van variabele omstandigheden of kwaliteit van reagentia cq. proefmaterialen. Dit komt vaak voor bij bio-assay situaties. In dat geval kan men ook bij iedere bepaling zowel een standaardreeks als een verdunningsreeks van het te onderzoeken monster meten. Vaak zullen de regressielijnen van het meetresultaat op de logaritmen van de dosering evenwijdig blijken te zijn. In dat geval is het verschil tussen de niveaus van de regressielijnen gedeeld door hun helling, een maat voor de relatieve sterkte van de standaard en het te meten monster.



- Interferentie-effecten kan men ook trachten te elimineren door de storende substantie zelf te meten of door metingen daarvoor te corrigeren. Men komt dan op het terrein van de multivariate calibratie.

8. Multivariate calibratie

Het gaat in deze basiscursus te ver om diep in te gaan op multivariate calibratie. Het basis idee is dat men meer dan 1 variabele meet (denk aan een aantal piekfrequenties bij NIR-bepalingen) om daarmee 1 (of meer) doelgrootheden vast te stellen. Het grote aantal piekfrequenties dient onder meer om allerlei storende invloeden te elimineren en het relevante signaal optimaal te versterken. Een algemeen probleem bij multivariate

calibratie, in verband met de identificeerbaarheid van de relaties bij een beperkt aantal te meten monsters, is beperking van het aantal verklarende variabelen. Dit kan op verschillende manieren gebeuren, onder meer:

- selectie van een beperkt aantal piekfrequenties,
- maken van lineaire combinaties van veel piekfrequenties op basis van principale componenten-analyse (PCA), die in principe gericht is op het representeren van de belangrijkste variatie in de set verklarende variabelen,
- idem op basis van partially least squares (PLS), die beoogt die combinaties te kiezen die de beste voorspelling van de te meten variabelen beloven. De lineaire combinaties worden vastgesteld op basis van de enkelvoudige correlaties tussen de te meten variabelen en de voorspellende variabelen.

Het ziet er naar uit dat het op dergelijke wijze mogelijk is de invloed van storende substanties in belangrijke mate te elimineren of te reduceren, bijvoorbeeld bij NIR-bepalingen. Het blijft echter noodzakelijk alert te zijn op het niet volledig elimineren van interferentie-effecten. In dat geval is de vaststelling van herhaalbaarheid en/of reproduceerbaarheid overeenkomstig de daarvoor gangbare normen niet afdoende voor het kwantificeren van de werkelijke meetfouten van de meetmethoden.

Referenties

Grubbs, F.E. - On estimating precision of measuring instruments and product variability. J. Amer. Statist. Assoc. 43, 243-264 (1948).

Grubbs, F.E. - Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. Technometrics 15, 53-66 (1973).

Russel, T.S. and Bradley, R.A. - One-way variances in a two-way classification. Biometrika 45, 11-129 (1958).

Jansen, A.A.M. - Vergelijken van meetinstrumenten. Mededeling IWIS-TNO, Wageningen, nr. 15 (1977).

HOOFDSTUK 20 : MULTIPELE LINEAIRE REGRESSIE

Regressie-analyse kan men ook toepassen om te onderzoeken hoe een responsvariabele y afhangt van meer dan één predictorvariabele. Hiertoe wordt aan de waarnemingen een model aangepast dat y verklaart uit alle predictoren tesamen. Als er m predictoren x_1, x_2, ..., x_m zijn dan ziet het eenvoudigste lineaire model voor de waarnemingen er als volgt uit:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + e$$

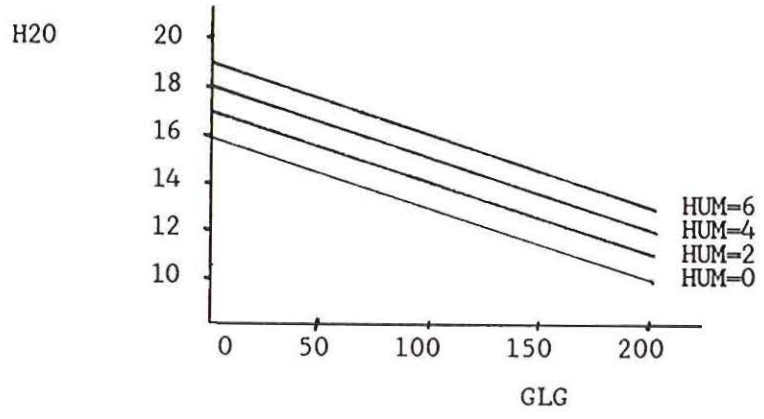
Hierin is $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$ het systematische (of verklarende) deel van het model en e heet het toevallige (of onverklaarde) deel van het model. De restterm e wordt weer normaal verdeeld verondersteld met verwachting 0 en variantie σ^2. De onbekende parameters β_0, β_1, ..., β_m heten de regressie-coëfficiënten en σ^2 heet de restvariantie.

20.1 Voorbeeld

Om te zien hoe zo'n model er uitziet beschouwen we het volgende voorbeeld waarin 2 predictorvariabelen de groei van fijnsparren verklaren. Als populatie beschouwen we bospercelen met fijnsparren op zandgronden. Als responsvariabele nemen we de hoogte in m na 20 jaar (H20). Als predictoren de gemiddeld laagste grondwaterstand (GLG, in cm beneden maaiveld) en het percentage humus in de bovengrond (HUM). Stel dat het volgende model geldt:

$$H20 = 16 - 0.03*GLG + 0.5*HUM + e$$

waarbij e een standaardafwijking heeft van 0.7. Dan kan men de verwachtingswaarde van H20 als volgt in een figuur weergeven



Voor een perceel met GLG=100 en HUM=2 zal de hoogte na 20 jaar naar verwachting gelijk zijn aan 14 m. De werkelijke boomhoogten op verschillende percelen met deze waarde van GLG en HUM zullen een standaardafwijking vertonen van 0.7 m. De figuur geeft dus alleen het systematische deel van het model weer. Het effect van de predictoren kan men als volgt beschrijven: verandering van GLG met 1 cm geeft een verandering in H20 van 0.03 m (als HUM gelijk blijft), verhoging van het humuspercentage met 1% levert een verhoging van H20 met 0.5 m (bij gelijkblijvende GLG).

Schatten van modelparameters

Meestal zijn de regressiecoefficienten en de restvariantie onbekend en willen we deze schatten uit waarnemingen die we hebben verricht aan n verschillende bospercelen. In feite verloopt deze schatting op dezelfde wijze als bij enkelvoudige regressie-analyse, nl. met behulp van de kleinste kwadratenmethode. Als we de waarnemingen aan het i^e perceel noteren als $H2O_i$, GLG_i en HUM_i dan schatten we de parameters β_0, β_1 en β_2 met die waarden waarvoor de kwadratensom

$$\sum_{i=1}^{n} (H2O_i - \beta_0 - \beta_1 * GLG_i - \beta_2 * HUM_i)^2$$

minimaal is. De schatters noteren we als $\hat{\beta}_0$, $\hat{\beta}_1$ en $\hat{\beta}_2$. Deze schatters hangen van de waarnemingen af en zijn dus stochastisch (en hebben dus een standaardfout se). De bovenstaande kwadratensom noemen we de restkwadratensom ($SS_{residual}$). Deze heeft n-3 vrijheidsgraden omdat we 3 vrijheidsgraden hebben gebruikt voor het schatten van de regressie-coëfficiënten. De gemiddelde restkwadratensom $MS_{residual}$ ($=SS_{residual} / (n-3)$) is een schatter voor σ^2.

Het percentage van de variantie van y dat door de regressie verklaard wordt is (evenals bij de enkelvoudige regressie) gelijk aan

$$R^2_{adj} = \frac{MS_{total} - MS_{residual}}{MS_{total}} \times 100\%$$

Als we weer $SS_{regressie}$ (hier met 2 vrijheidsgraden) definieren als SS_{totaal} - $SS_{residual}$ dan kan men met de volgende F-grootheid toetsen of het model al dan niet voorspellende waarde heeft

$$F = \frac{MS_{regressie}}{MS_{residual}}.$$

Om te zien of men parameter β_i uit het model kan weglaten (toets H_0: $\beta_i = 0$) kan men de t-verdeelde grootheid $\hat{\beta}_i / se(\hat{\beta}_i)$ beschouwen.
Eveneens kan men een betrouwbaarheidsinterval voor β_i opstellen:
$\hat{\beta}_i \pm t.se(\hat{\beta}_i)$, (t is de kritieke waarde van de t-verdeling met n-3 vrijheidsgraden).

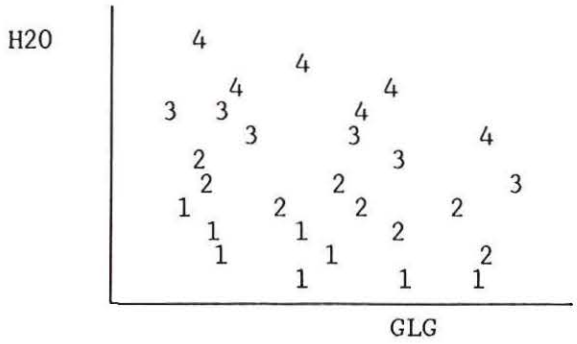
Met behulp van Genstat kan men deze resultaten verkrijgen via de volgende opdrachten

```
READ humus, glg, h20
MODEL h20
FIT humus, glg
```

Ook hierbij krijgt men weer melding van waarnemingen met een groot residu (niet passend bij het model) of met een hoge leverage (= invloedrijk, omdat de x-waarden zich aan de buitenkant van het experimenteergebied bevinden). Met behulp van RKEEP kan men de gestandaardiseerde residuen en de gefitte waarden bewaren zodat men ze in een grafiek tegen elkaar kan uitzetten om te zien of de variantie stabiel is (zie $ 18.3 en 18.4). Om te zien of het gebruikte model een adequate beschrijving geeft kan men een 3-dimensionale grafiek van de waarnemingen maken. Ook kan men de 3^e variabele zichtbaar maken via een groepsindeling naar grootte (bijvoorbeeld 4 klassen) en die met symbolen weergeven. In Genstat kan dit als volgt

```
SORT [INDEX = humus; GROUP = humklasse; NGROUPS = 4]
GRAPH  Y=h20; X=glg; SYMBOLS=humklasse
```

Als het gehanteerde model juist is dan verkrijgt men een grafiek die er bijvoorbeeld als volgt uitziet (de waarnemingen met het laagste humusgehaltekrijgen hier symbool 1)



Als de lijnen voor de 4 groepen niet parallel lijken dan zou men de productterm humus.glg in het model kunnen opnemen (er is dan interaktie tussen humus en glg). Zijn de 4 lijnen niet recht dan zou men kwadratische termen kunnen opnemen of men moet een niet-lineair model aanpassen. Als men vanuithet specifieke vakgebied al weet welk type model de samenhang goed beschrijft (bijvoorbeeld af te leiden uit een differentiaalvergelijking die redelijk lijkt) dan moet zo'n model uiteraard geprobeerd worden.

Tot slot van deze paragraaf nog enkele aanvullende opmerkingen. Uit regressies van y op de afzonderlijke variabelen kan men niet altijd iets afleiden over de kwaliteit van de regressie op alle predictorvariabelen tesamen. We geven hier 3 typische situaties aan:

a de door x_1 en x_2 verklaarde kwadraatsom is gelijk aan de som van de door x_1 en x_2 in afzonderlijke modellen verklaarde kwadraatsom. Ook de regressie-coefficienten zijn in het gezamelijke model gelijk aan die voor de afzonderlijke modellen (zie bijvoorbeeld opgave 21).

b de variabelen x_1 en x_2 verklaren gezamelijk niet veel meer dan ze in de afzonderlijke modellen ook al doen (zie bijvoorbeeld opgave 22).

c de variabelen x_1 en x_2 verklaren gezamelijk veel meer dan de som van de door x_1 en x_2 afzonderlijk verklaarde kwadraatsommen (zie bijvoorbeeld opgave 23).

Situatie a komt voor als x_1 en x_2 orthogonaal (dit is vaak voor bij opgezette proeven het geval), in situatie b noemen we x_1 en x_2 inwisselbaar (bijvoorbeeld als x_1 en x_2 sterk gecorreleerd zijn) en in situatie c noemen we ze elkaar aanvullend.

Het bovenstaande impliceert dat men op grond van afzonderlijke regressies geen uitspraak kan doen over welke predictoren een bijdrage leveren aan de voorspelling bij een multiple regressie. Omgekeerd geeft het aanpassen van alleen het volledige model ook geen uitsluitsel over welke variabelen weggelaten kunnen worden omdat de t-grootheid slechts aangeeft of de predictor nog nodig is voor de verklaring als het model de overige predictoren al bevat (in situatie b geven x_1 en x_2 beide een lage t-waarde in het volledige model!). Bij selectie van variabelen zullen we daarom in feite ook alle tussenliggende deelmodellen moeten bekijken (zie hiervoor $ 20.4).

20.2 Gebruik van matrix-notatie bij regressie-analyse

Indien gebruik wordt gemaakt van matrices en vectoren, dan kan men het regressiemodel, de schattingen van de modelparameters en hun eigenschappen op een bondige manier opschrijven. Omdat deze notatie vaak gebruikt wordt in de literatuur over regressie-analyse zullen we hier de belangrijkste resultaten in deze vorm weergeven. Bewijzen zullen we achterwege laten (zie hiervoor bv. Draper and Smith (1981)).
Met betrekking tot de matrixoperaties nemen we aan dat u kennis hebt van het vermenigvuldigen van matrices, het transponeren (= spiegelen) van matrices (A' is de gespiegelde van A) en van het berekenen van een inverse van een vierkante matrix (voor matrix A is inverse A^{-1} gedefinieerd door $AA^{-1} = A^{-1}A = I$, waarbij I de eenheidsmatrix is).

Voor een multiple regressiemodel met k verklarende variabelen $x_1, \ldots, x_k$ geldt:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

Als we aan n objecten waarnemingen hebben verricht dan kunnen we deze als volgt noteren:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + e_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + e_2$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + e_n$$

Deze n vergelijkingen kunnen we ook noteren als

$$y = X\beta + e$$

waarbij

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Hierbij geldt dus dat y een (nx1)-vector is met de waarnemingen, X is een (nxp)-matrix met de verklarende variabelen als kolommen (waarbij p=k+1), β is een (px1)- vector met de regressiecoefficienten en e is een (nx1)-vector met de toevallige afwijkingen van het model. De matrix X heet de <u>design matrix</u>.
Als we de kleinste kwadraten schatter $\hat{\beta}$ zoeken, dan zijn we op zoek naar die waarde van β waarvoor de uitdrukking

$$\text{kwadratensom} = (y - X\beta)'(y - X\beta)$$

minimaal is. Deze waarde noemen we $\hat{\beta}$.
Door de afgeleide naar β in $\hat{\beta}$ gelijk te stellen aan 0 vinden we dat voor $\hat{\beta}$ moet gelden:

$$X'X\hat{\beta} = X'y$$

Dus

$$\hat{\beta} = (X'X)^{-1}X'y$$

Verder geldt (door $\hat{\beta}$ in te vullen in de bovenstaande kwadratensom)

$$SS_{residual} = y'y - \hat{\beta}'X'y \qquad (\text{met } df = n-p = n-k-1)$$

en $\hat{\sigma}^2 = MS_{residual} = SS_{residual}/(n-p)$

Voor de covariantiematrix voor de parameterschatters $\hat{\beta}$ geldt

$$cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Het j^e diagonaalelement van $cov(\hat{\beta})$ is gelijk aan $var(\hat{\beta}_j)$, zodat voor de standaardfout van $\hat{\beta}_j$ geldt

$$se\ (\hat{\beta}_j) = \sqrt{MS_{residual} \cdot (X'X)^{-1}_{jj}}$$

Als we de vector met gefitte waarde noteren als $\hat{y}$ dan geldt:

$$\begin{aligned}\hat{y} &= X\hat{\beta} \\ &= X(X'X)^{-1}X'y \\ &= Hy\end{aligned}$$

waarbij we H definieren als $H = X(X'X)^{-1}X'$. H staat bekend als de "hat"-matrix, omdat die aangeeft hoe men $\hat{y}$ kan berekenen uit y. Uit de uitdrukking $\hat{y} = Hy$ zien we dat het i^e diagonaalelement van H (noem dit h_{ii}) aangeeft hoe groot de invloed is van de i^e waarneming op zijn eigen voorspelling. Als h_{ii} groot is dan is de invloed van de i^e waarneming op de ligging van het regressievlak groot. Het getal h_{ii} heet daarom de leverage (= hefboomwerking) van de i^e waarneming. Omdat $h_{ii} \geq 0$ en het gemiddelde over i gelijk is aan p/n kan men grofweg zeggen dat punten met een leverage groter dan 2p/n aan de buitenkant van het experimenteergebied liggen.

Voor de vector e met residuen geldt:

$$e = y - X\hat{\beta} = y - Hy = (I - H)y$$

waarbij I de eenheidsmatrix is.
Er geldt $cov(e) = (I - H)\sigma^2$, dus $var(e_i) = \sigma^2(1 - h_{ii})$. Het gestandaardiseerde residu voor de i^e waarneming is dus gelijk aan $e_i/\sqrt{\sigma^2(1 - h_{ii})}$ (dit zal derhalve met kans 95% tussen -2 en 2 liggen).

Als we een betrouwbaarheidsinterval voor de verwachte respons Ey willen opstellen in een willekeurig punt $(x_{01}, x_{02}, \ldots, x_{0k})$ dan definieren we eerst de vector x_0 als

$$x_0 = \begin{pmatrix} 1 \\ x_{01} \\ x_{02} \\ . \\ . \\ . \\ x_{0k} \end{pmatrix}$$

De schatter van de verwachte respons is dan gelijk aan $\hat{y}_0 = x_0'\hat{\beta}$ en de variantie ervan is

$$var(\hat{y}_0) = \sigma^2 x_0'(X'X)^{-1}x_0.$$

Een 95%-betrouwbaarheidsinterval voor $E\hat{y}_0$ is dus:

$$\hat{y}_0 \pm t_{n-p}\sqrt{MS_{residual} \cdot x_0'(X'X)^{-1}x_0}$$

waarbij t_{n-p} de kritieke waarde is van de Studentverdeling met n-p vrijheidsgraden.

Voor de voorspelling van een toekomstige waarneming bij x_0 is het betrouwbaarheidsinterval gelijk aan

$$\hat{y}_0 \pm t_{n-p}\sqrt{MS_{residual} \cdot (1 + x_0'(X'X)^{-1}x_0)}$$

Merk op dat X de instelpunten bevat waarop de regressievergelijking gebaseerd is en x_0 de x-waarden van het nieuwe te voorspellen object geeft.

Voor punten x_0 waarvoor geldt dat $x_0'(X'X)^{-1}x_0 > 2p/n$ moet men de voorspelling wantrouwen omdat x_0 dan buiten het experimenteergebied ligt en men dan dus bezig is met een mogelijk onverantwoorde extrapolatie.

20.3 Vergelijken van modellen

We zullen hier een algemeen bruikbare methode bespreken voor het vergelijken van regressiemodellen door middel van toetsen van hypothesen. Hierbij moet voor de twee te vergelijken modellen gelden dat het ene model een deelmodel is van het andere. Men toetst dan de nulhypothese H_0 dat beide modellen equivalent zijn tegen het alternatief H_1 dat het uitgebreide model een significant betere verklaring geeft dan het deelmodel.

Voorbeeld:
Men wil een model $Ey = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4$
vergelijken met $Ey = \beta_0 + \beta_1x_1 + \beta_2x_2$.
Dit komt overeen met de nulhypothese H_0: $\beta_3 = \beta_4 = 0$.

Algemener: neem aan dat het model p_0 parameters bevat onder H_0 en p_1 parameters onder H_1 (in het voorbeeld geldt $p_0 = 3$ en $p_1 = 5$). Onder beide modellen kunnen we de restkwadraatsommen uitrekenen:

onder H_0: $SS_{residual0}$ met $df_0 = n - p_0$
onder H_1: $SS_{residual1}$ met $df_1 = n - p_1$

Het blijkt dat onder H_0 de volgende grootheid (waarbij men voor p_1-p_0 ook kan lezen df_0-df_1):

$$F = [(SS_{residual0} - SS_{residual1}) / (p_1 - p_0)] / MS_{residual1}$$

F verdeeld is met (p_1-p_0) en $(n-p_1)$ vrijheidsgraden. We verwerpen H_0 dus voor grote waarden van deze grootheid (voor kritieke waarden zie tabel ..).

Dit komt overeen met het volgende : als de extra parameters van het uitgebreide model niet nodig zijn, dan is het verwachte verschil tussen $SS_{residual0}$ en $SS_{residual1}$ gelijk aan $(p_1-p_0).\sigma^2$. In het bovenstaande voorbeeld zal $SS_{residual0} - SS_{residual1}$ in verwachting dus gelijk zijn aan $2\sigma^2$ als $\beta_3=\beta_4=0$.

Deze methode hebben we al vaker toegepast, bijvoorbeeld op pag. 108 bij het vergelijken van het volledige model met het lege model $Ey= \beta_0$.

Verder moet opgemerkt worden dat de F-toets equivalent is aan de t-toets als het model onder H_0 slechts 1 parameter minder bevat dan onder H_1. Stel bijvoorbeeld dat men in het model $Ey = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4$ wil vergelijken met $Ey = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_4x_4$ (d.w.z. men wil toetsen H_0: $\beta_3=0$) dan is de bovengenoemde F-grootheid het kwadraat van de t-grootheid $\beta_3/se(\beta_3)$. Deze t-toets levert het effect van x_3 gegeven dat x_1, x_2 en x_4 al in het model zitten.

Tot slot zij nog opgemerkt dat bij multipele regressie niet alle deelmodellen tegen elkaar getoetst kunnen worden, maar alleen die tweetallen modellen waarvan het ene model het andere omvat. Bij 2 predictor-variabelen kan men dus wel de volgende vergelijkingen maken:

x_1, x_2 vs. x_1
x_1, x_2 vs. x_2
x_1, x_2 vs. -
x_1 vs. -
x_2 vs. -

Maar men kan niet toetsen of het model met alleen x_1 significant beter is dan een model met alleen x_2. Als bijvoorbeeld x_1 en x_2 inwisselbaar zijn dan kan men niet op statistische gronden besluiten welke van de twee voor de verklaring zorgt. Men kan dan voor de voorspelling de variabele kiezen die het eenvoudigst of het goedkoopst te meten is of die welke op theoretische gronden het best verdedigbaar is.

20.4 Selectie van variabelen.

Bij multipele regressie wil men een responsvariabele y verklaren c.q. voorspellen uit een aantal predictor variabelen. Het uiteindelijke model moet voldoen aan de volgende eisen:

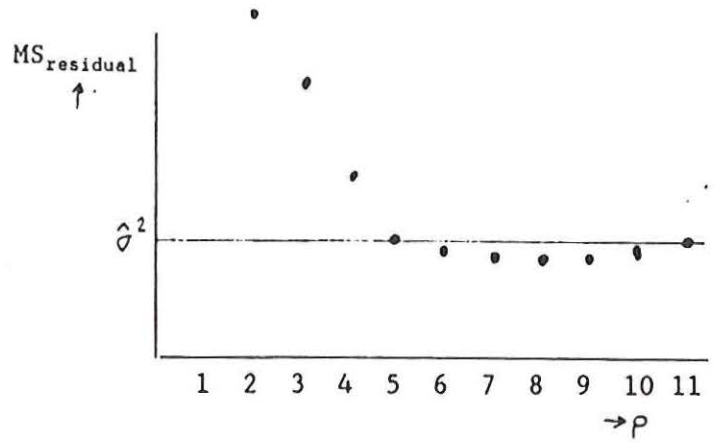
- -het model moet alle variabelen bevatten die bijdragen tot de voorspelling.
- -het aantal predictoren moet het liefst zo klein mogelijk zijn.

Deze eisen stellen we omdat de voorspelling zo nauwkeurig mogelijk moet zijn, terwijl men bij later gebruik van de voorspelformule niet meer variabelen wil meten dan nodig is. Omdat enerzijds de werkelijkheid beter

benaderd kan worden met een model met veel variabelen, maar anderzijds de schatting van het regressiemodel stabieler is als het geen overbodige variabelen bevat, zal men in de praktijk meestal tot een compromis moeten komen.

Men gaat hiervoor als volgt te werk. Men kiest alle predictor variabelen die mogelijk een verklaring kunnen geven van de respons (en indien nodig ook produkttermen en kwadraten om rekening te houden met interacties of niet-lineariteit). Dit noemt men het volledige model. Daarna probeert men via selectie van variabelen te komen tot een zo klein mogelijk model dat (bijna) evenveel verklaart als het volledig model (een model verklaart bijna even veel als de restvariantie nauwelijks groter is).
Voor een stabiel resultaat is het gewenst dat het aantal waarnemingen niet al te klein is (een vuistregel is: minimaal 5 maal het aantal te schatten parameters). Verder moet men uiteraard geen last hebben van uitbijters en invloedrijke punten (hoge residuen of leverages). Het is mogelijk dat er meerdere modellen als (bijna) gelijkwaardig te voorschijn komen (bijvoorbeeld als predictoren inwisselbaar zijn). In dat laatste geval kan men geen eenduidige uitspraak doen over de vraag welke predictoren of combinaties daarvan een bijdrage leveren aan de voorspelling.

Voor de selectie van variabelen moeten we in feite alle deelmodellen aanpassen en hun restvariantie berekenen (zie bovenaan pag. 110). Deze varianties vergelijken we met een goede schatter van de restvariantie van y. Hiervoor nemen we meestal de restvariantie $\hat{\sigma}^2$ van het volledige model (een alternatief is een variantieschatting uit herhalingen of uit naburige punten, zie $ 18.5). Om te zien hoeveel predictoren we nodig hebben zoeken we het beste model met 1 parameter, het beste model met 2 parameters, etc. Daarna zetten we in een grafiek de $MS_{residual}$ van deze modellen uit tegen het aantal paramaters p (inclusief β_0). Als er bijvoorbeeld 10 predictoren zijn en hiervan zijn er 4 nodig voor de voorspelling dan zal die grafiek er ongeveer als volgt uitzien



Vanaf p=5 blijft $MS_{residual}$ min of meer stabiel.
Modellen met een zo laag mogelijke p en $MS_{residual}$ niet veel groter dan σ^2 noemen we kandidaat-modellen. Deze modellen kan men gemakkelijker herkennen m.b.v. een andere maat die in de literatuur populair is, nl Mallow's C_p. Deze is gedefineerd als $C_p = SS_{residual}/\hat{\sigma}^2 - n + 2p$. Kandidaatmodellen zijn modellen waarvoor C_p ongeveer gelijk is aan p (zeg C_p kleiner dan p+2 of p+3). Van deze kandidaat-modellen zouden we nog kunnen controleren of elke parameter significant is (dit kan men zien aan de t-grootheden), zo niet dan kan men de minst significante weglaten, hetgeen een nieuwe kandidaat oplevert. Op deze wijze vindt men een of meer modellen die geschikt zijn voor de voorspelling.

We zullen dit illustreren aan het volgende voorbeeld (afkomstig van Hald, zie bijvoorbeeld Montgomery & Peck, 1982, pag 257). Merk overigens op dat in dit voorbeeld het aantal waarnemingen in feite wat klein is om er verregaande conclusies aan te verbinden.

• **Example** . Hald [1952] presents data concerning the heat evolved in calories per gram of cement (y) as a function of the amount of each of four ingredients in the mix; tricalcium aluminate (x_1), tricalcium silicate (x_2), tetracalcium alumino ferrite (x_3) and dicalcium silicate (x_4). The data are shown below. We will use this data to illustrate the all possible regressions approach to variable selection.

Observation, i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	'71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Aanpassing van alle modellen levert de volgende resultaten

Summary of all possible regression for Example

Number of Regressors in Model	p	Regressors in Model	$SS_{residual}$	R^2_{adj}	$MS_{residual}$	C_p
None	1	None	2715.7635	0	226.3136	442.92
1	2	x_1	1265.6867	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.45780	122.7073	198.10
2	3	x_1x_4	74.7621	0.96697	7.4762	5.50
2	3	x_2x_3	415.4427	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.61607	86.8880	138.23
2	3	x_3x_4	175.7380	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.97504	5.6485	3.50
3	4	$x_2x_3x_4$	73.8145	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.97356	5.9829	5.00

Kandidaatmodellen zijn x_1, x_2 en eventueel x_1, x_4 als redelijke tweede.

Het doorrekenen van alle modellen is een nogal rekenintensieve klus. Als men bijvoorbeeld 10 predictoren heeft dan moet men $2^{10} = 1024$ modellen aanpassen. Daarom zijn er ook allerlei andere methoden ontwikkeld voor selectie van variabelen, die op een of andere wijze het aantal door te rekenen modellen en daarmee ook de keuze beperken. Dit zijn o.a. (zie voor nog andere methoden Draper & Smith, 1981):

1. voorwaartse selectie van variabelen
2. achterwaartse selectie van variabelen
3. stapsgewijze selectie van variabelen

We zullen deze methoden illustreren aan het bovenstaande voorbeeld.

Bij <u>voorwaartse selektie</u> begint men met een leeg model en men voegt dan steeds een variabele toe aan het model, nl. degene die de grootste verbetering geeft (gegeven de variabelen die op dat moment al in het model zitten). Men stopt als de toevoeging niet meer significant is. In het voorbeeld wordt eerst x_4 in het model opgenomen, daarna x_1 en tot slot x_2.

Bij achterwaartse selektie begint men met het volledige model en men laat bij elke stap een variabele weg, nl. die met de laagste t-waarde. Men stopt op het moment dat het model alleen nog maar significante variabelen bevat. In het voorbeeld wordt eerst x_3 weggelaten en daarna x_4, zodat men x_1 en x_2 overhoudt.

Stapsgewijze regressie is een combinatie van voorwaartse en achterwaartse selektie. Bij elke stap wordt bekeken of door weglating resp. toevoeging van een variabele het model verbeterd kan worden. Bij zo'n stap wordt eerst gekeken of een niet-significante variabele kan worden weggelaten, zoniet dan wordt onderzocht of een variabele die op dat moment nog niet in het model zit een significante verbetering geeft. Bij stapsgewijze regressie kan men zowel met het lege model beginnen als met het volledige model (in feite kan men elk deelmodel als startpunt kiezen). In het voorbeeld levert stapsgewijze regressie het volgende resultaat:
-beginnend met het lege model: eerst wordt x_4 opgenomen, daarna x_1, dan x_2 en daarna wordt x_4 weggelaten, hetgeen het model oplevert met x_1 en x_2.
-beginnend met het volledige model: eerst wordt x_3 weggelaten en daarna x_4, zodat x_1 en x_2 overblijven.

Vergelijking van de methoden levert het volgende (zie ook Draper & Smith, 1981 of Montgomery & Peck, 1982):
-De beste methode is het vergelijken van alle modellen. Dit levert een overzicht van alle in aanmerking komende modellen. De keuze kan gemaakt worden op andere dan statistische gronden.
-Stapsgewijze selectie is minst slechte van de overige methoden, zeker als men die toepast beginnend met zowel het lege als met het volledige model. Voorwaartse en achterwaartse selectie zijn nog minder geschikt omdat predictoren die eenmaal zijn toegevoegd (resp. weggelaten) niet meer kunnen worden ingewisseld. Voor alle 3 methoden geldt echter dat men geen volledig overzicht krijgt van de eventuele andere goede modellen. Ze hoeven zelfs niet de aantrekkelijkste keuze uit de kandidaatmodellen op te leveren.

M.b.t. de statistische computerprogramma's kan het volgende worden opgemerkt. SAS en BMDP leveren mogelijkheden om alle modellen door te rekenen. In Genstat is hiervoor de procedure RSELECT beschikbaar (informatie hierover is verkrijgbaar bij de Groep Landbouwiskunde, DLO, Wageningen). Voor de bovenstaande gegevens van Hald werkt deze als volgt.

```
UNITS [NVALUES=13]
OPEN 'Hald.dat'; CHANNEL= 2
READ [CHANNEL=2] x1, x2, x3, x4, y
MODEL y
RSELECT [CRITERION= cp] !p(x1, x2, x3, x4)
STOP
```

De uitvoer ziet er als volgt uit

```
***** All Possible Subsets Regression *****

Response variate:  y

For each selected subset, the criteria, minimum tolerance and t-statistics
of included variables are printed. t-statistics are truncated at 99999.99.

Subsets with  2 explanatory variables

     R2     Adj      Cp   MinTol        x1        x2        x3        x4

  97.87   97.44    2.68   0.9478     12.10     14.44         -         -
  97.25   96.70    5.50   0.9398     10.40         -         -    -12.62
  93.53   92.23   22.37   0.9991         -         -     -6.35    -10.02
  84.70   81.64   62.44   0.9806         -      6.06     -3.44         -
  68.01   61.61  138.23   0.0534         -      0.42         -     -0.66

Subsets with  3 explanatory variables

     R2     Adj      Cp   MinTol        x1        x2        x3        x4

  98.23   97.64    3.02   0.0528     12.41      2.24         -     -1.37
  98.23   97.64    3.04   0.3076      8.29     14.85      1.35         -
  98.13   97.50    3.50   0.2719      4.70         -     -2.06    -14.43
  97.28   96.38    7.34   0.0411         -     -3.53     -9.85     -6.45

Subsets with  4 explanatory variables

     R2     Adj      Cp   MinTol        x1        x2        x3        x4

  98.24   97.36    5.00   0.0035      2.08      0.70      0.14     -0.20

Largest discrepancy observed for selection criterion is 0

8  STOP

******** End of job.  Maximum of 23390 data units used at line 7 (150464 left)
```

In de output worden de 10 beste modellen gegeven op basis van hun Cp-waarde (meer dan 10 modellen kan men vragen via de optie NUMBERBEST). Men ziet dat het model met x_1 en x_2 het kleinste model is waarvoor geldt dat Cp ongeveer gelijk is aan p. Verder bevat dit model geen niet-significante

termen (zie de t-waarden). Het model met x_1 en x_4 levert een redelijke tweede mogelijkheid.
Nadat men met RSELECT de modellen gekozen heeft kan men met FIT de regressiecoefficienten verkrijgen (deze zijn bij RSELECT weggelaten om een beknopte uitvoer te krijgen)

```
FIT x1, x2
FIT x1, x4
```

20.5 Regressie met kwalitatieve predictoren.

Tot nu toe hebben we kwantitatieve variabelen als predictor gebruikt in de regressie analyse. Kwalitatieve predictor variabelen zijn echter ook toegestaan. Neem het voorbeeld op pag 49 waarin men de opbrengst van de 3 uienrassen A, B en C wil voorspellen cq. vergelijken. Deze situatie waarin we tot nu variantieanalyse hebben toegepast kan men ook m.b.v. regressie analyse beschrijven. Hiervoor kan men het volgende model gebruiken.

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & . & . \\ 1 & . & . \\ 1 & 0 & . \\ 1 & 1 & . \\ 1 & 1 & . \\ 1 & . & . \\ 1 & . & . \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & . & . \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

De 2e en 3e kolom noemen we dummy-variabelen die aangeven of de experimentele eenheid ras B cq. C betreft. De verwachte opbrengst voor ras A is gelijk aan β_0, voor ras B is deze $\beta_0 + \beta_1$ en voor ras C is deze $\beta_0 + \beta_2$. De parameter β_1 geeft dus het verschil van ras B t.o.v. ras A en β_2 is het verschil tussen ras C en ras A. Via de methoden beschreven in $ 20.2 kan men schattingen voor de β's en hun se's berekenen. Verder geeft de F-grootheid $MS_{regressie}/MS_{residual}$ een toets voor verschillen tussen de rassen (H_0: $\beta_1=\beta_2=0$). In Genstat kan men dit model eenvoudig aanpassen via de opdrachten

```
UNITS [NVALUES=26]
FACTOR [LABELS= !t(A, B, C) ] ras
READ ras, opbrengst
MODEL opbrengst
FIT [PRINT= mod,acc,est] ras
```

De output volgt hieronder. Hierin worden de parameters β_0, β_1, β_2 aangeduid met constant, ras B en ras C.

```
***** Regression Analysis *****

 Response variate: opbrengs
     Fitted terms: Constant, ras


*** Accumulated analysis of variance ***

Change          d.f.        s.s.        m.s.        v.r.
+ ras              2      194.91       97.46        4.20
Residual          23      533.43       23.19

Total .           25      728.35       29.13


*** Estimates of regression coefficients ***

                estimate        s.e.           t
Constant           42.80        1.52       28.10
ras b              -3.30        2.15       -1.53
ras c               3.87        2.49        1.55
```

De ras-gemiddelden kan men krijgen via de opdracht:

```
PREDICT ras
```

FIT en PREDICT leveren echter geen sed-waarden. Via de volgende opdracht kan men echter wel zien welke rassen paarsgewijs verschillen:

```
RPAIR TREATFACT= !p(ras)
```

Dit levert een tabel met gemiddelden en verschillen van gemiddelden en een tabel met P-waarden. RPAIR is een procedure die aan Genstat kan worden toegevoegd (informatie hierover is verkrijgbaar bij de Groep Landbouwwiskunde, DLO, Wageningen)

Ook zijn meerdere kwalitatieve variabelen toegestaan als predictor-variabele bij regressie analyse.

In de bovenstaande situaties kan men zowel ANOVA als regressieanalyse gebruiken. Als beide gebruikt kunnen worden (in bijna alle keurig opgezette proeven) heeft ANOVA de voorkeur omdat ANOVA meteen alle benodigde uitvoer levert (o.a. gemiddelden en sed-waarden). ANOVA is echter niet toepasbaar op niet-gebalanceerde proeven en doorgaans ook niet op observationeel onderzoek. Met regressie kan men deze nog wel analyseren. Bedenk echter dat dan tengevolge van niet-orthogonaliteit de effecten van factoren niet meer

eenduidig zijn, maar kunnen afhangen van de volgorde waarin men de factoren in het model opneemt. Het verdient dus aanbeveling bij de proefopzet ervoor te zorgen dat te onderzoeken factoren zo goed mogelijk gebalanceerd zijn (dit geldt ook voor de opzet van steekproeven).

Tot slot is het ook mogelijk kwalitatieve en kwantitatieve predictoren tegelijkertijd in een regressiemodel te gebruiken. Een voorbeeld is een onderzoek waarbij men voor mestvarkens een voorspellingsmodel wil opstellen voor het vleespercentage op basis van de meting van de spekdikte op een 2-tal plaatsen van het dier (vleespercentage is een moeilijk te meten variabele terwijl spekdikte metingen snel kunnen worden uitgevoerd). Hiervoor heeft men van dieren van een aantal rassen zowel de spekdikten als het vleespercentage gemeten. Een van de vragen is of voor alle rassen dezelfde regressielijn gebruikt kan worden. Men kan hiervoor de volgende modellen aanpassen.

```
MODEL vlees%
FIT [PRINT=mod,acc,est] dikte1 + dikte2
FIT [PRINT=mod,acc,est] dikte1 + dikte2 + ras
FIT [PRINT=mod,acc,est] (dikte1 + dikte2 ) * ras
```

In het eerste model hangt de voorspelling niet af van ras. Het tweede model geeft voor elk ras een ander intercept, maar gemeenschappelijke regressiecoeficienten (hellingen) voor dikte1 en dikte2. In het derde model zijn voor elk ras zowel het intercept als de hellingen verschillend. (Opmerking: verder moet men in dit voorbeeld uiteraard ook nog onderzoeken of de beide spekdikten nodig zijn en of het verband niet-lineair is in dikte1 en dikte2).

Opgaven Verdiepingsmodule Toegepaste Statistiek, 29 mei 1989
PHLO-cursus Onderzoekstechniek

Opgave 20

Men wil een responsvariabele y voorspellen uit een predictor x. Hiertoe heeft men 16 waarnemingen uitgevoerd. Deze zijn als volgt:

x	y	x	y
0	6	4	31
0	3	5	29
1	12	5	27
1	15	6	28
2	18	6	31
2	19	7	33
3	23	7	27
3	20	8	33
4	25	8	29

De gegevens kan men copieren van file [PAOCSMP20.MODEL]OPGAVE20.DAT.
Maak een grafiek van y tegen x. Pas een rechtlijnig regressiemodel aan.
Toets of een rechte lijn past bij de waarnemingen.

Opgave 21

In een proef is onderzocht hoe de opbrengst van suikerbieten afhangt van de stikstof- en fosforbemesting. Hiertoe zijn 16 naast elkaar liggende veldjes gebruikt die elk een combinatie van N- en P-bemesting ontvingen. De waarnemingen zijn :

N	P	opbrengst	N	P	opbrengst
0	0	56	0	5	63
0	10	71	0	15	87
10	0	72	10	5	71
10	10	76	10	15	87
20	0	73	20	5	86
20	10	87	20	15	96
30	0	79	30	5	92
30	10	100	30	15	115

De gegevens kan men copieren van file [PAOCSMP20.MODEL]OPGAVE21.DAT.
Pas een multiple regressiemodel aan met de variabelen P en N. Pas ook modellen aan waarin de opbrengst uit N resp. P afzonderlijk verklaard wordt.
Maak een grafiek van de opbrengst tegen N waarbij P via symbolen wordt weergegeven (maak hiervoor eerst een groepsindeling van P via SORT).

Opgave 22

Sphaeropsis sapinea is een schimmelsoort die grote schade kan veroorzaken aan bossen met grove den. Om de invloed van NH3 en O3 op die aantasting te onderzoeken is de volgende inventarisatie verricht. Op een aantal plaatsen in Nederland met grove den opstanden is de gemiddelde belasting (in ug/m**3) van NH3 en O3 bepaald (uit overzichtskaarten van RIVM). Verder is de aantasting van dennen door Sphaeropsis gemeten op die plaatsen. Wat valt er te zeggen over de causaliteit van de effecten bij zo'n onderzoek? De gegevens zijn:

NH3	O3	aantasting
5.4	23	1.6
6.3	21	1.5
7.0	26	1.6
8.4	27	2.6
9.5	32	2.5
11.2	37	4.6
11.8	37	5.3
12.9	40	5.6
13.8	46	4.0
15.4	53	6.7
17.1	58	5.8
18.8	57	5.8
21.0	73	7.5
22.2	75	6.6 :

De gegevens kan men copieren van file [PAOCSMP20.MODEL]OPGAVE22.DAT.
Onderzoek m.b.v. regressie analyse of NH3 en/of O3 de verschillen in aantasting verklaren (pas eerst modellen aan voor de afzonderlijke predictoren en daarna gezamelijk). Zijn beide variabelen nodig in het model? Is het duidelijk welke van de twee variabelen voor de verklaring zorgt?
Maak een grafiek van de aantasting tegen NH3 waarbij O3 via symbolen wordt weergegeven (maak hiervoor eerst een groepsindeling van O3 via SORT). Maak een grafiek van NH3 tegen O3. Hoe zou de steekproef opgezet moeten worden om de effecten van NH3 en O3 onafhankelijk van elkaar te kunnnen schatten?
Bereken m.b.v. PREDICT de voorspelde aantasting (inclusief se) voor NH3=20 en O3=25. Waarom is deze voorspelling zo onnauwkeurig?

Opgave 23

Van een tiental appelmoes monsters is door een panel van 6 personen op sensorische wijze de zuurgraad bepaald (de 6 leden geven een oordeel op een bepaalde schaal, daarna worden de getallen gemiddeld). Tevens is van elk monster de hoeveelheid titreerbaar zuur en het suikergehalte bepaald. De waarnemingen zijn:

Monster	Titr. Zuur	Suiker	Sens. Zuur
1	5.2	15	1.8
2	8.3	18	2.8
3	6.4	19	1.6
4	7.8	20	2.3
5	5.4	12	2.4
6	7.4	15	2.9
7	8.5	24	2.2
8	6.3	16	2.2
9	6.1	11	2.8
10	7.4	23	1.7

De gegevens kan men copieren van file [PAOCSMP20.MODEL]OPGAVE23.DAT.
Geven titreerbaar zuur resp. suikergehalte afzonderlijk een goed verband met sensorisch zuur?
Geven ze gezamelijk een goed verband met sensorisch zuur?
Maak een grafiek waarin sensorisch zuur is uitgezet tegen titreerbaar zuur waarbij een groepsindeling naar suikergehalte als symbool gebruikt wordt.
Aanwijzing: Een groepsindeling, zeg suikerklas, is bijvoorbeeld te verkrijgen via:
SORT [INDEX=suikergehalte ; GROUP=suikerklas ; LIMITS=!(13,17,21)]

POPULATIES EN STEEKPROEVEN

Samenvatting van begrippen

Populatie		Steekproef	
Frequentieverdeling		Kansverdeling	
Gemiddelde	μ	Gemiddelde Verwachting	$\bar{x}$ $EX = \mu$
Variantie	σ^2	Variantie	s^2
Standaardafwijking	σ	Standaardafwijking	s
Aantal	N	Aantal	n

$$\sigma^2 = \frac{\Sigma (\mu - x_i)^2}{N}$$

$$s^2 = \frac{\Sigma (\bar{x} - x_i)^2}{n - 1}$$

Percentiel* z-as
standaard-normaal

Overschrijdingskans t-as
t - verdeling
(afhankelijk van df = n-1)

* Fractie onder/tussen/
boven grens

Statistische functies in VP-Planner

@COUNT(A1..A10)		aantal in geselecteerde blok
@SUM	( )	som
@AVG	( )	gemiddelde
@MIN	( )	minimum
@MAX	( )	maximum
@STD	( )	standaardafwijking van populatie
@STDS	( )	standaardafwijking van steekproef
@VAR	( )	variantie van populatie
@VARS	( )	variantie van steekproef

N.B. Lotus kent (kende?) niet de functies STDS en VARS, dus de functies die gelden voor een steekproef!

TERMEN EN DEFINITIES

Populatie (population)
Een ondubbelzinnig gedefiniëerde verzameling elementen waarop de conclusies van een statistisch onderzoek of van een keuring betrekking hebben. (NEN 3117, 4.1)

Steekproef; monster (sample)
Een of meer elementen die uit een populatie (partij of proces) zijn genomen en die zijn bedoeld om informatie te verschaffen over de populatie (partij of proces) of om als basis te dienen voor een beslissing over de populatie (partij of proces). (NEN 1132, 3.1.1)
N.B. In een steekproef kan een element al of niet herhaald optreden ('met of zonder teruglegging'). (NEN 3117, 4.10)

Steekproeftrekking, bemonstering (sampling)
Het samenstellen van een steekproef door trekking van elementen uit een populatie. (NEN 1132, 3.1.2)

Aselect trekken (random sampling)
Het trekken van een element uit een populatie met een kans die voor alle elementen van de populatie dezelfde is. (NEN 3117, 4.13)

Laboratoriummonster (laboratory sample)
Een monster dat bedoeld is voor onderzoek in een laboratorium. (NEN 1132, 3.1.22)
Een hoeveelheid voor onderzoek bestemd materiaal, in de vorm en toestand waarin het is afgeleverd bij het laboratorium. (ORA, naar ISO 78/2,1982)

Analysemonster (test sample)
Een monster dat, zonodig na voorbehandeling, voor analyse gereed is. (NEN 1132, 3.1.24)
Het deel van het laboratoriummonster, dat volgens van toepassing zijnde voorschriften zodanig is behandeld dat het voor de beoogde analyses kan worden gebruikt. (ORA, naar ISO 78/2,1982)

Analyseportie (test portion)
De hoeveelheid van het analysemonster die voor daadwerkelijke analyse wordt gebruikt. (ORA, naar ISO 78/2,1982)

Gemiddelde, b.v. $\bar{x}$ (mean)

De som van de waarnemingen in een steekproef gedeeld door hun aantal.

Mediaan (median)

Als het aantal waarnemingen oneven is: de waarde van de middelste van de naar grootte gerangschikte waarnemingen; als het aantal even is: elke waarde gelegen tussen de waarden van het middelste tweetal waarnemingen.
Opmerking. Bij een even aantal waarnemingen wordt de mediaan meestal gelijk gekozen aan de gemiddelde waarde van het middelste tweetal. (NEN 3117, 4.39)

Modus (mode)

De waarde met de hoogste frequentie in een frequentie-verdeling.
Opmerking. Het kan voorkomen dat de hoogste frequentie bij meer dan één waarde optreedt. In dat geval is de modus niet ondubbelzinnig vastgelegd. (NEN 3117, 4.

Variantie (variance)

Voor een populatie:

De som van de gekwadrateerde verschillen tussen waarnemingen en hun gemiddelde, gedeeld door het aantal waarnemingen n, in formule:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Voor een steekproef (*steekproefvariantie*);

De som van de gekwadrateerde verschillen tussen waarnemingen en hun gemiddelde, gedeeld door het aantal waarnemingen n-1, in formule:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

(NEN 3117, 4.43)

Standaardafwijking, σ (voor populatie) resp. s (voor steekproef) (standard deviation)

De vierkantswortel uit de variantie.
Opmerking. Het gebruik van de term *standaarddeviatie* wordt ontraden. (NEN 3117, 4.44)

Variatiecoëfficiënt (coefficient of variation)

(ook wel: *relatieve standaardafwijking*)

Het quotiënt van de standaardafwijking en de absolute waarde van het gemiddelde. (NEN 3117, 4.45)

In formule: $vc = \sigma/\mu = 100\ \sigma/\mu$ % resp.

$vc = s/v = 100\ s/v$ %.

Covariantie (covariane)

Een uitdrukking van de vorm

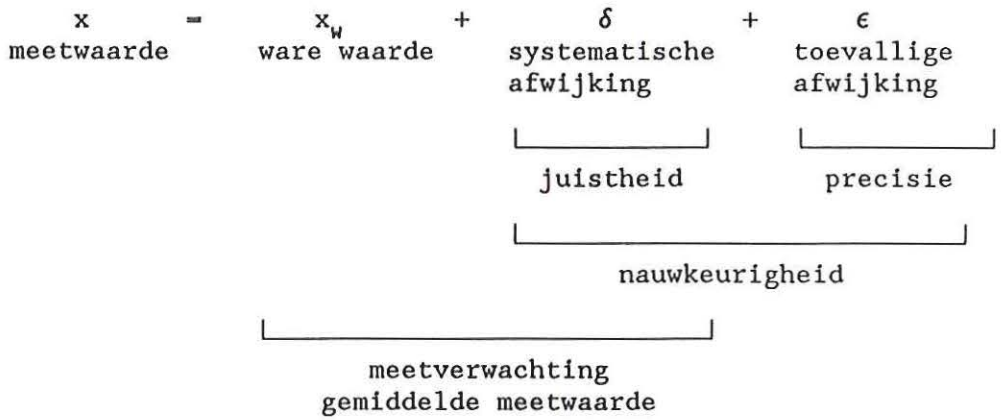
voor een populatie: $\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

voor een steekproef $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

(NEN 3117, 4.50)

NAUWKEURIGHEID VAN METINGEN

Statistisch model (Ontw.-NEN 3114):



Ware waarde (van een hoeveelheid) x_w

De waarde die een hoeveelheid karakteriseert die volledig gedefiniëerd is in de omstandigheden waaronder de hoeveelheid wordt geanalyseerd. (ISO/DP 3534/1, 3.2)

De exact gedefiniëerde waarde van een grootheid (Ontw.-NEN 3114 3.1)

Meetwaarde, x

Een door meting verkregen waarde (Ontw.-NEN 3114, 3.10)

Gemiddelde (meetwaarde), $\bar{x}$

Het rekenkundig gemiddelde van een rij meetwaarden (Ontw.-NEN 3114, 3.12)

Meetverwachting, μ

De waarde tot welke de gemiddelde meetwaarde nadert bij een toenemend aantal meetwaarden. Als een benadering (schatting) van μ wordt meestal het gemiddelde $\bar{x}$ gebruikt (Ontw.-NEN 3114, 3.14)

Toevallige afwijking (van een meetwaarde), ϵ

Het verschil tussen een meetwaarde en de meetverwachting (Ontw.-NEN 3114, 4.1)

Toevallige afwijking (van een waarnemingsresultaat) (bij een kwantitatieve analyse), ϵ

Het verschil tussen één individueel waarnemingsresultaat en het gemiddelde van een groot aantal waarnemingsresultaten (meetverwachting), verkregen met dezelfde methode voor hetzelfde homogene monster. (ORA)

Gemeten toevallige afwijking (van een meetwaarde): Het verschil tussen een meetwaarde en de gemiddelde meetwaarde. In formule: $x - \bar{x}$

Systematische afwijking, δ

Het verschil tussen de meetverwachting en de ware waarde.

(Ontw. NEN 3114, 5.1, ISO/DP 5725/1, 4.1.8: bias)

Het verschil tussen de gemiddelde waarde verkregen uit een groot aantal waarnemingsresultaten (= meetverwachting) en de ware waarde (Naar Ontw. NEN 3114, 5.1, ISO/DP 5725/1, 4.1.8: bias).

Waarnemingsresultaat, waargenomen waarde (observed value)

De waarde van een karakteristiek, verkregen als het resultaat van één enkele waarneming (ISO/DP 5725/1, 4.1.1)

Analyseresultaat (bij een kwantitatieve analyse) (test result)

De waarde van een karakteristiek verkregen door een gespecificeerde analysemethode uit te voeren (ISO/DP 5725/1, 4.1.2)

Opmerking- De analysemethode dient aan te geven dat één waarneming, dan wel een aantal individuele waarnemingen moet worden uitgevoerd. In het laatste geval wordt het gemiddelde en de standaardafwijking van het gemiddelde (of een andere geschikte functie, zoals de mediaan) opgegeven als analyseresultaat. Een analyseresultaat kan dus het resultaat zijn dat uit verschillende waargenomen waarden wordt berekend.

Analyseresultaat

Het resultaat van één enkele uitvoering van een werkwijze, beginnend met één analyseportie en eindigt met één resultaat. (ORA)

Nauwkeurigheid (van een analysemethode)

De mate waarin een met een bepaalde analysemethode verkregen waarnemingsresultaat de ware waarde benadert.

(Naar Ontw. NEN 3114, 8.1, ISO/DP 5725/1, 4.1.6: accuracy)

Nauwkeurigheid omvat juistheid en precisie.

Eng: Accuracy = trueness + presision (ISO/DIS 5725).

Deze begrippen worden veelal kwantitatief vastgelegd in omgekeerde maten, dus maten voor de onnauwkeurigheid, de onjuistheid (onzuiverheid) en

imprecisie. Aangezien grotere afwijkingen een geringere nauwkeurigheid inhouden, verdient dit de voorkeur (Ontw-NEN 3114, 8.2, opm. 4).

Onnauwkeurigheid (van een analysemethode)
De mate waarin een met een bepaalde analysemetode verkregen waarnemingsresultaat afwijkt van de ware waarde. (Ontw. NEN 3114, 8.2)

Precisie
De mate van overeenstemming tussen waarnemingsresultaten die worden verkregen door de werkwijze een aantal malen onder vastgelegde omstandigheden toe te passen. (87/410/EEG, ISO 3534-1977).
De mate van overeenstemming tussen meetwaarden bij herhaalde metingen. (Ontw. NEN3114, 4.11, ISO/DP 5725/1, 4.1.9: precision)

Juistheid
De mate van overeenstemming tussen de gemiddelde waarde verkregen uit een grote reeks waarnemingen en de ware waarde.
(ISO/DP 5725/1, 4.1.7: trueness)

Herhaalbaarheid, r (repeatability)
De mate van overeenstemming tussen opeenvolgende resultaten die worden verkregen met dezelfde methode bij identiek analysemateriaal en onder dezelfde omstandigheden (uitvoering door dezelfde persoon, met dezelfde apparatuur in hetzelfde laboratorium op hetzelfde tijdstip of met een korte tussentijd). De te gebruiken maat voor de herhaalbaarheid is de standaardafwijking, s_r, waarbij voor niet te kleine meetseries geldt, dat $r = 2{,}8\ s_r$. (ORA, naar NEN 3114 en ISO 5725-1986)

Binnen-laboratorium reproduceerbaarheid (within-laboratory reproducibility, intermediate reproducibility)
De mate van overeenstemming tussen resultaten die worden verkregen met dezelfde methode bij identiek monstermateriaal onder verschillende omstandigheden (uitvoering door verschillende personen, met dezelfde of verschillende apparatuur, in hetzelfde laboratorium en op verschillende tijdstippen). De te gebruiken maat voor de binnen-laboratorium reproduceerbaarheid is de standaardafwijking. (Naar ISO DP 5725-3).

Reproduceerbaarheid, R (reproduceerbaarheid)
De mate van overeenstemming tussen resultaten die worden verkregen met dezelfde methode bij identiek analysemateriaal onder verschillende omstandigheden (uitvoering door verschillende personen, in verschillende laboratoria, met verschillende apparatuur en op verschillende tijdstippen). De te gebruiken maat voor de reproduceerbaarheid is de standaardafwijking, s_R, waarbij voor niet te kleine meetseries geldt dat $R = 2{,}8\ s_R$. (Naar NEN 3114 en ISO 5725-1986).

Aantoonbaarheidsgrens (limit of detection)
Het laagste gehalte van de analyt in het analysemonster, dat volgens het voorschrift met een redelijke en/of vooraf vastgestelde statistische zekerheid kan worden bepaald. De aantoonbaarheidsgrens wordt weergegeven als een gehalte, bij voorkeur als massafractie - bij voorbeeld µg/kg of mg/kg (analyt/analysemonster)- met vermelding van de grootte van de analyseportie (in gram) die bij de analyse gebruikelijk is. De aantoonbaarheidsgrens is numeriek gelijk aan 3 keer de standaardafwijking van het gemeten gehalte van N representatieve blanc-monsters (N≥20), vermenigvuldigd met een factor 100 gedeeld door het gemiddelde terugvindingspercentage. (Naar: Beschikking 87/410/EEG, 89/610/EEG, EEG VI-1840-98, Rev.2)

Bepaalbaarheidsgrens (limit of quantification)
Het laagste gehalte van de analyt in het analysemonster, dat volgens het voorschrift met een redelijke en/of vooraf vastgestelde statistische zekerheid kan worden bepaald. De bepaalbaarheidsgrens wordt weergegeven als een gehalte, bij voorkeur als massafractie - bij voorbeeld µg/kg of mg/kg (analyt/analysemonster)- met vermelding van de grootte van de analyseportie (in gram) die bij de analyse gebruikelijk is. De bepaalbaarheidsgrens is numeriek gelijk aan 6 keer de standaardafwijking van het gemeten gehalte van N representatieve blanc-monsters (N≥20), vermenigvuldigd met een factor 100 gedeeld door het gemiddelde terugvindingspercentage. (Naar: Beschikking 87/410/EEG, 89/610/EEG, EEG VI-1840-98, Rev.2)

Positief (bij kwalitatief onderzoek)
De aanwezigheid van de analyt in het monster is volgens de gebruikte

methode bewezen, wanneer aan de criteria voor deze methode is voldaan. (Naar 87/410/EEG, II, 3.3)

Negatief (bij kwalitatief onderzoek)
Het resultaat van de analyse wordt als negatief beschouwd indien niet is voldaan aan de criteria van de gebruikte methode, of indien bij de analyse de aanwezigheid van de analyt in een concentratie groter dan de beslissingsgrens niet wordt aangetoond.
(Naar 87/410/EEG, II, 3.4)

Positief analyseresultaat (bij een kwantitatieve methode)
Het analyseresultaat voor de analyt in het monster volgens de werkwijze is gelijk aan of hoger dan het maximum-toegelaten gehalte plus n maal de standaardafwijking die overeenkomt met de voor de werkwijze maximaal toegelaten variatiecoëfficiënt.
Het gehalte van de analyt in het monster overschrijdt het maximum-toegelaten gehalte. (Naar EEG VI/1840/89, 2.2.1)

Negatief analyseresultaat (bij een kwantitatieve methode)
Het analyseresultaat voor de analyt in het monster volgens de werkwijze is lager dan het maximum-toegelaten gehalte plus n maal de standaardafwijking die overeenkomt met de voor de werkwijze maximaal toegelaten variatiecoëfficiënt.
Het gehalte van de analyt in het monster wordt geacht het maximum-toegelaten gehalte niet te overschrijden. (Naar EEG VI/1840/89, 2.2.2)

Gehalte (van een analyt in een matrix)

De fractie van de matrix die bestaat uit de analyt.

Opmerking 1. Dit is een dimensieloos getal. Er bestaat echter een gewoonte het gehalte op te geven als de differentiaal van twee massa-eenheden:

Gehalte

10^{-2}	10 g/kg, %
10^{-3}	mg/g g/kg 0,1%
10^{-4}	100 ppm
10^{-5}	10 ppm
10^{-6}	µg/g mg/kg ppm delen per miljoen, parts per million, ppm
10^{-7}	0,1 µg/g 0,1mg/kg 100ng/g 100µg/kg 0,1ppm 100ppb
10^{-8}	10ng/g 10µg/kg 0,01 µg/g 0,01mg/kg 10ppb 0,01ppm
10^{-9}	ng/g µg/kg ppb delen per miljard, parts per billion, ppb
10^{-12}	pg/g ng/kg ppt delen per biljoen, parts per trillion, ppt

Opmerking 2. De residue-analyse betreft gehalten beneden 10^{-3}.

(ORA)

Maximaal toegelaten variatiecoëfficiënt

De precisie, uitgedrukt als herhaalbaarheid, mag de volgende waarden voor de variatiecoëfficiënt niet overschrijden:

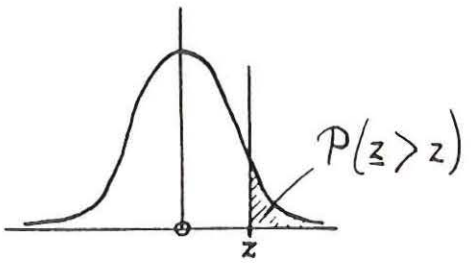
Fractie van de analyt in het monster, gemiddelde waarde	Variatiecoëfficiënt
meer dan 10^{-8} tot en met 10^{-7}	0,20
meer dan 10^{-7} tot en met 10^{-6}	0,15
meer dan 10^{-6}	0,10

TABLE I

Twenty-five hundred random digits.

	1	2	3	4	5	6	7	8	9	10	
1	48461	14952	72619	73689	52059	37086	60050	86192	67049	64739	1
2	76534	38149	49692	31366	52093	15422	20498	33901	10319	43397	2
3	70437	25861	38504	14752	23757	59660	67844	78815	23758	86814	3
4	59584	03370	42806	11393	71722	93804	09095	07856	55589	46020	4
5	04285	58554	16085	51555	27501	73883	33427	33343	45507	50063	5
6	77340	10412	69189	85171	29082	44785	83638	02583	96483	76553	6
7	59183	62687	91778	80354	23512	97219	65921	02035	59847	91403	7
8	91800	04281	39979	03927	82564	28777	59049	97532	54540	79472	8
9	12066	24817	81099	48940	69554	55925	48379	12866	51232	21580	9
10	69907	91751	53512	23748	65906	91385	84983	27915	48491	91068	10
11	80467	04873	54053	25955	48518	13815	37707	68687	15570	08890	11
12	78057	67835	28302	45048	56761	97725	58438	91528	24645	18544	12
13	05648	39387	78191	88415	60269	94880	58812	42931	71898	61534	13
14	22304	39246	01350	99451	61862	78688	30339	60222	74052	25740	14
15	61346	50269	67005	40442	33100	16742	61640	21046	31909	72641	15
16	66793	37696	27965	30459	91011	51426	31006	77468	61029	57108	16
17	86411	48809	36698	42453	83061	43769	39948	87031	30767	13953	17
18	62098	12825	81744	28882	27369	88183	65846	92545	09065	22655	18
19	68775	06261	54265	16203	23340	84750	16317	88686	86842	00879	19
20	52679	19595	13687	74872	89181	01939	18447	10787	76246	80072	20
21	84096	87152	20719	25215	04349	54434	72344	93008	83282	31670	21
22	63964	55937	21417	49944	38356	98404	14850	17994	17161	98981	22
23	31191	75131	72386	11689	95727	05414	88727	45583	22568	77700	23
24	30545	68523	29850	67833	05622	89975	79042	27142	99257	32349	24
25	52573	91001	52315	26430	54175	30122	31796	98842	37600	26025	25
26	16586	81842	01076	99414	31574	94719	34656	80018	86988	79234	26
27	81841	88481	61191	25013	30272	23388	22463	65774	10029	58376	27
28	43563	66829	72838	08074	57080	15446	11034	98143	74989	26885	28
29	19945	84193	57581	77252	85604	45412	43556	27518	90572	00563	29
30	79374	23796	16919	99691	80276	32818	62953	78831	54395	30705	30
31	48503	26615	43980	09810	38289	66679	73799	48418	12647	40044	31
32	32049	65541	37937	41105	70106	89706	40829	40789	59547	00783	32
33	18547	71562	95493	34112	76895	46766	96395	31718	48302	45893	33
34	03180	96742	61486	43305	34183	99605	67803	13491	09243	29557	34
35	94822	24738	67749	83748	59799	25210	31093	62925	72061	69991	35
36	34330	60599	85828	19152	68499	27977	35611	96240	62747	89529	36
37	43770	81537	59527	95674	76692	86420	69930	10020	72881	12532	37
38	56908	77192	50623	41215	14311	42834	80651	93750	59957	31211	38
39	32787	07189	80539	75927	75475	73965	11796	72140	48944	74156	39
40	52441	78392	11733	57703	29133	71164	55355	31006	25526	55790	40
41	22377	54723	18227	28449	04570	18882	00023	67101	06895	08915	41
42	18376	73460	88841	39602	34049	20589	05701	08249	74213	25220	42
43	53201	28610	87957	21497	64729	64983	71551	99016	87903	63875	43
44	34919	78901	59710	27396	02593	05665	11964	44134	00273	76358	44
45	33617	92159	21971	16901	57383	34262	41744	60891	57624	06962	45
46	70010	40964	98780	72418	52571	18415	64362	90636	38034	04909	46
47	19282	68447	35665	31530	59832	49181	21914	65742	89815	39231	47
48	91429	73328	13266	54898	68795	40948	80808	63887	89939	47938	48
49	97637	78393	33021	05867	86520	45363	43066	00988	64040	09803	49
50	95150	07625	05255	83254	93943	52325	93230	62668	79529	65964	50

Tabel II



Tabel van de standaardnormale verdeling.

Aangegeven is $P(z \geq z)$. Dus $P(z \geq 1{,}5) = 0{,}0668$.

	Tweede decimaal van z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0,0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0,1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0,2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0,3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0,4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0,5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0,6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0,7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0,8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0,9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1,0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1,1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1,2	.1151	.1131	.1112	.1093	.1074	.1056	.1038	.1020	.1002	.0985
1,3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1,4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0703	.0694	.0681
1,5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1,6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1,7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1,8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1,9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2,0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2,1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2,2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2,3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2,4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2,5	.0062	0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2,6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2,7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2,8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2,9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3,0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

Tabel III

De *t*-verdeling

Aantal vrijheidsgraden	rechteroverschrijdingskans				
	.1	.05	025	.01	.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5 841
4	1.533	2.132	2.776	3.747	4 604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Tabel IV

Student's t-Distribution

NUMBER OF OBSERVATIONS FOR *t*-TEST OF MEAN

Single-sided test / Double-sided test	Level of *t*-test																				
	$\alpha = 0.005$ / $\alpha = 0.01$					$\alpha = 0.01$ / $\alpha = 0.02$					$\alpha = 0.025$ / $\alpha = 0.05$					$\alpha = 0.05$ / $\alpha = 0.1$					
$\beta =$	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	
0.05																					0.05
0.10																					0.10
0.15																				122	0.15
0.20										139					99					70	0.20
0.25					110					90				128	64				139	101	45	0.25
0.30				134	78				115	63			119	90	45		122	97	71	32	0.30
0.35			125	99	58			109	85	47		109	88	67	34		90	72	52	24	0.35
0.40		115	97	77	45		101	85	66	37	117	84	68	51	26	101	70	55	40	19	0.40
0.45		92	77	62	37	110	81	68	53	30	93	67	54	41	21	80	55	44	33	15	0.45
0.50	100	75	63	51	30	90	66	55	43	25	76	54	44	34	18	65	45	36	27	13	0.50
0.55	83	63	53	42	26	75	55	46	36	21	63	45	37	28	15	54	38	30	22	11	0.55
0.60	71	53	45	36	22	63	47	39	31	18	53	38	32	24	13	46	32	26	19	9	0.60
0.65	61	46	39	31	20	55	41	34	27	16	46	33	27	21	12	39	28	22	17	8	0.65
0.70	53	40	34	28	17	47	35	30	24	14	40	29	24	19	10	34	24	19	15	8	0.70
0.75	47	36	30	25	16	42	31	27	21	13	35	26	21	16	9	30	21	17	13	7	0.75
0.80	41	32	27	22	14	37	28	24	19	12	31	22	19	15	9	27	19	15	12	6	0.80
0.85	37	29	24	20	13	33	25	21	17	11	28	21	17	13	8	24	17	14	11	6	0.85
0.90	34	26	22	18	12	29	23	19	16	10	25	19	16	12	7	21	15	13	10	5	0.90
Value of 0.95	31	24	20	17	11	27	21	18	14	9	23	17	14	11	7	19	14	11	9	5	0.95
$\Delta = \frac{\mu - \mu_0}{\sigma}$ 1.00	28	22	19	16	10	25	19	16	13	9	21	16	13	10	6	18	13	11	8	5	1.00
1.1	24	19	16	14	9	21	16	14	12	8	18	13	11	9	6	15	11	9	7		1.1
1.2	21	16	14	12	8	18	14	12	10	7	15	12	10	8	5	13	10	8	6		1.2
1.3	18	15	13	11	8	16	13	11	9	6	14	10	9	7		11	8	7	6		1.3
1.4	16	13	12	10	7	14	11	10	9	6	12	9	8	7		10	8	7	5		1.4
1.5	15	12	11	9	7	13	10	9	8	6	11	8	7	6		9	7	6			1.5
1.6	13	11	10	8	6	12	10	9	7	5	10	8	7	6		8	6	6			1.6
1.7	12	10	9	8	6	11	9	8	7		9	7	6	5		8	6	5			1.7
1.8	12	10	9	8	6	10	8	7	7		8	7	6			7	6				1.8
1.9	11	9	8	7	6	10	8	7	6		8	6	6			7	5				1.9
2.0	10	8	8	7	5	9	7	7	6		7	6	5			6					2.0
2.1	10	8	7	7		8	7	6	6		7	6				6					2.1
2.2	9	8	7	6		8	7	6	5		7	6				6					2.2
2.3	9	7	7	6		8	6	6			6	5				5					2.3
2.4	8	7	7	6		7	6	6			6										2.4
2.5	8	7	6	6		7	6	6			6										2.5
3.0	7	6	6	5		6	5	5			5										3.0
3.5	6	5	5			5															3.5
4.0	6																				4.0

Tabel V

Student's t-Distribution

NUMBER OF OBSERVATIONS FOR *t*-TEST OF DIFFERENCE BETWEEN TWO MEANS

	Level of *t*-test																				
Single-sided test	$\alpha = 0.005$					$\alpha = 0.01$					$\alpha = 0.025$					$\alpha = 0.05$					
Double-sided test	$\alpha = 0.01$					$\alpha = 0.02$					$\alpha = 0.05$					$\alpha = 0.1$					
$\beta =$	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	
																					0.05
																					0.10
																					0.15
																				137	0.20
															124					88	0.25
										123					87					61	0.30
					110					90					64				102	45	0.35
					85					70				100	50			108	78	35	0.40
				118	68				101	55			105	79	39		108	86	62	28	0.45
				96	55			106	82	45		106	86	64	32		88	70	51	23	0.50
			101	79	46		106	88	68	38		87	71	53	27	112	73	58	42	19	0.55
		101	85	67	39		90	74	58	32	104	74	60	45	23	89	61	49	36	16	0.60
		87	73	57	34	104	77	64	49	27	88	63	51	39	20	76	52	42	30	14	0.65
	100	75	63	50	29	90	66	55	43	24	76	55	44	34	17	66	45	36	26	12	0.70
	88	66	55	44	26	79	58	48	38	21	67	48	39	29	15	57	40	32	23	11	0.75
	77	58	49	39	23	70	51	43	33	19	59	42	34	26	14	50	35	28	21	10	0.80
	69	51	43	35	21	62	46	38	30	17	52	37	31	23	12	45	31	25	18	9	0.85
	62	46	39	31	19	55	41	34	27	15	47	34	27	21	11	40	28	22	16	8	0.90
Value of	55	42	35	28	17	50	37	31	24	14	42	30	25	19	10	36	25	20	15	7	0.95
$\Delta = \frac{\mu_z - \mu}{\sigma}$	50	38	32	26	15	45	33	28	22	13	38	27	23	17	9	33	23	18	14	7	1.00
	42	32	27	22	13	38	28	23	19	11	32	23	19	14	8	27	19	15	12	6	1.1
	36	27	23	18	11	32	24	20	16	9	27	20	16	12	7	23	16	13	10	5	1.2
	31	23	20	16	10	28	21	17	14	8	23	17	14	11	6	20	14	11	9	5	1.3
	27	20	17	14	9	24	18	15	12	8	20	15	12	10	6	17	12	10	8	4	1.4
	24	18	15	13	8	21	16	14	11	7	18	13	11	9	5	15	11	9	7	4	1.5
	21	16	14	11	7	19	14	12	10	6	16	12	10	8	5	14	10	8	6	4	1.6
	19	15	13	10	7	17	13	11	9	6	14	11	9	7	4	12	9	7	6	3	1.7
	17	13	11	10	6	15	12	10	8	5	13	10	8	6	4	11	8	7	5		1.8
	16	12	11	9	6	14	11	9	8	5	12	9	7	6	4	10	7	6	5		1.9
	14	11	10	8	6	13	10	9	7	5	11	8	7	6	4	9	7	6	4		2.0
	13	10	9	8	5	12	9	8	7	5	10	8	6	5	3	8	6	5	4		2.1
	12	10	8	7	5	11	9	7	6	4	9	7	6	5		8	6	5	4		2.2
	11	9	8	7	5	10	8	7	6	4	9	7	6	5		7	5	5	4		2.3
	11	9	8	6	5	10	8	7	6	4	8	6	5	4		7	5	4	4		2.4
	10	8	7	6	4	9	7	6	5	4	8	6	5	4		6	5	4	3		2.5
	8	6	6	5	4	7	6	5	4	3	6	5	4	4		5	4	3			3.0
	6	5	5	4	3	6	5	4	4		5	4	4	3		4	3				3.5
	6	5	4	4		5	4	4	3		4	4	3			4					4.0

Tabel VI a

UPPER 1 PERCENTAGE POINTS OF THE RATIO OF THE LARGEST TO THE SUM OF k INDEPENDENT ESTIMATES OF VARIANCE, EACH OF WHICH IS BASED ON n OBSERVATIONS

k \ n	2	3	4	5	6	7	8	9	10	11	17	37	145	∞
2	0.9999	0.9950	0.9794	0.9586	0.9373	0.9172	0.8988	0.8823	0.8674	0.8539	0.7949	0.7067	0.6062	0.5000
3	0.9933	0.9423	0.8831	0.8335	0.7933	0.7606	0.7335	0.7107	0.6912	0.6743	0.6059	0.5153	0.4230	0.3333
4	0.9676	0.8643	0.7814	0.7212	0.6761	0.6410	0.6129	0.5897	0.5702	0.5536	0.4884	0.4057	0.3251	0.2500
5	0.9279	0.7885	0.6957	0.6329	0.5875	0.5531	0.5259	0.5037	0.4854	0.4697	0.4094	0.3351	0.2644	0.2000
6	0.8828	0.7218	0.6258	0.5635	0.5195	0.4866	0.4608	0.4401	0.4229	0.4084	0.3529	0.2858	0.2229	0.1667
7	0.8376	0.6644	0.5685	0.5080	0.4659	0.4347	0.4105	0.3911	0.3751	0.3616	0.3105	0.2494	0.1929	0.1429
8	0.7945	0.6152	0.5209	0.4627	0.4226	0.3932	0.3704	0.3522	0.3373	0.3248	0.2779	0.2214	0.1700	0.1250
9	0.7544	0.5727	0.4810	0.4251	0.3870	0.3592	0.3378	0.3207	0.3067	0.2950	0.2514	0.1992	0.1521	0.1111
10	0.7175	0.5358	0.4469	0.3934	0.3572	0.3308	0.3106	0.2945	0.2813	0.2704	0.2297	0.1811	0.1376	0.1000
12	0.6528	0.4751	0.3919	0.3428	0.3099	0.2861	0.2680	0.2535	0.2419	0.2320	0.1961	0.1535	0.1157	0.0833
15	0.5747	0.4069	0.3317	0.2882	0.2593	0.2386	0.2228	0.2104	0.2002	0.1918	0.1612	0.1251	0.0934	0.0667
20	0.4799	0.3297	0.2654	0.2288	0.2048	0.1877	0.1748	0.1646	0.1567	0.1501	0.1248	0.0960	0.0709	0.0500
24	0.4247	0.2871	0.2295	0.1970	0.1759	0.1608	0.1495	0.1406	0.1338	0.1283	0.1060	0.0810	0.0595	0.0417
30	0.3632	0.2412	0.1913	0.1635	0.1454	0.1327	0.1232	0.1157	0.1100	0.1054	0.0867	0.0658	0.0480	0.0333
40	0.2940	0.1915	0.1508	0.1281	0.1135	0.1033	0.0957	0.0898	0.0853	0.0816	0.0668	0.0503	0.0363	0.0250
60	0.2151	0.1371	0.1069	0.0902	0.0796	0.0722	0.0668	0.0625	0.0594	0.0567	0.0461	0.0344	0.0245	0.0167
120	0.1225	0.0759	0.0585	0.0489	0.0429	0.0387	0.0357	0.0334	0.0316	0.0302	0.0242	0.0178	0.0125	0.0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabel VI b

UPPER 5 PERCENTAGE POINTS OF THE RATIO OF THE LARGEST TO THE SUM OF k INDEPENDENT ESTIMATES OF VARIANCE, EACH OF WHICH IS BASED ON n OBSERVATIONS

k \ n	2	3	4	5	6	7	8	9	10	11	17	37	145	∞
2	0.9985	0.9750	0.9392	0.9057	0.8772	0.8534	0.8332	0.8159	0.8010	0.7880	0.7341	0.6602	0.5813	0.5000
3	0.9669	0.8709	0.7977	0.7457	0.7071	0.6771	0.6530	0.6333	0.6167	0.6025	0.5466	0.4748	0.4031	0.3333
4	0.9065	0.7679	0.6841	0.6287	0.5895	0.5598	0.5365	0.5175	0.5017	0.4884	0.4366	0.3720	0.3093	0.2500
5	0.8412	0.6838	0.5981	0.5441	0.5065	0.4783	0.4564	0.4387	0.4241	0.4118	0.3645	0.3066	0.2513	0.2000
6	0.7808	0.6161	0.5321	0.4803	0.4447	0.4184	0.3980	0.3817	0.3682	0.3568	0.3135	0.2612	0.2119	0.1667
7	0.7271	0.5612	0.4800	0.4307	0.3974	0.3726	0.3535	0.3384	0.3259	0.3154	0.2756	0.2278	0.1833	0.1429
8	0.6798	0.5157	0.4377	0.3910	0.3595	0.3362	0.3185	0.3043	0.2926	0.2829	0.2462	0.2022	0.1616	0.1250
9	0.6385	0.4775	0.4027	0.3584	0.3286	0.3067	0.2901	0.2768	0.2659	0.2568	0.2226	0.1820	0.1446	0.1111
10	0.6020	0.4450	0.3733	0.3311	0.3029	0.2823	0.2666	0.2541	0.2439	0.2353	0.2032	0.1655	0.1308	0.1000
12	0.5410	0.3924	0.3264	0.2880	0.2624	0.2439	0.2299	0.2187	0.2098	0.2020	0.1737	0.1403	0.1100	0.0833
15	0.4709	0.3346	0.2758	0.2419	0.2195	0.2034	0.1911	0.1815	0.1736	0.1671	0.1429	0.1144	0.0889	0.0667
20	0.3894	0.2705	0.2205	0.1921	0.1735	0.1602	0.1501	0.1422	0.1357	0.1303	0.1108	0.0879	0.0675	0.0500
24	0.3434	0.2354	0.1907	0.1656	0.1493	0.1374	0.1286	0.1216	0.1160	0.1113	0.0942	0.0743	0.0567	0.0417
30	0.2929	0.1980	0.1593	0.1377	0.1237	0.1137	0.1061	0.1002	0.0958	0.0921	0.0771	0.0604	0.0457	0.0333
40	0.2370	0.1576	0.1259	0.1082	0.0968	0.0887	0.0827	0.0780	0.0745	0.0713	0.0595	0.0462	0.0347	0.0250
60	0.1737	0.1131	0.0895	0.0765	0.0682	0.0623	0.0583	0.0552	0.0520	0.0497	0.0411	0.0316	0.0234	0.0167
120	0.0998	0.0632	0.0495	0.0419	0.0371	0.0337	0.0312	0.0292	0.0279	0.0266	0.0218	0.0165	0.0120	0.0083
∞	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabel VII

Section six - Statistical tables

22 Critical values for Cochran's test (see 14.3)

p = the number of laboratories at a given level
n = the number of results per cell (see 14.3.3)

p	n = 2		n = 3		n = 4		n = 5		n = 6	
	1 %	5 %	1 %	5 %	1 %	5 %	1 %	5 %	1 %	5 %
2	-	-	0,995	0,975	0,979	0,939	0,959	0,906	0,937	0,877
3	0,993	0,967	0,942	0,871	0,883	0,798	0,834	0,746	0,793	0,707
4	0,968	0,906	0,864	0,768	0,781	0,684	0,721	0,629	0,676	0,590
5	0,928	0,841	0,788	0,684	0,696	0,598	0,633	0,544	0,588	0,506
6	0,883	0,781	0,722	0,616	0,626	0,532	0,564	0,480	0,520	0,445
7	0,838	0,727	0,664	0,561	0,568	0,480	0,508	0,431	0,466	0,397
8	0,794	0,680	0,615	0,516	0,521	0.438	0,463	0,391	0,423	0,360
9	0,754	0,638	0,573	0,478	0,481	0,403	0,425	0,358	0,387	0,329
10	0,718	0,602	0,536	0,445	0,447	0,373	0,393	0,331	0,357	0,303
11	0,684	0,570	0,504	0,417	0,418	0,348	0,366	0,308	0,332	0,281
12	0,653	0,541	0,475	0,392	0,392	0,326	0,343	0,288	0,310	0,262
13	0,624	0,515	0,450	0,371	0,369	0,307	0,322	0,271	0,291	0,243
14	0,599	0,492	0,427	0,352	0,349	0,291	0,304	0,255	0,274	0,232
15	0,575	0,471	0,407	0,335	0,332	0,276	0,288	0,242	0,259	0,220
16	0,553	0,452	0,388	0,319	0,316	0,262	0,274	0,230	0,246	0,208
17	0,532	0,434	0,372	0,305	0,301	0,250	0,261	0,219	0,234	0,198
18	0,514	0,418	0,356	0,293	0,288	0,240	0,249	0,209	0,223	0,189
19	0,496	0,403	0,343	0,281	0,276	0,230	0,238	0,200	0,214	0,181
20	0,480	0,389	0,330	0,270	0,265	0,220	0,229	0,192	0,205	0,174
21	0,465	0,377	0,318	0,261	0,255	0,212	0,220	0,185	0,197	0,167
22	0,450	0,365	0,307	0,252	0,246	0,204	0,212	0,178	0,189	0,160
23	0,437	0,354	0,297	0,243	0,238	0,197	0,204	0,172	0,182	0,155
24	0,425	0,343	0,287	0,235	0,230	0,191	0,197	0,166	0,176	0,149
25	0,413	0,334	0,278	0,228	0,222	0,185	0,190	0,160	0,170	0,144
26	0,402	0,325	0,270	0,221	0,215	0,179	0,184	0,155	0,164	0,140
27	0,391	0,316	0,262	0,215	0,209	0,173	0,179	0,150	0,159	0,135
28	0,382	0,308	0,255	0,209	0,202	0,168	0,173	0,146	0,154	0,131
29	0,372	0,300	0,248	0,203	0,196	0,164	0,168	0,142	0,150	0,127
30	0,363	0,293	0,241	0,198	0,191	0,159	0,164	0,138	0,145	0,124
31	0,355	0,286	0,235	0,193	0,186	0,155	0,159	0,134	0,141	0,120
32	0,347	0,280	0,229	0,188	0,181	0,151	0,155	0,131	0,138	0,117
33	0,339	0,273	0,224	0,184	0,177	0,147	0,151	0,127	0,134	0,114
34	0,332	0,267	0,218	0,179	0,172	0,144	0,147	0,124	0,131	0,111
35	0,325	0,262	0,213	0,175	0,168	0,140	0,144	0,121	0,127	0,108
36	0,318	0,256	0,208	0,172	0,165	0,137	0,140	0,118	0,124	0,106
37	0,312	0,251	0,204	0,168	0,161	0,134	0,137	0,116	0,121	0,103
38	0,306	0,246	0,200	0,164	0,157	0,131	0,134	0,113	0,119	0,101
39	0,300	0,242	0,196	0,161	0,154	0,129	0,131	0,111	0,116	0,099
40	0,294	0,237	0,192	0,158	0,151	0,126	0,128	0,108	0,114	0,097

Tabel VIII

Critical values for Dixon's outlier test[1]

Test criterion[2]	H	Critical values 5 %	Critical values 1 %
$Q_{10} = \frac{z(2) - z(1)}{z(H) - z(1)}$ or $\frac{z(H) - z(H-1)}{z(H) - z(1)}$ whichever is the greater	3	0,970	0,994
	4	0,829	0,926
	5	0,710	0,821
	6	0,628	0,740
	7	0,569	0,680
$Q_{11} = \frac{z(2) - z(1)}{z(H-1) - z(1)}$ or $\frac{z(H) - z(H-1)}{z(H) - z(2)}$ whichever is the greater	8	0,608	0,717
	9	0,504	0,672
	10	0,530	0,635
	11	0,502	0,605
	12	0,479	0,579
$Q_{22} = \frac{z(3) - z(1)}{z(H-2) - z(1)}$ or $\frac{z(H) - z(H-2)}{z(H) - z(3)}$ whichever is the greater	13	0,611	0,697
	14	0,586	0,670
	15	0,565	0,647
	16	0,546	0,627
	17	0,529	0,610
	18	0,514	0,594
	19	0,501	0,580
	20	0,489	0,567
	21	0,478	0,555
	22	0,468	0,544
	23	0,459	0,535
	24	0,451	0,526
	25	0,443	0,517
	26	0,436	0,510
	27	0,429	0,502
	28	0,423	0,495
	29	0,417	0,489
	30	0,412	0,483
	31	0,407	0,477
	32	0,402	0,472
	33	0,397	0,467
	34	0,393	0,462
	35	0,388	0,458
	36	0,384	0,454
	37	0,381	0,450
	38	0,377	0,446
	39	0,374	0,442
	40	0,371	0,438

1) This is R. S. Gardner's version of Dixon's test as published in table 16.[3] This version applies when it is not known at which end of a series of data an outlier may occur.

Tabel IX

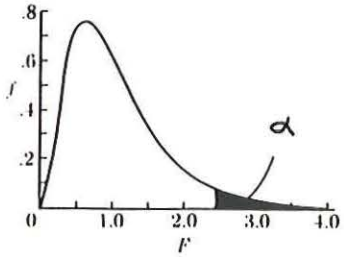
23 Critical values for Grubbs' test (see 14.4)

p is the number of laboratories at a given level.

p	One largest or One smallest		Two largest or Two smallest	
	Upper 1 %	Upper 5 %	Lower 1 %	Lower 5 %
3	1,155	1,155	-	-
4	1,496	1,481	,0000	,0002
5	1,764	1,715	,0018	,0090
6	1,973	1,887	,0116	,0349
7	2,139	2,020	,0308	,0708
8	2,274	2,126	,0563	,1101
9	2,387	2,215	,0851	,1492
10	2,482	2,290	,1150	,1864
11	2,564	2,355	,1448	,2213
12	2,636	2,412	,1738	,2537
13	2,699	2,462	,2016	,2836
14	2,755	2,507	,2280	,3112
15	2,806	2,549	,2530	,3367
16	2,852	2,585	,2767	,3603
17	2,894	2,620	,2990	,3822
18	2,932	2,651	,3200	,4025
19	2,968	2,681	,3398	,4214
20	3,001	2,709	,3585	,4391
21	3,031	2,733	,3761	,4556
22	3,060	2,758	,3927	,4711
23	3,087	2,781	,4085	,4857
24	3,112	2,802	,4234	,4994
25	3,135	2,822	,4376	,5123
26	3,157	2,841	,4510	,5245
27	3,178	2,859	,4638	,5360
28	3,199	2,876	,4759	,5470
29	3,218	2,893	,4875	,5574
30	3,236	2,908	,4985	,5672
31	3,253	2,924	,5091	,5766
32	3,270	2,938	,5192	,5856
33	3,286	2,952	,5288	,5941
34	3,301	2,965	,5381	,6023
35	3,316	2,979	,5469	,6101
36	3,330	2,991	,5554	,6175
37	3,343	3,003	,5636	,6247
38	3,356	3,014	,5714	,6316
39	3,369	3,025	,5789	,6382
40	3,381	3,036	,5862	,6445

Tabel X

$F_{\alpha}[\nu_{teller}, \nu_{noemer}]$



ν_{teller}

ν_{noemer}

	α	1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	50	60	120	∞	α
1	.05	161	199	216	225	230	234	237	239	241	241	243	244	246	248	249	250	251	252	252	253	254	.05
	.025	648	800	864	900	922	937	948	957	963	969	973	977	985	993	997	1000	1010	1010	1010	1010	1020	.025
	.01	4050	5000	5400	5620	5760	5860	5930	5980	6020	6060	6080	6110	6160	6210	6230	6260	6290	6300	6310	6340	6370	.01
2	.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5	.05
	.025	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5	39.5	39.5	.025
	.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5	.01
3	.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.70	8.66	8.64	8.62	8.59	8.58	8.57	8.55	8.53	.05
	.025	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.3	14.2	14.1	14.1	14.0	14.0	14.0	13.9	13.9	.025
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.3	26.2	26.1	.01
4	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.63	.05
	.025	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.66	8.56	8.51	8.46	8.41	8.38	8.36	8.31	8.26	.025
	.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.7	13.6	13.5	.01
5	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68	4.62	4.56	4.53	4.50	4.46	4.44	4.43	4.40	4.36	.05
	.025	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.43	6.33	6.28	6.23	6.18	6.14	6.12	6.07	6.02	.025
	.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.99	9.89	9.72	9.55	9.47	9.38	9.29	9.24	9.20	9.11	9.02	.01
6	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.94	3.87	3.84	3.81	3.77	3.75	3.74	3.70	3.67	.05
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.27	5.17	5.12	5.07	5.01	4.98	4.96	4.90	4.85	.025
	.01	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.56	7.40	7.31	7.23	7.14	7.09	7.06	6.97	6.88	.01
7	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.77	3.73	3.68	3.64	3.60	3.57	3.51	3.44	3.41	3.38	3.34	3.32	3.30	3.27	3.23	.05
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.89	4.82	4.76	4.71	4.67	4.57	4.47	4.42	4.36	4.31	4.27	4.25	4.20	4.14	.025
	.01	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.31	6.16	6.07	5.99	5.91	5.86	5.82	5.74	5.65	.01
8	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.22	3.15	3.12	3.08	3.04	3.02	3.01	2.97	2.93	.05
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.25	4.20	4.10	4.00	3.95	3.89	3.84	3.80	3.78	3.73	3.67	.025
	.01	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.52	5.36	5.28	5.20	5.12	5.07	5.03	4.95	4.86	.01
9	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.01	2.94	2.90	2.86	2.83	2.81	2.79	2.75	2.71	.05
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.77	3.67	3.61	3.56	3.51	3.47	3.45	3.39	3.33	.025
	.01	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	4.96	4.81	4.73	4.65	4.57	4.52	4.48	4.40	4.31	.01
10	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.85	2.77	2.74	2.70	2.66	2.64	2.62	2.58	2.54	.05
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.67	3.62	3.52	3.42	3.37	3.31	3.26	3.22	3.20	3.14	3.08	.025
	.01	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.56	4.41	4.33	4.25	4.17	4.12	4.08	4.00	3.91	.01

	α	1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	50	60	120	∞	α
11	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.72	2.65	2.61	2.57	2.53	2.51	2.49	2.45	2.40	.05
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.48	3.43	3.33	3.23	3.17	3.12	3.06	3.02	3.00	2.94	2.88	.025
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.25	4.10	4.02	3.94	3.86	3.81	3.78	3.69	3.60	.01
12	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.62	2.54	2.51	2.47	2.43	2.40	2.38	2.34	2.30	.05
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.18	3.07	3.02	2.96	2.91	2.87	2.85	2.79	2.72	.025
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.01	3.86	3.78	3.70	3.62	3.57	3.54	3.45	3.36	.01
15	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.40	2.33	2.39	2.25	2.20	2.18	2.16	2.11	2.07	.05
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.86	2.76	2.70	2.64	2.59	2.55	2.52	2.46	2.40	.025
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.52	3.37	3.29	3.21	3.13	3.08	3.05	2.96	2.87	.01
20	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.20	2.12	2.08	2.04	1.99	1.97	1.95	1.90	1.84	.05
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.57	2.46	2.41	2.35	2.29	2.25	2.22	2.16	2.09	.025
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.09	2.94	2.86	2.78	2.69	2.64	2.61	2.52	2.42	.01
24	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.11	2.03	1.98	1.94	1.89	1.86	1.84	1.79	1.73	.05
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.44	2.33	2.27	2.21	2.15	2.11	2.08	2.01	1.94	.025
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.89	2.74	2.66	2.58	2.49	2.44	2.40	2.31	2.21	.01
30	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.01	1.93	1.89	1.84	1.79	1.76	1.74	1.68	1.62	.05
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.31	2.20	2.14	2.07	2.01	1.97	1.94	1.87	1.79	.025
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.90	2.84	2.70	2.55	2.47	2.39	2.30	2.25	2.21	2.11	2.01	.01
40	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.04	1.92	1.84	1.79	1.74	1.69	1.66	1.64	1.58	1.51	.05
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.18	2.07	2.01	1.94	1.88	1.83	1.80	1.72	1.64	.025
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.52	2.37	2.29	2.20	2.11	2.06	2.02	1.92	1.80	.01
60	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.84	1.75	1.70	1.65	1.59	1.56	1.53	1.47	1.39	.05
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.06	1.94	1.88	1.82	1.74	1.70	1.67	1.58	1.48	.025
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.35	2.20	2.12	2.03	1.94	1.88	1.84	1.73	1.60	.01
120	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.87	1.83	1.75	1.66	1.61	1.55	1.50	1.46	1.43	1.35	1.25	.05
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.10	2.05	1.95	1.82	1.76	1.69	1.61	1.56	1.53	1.43	1.31	.025
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.19	2.03	1.95	1.86	1.76	1.70	1.66	1.53	1.38	.01
∞	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.67	1.57	1.52	1.46	1.39	1.35	1.32	1.22	1.00	.05
	.025	5.02	3.69	3.11	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.99	1.94	1.83	1.71	1.64	1.57	1.48	1.43	1.39	1.27	1.00	.025
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18	2.04	1.88	1.79	1.70	1.59	1.52	1.47	1.32	1.00	.01

Gegevens over ISO 5725

Formules

standaardafwijking s_i (voor level j) (formule 1)

$$s_i = \sqrt{\frac{1}{n_i - 1} \sum_{k=1}^{n_i} (y_{ik} - y_i)^2}$$

standaardafwijking s_i (voor level j) (formule 2)

$$s_i = \sqrt{\frac{1}{n_i - 1} \left[\sum_{k=1}^{n_i} y_{ik}^2 - \frac{1}{n_i} \left(\sum_{k=1}^{n_i} y_{ik} \right)^2 \right]}$$

Cochran's testwaarde C
gegeven p standaard afwijkingen s, berekend over n replica testresultaten

$$C = \frac{s^2_{max}}{\sum_{i=1}^{p} s^2_i} \qquad \text{(formule 3)}$$

Dixon's testwaarde Q

H = aantal testwaarden, gebruikt voor Dixon's test

gegeven een set waarden z(h), h = 1,2, ..., H; geordend in opklimmende grootte:

H	testwaarde		
3-7	Q = de grootste waarde van:	$\frac{z(2) - z(1)}{z(H) - z(1)}$	(formule 4a)
	of :	$\frac{z(H) - z(H-1)}{z(H) - z(1)}$	(formule 4b)
8-12	Q = de grootste waarde van:	$\frac{z(2) - z(1)}{z(H-1) - z(1)}$	(formule 5a)
	of :	$\frac{z(H) - z(H-1)}{z(H) - z(2)}$	(formule 5b)
13 of meer	Q = de grootste waarde van:	$\frac{z(3) - z(1)}{z(H-2) - z(1)}$	(formule 6a)
	of :	$\frac{z(H) - z(H-2)}{z(H) - z(3)}$	(formule 6b)

herhaalbaarheids variantie s^2_r (voor level j) (formule 7)

$$s^2_r = \frac{\sum_{i=1}^{p} (n_i - 1)\, s^2_i}{\left[\sum_{i=1}^{p} n_i\right] - p}$$

tussen-laboratorium variantie s^2_L (voor level j) (formule 8)

$$s^2_L = \frac{\frac{1}{p-1}\left[\sum_{i=1}^{p} n_i (\bar{y}_i - \bar{\bar{y}})^2\right] - s^2_r}{\bar{\bar{n}}}$$

waarin (t.b.v. form. 8): (formule 9)

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{p} n_i \bar{y}_i}{\sum_{i=1}^{p} n_i}$$

en (t.b.v. form. 8): (formule 10)

$$\bar{\bar{n}} = \frac{1}{p-1}\left[\sum_{i=1}^{p} n_i - \frac{\sum_{i=1}^{p} n^2_i}{\sum_{i=1}^{p} n_i}\right]$$

$r = b * m$ (TYPE I)

$$b = \frac{\sum_{1}^{i} \left[\frac{r_i}{m_i} \right]}{q} \qquad \text{(formule 11)}$$

waarin q is aantal niveau's

$r = a + b * m$ (TYPE II)

$$a = \frac{T_3 * T_4 - T_2 * T_5}{T_1 * T_3 - T^2_2} \qquad \text{(formule 12)}$$

$$b = \frac{T_1 * T_5 - T_2 * T_4}{T_1 * T_3 - T^2_2} \qquad \text{(formule 13)}$$

$\log (r) = c + d * \log (m)$ (TYPE III)

$$c = \frac{G_2 * G_3 - G_1 * G_4}{q * G_2 - G^2_1} \qquad \text{(formule 14)}$$

$$d = \frac{q * G_4 - G_1 * G_3}{q * G_2 - G^2_1} \qquad \text{(formule 15)}$$

waarin q = aaantal niveau's

14.2 Outlying results

14.2.1 The following practice is recommended for dealing with outliers

a) The tests recommended in the following paragraphs are applied in combination with the following procedure:

$P > 5\ \%$, i.e. the test statistic is less than its 5 % critical value; the item tested is accepted as correct.

$5\ \% \geq P \geq 1\ \%$, i.e. the test statistic lies between its 5 % and 1 % critical values; the item tested is called a straggler and indicated by a single asterisk.

$P < 1\ \%$, i.e. the test statistic is greater than its 1 % critical value; the item is called a statistical outlier and is indicated by a double asterisk.

where P is the probability of the observed value of the test statistic.

The 5 % and 1 % critical values for the different tests are given in section six.

b) It is next investigated whether the stragglers and/or statistical outliers can be explained by some technical error, for example a slip in performing the test, an error in computation, a simple clerical error in transcribing a test result or the analysis of the wrong sample. Where the error was one of the computation or transcription type, the suspect test result should be replaced by the correct value; where the error was from analysing a wrong sample, the test result should be placed in its correct cell. After such correction has been made, the examination for stragglers or outliers should be repeated. If the explanation of the technical error is such that it proves impossible to replace the suspect test result then it should be discarded as a 'genuine' outlier that does not belong to the experiment proper.

c) When any stragglers and/or statistical outliers remain that have not been explained or rejected as belonging to an outlying laboratory, the stragglers are retained as correct items and the statistical outliers are discarded unless the statistician for good reason decides to retain them.

d) When the data for a cell has been rejected for table B under the above procedure, then the corresponding data must be rejected for table C, and vice-versa.

14.2.2 The tests given in 14.3 to 14.5 are of two types. Cochran's test is a test of the within-laboratory variabilities, and should be applied first, and any necessary action taken, with repeated tests if necessary. The other test, (Grubbs') is primarily a test of between-laboratory variability, and can also be used (if $n > 2$) where Cochran's test has raised suspicions as to whether the high within-laboratory variation was attributable to only one of the results in the cell.

14.3 Cochran's test

14.3.1 This standard assumes that between laboratories only small differences exist in the within-laboratory variances. Experience, however, shows that this is not always the case, so that a test has been included here to test the validity of this assumption. Several tests could be used for this purpose, but Cochran's test has been chosen for this standard.

14.3.2 Given a set of p standard deviations s_i, all computed from the same number n of replicate test results, Cochran's criterion is:

$$C = \frac{s_{max.}^2}{\sum_{i=1}^{p} s_i^2} \qquad (9)$$

In the case of two replicates the ranges w_i can be used instead of the standard deviations s_i, and Cochran's criterion then becomes:

$$C = \frac{w_{max.}^2}{\sum_{i=1}^{p} w_i^2}$$

In these expressions $s_{max.}$ and $w_{max.}$ stand for the highest values in the set. If the test is significant, $s_{max.}$ (or $w_{max.}$) is classified as a straggler or statistical outlier according to the procedure of 14.2.1 a).

Critical values for Cochran's criterion at the 5 % and 1 % levels are given for $p = 2$ to 40 and $n = 2$ to 6 in section six.

Cochran's test must be applied to table C of figure 1 at each level separately.

14.3.3 Cochran's criterion applies strictly only when all the standard deviations are derived from the same number (n) of test results obtained under conditions of repeatability. In actual cases this number may vary owing to missing or discarded data. This standard assumes, however, that in a properly organized experiment such variations in the number of test results per cell will be limited and can be ignored, and therefore Cochran's criterion is applied using for n the number of results occurring in the majority of cells.

14.3.4 Cochran's criterion tests only the highest value in a set of standard deviations or ranges and is therefore a one-sided outlier test. Variance heterogeneity may also, of course, manifest itself in some of the standard deviations being comparatively too low. However, small values of standard deviation or range may be very strongly influenced by the degree of rounding of the original test results and are for that reason not very reliable. In addition, it seems unreasonable to reject the data from a laboratory because it has accomplished a higher precision in its test results than the other laboratories. Hence Cochran's criterion is considered adequate.

14.3.5 A critical examination of table C of figure 1 may sometimes reveal that the standard deviations for a particular laboratory are at all or at most levels lower than those for other laboratories. This may indicate that the laboratory works with a lower repeatability than the other laboratories, which in turn may be caused either by better technique and equipment or by a modified or incorrect application of the standard test method. If this occurs it should be reported to the panel, which should then decide whether the point is worthy of a more detailed investigation. (An example of this is laboratory 2 in the experiment detailed in clause 19.)

14.3.6 If the highest standard deviation is classed as an outlier, then the value should be omitted and Cochran's test repeated on the remaining values. This process can be repeated but it may lead to excessive rejections when, as is sometimes the case, the underlying assumption of normality is not sufficiently well approximated to. The repeated application of Cochran's test is here proposed only as a helpful tool in view of the lack of a statistical test designed for testing several outliers together. Cochran's test is not designed for this purpose and great caution should be exercised in drawing conclusions. When two or three laboratories give results having high standard deviations, particularly if this is within only one of the levels, conclusions from Cochran's test should be examined carefully. On the other hand, if several stragglers and/or statistical outliers are found at different levels within one laboratory, this may be a strong indication that the laboratory's within-laboratory variance is exceptionally high, and the whole of the data from that laboratory should be rejected.

14.4 Grubbs' test

14.4.1 Given a set of data g_i for $i = 1,2 \ldots, p$ arranged in ascending order, then to determine whether the largest observation is an outlier using Grubbs' test, compute the statistic

$$G_p = (g_p - \bar{g})/S$$

where

$$\bar{g} = \sum_{i=1}^{p} g_i/p$$

and

$$S = \sqrt{\left[\sum_{i=1}^{p} (g_i - \bar{g})^2/(p-1) \right]}$$

14.4.2 To test the significance of the {smallest / largest} observation, then compute

$$G_1 = (\bar{g} - g_1)/S \qquad G_p = (g_p - \bar{g})/S$$

Stragglers and outliers give rise to values of G_p and G_1 which exceed the tabulated values.

14.4.3 ***Two outlying observations***. To test whether the two largest observations may be outliers, compute

$$G = S^2_{p-1,p} / S^2_0$$

where

$$S^2_0 = \sum_{i=1}^{p} (g_i - \bar{g})^2$$

and

$$S^2_{p-1,p} = \sum_{i=1}^{p-2} (g_i - \bar{g}_{p-1,p})^2$$

and

$$\bar{g}_{p-1,p} = \sum_{i=1}^{p-2} g_i/(p - 2)$$

Alternatively, to test the two smallest observations, compute

$$G = S^2_{1,2}/S^2_0$$

where

$$S^2_{1,2} = \sum_{i=3}^{p} (g_i - \bar{g}_{1,2})^2$$

and

$$\bar{g}_{1,2} = \sum_{i=3}^{p} g_i/(p - 2)$$

Critical values for Grubbs' test are given in section six. Stragglers and outliers give rise to values of G which are smaller than the tabulated values.

14.4.4 When analysing a precision experiment Grubbs' test can be applied to:

a) the cell averages (table B) for a given level j, in which case $g_i = \bar{y}_{ij}$; and $p = p_j$: j being fixed.

b) a single test result within a cell, where Cochran's test has shown the cell range or standard deviation to be suspect.

DIXON

Resultaten rangschikken:

$z(1)$ $z(2)$ $z(3)$ ..... $z(N-2)$ $z(N-1)$ $z(N)$

laagste hoogste

Kritische waarde : Q

$N = 3 \dots 7$

Q_{10} is grootste van

$$\frac{z(2) - z(1)}{z(N) - z(1)} \qquad \frac{z(N) - z(N-1)}{z(N) - z(1)}$$

$N = 8 \dots 12$

Q_{11} is grootste van

$$\frac{z(2) - z(1)}{z(N-1) - z(1)} \qquad \frac{z(N) - z(N-1)}{z(N) - z(2)}$$

$N = 13 \dots$

Q_{22} is grootste van

$$\frac{z(3) - z(1)}{z(N-2) - z(1)} \qquad \frac{z(N) - z(N-2)}{z(N) - z(3)}$$

GRUBB TOETS

p laboratoria
Resultaten, van laag → hoog

$$g_1 \quad g_2 \quad g_3 \quad \cdots \quad g_i \quad \cdots \quad g_{p-2} \quad g_{p-1} \quad g_p$$

$$G_p = \frac{|g_p - \bar{g}|}{S}$$

$$\bar{g} = \sum_{i=1}^{p} \frac{g_i}{p}$$

$$S = \sqrt{\frac{\sum_{i=1}^{p} (g_i - \bar{g})^2}{n-1}}$$

Laagste toetsen

$$G_{f1} = \frac{\bar{g} - g_1}{S}$$

Hoogste toetsen

$$G_{fp} = \frac{g_p - \bar{g}}{S}$$

indien $>$ kritische waarde in Tabel

1% – uitbijter

5% – afzwaaier

$$g_1 \quad g_2 \quad g_3 \cdots g_i \cdots g_{p-2} \quad g_{p-1} \quad g_p$$

Twee laagsten toetsen:

$$G_{1,2} = \frac{S^2_{1,2}}{S_0^2}$$

$$S_0^2 = \sum_{i=1}^{p} (g_i - \bar{g})^2$$

$$S^2_{1,2} = \sum_{i=3}^{p} (g_i - \bar{g}_{1,2})^2$$

$$\bar{g}_{1,2} = \sum_{i=3}^{p} \frac{g_i}{p-2}$$

Twee hoogsten toetsen:

$$G_{p-1,p} = \frac{S^2_{p-1,p}}{S_0^2}$$

$$S^2_{p-1,p} = \sum_{i=1}^{p-2} (g_i - \bar{g}_{p-1,p})^2$$

$$\bar{g}_{p-1,p} = \sum_{i=1}^{p-2} \frac{g_i}{p-2}$$

Uitwerking vraagstuk 1.5

BIJLAGE 4

Opdrachtfile voor SPSS: SOM15.INC
Resultaten in : SOM15.LIS

SOM15.LIS met daarin SOM15.INC, overtolligheden verwijderd.

```
          SPSS/PC+ The Statistical Package for IBM PC                    5/2/90
|INC 'som15.inc'
|Set screen = off.
|
|Get file = 'vm.sys'.
The SPSS/PC+ system file is read from
    file vm.sys
|Sample 5 from 100.
       Er worden 5 willekeurige monsters getrokken.
|Compute groep = 1.
       Een formeel commando, dat SPSS nodig heeft voor Aggregate
       Mogelijk is b.v. groep = rood, wit, blauw, als deze tevoren
       als variabelen zijn gedefinieerd.
|Aggregate outfile = */break=groep
       Als hier gekozen was groep = rood, werd het volgende
       alleen uitgevoerd voor de variabele rood.
| /gem = mean(vleugel)
| /stat =sd(vleugel).
       Gem.vleugel en s.a.vleugel worden berekend en opgeslagen
       in file * (algemene opdracht in SPSS). Alle overige
       gegevens zijn door Aggregate vanaf nu weg.
| /Save outfile = 'res1.sys'.
       Gem en s.a. van vleugels worden overgeschreven in RES1.SYS.

       Deze procedure moet nog 4 keer herhaald worden.
|Get file = 'vm.sys'.
|Sample 5 from 100.
|Compute groep = 1.
|Aggregate outfile = 'res2.sys'/break=groep
       De opbergfile kan korter meteen hier worden opgegeven
| /gem = mean(vleugel)
| /stat =sd(vleugel).
A system file will be written to the file designated
     by res2.sys
      6 variables (including system variables) will be saved.
      0 variables have been dropped.
|
|Get file = 'vm.sys'.
|Variable labels nr 'nummer'
| /vleugel 'vleugellengte'.
|Sample 5 from 100.
|Compute groep = 1.
|Aggregate outfile = 'res3.sys'/break=groep
| /gem = mean(vleugel)
| /stat =sd(vleugel).

|Get file = 'vm.sys'.
|Variable labels nr 'nummer'
| /vleugel 'vleugellengte'.
|Sample 5 from 100.
|Compute groep = 1.
|Aggregate outfile = 'res4.sys'/break=groep
| /gem = mean(vleugel)
| /stat =sd(vleugel).

|Get file = 'vm.sys'.
|Variable labels nr 'nummer'
| /vleugel 'vleugellengte'.
|Sample 5 from 100.
|Compute groep = 1.
|Aggregate outfile = 'res5.sys'/break=groep
| /gem = mean(vleugel)
| /stat =sd(vleugel).
       Bijeenvoegen van de afzonderlijke files,
       waarin 5 keer een gemiddelde en s.a. staan:
|Join add file = 'res1.sys'
| /file = 'res2.sys'
| /file = 'res3.sys'
| /file = 'res4.sys'
| /file = 'res5.sys'.
```

```
|List.
        We willen het resultaat zien:

   GROEP      GEM     STAT

    1.00    44.20     2.17
    1.00    46.40     3.13
    1.00    44.60     3.44
    1.00    47.40     2.61
    1.00    43.20     4.32

Number of cases read =       5    Number of cases listed =       5

        Opbergen in file RESTOT.SYS:
|Save outfile = 'restot.sys'.
The SPSS/PC+ system file is written to
    file restot.sys
      6 variables (including system variables) will be saved.
      0 variables have been dropped.
      5 out of      5 cases have been saved.

        Nu voor de 5 gemiddelden = variabele gem het gemiddelde
        (=in SPSS descriptive statistic 1), de s.a. (=%) en
        variantie (=6) berekenen:
|Descriptive var = gem / stat = 1,5,6.

Number of Valid Observations (Listwise) =         5.00

Variable      Mean     Std Dev   Variance      N  Label

GEM          45.16        1.71        2.91      5

        Nu hetzelfde voor 45 uitslagen, dus
        5 keer steekproef van 45 elementen nemen.
        Bovenstaande opdrachten kopieren, alleen 5 vervangen door 45:

|Get file = 'vm.sys'.
|Sample 45 from 100.
|Compute groep = 1.
|Aggregate outfile = */break=groep
 /gem = mean(vleugel)
 /stat =sd(vleugel).
A new (AGGREGATED) active file has replaced the existing active file.
It contains    6 variables (including system variables).
|Save outfile = 'res1.sys'.

        Enzovoorts voor Res2.sys ...Res5.sys.
        Weer samenvoegen:

|Join add file = 'res1.sys'
 /file = 'res2.sys'
 /file = 'res3.sys'
 /file = 'res4.sys'
 /file = 'res5.sys'.

|List.

   GROEP      GEM     STAT

    1.00    45.27     4.10
    1.00    45.67     3.94
    1.00    45.27     3.68
    1.00    45.16     3.94
    1.00    45.78     3.55

|Save outfile = 'restot.sys'.

        Weer gemiddelde, s.a. en variantie van deze gemiddelden
        berekenen:

|Descriptive var = gem / stat = 1,5,6.

Variable      Mean     Std Dev   Variance      N  Label

GEM          45.43         .28         .08      5

|Exit.
End of Include file.
```

3.2 UITWERKING

	Waarde	s	x + dx	y + dy	z + dz
Lengte x	3	.1	3.1	3	3
Breedte y	2.5	.1	2.5	2.6	2.5
Hoogte z	1.2	.1	1.2	1.2	1.3
Volume	9		9.3	9.36	9.75
Toename			.3	.36	.75
Variantie			.7821		
Stand.afw.			.8843642		
Oppervlak	28.2		28.94	29.04	29.3
Toename			.74	.84	1.1
Variantie			2.4632		
Stand.afw.			1.569459		

3.3

UITWERKING

Monster 1

BEREKENING VAN DE NAUWKEURIGHEID VAN HET FEDERGETAL

		1e meting	2e meting	3e meting	Gemiddel	St.afw.
Vocht	w	55.10	55.05	55.48	55.210	.235
Vet	v	23.10	22.89	22.71	22.900	.148
As	a	3.09	3.07		3.080	.014
Zetmeel	z	5.58	5.62		5.600	.028

Federgetal F = W/ (100-W-V-A-Z)	4.179	
Noemer N = 100-W-V-A-Z	13.210	
Variantie noemer sN^2= sW^2 + sV^2 + sA^2 + sZ^2		.078
Variantie F sF^2 = F^2 (sW^2/W^2 + sN^2/N^2)		.001
St.afw. F sF = SQRT sF^2		.031

Monster 3

BEREKENING VAN DE NAUWKEURIGHEID VAN HET FEDERGETAL

		1e meting	2e meting	3e meting	Gemiddel	St.afw.
Vocht	w	58.30	55.90	61.70	58.633	2.914
Vet	v	18.10	20.60	16.30	18.333	1.768
As	a	3.05	3.08	2.72	2.950	.021
Zetmeel	z	4.90	4.50	3.80	4.400	.283

Federgetal F = W/ (100-W-V-A-Z)	3.739	
Noemer N = 100-W-V-A-Z	15.683	
Variantie noemer sN^2 = sW^2 + sV^2 + sA^2 + sZ^2		11.699
Variantie F sF^2 = F^2 *(sW^2/W^2 + sN^2/N^2)		7.812
St.afw. F sF = SQRT sF^2		2.795

Hoofdstuk 14.

Antwoorden huiswerkvraag

De in het Gemeenschappelijk Onderzoek beoordeelde analysemethode voor natamycine bezit een herhaalbaarheid en reproduceerbaarheid die beide afhankelijk zijn van het bepalingsniveau.
De relatie tussen het niveau en resp. r en R wordt het best beschreven volgens model type II.

De berekende relaties zijn:

$r = 0{,}533 + 0{,}125 * m$

$R = 1{,}086 + 0{,}387 * m$

antwoord vraag a
Op het gemiddelde niveau van beide uitslagen (16) bedraagt de herhaalbaarheid volgens model II: 2,53.
Het verschil tussen beide uitslagen is 1,96. Dit verschil is kleiner dan de herhaalbaarheid, dus behoeft de analyse niet te worden herhaald.

antwoord vraag b
Op het gemiddelde niveau van beide uitslagen (80) bedraagt de reproduceerbaarheid volgens model II: 32,05.
Het verschil tussen beide uitslagen is 40,01 wat groter is dan de reproduceerbaarheid. Beide uitslagen stemmen derhalve niet overeen: herhaling moet uitwijzen welke van beide (of beide!) laboratoria een onjuiste uitslag heeft geleverd.

OPZET VAN DE CURSUS 'TOEGEPASTE STATISTIEK IN DE ANALYTISCHE CHEMIE (CHEMOMETRIE)'

dr W.G. de Ruig en dr ir A.B. Cramwinckel

Doelen:

1. Begrip krijgen voor het relatieve karakter van meten (middag 1).

2. Zicht krijgen op de noodzaak om met modellen te werken. Begrip krijgen voor de consequenties, die het werken met modellen met zich meebrengen (middag 2).

3. Het kunnen aangeven in hoeverre metingen op verschillende niveau's kunnen plaatsvinden en op welke wijze een verzameling metingen gekarakteriseerd kan worden (middag 3).

4. Het kunnen onderscheiden van het verwerken van gegevens als gevolg van twee hoofdtaken van het RIKILT: het meten van monsters op basis van goed- resp. afkeuren en het verbeteren van methoden en technieken (middag 4).

5. Het kunnen onderscheiden van twee belangrijke onderzoekstechnieken: het leggen van verbanden zoals ijklijnen maken, vergelijken van methoden dmv. correlatie, lineaire regressie en multiple regressie en het nagaan of er systematische niveauverschillen bestaan tussen bijv. verschillende ontsluitingsmethoden en detectiemethoden dmv variantieanalyse (middag 5, 6 en 7).

6. Het kunnen opzetten van een experiment om niveauverschillen te kunnen onderzoeken (middag 6).

7. Het kunnen opzetten van een experiment om verbanden te kunnen onderzoeken (middag 7).

8. Het kunnen bespreken van de belangrijkste experimentele onderdelen:
   - het kunnen omschrijven van de doelen van een experiment (sneller, goedkoper, betrouwbaarder etc),
   - het kunnen opzetten van een efficient experiment in relatie tot de gewenste statistische toetsen,
   - het kunnen omschrijven van het onderzoeksmodel,
   - het kennen van de verschillende invloeden op uitkomsten,
   - het begrip krijgen voor spreidingen in uitkomsten,
   - verantwoorde uitspraken kunnen doen over de verzamelde gegevens naar aanleiding van het doel van het experiment.

5. Het kunnen hanteren van het statistische pakket SPSS/PC: het invoeren van data, het plotten van gegevens, het beschrijven van de data, het analyseren van de data (middag 1, 2 en 3).

6. Met SPSS/PC een variantieanalyse kunnen uitvoeren ter toetsing van interacties en effecten. Verder een regressieanalyse kunnen uitvoeren ter toetsing van een verband (middag 5, 6 en 7).

7. Specifieke toetsen kunnen uitvoeren (Programma ISO 5725 van De Vries)

Middag 1. Wat is meten?
Uitleg over de opzet van de cursus en toelichting op de achtergronden van de docenten.
Korte uitleg van de mogelijkheden van statistiek.
Wat is (chemisch) meten?
Welke faktoren beinvloeden uitkomsten?
Toevallige en systematische fouten.
Wat houdt het doen van uitspraken in op basis van metingen?
Introductie van SPSS/PC.
Oefening: Het opstarten van SPSS. Een file interactief kunnen inlezen en met behulp van een opdrachtfile.

Middag 2. Nadere uitwerking van de basis thema's van de cursus. Het doen van uitspraken op basis van een meting.
Oefening: het invoeren van gegevens in SPSS, het inlezen van reeds ingevoerde gegevens als ASCII-file of als spreadsheet-gegevens.

Middag 3. Nadere uitwerking van de basis thema's van de cursus. Het opzetten van een experiment.
Oefening: het beschrijven van de data: plot, examine, describe, frequencies.

Middag 4. Nadere uitwerking van de basis thema's van de cursus. Toelichting op het begrip variantie en variantieanalyse. Nut van variantieanalyse voor het onderzoeken van interacties en het schatten van effecten.
Oefening: Means, oneway en anova.

Middag 5. Nadere uitwerking van de basis thema's van de cursus. Bespreking van de variantieanalyse. Introductie van het begrip regressieanalyse voor het onderzoeken van verbanden.
Oefening: (Meervoudige) regressieanalyse.

Middag 6. Bespreking van de regressieanalyse. Evaluatie van de cursus. Inventarisatie van problemen die niet behandeld zijn. Nut van de cursus. Bijstelling van het programma?

BIJLAGE 6

EVALUATIE VAN DE CURSUS STATISTIEK IN DE CHEMOMETRIE 1990

A.B. Cramwinckel, W.G. de Ruig, A.A.M. Jansen en D.M. van Mazijk-Bokslag

INHOUD

RIKILT, Wageningen, 11 oktober 1990.

1 INLEIDING

In de eerste helft van 1990 is ten behoeve van RIKILT-medewerkers een cursus Statistiek in de Chemometrie opgezet. Hierbij is voor de opzet en uitvoering van de cursus dankbaar gebruik gemaakt van de ervaringen van de Groep Landbouwwiskunde op dit gebied. Gesprekken hierover zijn gevoerd met Jansen en Oude Voshaar, die beiden ook een belangrijk deel van de cursus voor hun rekening hebben genomen. In deze nota worden de ervaringen met deze cursus beschreven. Eerst wordt er een overzicht gegeven van de deelnemers en de dagen dat er cursus is gegeven. Vervolgens komen de deelnemers met hun oordeel over de cursus aan de beurt. Na afloop van de cursus is een evaluatieformulier uitgedeeld. De meeste cursisten hebben het formulier ook ingevuld ingeleverd. Deze nota wordt besloten met enkele opmerkingen en aanbevelingen.

2 DE DEELNEMERS AAN DE CURSUS

AC	1.	Ab van Polanen
	2.	Mieke Tusveld
	3.	Jaap Driessen (gedeelte meegemaakt)
	4.	John Labrijn (gedeelte meegemaakt)
MNT	5.	Jean Slangen
	6.	Dini Venema
TOX	7.	Marcel Mengelers
BFA	8.	Willem Haasnoot
DGM	9.	Henk Keukens
	10.	Ruud Binnendijk
S&V	11.	Dick Wolters
MICROB	12.	Peter Herben
ACON	13.	Jan Horstman

3 DAGEN EN PLAATS VAN DE CURSUS

De cursus heeft hoofdzakelijk plaats gevonden op de TFDL. Daar was een leslokaal met PC's gehuurd, geschikt om zowel het theoretische gedeelte als ook het praktische gedeelte te geven.

Lesdata	Plaats	Tijd	Docenten
20-04	TFDL	13.30-17.00	1,2
27-04	TFDL	13.30-17.00	1,2
11-05	TFDL	13.30-17.00	1,2
18-05	TFDL	9.00-17.00	1,2,3
22-05	RIKILT	13.30-16.00	1
1-06	TFDL	9.00-17.00	2,4,5
8-06	TFDL	9.00-17.00	2,4,5

1= dr. W. G. de Ruig
2= dr. ir. A. B. Cramwinckel en D. M. van Mazijk-Bokslag
3= drs. P. H. U. de Vries
4= ir. A. A. M. Jansen
5= drs. J. H. Oude Voshaar

4 DE INHOUD VAN DE CURSUS

De cursus bestond uit een theoretisch en uit een praktisch gedeelte. De theorie omvat inleiding in statistische begrippen en achtergronden (De Ruig), calibratie (Jansen) en variantie analyse/(multiple) regressie (Oude Voshaar). Het praktische gedeelte omvatte het leren omgaan met SPSS m.b.t. data-invoer, beschrijvende statistiek, variantie analyse en (multiple) regressie analyse (Cramwinckel en Van Mazijk-Bokslag). Voor beide onderdelen is een handleiding samengesteld.

De stof van het theoretische gedeelte was als volgt over de lesdagen verdeeld.

Dag 1 Hoofdstuk 1 t/m 5
Dag 2 Hoofdstuk 6 t/m 9
Dag 3 Hoofdstuk 10 t/m 13
Dag 4 Hoofdstuk 14
Dag 5 Hoofdstuk 15 t/m 17
Dag 6 Hoofdstuk 18 t/m 19
Dag 7 Hoofdstuk 20

5 BEOORDELING VAN DE CURSUS DOOR DE DEELNEMERS

Tien deelnemers hebben het evaluatieformulier ingevuld. De resultaten hiervan staan hieronder. De getallen zijn percentages:

	Niveau			Duidelijkheid		
	Te laag	Goed	Te hoog	Niet	Gaat	Heel
THEORETISCH GEDEELTE						
De Ruig	0	100	0	0	10	90
De Vries	0	100	0	0	32	68
Oude Voshaar	0	100	0	0	10	90
Jansen	0	68	32	0	78	22
PRAKTISCH GEDEELTE (Cramwinckel)						
Datainvoer	0	100	0	10	60	30
Beschrijvende stat.	0	100	0	0	70	30
ISO 5725 (De Vries)	12	88	0	10	21	68
Variantie analyse	0	100	0	0	38	62
Regressie analyse	0	100	0	0	38	62

In totaal zijn 10 evaluatieformulieren ontvangen

Op de formulieren kon men opmerkingen over de cursus kwijt. Hier volgt een overzicht op basis van de gegeven rubrieken:

* Onderwerpen die ik heb gemist, of die voor mij te summier behandeld zijn:

- Hoe ga je om met de interpretatie van de resultaten?
- n.v.t. (3 x)
- Uitgebreid overzicht van gebruikte symbolen en begrippen
- Inhoudsopgave theoretisch gedeelte
- De start met SPSS was voor mij erg verwarrend. Dit was een stuk beter geweest wanneer direct aan het begin een tabelletje had gestaan van de te typen handelingen zoals de opdrachtfile ook helemaal uitgeschreven was. Wat je moet doen is: ne *.inc (opdracht file maken); spss *.inc (opdracht file uitvoeren) en ne spss.lis (resultaten bekijken).
- Toepassingen en uitwerking variantie-analyse en regressie-analyse kon uitvoeriger.
- Praktische voorbeelden hadden er wat mij betreft meer naar voren moeten komen en bijv. klassikaal uitgewerkt moeten worden. Bijv. ISO 5725 gebruik ik in de praktijk nooit en aan de hand van voorbeelden gaat het dan wat meer spreken.
- Het gedeelte m.b.t. standaarddeviatie, SED, SEM vond ik wat te summier.
- Verwerking gegevens rondzendmonsters met referentiemonsters of lab.
- Data invoer, definiëren variabelen e.d. Communicatie met andere pro gramma's (Slidewrite i.v.m. plotter; Lotus i.v.m. data; ASCII i.v.m. data).

* Totaal oordeel over de cursus. De cursus was voor mij:

- nuttig (2 x)
- heel leerzaam
- goed
- te uitgebreid. Zaken zoals variantie-analyse en multiple regressie-analyse zijn onderdelen die m.i. te specialistisch zijn en te weinig worden toegepast binnen het lab. Of worden zo weinig toegepast zodat beter 'n specialist kan worden geconsulteerd.
- De opzet van de tweede helft van de cursus was mijns inziens beter ('s middags praktijk). Zeker in het begin was begeleiding met het werken met SPSS nodig. Alléén oefenen kost dan relatief veel meer tijd, als het er al van komt.
- Zeer interessant. Hoewel wij op de afdeling geen ringonderzoek uitvoeren lijkt mij dit aspect voor het RIKILT zeer belangrijk. Voor ons ligt de nadruk meer op de variantie- en regressie-analyse.
- Over het algemeen was de cursus erg duidelijk, mede dankzij de syllabus. Ik had het echter zeer op prijs gesteld als we de te behandelen hoofdstukken eerder hadden ontvangen, zodat we het alvast konden bestuderen. Wellicht waren er dan ook meer gerichte vragen gesteld.
- Ik heb er helaas te weinig tijd aan kunnen besteden (voor en na de lessen).
- Boeiend en veelal verhelderend, drempelverlagend m.b.t. SPSS. ISO 5725 sprak mij minder aan (werk ik niet mee). Idem multiple regressie.
- Duidelijk, overzichtelijk, goede volgorde, alleen problemen met het gebruik van SPSS op de afdeling i.v.m. oefenen.

* Nadere toelichting, suggesties, opmerkingen:

- Bij het praktische gedeelte meer uitgewerkte voorbeelden erbij.
- Week van te voren nieuwe lessen op schrift.
- Aan de hand van voorbeeld SPSS leren, daarna duidelijk maken wat ermee moet. Op papier.
- Misschien is het een goed idee om in de opdrachtfile als 'REM-statement' de betekenis van de 'stenografische' opdrachten te geven.
- n.v.t. (1 x)
- De vragen over het niveau en de duidelijkheid van de cursus zijn niet zuiver (dualistisch). Een antwoord op deze vragen zegt behalve over de cursus ook iets over de cursist(e).
  Voor het geval deze cursus nogmaals wordt herhaald is het misschien zinvol om meer voorbeelden van het RIKILT te gebruiken. Verder vereist het leren omgaan met SPSS (zoals trouwens met elk 'nieuw' programma) veel tijd als het gaat om datainvoer en het gebruik maken van de mogelijkheden die zo'n programma biedt. Als het niet mogelijk is om van de cursist huiswerk in deze te eisen dan is het misschien verstandiger om in korter tijdsbestek dan 1x week, meer uren te besteden aan SPSS. Te vaak komt het voor, als je slechts 1 x per week gebruik maakt van 'n

programma, dat het maanden duurt voordat je er zelfstandig mee om kunt gaan.

- Voor wat betreft het computerprogramma SPSS, moet ik zeggen dat het me in het begin totaal niet aansprak, maar naarmate de cursus vorderde, ik toch een duidelijk beeld heb gekregen van wat je er zoal mee kunt doen. Of ik er in de toekomst ook daadwerkelijk mee ga werken is voor mij nog de vraag, te meer omdat ik voor mijn werk (NIRS) gebruik maak van specifieke software.
  Praktisch gedeelte SPSS: De oefeningen zoals gepland in het cursusboek lijken goed aan te sluiten op de 's morgens behandelde theorie. Jammer dat er (voor mijn gevoel) niets van terecht is gekomen. Vooral in het begin is er te weinig tijd aan SPSS besteed, waardoor het een en ander behoorlijk chaotisch op mij over kwam. SPSS-uitleg zou wat concreter kunnen: duidelijk probleem stellen en stap voor stap naar antwoord toe werken (de laatste cursusdagen ging dit prima).
  Algemeen oordeel: goed.
- Met name het praktisch gedeelte is in het begin nogal verwarrend voor iemand die nooit met SPSS heeft gewerkt. Het cursusmateriaal zou vóór de aanvang van de cursus moeten worden uitgedeeld (minstens een maand)!
- Opmerkingen: Enthousiast team van 'leraren'. Goed voorbereid en uitgewerkt waardoor er een goed naslagwerk is ontstaan. Het praktisch gedeelte vond ik met name in het begin wat rommelig. Opvallend dat het RIKILT alleen het gemiddeld gevonden resultaat 'de deur uitdoet', terwijl het dus ook anders kan.
  Suggesties: Wat meer praktijk gerichte voorbeelden op chemisch gebied, bijv. door de cursisten te vragen praktijkgerichte problemen of voorbeelden te laten inleveren. Wat meer aandacht besteden aan de diverse standaarddeviaties, betrouwbaarheidsintervallen.
- Gedemonstreerd programma ISO 5725 onvoldoende beveiligd. Kan dit ook met SPSS?
- Storend was het gegoochel met lesdata. Dit soort cursus is voor herhaling vatbaar.
- Misschien toch handiger om die mensen de cursus te laten volgen die op grond van hun functie c.q. taak er het meest mee te maken hebben. Daarbij denk ik aan het theoretisch gedeelte aangezien de conclusies in een RIKILT rapport toch allemaal volgens eenzelfde lijn zouden moeten worden uitgevoerd. Mogelijk ook een lichtere cursus voor diegene die alleen te maken hebben met data invoer en/of gegevensverwerking t.b.v. een labassistent of afdelingshoofd en zijn/haar rapportage. Wat ook gebleken is dat er vooral behoefte bestaat aan voorbeelden uit de praktijk. Misschien dat hiervoor de afdelingen benaderd kunnen worden.

6 OPMERKINGEN EN AANBEVELINGEN

De cursus is door de docenten als zeer nuttig ervaren. Het gemiddelde niveau van de cursisten wat betreft statistische vaardigheden lijkt niet erg hoog te zijn en is tamelijk variabel. Dat komt mogelijk omdat er op