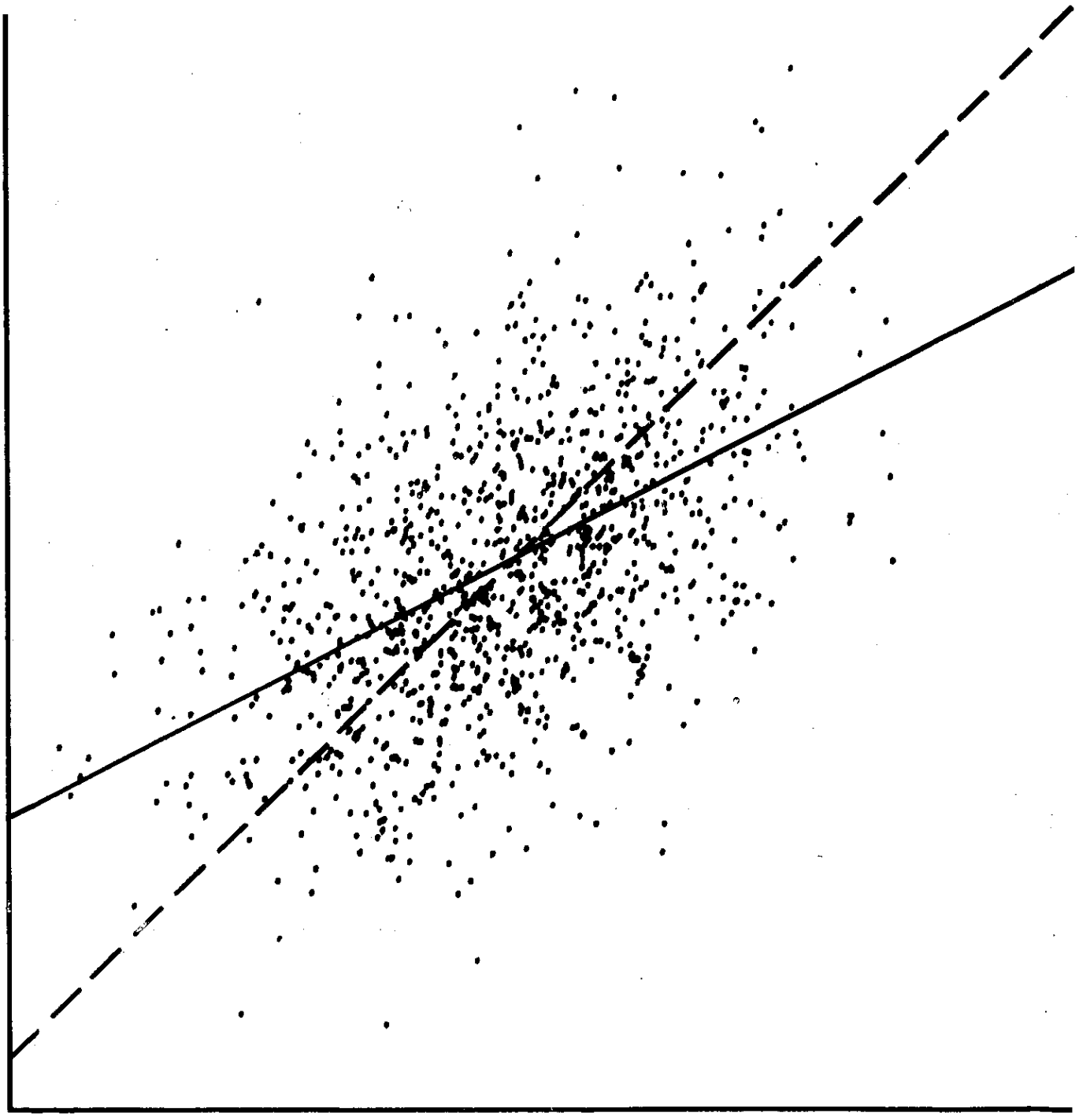
Technische Hogeschool
Eindhoven

Dictaatnummer 2.230
Prijs f. 9,00

Onderafdeling der Wiskunde en Informatica

Toegepaste Statistiek

drs. A.J. Bosch
prof.dr. R. Doornbos
dr.ir. H.N. Linssen
J.Th.M. Wijnen



TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

TOEGEPASTE STATISTIEK

door

Drs. A.J. Bosch

Prof. Dr. R. Doornbos

Dr. Ir. H.N. Linssen

J.Th.M. Wijnen

Voorjaarssemester 1982

Inhoudsbeschrijving

TOEGEPASTE STATISTIEK

A.J. Bosch, R. Doornbos, H.N. Linssen, J.Th.M. Wijnen

Voorjaarssemester 1982

JdG, 24 Juli 2005

TECHNISCHE HOGESCHOOL EINDHOVEN

Onderafdeling der Wiskunde en Informatica

TOEGEPASTE STATISTIEK

door

drs. A.J. Bosch
prof. dr. R. Doornbos
J.Th.M. Wijnen
dr. ir. H.N. Linssen

Voorjaarssemester 1982

Inhoud.

Literatuur.

[1] H.L. Alder and E.B. Roessler, Introduction to Probability and Statistics, 5th Edition, Freeman, 1972.

[2] K.A. Brownlee, Statistical Theory and Methodology in Science and Engineering, Wiley, 1960.

[3] H.A. Halstead, Introduction to Statistical Methods, MacMillan, 1969.

1. De toepassingsgebieden van de statistiek.

Statistiek is een hulpwetenschap die toepassing vindt bij het verzamelen, analyseren, weergeven en interpreteren van waarnemingen op vele gebieden. De toegepaste beginselen zijn steeds dezelfde, de toegepaste methoden zijn echter in hoge mate afhankelijk van de aard van de waarnemingen en van het vakgebied.
Essentieel is dat ervan wordt uitgegaan dat alle waarnemingen althans gedeeltelijk van het toeval afhankelijk zijn. De taak van de statistiek is onder meer de invloed van dit toevalselement op de juiste wijze in de interpretatie van de uitkomsten te verwerken.

Voorbeelden van toepassingsgebieden:

Landbouw	: Proefopzetten waarbij de variaties in bodemstructuur en weersomstandigheden zo goed mogelijk worden uitgeschakeld. Steekproefonderzoek voor het schatten van de grootte van de veestapel, het te velde staande gewas, enz.
Geneeskunde	: Onderzoek naar het effect van medicamenten; proeven met "placebo" als controle. Epidemiologisch onderzoek naar verspreiding en oorzaken van epidemieën (bv. de invloed van roken op de gezondheid).
Biologie Farmacologie	: Proeven met planten en dieren, waarbij grote toevallige variaties optreden tussen de experimentele eenheden. Het toetsen van farmaceutische preparaten en het opstellen van normen.
Technologie Industrie	: Proeven onder technologische omstandigheden. Kwaliteitsbeheersing, vaststellen van toleranties. Marktonderzoek en marktvoorspellingen.
Sociologie	: Steekproefonderzoek door middel van enquêtes. Problemen: een juiste vraagstelling, "non-response", omvang en opzet van de steekproef. Het toepassen van multivariate technieken.

Econometrie : Het waarnemen van economische veranderingen, het construeren van modellen die de onderlinge samenhang beschrijven, het toetsen van die modellen.

Psychometrie : Het interpreteren van de resultaten van psychologische tests met behulp van multivariate methoden.

Natuurwetenschappen: Hier speelt de theorie van de foutenvoortplanting een belangrijke rol.

2. Bewerking van een serie waarnemingen. Beschrijvende statistiek.

2.1. Inleiding.

Wij beschouwen in dit hoofdstuk één enkele serie waarnemingen. We nemen daarbij aan dat we te maken hebben met een aselecte steekproef uit een normale verdeling met onbekende parameters. Deze onderstelling wordt in de praktijk vaak al dan niet stilzwijgend gemaakt, zonder dat dit gemotiveerd wordt. In principe moet echter altijd een motivering bestaan ook al wordt zij niet vermeld. In dit hoofdstuk laten wij de motivering rusten: wij komen hier later in een apart hoofdstuk op terug.
Het statistische probleem is nu schattingen te vinden voor de onbekende parameters en de betrouwbaarheid van die schattingen vast te leggen. In de volgende paragrafen wordt een en ander uiteengezet.
In de laatste paragraaf van dit hoofdstuk wordt aandacht besteed aan de beschrijvende statistiek; het duidelijk en overzichtelijk weergeven van waarnemingen en analyseresultaten is een zeer belangrijk facet van het werk van de statisticus.

2.2. Een lijst van steekproefgrootheden en symbolen.

Een enkele aselecte steekproef uit een normale verdeling kan afdoende worden samengevat door het aantal waarnemingen, n, het gemiddelde, $\bar{y}$, en de steekproefvariantie, s^2, of standaardafwijking, s. Verschillende andere steekproefgrootheden spelen echter een rol in groter verband. Hieronder volgt een lijst van grootheden, die we nog zullen toepassen, en de bijbehorende symbolen.

Tabel 2.2.1. Grootheden en symbolen.

1) de seriegrootte of steekproefomvang	n
2) de individuele waarneming	y_i $(i = 1,\ldots,n)$
3) de naar opklimmende grootte gerangschikte waarneming	$y_{(i)}$
4) de som der waarnemingen	$y_\cdot = \Sigma y_i$
5) het gemiddelde	$\bar{y}_\cdot$ of $\bar{y}$
6) de ruwe kwadratensom	Σy_i^2
7) de correctieterm	$y_\cdot^2/n = (y_\cdot)^2/n$
8) de (gereduceerde) kwadratensom	$KS = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{y_\cdot^2}{n}$
9) het aantal vrijheidsgraden	$\nu = n - 1$
10) de steekproefvariantie	$s^2 = KS/\nu$, of $\hat{\sigma}^2$
11) de steekproefstandaardafwijking	$s = \sqrt{s^2}$ of $\hat{\sigma}$
12) de steekproefvariantie van het gemiddelde	$s_{\bar{y}}^2 = s^2/n$
13) de steekproefstandaardafwijking van het gemiddelde	$s_{\bar{y}} = s/\sqrt{n}$
14) de mediaan = de middelste waarneming voor n oneven,	$M = y_{(\frac{1}{2}n+\frac{1}{2})}$
= het gemiddelde van het middelste paar voor n even	$M = \frac{1}{2}[y_{(\frac{1}{2}n)} + y_{(\frac{1}{2}n+1)}]$
15) de spreidingsbreedte of range	$R = y_{(n)} - y_{(1)}$
16) de bovengrens voor een afrondingsinterval	a_{max}
17) de klassebreedte	b
18) de frequentie in een klasse i = het aantal waarnemingen in die klasse	f_i

2.3. Codering en afronding.

2.3.1. Codering.

Om rekenwerk te besparen is het vaak handig de waarnemingen te coderen. Wij doen dit door de waarneming y_i te vervangen door $y_i^* = py_i - q$, waarin p en q constante getallen zijn.
Wanneer $\underline{y}$ een stochastische variabele is met $\mathcal{E}\underline{y} = \mu$ en var $\underline{y} = \sigma^2$, dan geldt:

$$\mathcal{E}\underline{y}^* = \mathcal{E}(p\underline{y} - q) = p\mathcal{E}\underline{y} - q = p\mu - q$$

en

$$\text{var } \underline{y}^* = \text{var}(p\underline{y} - q) = p^2 \text{ var } \underline{y} = p^2\sigma^2 .$$

Voor de steekproefgrootheden $\bar{y}$ en s^2 gelden soortgelijke formules. (Ga dat na!)
Het is van belang na het rekenwerk de resultaten weer te decoderen, omdat we per slot van rekening geïnteresseerd zijn in gemiddelde en variantie van de oorspronkelijke waarnemingen.
In de paragrafen 2.4 en 2.5 wordt het bovenstaande met behulp van een numeriek voorbeeld geïllustreerd.

2.3.2. Afronding.

Vaak bevatten de waarnemingen meer cijfers dan zin heeft gezien de bereikte nauwkeurigheid van de metingen. Een afrondingsinterval

$$a < \tfrac{1}{2}\sigma \qquad (2.3.2.1)$$

is altijd toelaatbaar. De motivering is dat door afronding toevallige fouten, $\underline{e}$, worden gemaakt met een verdeling

$$f(e) = \frac{1}{a} , \quad -\tfrac{1}{2}a < e < \tfrac{1}{2}a$$

zodat

$$\mathcal{E}\underline{e} = 0 , \quad \sigma_{\underline{e}}^2 = \frac{1}{12} a^2 .$$

Is de variantie van de waarnemingen vóór afronding σ^2, dan wordt de variantie na afronding volgens regel (2.3.2.1):

$$\sigma_1^2 = \sigma^2 + \sigma_{\underline{e}}^2 = \sigma^2 + \frac{1}{12} a^2 < \sigma^2(1 + \frac{1}{48}) .$$

De variantie neemt dus hoogstens met 2% toe en dit is toelaatbaar gezien de vereenvoudiging die door een afronding wordt bewerkstelligd.
Is, zoals doorgaans, σ onbekend dan kan

$$a < \tfrac{1}{2}s \qquad (2.3.2.2)$$

of

$$a < R/2\sqrt{n} \ , \quad \text{mits } n \leq 10 \qquad (2.3.2.3)$$

worden genomen. Deze laatste formule is vooral van nut als s nog niet is berekend.
Is $n > 10$ dan kan (2.3.2.3) worden toegepast op een deelserie van 10 waarnemingen (deze deelserie moet aselect worden gekozen) of men kan uit de serie k deelseries van r waarnemingen afsplitsen ($kr \leq n$) en dan

$$a < \bar{R}/2\sqrt{r} \qquad (2.3.2.4)$$

kiezen, waarbij $\bar{R}$ de gemiddelde spreidingsbreedte is, berekend uit de k deelseries.
(2.3.2.2) en (2.3.2.3) gelden voor de originele waarnemingen. Voor het gemiddelde $\bar{y}$ kan een afronding

$$a < s/2\sqrt{n} \quad \text{of} \quad a < R/2n \qquad (2.3.2.5)$$

worden toegepast.
Als toelaatbare afronding voor s gebruiken we

$$a < s/\sqrt{8\nu} \ ,$$

waarbij ν het aantal vrijheidsgraden is behorende bij de schatting s^2.
Verder verdient het aanbeveling steeds op een geheel aantal decimalen af te ronden, dus niet op even getallen of veelvouden van 5. En waarnemingen die op een 5 eindigen worden afgerond op een even cijfer; dit om systematische afrondingen naar boven of naar beneden te vermijden.
Ter illustratie volgen hierna twee numerieke voorbeelden.

2.4. De numerieke bewerking van een kleine serie waarnemingen.

2.4.1. De berekening van de kwadratensom.

Alvorens een voorbeeld uit te werken willen wij de aandacht vestigen op het volgende. Voor de kwadratensom, KS, geldt

$$KS = \Sigma (y_i - \bar{y})^2 = \Sigma y_i^2 - n\bar{y}^2 \ , \qquad (2.4.1.1)$$

of ook

$$\boxed{KS = \Sigma y_i^2 - (\Sigma y_i)^2/n = \Sigma (y_i - c)^2 - \{\Sigma (y_i - c)\}^2/n} \qquad (2.4.1.2)$$

(2.4.1.1) is vooral theoretisch van belang, maar is niet geschikt voor berekening, omdat $\bar{y}$ vaak een oneindig voortlopende breuk is. Door schijnbaar verwaarloosbare afrondingen van $\bar{y}$ kunnen dan ernstige numerieke fouten in de KS optreden.

Alléén de formules (2.4.1.2) komen daarom voor berekening van de KS in aanmerking.

De codering van y door een constante c af te trekken dient om de berekening zo eenvoudig mogelijk te maken. Bij gebruik van een computer is dat niet nodig, bij gebruik van een tafelrekenmachine heeft het zijn nut. In dit laatste geval is het vaak gemakkelijk c zó te kiezen dat alle bedragen $(y_i - c) \geq 0$ zijn; de sommen $\Sigma (y_i - c)$ en $\Sigma (y_i - c)^2$ kunnen dan tegelijk worden uitgerekend.
Bij gebruik van een rekenmachine is afronding, wat de berekening betreft, niet nodig; wanneer echter de berekening uit het hoofd moet geschieden, dan heeft afronding volgens 2.3.2 grote voordelen, zoals hieronder wordt gedemonstreerd. Ook blijkt dan dat de verschillen in $\bar{y}$, s^2 en s berekend zonder en met afronding zó gering zijn dat ze tegenover de statistische betrouwbaarheid van de gegevens kunnen worden verwaarloosd.

2.4.2. Berekeningen met en zonder afronding.

Tabel 2.4.2.1. Voorbeeld.

	c = 920		afgerond	
y_i	$y_i' = y_i - c$	$(y_i')^2$	$y_i'' = 10^{-1}(y_i - c)$	$(y_i'')_2$
887	-33	1089	-3	9
964	44	1936	4	16
939	19	361	2	4
987	67	4489	7	49
925	5	25	0	0
946	26	676	3	9
920	0	0	0	0
	$\Sigma y_i' = 128$	$\Sigma(y_i')^2 = 8576$	$\Sigma y_i'' = 13$	$\Sigma(y_i'')^2 = 87$

Afronding op tientallen is toegestaan. Immers

$$R = 987 - 887 = 100$$

dus

$$a_{max} = R/2\sqrt{n} = 100/2\sqrt{7} = 19 \ .$$

Tabel 2.4.2.2. Berekeningen.

zonder afronding		met afronding	
$\Sigma y_i'$	= 128	$\Sigma y_i''$	= 13
$\Sigma(y_i')^2$	= 8576	$\Sigma(y_i'')^2$	= 87.0
$(\Sigma y')^2/n$	= 2341	$(\Sigma y_i'')^2/n$	= 24.1
KS	= 6235	10^{-2} KS	= 62.9
$s^2 = KS/(n-1)$	= 1039	s^2	= 1048
$s_{\bar{y}}^2 = s^2/n$	= 148.4	$s_{\bar{y}}^2$	= 149.7

Tabel 2.4.2.3. Samenvatting.

zonder afronding	met afronding
$n = 7$	$n = 7$
$\bar{y} = c + \frac{\Sigma y_i'}{7} = 938.3$	$\bar{y} = c + 10\,\frac{\Sigma y_i''}{7} = 938.6$
$s = 32.2$	$s = 32.4$
$s_{\bar{y}} = 12.2$	$s_{\bar{y}} = 12.2$

De toelaatbare afronding voor $\bar{y}$ is

$$a_{max} = s/2\sqrt{n} = \frac{32.4}{2\sqrt{7}} = 6$$

zodat decimalen weggelaten kunnen worden.
Voor s vinden we

$$a_{max} = \frac{s}{\sqrt{8\nu}} = \frac{32.4}{\sqrt{8 \times 6}} = 4 \ .$$

De waarnemingsreeks kunnen we derhalve voldoende beschrijven met

$$n = 7 \ , \quad \bar{y} = 938 \ , \quad s = 32 \ .$$

2.5. De numerieke bewerking van een grote serie waarnemingen.

2.5.1. Inleiding.

Bij een grote waarnemingsreeks worden de waarnemingen zonder ze eerst af te ronden ingedeeld in gelijke intervallen of klassen. Het aantal klassen kiezen we ongeveer $\sqrt{n}$, maar in ieder geval > 5 en < 16. De klassegrenzen dienen zo bepaald te worden dat iedere waarneming ondubbelzinnig in één klasse thuishoort. De waarnemingen in één klasse worden dan geacht samen te vallen met het klassemidden.
We kiezen een klassemidden c als nulpunt. Als dan y_i het klassemidden is van klasse i en b de klassebreedte, dan coderen we

$$y_i' = \frac{1}{b}\,(y_i - c) \ .$$

Daarmee zijn de klassen genummerd: ...,-2,-1,0,1,2,... .

2.5.2. Het indelen van de waarnemingen in klassen.

In verband met een grootscheeps onderzoek in de Verenigde Staten op het gebied van hart- en vaatziekten werd in 1950 van een groot aantal hartpatiënten o.a. het cholesterolgehalte in het bloed bepaald.
De waarnemingen in de volgende tabel vormen een steekproef uit deze gegevens.

Tabel 2.5.2.1. Het cholesterolgehalte in het bloed van hartpatiënten in mg per 100 cc. (mg%) (The Los Angeles Heart Study, 1950).

311	283	310	294	206	193	365	420	156	246
244	328	370	311	224	273	274	353	218	214
277	185	365	260	173	474	256	286	282	222
403	252	304	264	266	290	312	239	254	219
336	264	322	178	187	305	302	253	327	258

Voor het aantal waarnemingen geldt: $n = 50$. De range of spreidingsbreedte $R = 474 - 156 = 318$. We kiezen dan het aantal klassen gelijk aan 7 ($\approx \sqrt{50}$) en de klassebreedte wordt dan $b \approx \frac{318}{7}$. We kiezen bijvoorbeeld $b = 50$. We komen dan tot de volgende klasse-indeling:

151 - 200, 201 - 250, 251 - 300, 301 - 350, 351 - 400, 401 - 450, 451 - 500.

Iedere waarneming ligt nu in precies één van deze klassen. We gaan nu tellen hoeveel waarnemingen er in iedere klasse zitten met behulp van een zogenaamde frequentietabel; het aantal waarnemingen in klasse i is f_i ($i = 1,\ldots,7$).

Tabel 2.5.2.2. Frequentietabel van het cholesterolgehalte van het bloed van hartpatiënten (The Los Angeles Heart Study, 1950).

i	klassegrenzen in mg%	klassemidden y_i in mg%	turfstaat	f_i
1	151 - 200	175.5	𝍸 /	6
2	201 - 250	225.5	𝍸 ////	9
3	251 - 300	275.5	𝍸 𝍸 𝍸 //	17
4	301 - 350	325.5	𝍸 𝍸 /	11
5	351 - 400	375.5	////	4
6	401 - 450	425.5	//	2
7	451 - 500	475.5	/	1
				n = 50

2.5.3. Berekeningen.

De formules voor de berekening van $\bar{y}$ en s^2 zijn:

$$\bar{y} = c + b \frac{\Sigma f_i y_i'}{\Sigma f_i}$$

$$KS = b^2 \{\Sigma f_i y_i'^2 - \frac{(\Sigma f_i y_i')^2}{\Sigma f_i}\}$$

$$s^2 = \frac{KS}{\Sigma f_i - 1} = \frac{KS}{n-1} .$$

In dit voorbeeld is $\Sigma f_i = n = 50$, $b = 50$ en we kiezen $c = 275.5$.

Tabel 2.5.3.1. Berekening van gemiddelde en variantie van een steekproef.

f_i	$y_i' = \frac{1}{50}(y_i - 275.5)$	$f_i y_i'$	$f_i y_i'^2$	
6	-2	-12	24	$\bar{y} = 275.5 + 50 \frac{8}{50} = 283.5$
9	-1	- 9	9	
17	0			
11	1	11	11	$KS = (50)^2\{94 - \frac{(8)^2}{50}\} = 231800$
4	2	8	16	
2	3	6	18	$s^2 = \frac{231800}{49} = 4730.6$
1	4	4	16	
50	Σ	8	94	$s = \sqrt{4730.6} = 68.8$, $\nu = 49$

De toelaatbare afronding voor $\bar{y}$ is

$$a_{max} = \frac{s}{2\sqrt{n}} = \frac{68.8}{2\sqrt{50}} = 4.9$$

zodat decimalen weggelaten kunnen worden.

En voor s is

$$a_{max} = \frac{s}{\sqrt{8\nu}} = \frac{68.8}{\sqrt{8.49}} = 3.5 .$$

We kunnen deze reeks van 50 waarnemingen dan beschrijven met

$$n = 50 \ , \quad \bar{y} = 284 \ , \quad s = 69 \ .$$

Bij het vaststellen van een klasse-indeling heeft men nog al wat vrijheid. Kiest men een andere klasse-indeling (daarbij de gegeven vuistregels in aanmerking nemend), dan vindt men ongeveer dezelfde uitkomsten.

2.6. Tabellen, grafieken en histogrammen.

2.6.1. Inleiding.

Het deel van de statistiek dat zich bezighoudt met het weergeven en samenvatten van waarnemingen heet beschrijvende statistiek. Met name leert de beschrijvende statistiek hoe grafieken en tabellen moeten worden ingericht om de waarnemingen doelmatig te kunnen visualiseren.

2.6.2. Tabellen.

Het doel van een tabel is vooral de waarnemingen samenvattend weer te geven, zodanig dat vergelijking van de gegevens naar verschillende gezichtspunten mogelijk is. Hoe de tabel moet worden ingericht hangt derhalve niet alleen af van de waarnemingen zelf, maar ook van wat men duidelijk wil maken aan de lezer. Als gevolg daarvan is het niet mogelijk eenduidige voorschriften te geven voor alle voorkomende situaties. We geven hieronder daarom puntsgewijze enige richtlijnen voor het samenstellen van tabellen.

a) Het opschrift (of onderschrift) van een tabel moet beknopt en volledig de inhoud beschrijven. Ook de bron moet worden vermeld.
b) De kop van een kolom of kolomgroep moet een beknopte omschrijving bevatten van de inhoud van de kolom of kolomgroep met vermelding der eenheden. Ditzelfde geldt uiteraard ook voor de regels van een tabel.
c) De tabel mag niet te gedetailleerd zijn.
d) Gebruik bij wat grotere tabellen kolom- en regelgroepen.
e) Als kolommen en regels genummerd zijn wordt verwijzing naar de tabel gemakkelijker.
f) Rond getallen van meer dan 4 cijfers af.
g) Bij percentages moeten steeds de aantallen worden vermeld.
h) Zorg dat te vergelijken getallen in aan elkaar grenzende velden staan.

i) De volgende notatie wordt toegepast:

. : gegeven ontbreekt
- : gegeven is exact nul
0 : gegeven is te klein om in de gegeven eenheid te worden uitgedrukt
* : gegeven is voorlopig
blanco: gegeven kan logischerwijze niet voorkomen.

Een voorbeeld is frequentietabel 2.5.2.2. (In een rapport laat men de turfstaat meestal weg.) Deze tabel geeft de waarnemingen veel overzichtelijker weer dan tabel 2.5.2.1.

We geven hieronder nog een voorbeeld van een goede manier van tabelleren.

Tabel 2.6.2.1. Luchtvlootcapaciteit van de KLM.

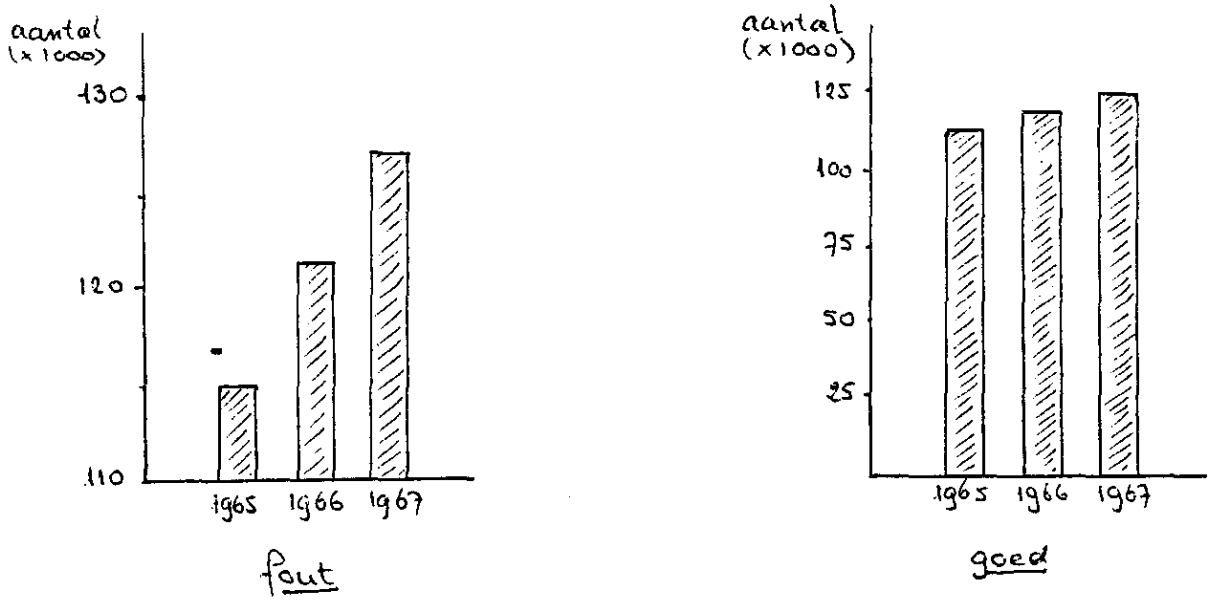
jaar	vliegtuigen in gebruik (jaargemiddelde)	gemiddeld per vliegtuig				totaal		
		aantal vlieguren	snelheid in km. per uur	laadvermogen in tonnen	aantal zitplaatsen	plaatskm	tonkm produktie	vlieguren
						mln		× 1000
1958	.	.	369	6,8	.	.	475	188
1963	60,2	2270	496	12,1	101	6472	855	142
1965	39,6	2700	567	15,1	116	6884	997	117
1966	38,7	2910	585	15,5	118	7704	1119	123
1967	39,5	2920	605	15,8	121	8490	1239	129

Bron: jaarverslag KLM.

2.6.3. Grafieken.

Een goede grafiek is veel meer nog dan een tabel het middel bij uitstek om informatie duidelijk over te brengen. Weliswaar worden details nog meer naar de achtergrond geschoven, maar daar staat tegenover dat essenties zoals orderelaties, toename en afname, maxima en minima veel duidelijker tot uitdrukking komen.
Een nadeel is dat men door ondeskundig of boosaardig gebruik van grafieken gemakkelijk een heel verkeerde indruk kan wekken, zoals onderstaand fictief voorbeeld laat zien.

Figuur 2.6.3.1. Het aantal voltooide woningen in de jaren 1965, 1966 en 1969.

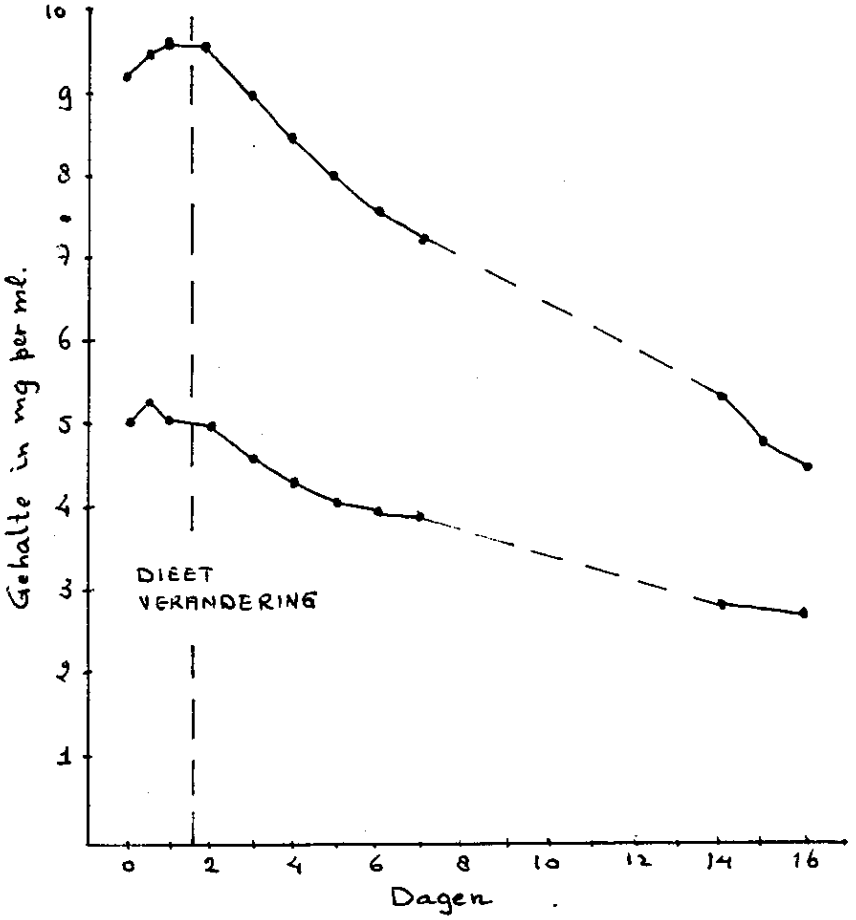


Er zijn vele soorten van grafieken, waarvan wij alleen de belangrijkste kort bespreken.

Het lijndiagram is verreweg de meest toegepaste vorm van grafisch weergeven. Het verband tussen de afhankelijke variabele (uitgezet langs de ordinaat) en de onafhankelijke variabele (uitgezet langs de abscis) wordt dan voorgesteld door een lijn.

In figuur 2.6.3.2 wordt het verloop van een grootheid in de tijd weergegeven. Elke waarneming wordt door een punt weergegeven. Door de punten te verbinden ontstaat dan een gebroken lijn, die globaal het verloop accentueert. Als in een regelmatige reeks enkele punten ontbreken is het wenselijk dit aan te geven met een stippellijn.

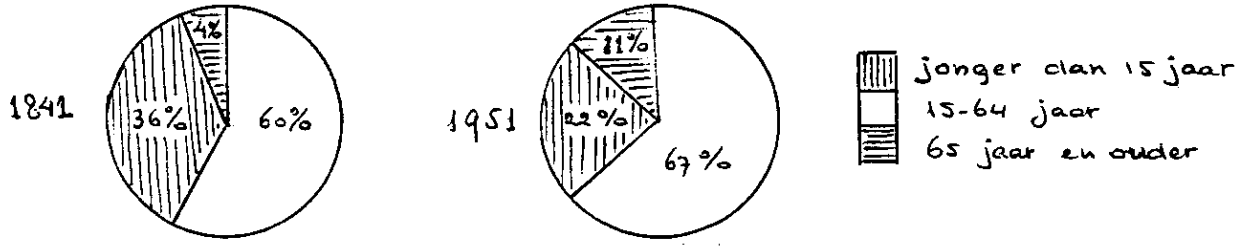
Figuur 2.6.3.2. Verloop van de serum-cholesterolgehalten bij twee mannelijke lijders aan familiaire hypercholesteraemie, vóór en na de overgang naar een vetvrij vegetarisch dieet, geheel vrij van cholesterol. Voor de dieetverandering was de cholesterolopname reeds betrekkelijk laag. Bron: Keys, Science 112 (1950), 79-81.



Eveneens veel toegepast wordt het staafdiagram. Een voorbeeld hiervan is figuur 2.6.3.1.

Om een verdeling van een specifieke populatie te illustreren wordt vaak gebruik gemaakt van een cirkeldiagram. Bij vergelijken van populaties moeten dan cirkels met gelijke stralen genomen worden. Hier volgt een voorbeeld.

Figuur 2.6.3.3. De leeftijdsverdeling van de bevolking in Engeland en Wales bij de volkstellingen van 1841 en 1951.



Bron: Registrar General's Review.

Bij beeldstatistieken wordt de eenheid (meestal betreft het hoeveelheden of aantallen) voorgesteld door een toepasselijk figuurtje. Het aantal figuurtjes geeft dan het aantal eenheden aan. Een foutieve toepassing is het, wanneer de grootte van het figuurtje een hoeveelheid voorstelt, omdat oppervlakte een tweedimensionale en hoeveelheid een eendimensionale grootheid is. Deze diagrammen zijn daarom niet aan te bevelen.
Een combinatie tenslotte van landkaart en statistisch diagram heet cartogram.

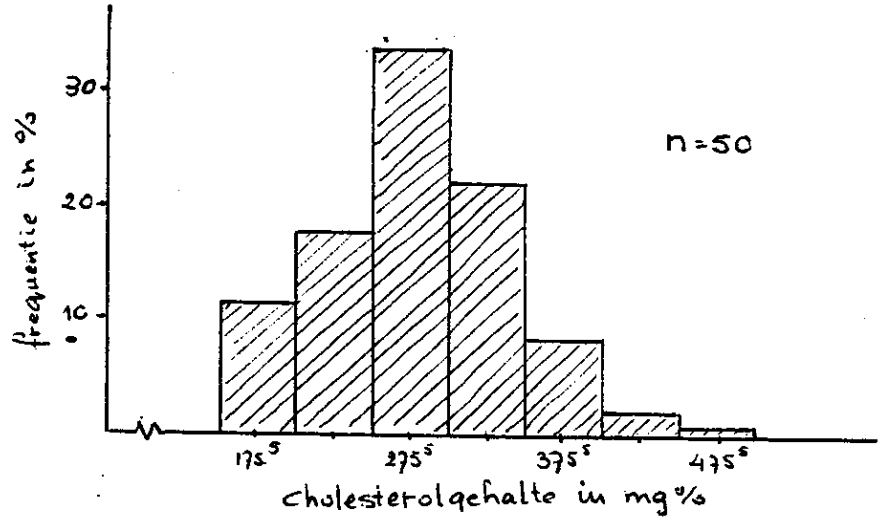
Voorbeelden van grafieken kan men in overvloed vinden in kranten, tijdschriften, jaarverslagen, rapporten van het CBS (Centraal Bureau voor de Statistiek), enz.
Ook voor het maken van grafieken gelden een aantal algemene richtlijnen:

a) Het opschrift (of onderschrift) moet kort de inhoud beschrijven en een bronvermelding bevatten.
b) Langs de assen moeten grootheden en eenheden worden vermeld.
c) De schaalverdeling moet duidelijk zijn en dus niet te gedetailleerd.
d) De verhouding van de schaaldelen van abscis en ordinaat moet zo gekozen worden dat een juiste indruk van het verschijnsel ontstaat.
e) De schaalverdeling moet in beginsel in de oorsprong beginnen. Als hiervan afgeweken wordt, dan moeten zogenaamde breuk- of scheurlijnen worden toegepast.
f) Een grafiek moet simpel en overzichtelijk zijn en mag dus niet teveel grootheden bevatten.

2.6.4. Histogrammen.

Een histogram is een staafdiagram van een frequentieverdeling. Langs de abscis worden de klassemiddens uitgezet, langs de ordinaat de frequenties (bij voorkeur in procenten).
Figuur 2.6.4.1 geeft het histogram behorend bij de frequentieverdeling van tabel 2.5.2.2.

Figuur 2.6.4.1. Histogram van het cholesterolgehalte van het bloed van hartpatiënten (bron: The Los Angeles Heart Study, 1950).

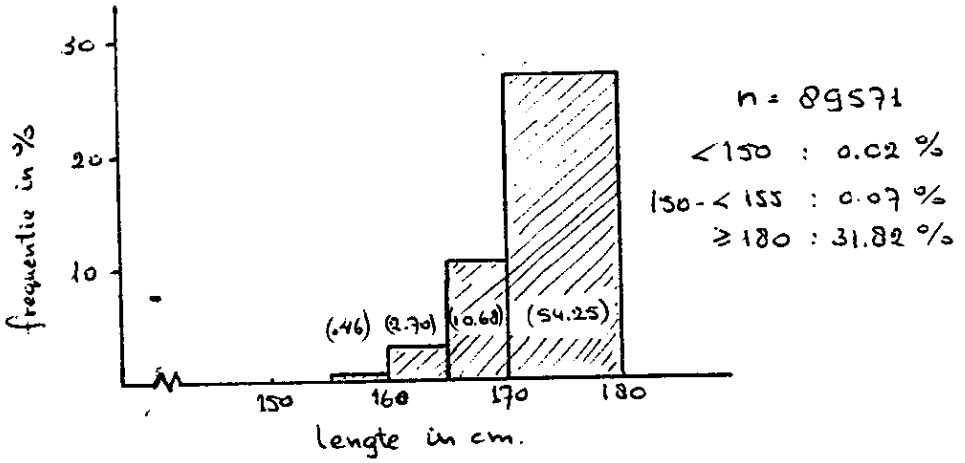


Voor de constructie van histogrammen gelden nog een aantal extra regels.

a) Vermeld steeds het totale aantal waarnemingen.
b) Ten behoeve van onderlinge vergelijkbaarheid verdient vermelding van de frequentie in % de voorkeur.
c) Het oppervlak van de kolom boven een klasse-interval is evenredig met de frequentie: dit is van belang als niet alle klassen even breed zijn (zie figuur 2.6.4.2).
d) Onvolledig gedefinieerde klassen (≥ of <) worden niet getekend; de frequenties worden apart vermeld.

We geven ter illustratie nog een voorbeeld van een histogram met zowel ongelijke klassebreedten als onvolledig gedefinieerde klassen.

Figuur 2.6.4.2. De lengte van rekruten in 1963 (bron: Statistisch Zakboek 1964).



Het tekenen van histogrammen is ontzettend nuttig. Eigenschappen van steekproeven, die bij louter rekenen onopgemerkt blijven, worden door het maken van een histogram direkt zichtbaar. Men denke maar aan o.a. tweetoppigheid, afgeknotte verdelingen (wijst op sortering).

3. Het toetsen van hypothesen en de constructie van betrouwbaarheidsintervallen.

3.1. Inleiding.

Het toetsen van hypothesen is een algemene methode in de toegepaste statistiek. Het principe luidt: men stelt een model op voor de wijze waarop men meent dat de waarnemingen tot stand zijn gekomen en men gaat na of de gerealiseerde waarnemingen op grond van dit model redelijk verklaarbaar zijn. Blijkt het dat de waarnemingen, vanuit dit model gezien, in één of meer opzichten in hoge mate onwaarschijnlijk zijn, dan wordt het model verworpen. Het is een samengestelde operatie waarbij onder meer de volgende stappen kunnen worden onderscheiden.

a) Het formuleren van de vragen waarop een antwoord wordt verwacht.
b) Het formuleren van een model met inbegrip van onderstellingen.
c) Het formuleren van nulhypothesen en alternatieve hypothesen.
d) De keuze van een toetsingsgrootheid.
e) De afleiding van de verdeling van de toetsingsgrootheid.
f) Het eventueel vervangen van de exacte verdeling van de toetsingsgrootheid door een benaderende verdeling.
g) Het beslissen hoeveel waarnemingen moeten worden verricht.
h) Beslissen of één- of tweezijdig moet worden getoetst en het kiezen van de onbetrouwbaarheidsdrempel.
i) Het uitvoeren van de waarnemingen.
j) Het berekenen van de gerealiseerde toetsingsgrootheid en nagaan of de gevonden waarde in het kritieke gebied ligt.
k) Alternatief: het berekenen van de gerealiseerde toetsingsgrootheid en van de bijbehorende éénzijdige overschrijdingskans.
l) Alternatief: het construeren van een betrouwbaarheidsinterval als de nulhypothese de waarde van één parameter betreft.
m) Het trekken van conclusies.

3.2. Het formuleren van de vragen.

De vragen waarop men op grond van bestaande of nog te verrichten waarnemingen een antwoord zoekt worden vaak onduidelijk geformuleerd. Het vergt dan een inleidende discussie teneinde tot een duidelijke statistische formulering van het probleem te komen.

3.3. Het model.

De statistische modellen die worden toegepast en die steeds één of meer stochastische variabelen bevatten zijn zeer gevarieerd en hangen af van het probleem waarom het gaat. Zelfs wanneer we met één enkele reeks waarnemingen te doen hebben zijn vele modellen mogelijk: een model geeft o.a. aan uit welke verdeling de steekproef is genomen. Daarnaast bevat een model nog andere onderstellingen, die vaak niet worden getoetst. Bij één enkele reeks waarnemingen komen de volgende onderstellingen in aanmerking.

3.3.1. Onafhankelijkheid.

De eerste onderstelling luidt: de waarnemingen vormen een aselecte steekproef van een stochastische variabele, of in andere woorden, we hebben te maken met een serie onderling onafhankelijke waarnemingen (zie Wiskunde 31-49).

Voor kleine reeksen waarnemingen is deze onderstelling niet efficiënt te toetsen.

Zijn de waarnemingen nog niet uitgevoerd dan moeten bij het verrichten van de waarnemingen zonodig maatregelen worden genomen om de onderlinge onafhankelijkheid te waarborgen.

Zijn de waarnemingen reeds uitgevoerd, dan is het zaak zorgvuldig te informeren hoe dat is geschied.

Het toepassen van statistische interpretatiemethoden op waarnemingen waarvan niet bekend is hoe zij werden verkregen kan gemakkelijk tot onjuiste conclusies leiden.

De onderstelling van een aselecte steekproef houdt tevens in dat de waarnemingen onafhankelijk moeten zijn van uitwendige storende factoren en ook hierop moet acht worden geslagen.

Soms kan men uit de volgorde waarin de waarnemingen zijn verricht of zijn genoteerd concluderen dat aan de onderstelling van onafhankelijkheid niet is voldaan. Een kritische instelling is daarom altijd gewenst.

Voorbeelden: tijdreekseffect, systematisch verloop, systematische rangschikking.

3.3.2. De vorm van de verdeling.

Een tweede onderstelling betreft meestal de vorm van de verdeling. Dit leidt tot bijvoorbeeld de volgende modellen:

$\underline{x}_i \sim N(\mu,\sigma)$ of $\underline{x}_i = \mu + \underline{u}_i\sigma$,

d.w.z. de waarnemingen zijn afkomstig uit een normale verdeling met verwachting μ en standaardafwijking σ (of variantie σ^2). Meestal gebruikt men $N(\mu,\sigma)$, sommigen gebruiken echter $N(\mu,\sigma^2)$. Voor een steekproef uit een standaard-normale of N(0,1)-verdeling gebruikt men vaak de notatie $\underline{u}_i$ $(i = 1,\ldots,n)$.

$\underline{x}_i \sim EXP(\lambda)$:

de waarnemingen zijn afkomstig uit een exponentiële verdeling met $\mathcal{E}\underline{x} = \frac{1}{\lambda}$.

$\underline{x}_i \sim f(x)$, $f(x)$ continu:

de waarnemingen zijn afkomstig uit een willekeurige continue verdeling.

$\underline{x}_i \sim HG(N,M,n)$:

de waarnemingen zijn afkomstig uit een hypergeometrische verdeling met parameters M, N en n.

$\underline{x}_i \sim BN(n,p)$:

de waarnemingen zijn afkomstig uit een binomiale verdeling met parameters n en p.

$\underline{x}_i \sim PS(\mu)$:

de waarnemingen zijn afkomstig uit een Poisson-verdeling met verwachting μ.

De vorm van de verdeling (niet de waarde van de parameters) wordt vaak zonder toetsing als juist aanvaard op grond van een algemene ervaring of van de wijze waarop de waarnemingen werden uitgevoerd.
Wij zullen evenwel later op het toetsen van dit soort onderstellingen terugkomen (hoofdstukken 5 en 7).
Hebben we met meer dan één reeks waarnemingen te maken, dan zijn gecompliceerde modellen vereist, zoals nog zal blijken.

3.4. De hypothesen.

De onderstellingen, die moeten worden getoetst om de gestelde vragen te kunnen beantwoorden, worden nulhypothesen (H_0) genoemd. Bij verwerpen van de nulhypothesen worden de daarbij behorende alternatieve hypothesen (H_1) aanvaard.

Verreweg de meeste toetsen betreffen nulhypothesen over de waarde van populatieparameters onder de onderstellingen dat de steekproef aselect is en dat de vorm van de verdeling bekend is. Men toetst dan bijv. waarden voor μ en/of σ bij een normale verdeling of van p bij een binomiale verdeling. Zeer vaak wordt een normale verdeling ondersteld. Dit kan gerechtvaardigd zijn omdat men uit ervaring met soortgelijke problemen weet dat de verdeling althans in zeer goede benadering normaal was; maar ook omdat toetsen gebaseerd op de normale verdeling veelal zg. "robust" zijn, d.w.z. dat de verdeling van de toetsingsgrootheid slechts in geringe mate beïnvloed wordt door afwijkingen van normaliteit. Een voorbeeld is de centrale limietstelling die leert dat, mits het aantal waarnemingen niet te klein is, een gemiddelde $\bar{x}$ in zeer goede benadering een normale verdeling bezit, ook al geldt dit niet voor de individuele waarnemingen (vergelijk: som van 4 worpen met een dobbelsteen).
Twijfelt men aan de juistheid van de onderstelling dat men met een normale verdeling mag rekenen, dan kan men òf de normaliteit toetsen (zie hoofdstuk 5) òf van zg. verdelingsvrije toetsen (zie hoofdstuk 12) gebruik maken, waarbij alleen wordt ondersteld dat de verdeling continu is, maar de vorm van de verdeling er niet toe doet.

3.5. De toetsingsgrootheid.

Eenzelfde hypothese kan vaak met behulp van verschillende toetsingsgrootheden worden getoetst. De keuze van de toetsingsgrootheid wordt mede bepaald door de mate waarin de toets onderscheidend is voor de alternatieve hypothesen. Dit laatste betekent, dat de kans op verwerpen van de nulhypothese (d.i. de kans dat de toetsingsgrootheid in het kritieke gebied valt) groter moet zijn voor de alternatieve hypothesen dan voor de nulhypothese, en wel des te groter naarmate de alternatieve hypothese sterker van de nulhypothese afwijkt.
In sommige gevallen bestaat er een meest onderscheidende toets, speciaal bij éénzijdige toetsing. In de praktijk wordt echter soms aan een niet-optimale toets de voorkeur gegeven terwille van de eenvoud der berekeningen.
Ook de keuze van het kritieke gebied wordt bepaald door de eis dat het onderscheidend vermogen zo groot mogelijk is.
De afleiding van de verdeling van toetsingsgrootheden onder de voorwaarde dat aan de onderstellingen en de nulhypothese is voldaan, is zuiver een kwestie van toepassing van de kansrekening.

Van de gangbare toetsingsgrootheden zijn de verdelingen (of benaderingen) bekend en getabelleerd. Bij het uitvoeren van de toets behoeven we alleen die tabellen te hanteren (Statistisch Compendium).
Soms is een wiskundige afleiding te gecompliceerd en maakt men gebruik van simulatie, d.w.z. het kunstmatig genereren van bijv. 1000 steekproeven en daarmee corresponderend 1000 realisaties van de toetsingsgrootheid, waaruit men empirisch een benadering van de verdeling vindt.

3.6. Het aantal waarnemingen.

De keuze van het aantal waarnemingen wordt vaak gedikteerd door de technische mogelijkheden en uitvoerbaarheid. In beginsel kan het aantal waarnemingen ook worden voorgeschreven aan de hand van eisen die aan het onderscheidingsvermogen van een toets worden gesteld. Algemeen neemt dit onderscheidingsvermogen toe met toenemende n. Men loopt dan echter wel het risico dat men teveel waarnemingen verricht omdat men aan het onderscheidingsvermogen te hoge eisen stelt.

3.7. Het toetsen.

Aan het toetsen gaat vooraf het beslissen of één- of tweezijdig moet worden getoetst. De keuze tussen één- en tweezijdig toetsen hangt af van de vraagstelling en van het doel van het experiment. In veel situaties is de keuze helemaal niet duidelijk. Van belang is wel dat een bepaalde keuze voldoende gemotiveerd kan worden. Deze motivatie is van niet-statistische aard.
Ook de onbetrouwbaarheidsdrempel, α, moet tevoren worden vastgesteld. De keuze is vrij willekeurig: $\alpha = 0.05$ en $\alpha = 0.01$ zijn de meest ingeburgerde waarden.
Nu kan het kritieke gebied worden bepaald en de nulhypothese wordt verworpen als de berekende toetsingsgrootheid in dit kritieke gebied ligt.
Een andere manier om te toetsen is het berekenen van de éénzijdige overschrijdingskans, P, van de gevonden waarde van de toetsingsgrootheid. De nulhypothese wordt dan verworpen wanneer

$P \leq \alpha$ bij een eenzijdige toets

of

$P \leq \frac{1}{2}\alpha$ bij een tweezijdige toets.

Op deze wijze toetsen heeft het voordeel dat de lezer zijn eigen oordeel kan vellen, als de keuze van α en tussen één- en tweezijdig wordt opengelaten.

Als het gaat om het toetsen van één parameter dan verdient, waar mogelijk, een betrouwbaarheidsinterval de voorkeur. Het toetsen van een gegeven waarde voor die parameter komt dan neer op het vaststellen of die waarde al dan niet in dit betrouwbaarheidsinterval ligt. Doch het interval geeft tevens een beeld van alle hypothesen die door de toets niet zullen worden verworpen en dus van de speelruimte in interpretatie die de beschikbare waarnemingen toelaten.
Een betrouwbaarheidsinterval wordt gedefinieerd als volgt:

Een tweezijdig betrouwbaarheidsinterval met betrouwbaarheid $(1-\alpha)$ is een interval $\theta_\ell < \theta < \theta_r$, dat al die waarden van de onbekende parameter θ bevat die bij een tweezijdige toets met onbetrouwbaarheidsdrempel α niet worden verworpen.

Uit deze definitie kan de volgende eigenschap worden afgeleid:

Een tweezijdig betrouwbaarheidsinterval met betrouwbaarheid $(1-\alpha)$ is de éénmalige realisatie van een stochastisch interval $(\underline{\theta}_\ell, \underline{\theta}_r)$, zodanig dat

$$P(\underline{\theta}_\ell > \theta) \leq \tfrac{1}{2}\alpha \quad \text{èn} \quad P(\underline{\theta}_r < \theta) \leq \tfrac{1}{2}\alpha$$

en dus

$$P(\underline{\theta}_\ell < \theta < \underline{\theta}_r) \geq (1-\alpha)$$

wanneer θ de ware waarde is van de onbekende parameter.

Een éénzijdig betrouwbaarheidsinterval wordt op analoge wijze gedefinieerd.

3.8. Het trekken van conclusies.

De toets van een hypothese en de praktische conclusies die men aan de uitslag van die toets verbindt moet men duidelijk gescheiden houden. Men kan nooit toetsen of een parameter een bepaalde waarde bezit. Men kan alleen door een toets vaststellen dat de waarnemingen niet in strijd zijn met de hypothese dat die parameter een bepaalde waarde heeft. Men kan dus eventueel die waarde als een nuttige werkhypothese aanhouden. Dit laatste is echter een afzonderlijke beslissing en niet het resultaat van de uitgevoerde toets. Wanneer nu een toets met onbetrouwbaarheid α een nulhypothese niet verwerpt mag men daaruit niet concluderen dat "de nulhypothese juist is" of dat "de nulhypothese juist is met kans $(1-\alpha)$". Dergelijke uitspraken zijn principieel fout!

4. Waarnemingen uit normale verdelingen.

4.1. Inleiding.

De normale verdeling speelt in de statistiek een bijzonder belangrijke rol. In een groot aantal praktijkgevallen wordt ondersteld dat een serie waarnemingen, x_i, $i = 1,2,\ldots,n$, kan worden opgevat als een steekproef uit een normale verdeling. We noteren dit als

$$\underline{x}_i = \mu + \underline{u}_i\sigma \ .$$

In dit model zijn μ en σ^2 de parameters van de normale verdeling; $\underline{u}_i$ is een stochastische variabele die normaal verdeeld is met $\mathcal{E}\underline{u}_i = 0$ en $\text{var}\ \underline{u}_i = 1$. (M.a.w. $\underline{u}_i$ is standaard-normaal verdeeld.)
Bij alle in dit hoofdstuk te behandelen toetsen is dit een van de onderstellingen: men neemt aan dat er aan voldaan is zonder dit te toetsen.

4.2. Eén serie waarnemingen.

Als we een aselecte steekproef nemen uit een normale verdeling kunnen we de waarnemingen, $\underline{x}_i$, $i = 1,2,\ldots,n$, beschrijven met het model:

$$\underline{x}_i = \mu + \underline{u}_i\sigma \ .$$

We kunnen nu twee situaties onderscheiden:

1) de variantie σ^2 is bekend,
2) de variantie σ^2 is niet bekend.

4.2.1. Een normale verdeling met bekende variantie.

4.2.1.1. De enige onbekende parameter is dan de verwachting μ. De nulhypothesen: $\mu = \mu_0$, $\mu \geq \mu_0$ en $\mu \leq \mu_0$ kunnen dan worden getoetst en een betrouwbaarheidsinterval voor μ kan worden geconstrueerd op grond van

$$\underline{u} = \frac{\bar{\underline{x}} - \mu}{\sigma/\sqrt{n}}$$

en tabel S.C. 1.1. Immers, als $\underline{x}$ normaal verdeeld is met parameters μ en σ^2 dan is $\bar{\underline{x}} = \frac{\Sigma\underline{x}_i}{n}$ normaal verdeeld met verwachting μ en variantie σ^2/n.

De onderstelling is: de waarnemingen vormen een aselecte steekproef uit een normale verdeling met bekende variantie.

4.2.1.2. In een laboratorium is het zwavelgehalte bepaald van steenkool volgens methode M_1. De waarnemingen zijn

3.18; 3.20; 3.22; 3.14; 3.09; 3.10; 3.10.

Na evidente codering luidt de serie:

18 20 22 14 9 10 10 (4.2.1.2.1)

waaruit volgt voor de gecodeerde waarnemingen:

$n = 7$, $\bar{x} = 14.7$.

Voor $\sigma^2 = 30$ en H_0: $\mu = 20.0$, H_1: $\mu \neq 20.0$ geldt dan

$$u = \frac{14.7 - 20.0}{\sqrt{30/7}} = -2.56 .$$

De bijbehorende tweezijdige overschrijdingskans bedraagt dan

$2 \times (1 - 0.9948) = 0.0104$

De waarnemingen zijn derhalve in strijd met de nulhypothese.

De grenzen van een tweezijdig betrouwbaarheidsinterval zijn

$$\bar{x} \pm u_{(\frac{1}{2}\alpha)} \times \sigma/\sqrt{n} ,$$

bijv.

$$14.7 \pm 1.96 \times \sqrt{\frac{30}{7}} ; \quad 1 - \alpha = 0.95$$

dus

$10.6 < \mu < 18.8$. (Wat wordt het betr. interval voor de μ van de orig. waarnemingen?)

4.2.2. Een normale verdeling met onbekende variantie.

4.2.2.1. De variantie σ^2 moet nu ook uit de waarnemingen worden geschat. Hierbij kunnen de nulhypothesen $\sigma^2 = \sigma_0^2$, $\sigma^2 \geq \sigma_0^2$, $\sigma^2 \leq \sigma_0^2$ worden opgesteld. Deze hypothesen worden getoetst op grond van

$$\frac{\nu s^2}{\sigma^2} = \frac{KS}{\sigma^2} \simeq \chi_\nu^2 .$$

Hierin is $\underline{\chi}^2_\nu$ een stochastische variabele met een verdeling (chi-kwadraatverdeling geheten) die alleen afhangt van ν (= het aantal vrijheidsgraden bij s^2). Tabel S.C. 3.1 geeft de kritieke waarden van deze verdeling.
De onderstelling luidt nu: de waarnemingen vormen een aselecte steekproef uit een normale verdeling.

4.2.2.2. Voor de serie waarnemingen (4.2.1.2.1) geldt:

$$KS = 169.4 \ , \quad s^2 = 28.2 \ , \quad s = 5.3 \ , \quad \nu = 6 \ .$$

Voor het toetsen van H_0: $\sigma^2 = 30$ tegen H_1: $\sigma^2 \neq 30$ geldt dan

$$\chi^2_6 = \frac{169.4}{30} = 5.65 \ .$$

De kritieke waarden bij $\alpha = 0.05$ (tweezijdig) zijn

$$\chi^2_6(1 - \tfrac{1}{2}\alpha) = 1.24 \quad \text{en} \quad \chi^2_6(\tfrac{1}{2}\alpha) = 14.4 \ .$$

De nulhypothese kan dus niet worden verworpen.
Een betrouwbaarheidsinterval ($\alpha = 0.05$) wordt gevonden uit

$$P(1.24 < \underline{\chi}^2_6 < 14.4) = 0.95$$

of

$$P(\frac{KS}{14.4} < \sigma^2 < \frac{KS}{1.24}) = 0.95 \ ,$$

zodat met 95% betrouwbaarheid

$$11.8 < \sigma^2 < 136.6 \quad \text{en} \quad 3.4 < \sigma < 11.7 \ .$$

Met behulp van tabel S.C. 3.2 wordt dit laatste resultaat ook gevonden, nl.

$$3.4 = 0.64 \times 5.3 < \sigma < 2.20 \times 5.3 = 11.7 \ ; \quad (1 - \alpha) = 0.95 \ .$$

4.2.2.3. Voor de <u>nulhypothesen</u>: $\mu = \mu_0$, $\mu \leq \mu_0$, $\mu \geq \mu_0$ kunnen we nu geen gebruik maken van de standaardnormale verdeling omdat de variantie niet bekend is.
Echter geldt

$$\frac{\underline{\bar{x}} - \mu}{\underline{s}/\sqrt{n}} \approx \underline{t}_\nu \ .$$

De stochastische variabele $\underline{t}_\nu$ heeft een symmetrische verdeling, die alleen

afhangt van het aantal vrijheidsgraden, ν, bij $\underline{s}^2$. De kritieke waarden van de t-verdeling (van Student) zijn te vinden in tabel S.C. 2.1.
Ook hier luidt de onderstelling dat de waarnemingen aselecte trekkingen zijn uit een normaal verdeelde populatie.

4.2.2.4. Uit tabel S.C. 2.1 blijkt

$$P(-2.45 < \underline{t}_6 < 2.45) = 0.95 \ .$$

Voor de serie waarnemingen (4.1.1.2.1) en H_0: $\mu = 20.0$ geldt

$$t_6 = \frac{14.7 - 20.0}{5.3/\sqrt{7}} = -2.64 \quad (< -2.45) \ ,$$

zodat ook nu de waarnemingen in strijd blijken met de nulhypothese.
De grenzen van een 95%-betrouwbaarheidsinterval worden nu:

$$14.7 \pm 2.45 \times 5.3/\sqrt{7},$$

zodat

$$9.8 < \mu < 19.6 \ ; \quad (1-\alpha) = 0.95 \ .$$

4.3. Twee series waarnemingen.

We beschouwen nu de situatie dat er twee series waarnemingen, $\underline{x}_{1j}$, $j = 1,2,\ldots,n_1$, en $\underline{x}_{2j}$, $j = 1,2,\ldots,n_2$, zijn, elk afkomstig uit een normale verdeling. Hierbij zijn vier modellen denkbaar:

$$\underline{x}_{ij} = \mu + \underline{u}_{ij}\sigma \ , \tag{4.3.1}$$

$$\underline{x}_{ij} = \mu_i + \underline{u}_{ij}\sigma \ , \tag{4.3.2}$$

$$\underline{x}_{ij} = \mu + \underline{u}_{ij}\sigma_i \ , \tag{4.3.3}$$

$$\underline{x}_{ij} = \mu_i + \underline{u}_{ij}\sigma_i \ . \tag{4.3.4}$$

Hierbij is steeds $i = 1,2$ en $j = 1,\ldots,n_i$.
Bij model (4.3.1) vormen de beide series samen één steekproef uit een normaal verdeelde populatie en deze situatie is in de vorige paragraaf reeds besproken.
De overige drie modellen betreffen het vergelijken van twee gemiddelden en/of twee varianties.

4.3.1. Normale verdeling met bekende varianties.

4.3.1.1. We onderstellen eerst dat $\sigma_1^2 = \sigma_2^2 = \sigma^2$. De nulhypothesen zijn dan: $\mu_1 = \mu_2$, $\mu_1 \leq \mu_2$, $\mu_1 \geq \mu_2$. Toetsen van deze hypothesen en het construeren van een betrouwbaarheidsinterval voor $\mu_1 - \mu_2$ geschiedt dan op grond van

$$\underline{u} \simeq \frac{(\bar{\underline{x}}_1 - \bar{\underline{x}}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

en tabel S.C. 1.1.

Immers is

$$\text{var}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2) = \text{var}\ \bar{\underline{x}}_1 + \text{var}\ \bar{\underline{x}}_2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) .$$

De bijbehorende onderstelling is: de waarnemingen zijn onafhankelijke, aselecte steekproeven uit normaal verdeelde populaties.

4.3.1.2. We beschouwen naast de serie waarnemingen (4.2.1.2.1) nog de volgende serie bepalingen van het zwavelgehalte, nu volgens een andere methode M_2:

3.20; 3.19; 3.18; 3.27; 3.24.

Deze serie luidt na codering

20 19 18 27 24 (4.3.1.2.1)

waaruit volgt

$$n_2 = 5 , \quad \bar{x}_2 = 21.6 .$$

Voor serie (4.2.1.2.1) was

$$n_1 = 7 , \quad \bar{x}_1 = 14.7 .$$

Voor $\sigma^2 = 30$ en H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$, geldt

$$u = \frac{14.7 - 21.6}{\sqrt{(\frac{1}{7} + \frac{1}{5}) \times 30}} = -2.15 ,$$

zodat de nulhypothese moet worden verworpen ($\alpha = 0.05$).

De 95%-betrouwbaarheidsgrenzen zijn

$$\bar{x}_1 - \bar{x}_2 \pm u_{(\frac{1}{2}\alpha)}\,\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} ,$$

dus

$$-6.9 \pm 1.96 \times \sqrt{(\tfrac{1}{7} + \tfrac{1}{5})30} \ ,$$

zodat

$$-13.2 < \mu_1 - \mu_2 < -0.6 \ ; \quad (1-\alpha) = 0.95 \ .$$

4.3.1.3. Zijn beide varianties bekend maar ongelijk, dan worden de nulhypothesen: $\mu_1 = \mu_2$, $\mu_1 \leq \mu_2$, $\mu_1 \geq \mu_2$ getoetst op grond van

$$\underline{u} \approx \frac{(\bar{\underline{x}}_1 - \bar{\underline{x}}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \ .$$

Nu geldt immers

$$\text{var}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2) = \text{var}\ \bar{\underline{x}}_1 + \text{var}\ \bar{\underline{x}}_2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \ .$$

De onderstelling is: de waarnemingen zijn onafhankelijke aselecte steekproeven uit normaal verdeelde populaties.

4.3.1.4. Voor $\sigma_1^2 = 30$, $\sigma_2^2 = 20$ en H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$, wordt

$$u = \frac{14.7 - 21.6}{\sqrt{\frac{30}{7} + \frac{20}{5}}} = -2.40$$

zodat H_0 moet worden verworpen ($\alpha = 0.05$).
De grenzen van het betrouwbaarheidsinterval zijn

$$-6.9 \pm 1.96 \sqrt{\frac{30}{7} + \frac{20}{5}}$$

dus

$$-12.5 < \mu_1 - \mu_2 < -1.3 \ ; \quad (1-\alpha) = 0.95 \ .$$

4.3.2. Normale verdelingen met onbekende varianties.

4.3.2.1. Omdat σ_1^2 en σ_2^2 niet bekend zijn moeten ze worden geschat uit de waarnemingen. De eerste vraag die zich dan voordoet is: zijn beide varianties gelijk? Het toetsen van de nulhypothesen: $\sigma_1^2 = \sigma_2^2$, $\sigma_1^2 \leq \sigma_2^2$, $\sigma_1^2 \geq \sigma_2^2$ en de constructie van betrouwbaarheidsintervallen voor $\frac{\sigma_1^2}{\sigma_2^2}$ zijn gebaseerd op:

$$\frac{\underline{s}_1^2}{\sigma_1^2} \approx \frac{\underline{\chi}^2_{\nu_1}}{\nu_1}, \quad \frac{\underline{s}_2^2}{\sigma_2^2} \approx \frac{\underline{\chi}^2_{\nu_2}}{\nu_2} \quad \text{en} \quad \frac{\underline{s}_1^2}{\sigma_1^2} : \frac{\underline{s}_2^2}{\sigma_2^2} \approx \frac{\underline{\chi}^2_{\nu_1}}{\nu_1} : \frac{\underline{\chi}^2_{\nu_2}}{\nu_2} \approx \underline{F}^{\nu_1}_{\nu_2} .$$

$\underline{F}^{\nu_1}_{\nu_2}$ is het quotiënt van twee onafhankelijke stochastische variabelen, beide met een $\underline{\chi}^2_\nu/\nu$-verdeling. De bovenindex, ν_1, hoort steeds bij de teller, de onderindex, ν_2, bij de noemer.

Uit de definitie van de F-verdeling volgt direkt

$$F^{\nu_1}_{\nu_2}(\alpha) = \frac{1}{F^{\nu_2}_{\nu_1}(1-\alpha)} .$$

Voor $\alpha = 0.05$, $\alpha = 0.025$, $\alpha = 0.01$ en $\alpha = 0.005$ zijn rechts-éénzijdige kritieke waarden van de F-verdeling getabelleerd in S.C. 4.1 t/m 4.4.
Zo is bijv.

$$P(\underline{F}^4_6 < 4.53) = 0.95 ,$$

$$P(\underline{F}^4_6 < \frac{1}{6.16}) = 0.05 , \quad \text{(zie tabel S.C. 4.1)}$$

$$P(\frac{1}{9.20} < \underline{F}^4_6 < 6.23) = 0.95 \quad \text{(zie tabel S.C. 4.2).}$$

<u>Belangrijk is</u>:

H_0: $\sigma_1^2 \le \sigma_2^2$ wordt verworpen als $\frac{s_1^2}{s_2^2} > F^{\nu_1}_{\nu_2}(\alpha)$;

H_0: $\sigma_1^2 \ge \sigma_2^2$ wordt verworpen als $\frac{s_2^2}{s_1^2} > F^{\nu_2}_{\nu_1}(\alpha)$;

H_0: $\sigma_1^2 = \sigma_2^2$ wordt verworpen als $\frac{s_1^2}{s_2^2} > F^{\nu_1}_{\nu_2}(\frac{1}{2}\alpha)$,

of als $\frac{s_2^2}{s_1^2} > F^{\nu_2}_{\nu_1}(\frac{1}{2}\alpha)$.

Een $100(1-\alpha)\%$-betrouwbaarheidsinterval voor $\frac{\sigma_1^2}{\sigma_2^2}$ is

$$\frac{1}{F^{\nu_1}_{\nu_2}(\frac{1}{2}\alpha)} \times \frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < F^{\nu_2}_{\nu_1}(\frac{1}{2}\alpha) \times \frac{s_1^2}{s_2^2} .$$

Bij toepassing van deze formules moet erom worden gedacht dat de waarde van α vermeld boven de tabellen S.C. 4.1 t/m 4.4 de éénzijdige onbetrouwbaarheid is. $\frac{1}{2}\alpha$ in bovenstaande formules moet dus gelijk worden gesteld aan α boven de tabellen!
De onderstelling bij deze toetsen is: de waarnemingen zijn onafhankelijke, aselecte steekproeven uit normaal verdeelde populaties.

4.3.2.2. Voor de beide series waarnemingen (4.2.1.2.1) en (4.3.1.2.1) geldt:

$$KS_1 = 169.4 \ , \quad s_1^2 = 28.2 \ , \quad s_1 = 5.3 \ , \quad \nu_1 = 6$$

resp.

$$KS_2 = 57.2 \ , \quad s_2^2 = 14.3 \ , \quad s_2 = 3.8 \ , \quad \nu_2 = 4 \ .$$

Toepassing van het bovenstaande op deze waarnemingen geeft H_0: $\sigma_1^2 = \sigma_2^2$ wordt niet verworpen bij $\alpha = 0.05$, want

$$\frac{s_1^2}{s_2^2} = \frac{28.2}{14.3} = 1.97 < F_4^6(\tfrac{1}{2}\alpha = 0.025) = 9.20 \quad \text{en} \quad \frac{s_2^2}{s_1^2} < 1$$

(zie tabel S.C. 4.2).
Vanwege de inrichting der F-tabellen (alleen F-waarden groter dan 1 komen hierin voor) onderzoekt men alleen quotiënten met de grootste variantieschatting in de teller.
De modellen (4.3.3) en (4.3.4) kunnen dus buiten beschouwing worden gelaten.
Opgemerkt moet worden dat in vele praktijkgevallen de hypothese H_0: $\sigma_1^2 = \sigma_2^2$ als onderstelling wordt aanvaard zonder haar te toetsen (zie 4.3.2.3). Dit voorbeeld laat zien dat bij kleine series waarnemingen de toets alleen tot verwerping leidt als de verhouding s_1^2/s_2^2 of s_2^2/s_1^2 een zeer hoge waarde heeft. Uit de waarnemingen zelf kan men al vaak zien dat dit er niet in zit.
Het 95%-betrouwbaarheidsinterval luidt

$$\frac{1}{9.20} \times \frac{28.2}{14.3} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{28.2}{14.3} \times 6.23$$

of

$$0.22 < \frac{\sigma_1^2}{\sigma_2^2} < 12.3 \; ; \quad (1-\alpha) = 0.95 \; .$$

4.3.2.3. We beschouwen het toetsen van de nulhypothesen: $\mu_1 = \mu_2$, $\mu_1 \leq \mu_2$, $\mu_1 \geq \mu_2$ onder de onderstelling dat $\sigma_1^2 = \sigma_2^2$. Nu zijn s_1^2 en s_2^2 onafhankelijke schattingen van eenzelfde σ^2 en zij kunnen worden gecombineerd tot één schatting:

$$s^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2} = \frac{KS_1 + KS_2}{\nu_1 + \nu_2}$$

met $(\nu_1 + \nu_2)$ vrijheidsgraden.
Dit combineren van meerdere onafhankelijke schattingen voor σ^2 tot één schatting noemt men wel "poolen". Zo'n "gepoolde" schatting is gelijkwaardig met een s^2 die bepaald is uit één enkele reeks waarnemingen met hetzelfde aantal vrijheidsgraden.
De toets van de nulhypothesen berust nu op

$$\underline{t}_\nu = \frac{(\bar{\underline{x}}_1 - \bar{\underline{x}}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \; ,$$

waarbij

$$\nu = \nu_1 + \nu_2 = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2 \; .$$

De bij deze toets behorende onderstelling luidt: de waarnemingen zijn onafhankelijke aselecte steekproeven uit normale verdelingen met gelijke varianties.

4.3.2.4. Als we aannemen dat de zwavelgehaltebepaling bij beide methoden M_1 en M_2 dezelfde variantie σ^2 heeft, dan is een schatting van deze σ^2

$$s^2 = \frac{KS_1 + KS_2}{\nu_1 + \nu_2} = \frac{169.4 + 57.2}{6 + 4} = 22.7 \; ,$$

zodat $s = 4.76$, terwijl $\nu = 6 + 4 = 10$.
Deze aanname zullen we in het algemeen zonder meer aanvaarden. In 4.3.2.2 is overigens gebleken dat deze aanname gerechtvaardigd is.
H_0: $\mu_1 = \mu_2$ wordt nu getoetst tegen H_1: $\mu_1 \neq \mu_2$ met

$$t_{10} = \frac{14.7 - 21.6}{4.76\sqrt{\frac{1}{7} + \frac{1}{5}}} = -2.48 \; .$$

$t_{10}(\alpha = 0.05$, tweezijdig) = 2.23, dus de gevonden waarde $t_{10} = -2.48$ is net significant hetgeen twijfel aan de juistheid van de nulhypothese rechtvaardigt.
De grenzen van een 95%-betrouwbaarheidsinterval zijn

$$-6.9 \pm 2.23 \times 4.76\sqrt{\frac{1}{7} + \frac{1}{5}}$$

zodat

$$-13.1 < \mu_1 - \mu_2 < -0.7 \; ; \quad (1 - \alpha) = 0.95 \; .$$

4.3.2.5. Hoe groter het aantal vrijheidsgraden, ν, is, des te kleiner is de waarde van t_ν. $\lim_{\nu \to \infty} \underline{t}_\nu = \underline{u}$, en, zoals blijkt uit tabel S.C. 2.1, wijkt t voor $\nu > 30$ nog maar weinig af van u. In de praktijk vervangt men voor $\nu > 30$ de t-verdeling door de u-verdeling.
Voor het toetsen van H_0: $\mu_1 = \mu_2$ als $\sigma_1^2 \neq \sigma_2^2$ (beide varianties onbekend) is een toets gebaseerd op

$$\underline{u} \approx \frac{(\underline{\bar{x}}_1 - \underline{\bar{x}}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\underline{s}_1^2}{n_1} + \frac{\underline{s}_2^2}{n_2}}}$$

een voor de hand liggende en goede methode, mits de series bestaan uit 30 of meer waarnemingen. Voor kleinere series levert dit geval theoretische moeilijkheden op waarvoor geen algemeen aanvaarde oplossing bestaat. Dit geval komt in de praktijk echter weinig voor en we laten het hier buiten beschouwing.

4.3.3. Paren waarnemingen.

Een steekproef bestaande uit een reeks van paren waarnemingen wordt beschreven met het model:

$$\underline{x}_{ij} = \mu_{ij} + \underline{u}_{ij}\sigma_i \; , \quad i = 1,2,; \; j = 1,\ldots,n \; .$$

Hierin zijn de $\underline{u}_{ij}$ onderling onafhankelijk N(0,1) verdeeld. We zijn geinteresseerd in de nulhypothesen $H_0 : \mu_{1j} = \mu_{2j}$, $\mu_{1j} \leq \mu_{2j}$, $\mu_{1j} \geq \mu_{2j}$, $j = 1,\ldots,n$.

Zij $\underline{d}_j = \underline{x}_{1j} - \underline{x}_{2j}$, $j = 1,\ldots,n$. Onder H_0 geldt dan $\underline{d}_i \sim N(0, \sigma_1^2 + \sigma_2^2)$ en H_0 kan worden vervangen door de equivalente nulhypothesen
$H_0 : \xi_{\underline{d}} = 0$, $\xi_{\underline{d}} \leq 0$, $\xi_{\underline{d}} \geq 0$.
Als σ_1^2 en σ_2^2 bekend zijn en dus ook $\sigma_1^2 + \sigma_2^2$ kan een u-toets worden toegepast. Meestal zal dit niet het geval zijn en moet var $\underline{d}$ worden geschat door:

$$s_{\underline{d}}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (d_j - \bar{d})^2 \ , \quad \nu = n-1 \text{ en}$$

dan is de toetsingsgrootheid

$$\frac{\bar{d}}{s_{\underline{d}} / \sqrt{n}} \simeq \underline{t}_{n-1} \ .$$

3.3.1. Van een groot aantal hartpatiënten is zowel in 1950 als in 1962 o.a. de (systolische) bloeddruk bepaald. In onderstaande tabel zijn van 10 van deze patiënten de resultaten gegeven.

Tabel 4.3.3.1.1. De systolische bloeddruk (mg Hg) van 10 hartpatiënten in 1950 en in 1962.

patiënt nr.	1950	1962	verschil d_j
1	115	144	-29
2	128	148	-20
3	110	166	-56
4	120	108	12
5	112	124	-12
6	110	126	-16
7	134	168	-34
8	120	136	-16
9	130	118	12
10	120	118	2

Bron: The Los Angeles Heart Study.

Voor de verschillen d_j, $j = 1,\ldots,10$, geldt dan

$$\Sigma d_j = -157 \qquad \Sigma d_j^2 = 6481 \qquad KS_{\underline{d}} = 4016.1$$

$$\bar{d} = -15.7 \qquad s_{\underline{d}}^2 = 446.23 \qquad S_{\underline{d}} = 21.1 \qquad \nu = 9 \ .$$

De toetsingsgrootheid wordt voor H_0: $\mathcal{E}\underline{d} = 0$:

$$t_9 = \frac{-15.7}{21.1/\sqrt{10}} = -2.35 \ .$$

Uit tabel S.C. 2.1 volgt

$$P\{-2.26 < \underline{t}_9 < 2.26\} = 0.95 \ .$$

De gevonden waarde $t_9 = -2.35$ is dus juist significant ($\alpha = 0.05$).

Een betrouwbaarheidsinterval voor $\xi_{\underline{d}}$ wordt gevonden uit

$$\bar{d} \pm t_9(\alpha = 0.025)\ s_{\underline{d}}/\sqrt{n} \ ,$$

zodat

$$-30.8 < \xi_{\underline{d}} < -0.6 \ ; \quad (1-\alpha) = 0.95 \ .$$

4.4. De χ^2-, F- en t-verdeling.

In dit hoofdstuk zijn achtereenvolgens de χ^2-verdeling, de t-verdeling en de F-verdeling geïntroduceerd.

4.4.1. De $\underline{\chi}^2_\nu$-stochastiek is per definitie isomoor met (d.w.z. bezit dezelfde verdeling als) de som der kwadraten van ν onafhankelijke standaardnormaal verdeelde stochastieken $\underline{u}_i$, dus

$$\underline{\chi}^2_\nu \simeq \sum_{i=1}^{\nu} \underline{u}_i^2 \ .$$

Uit deze definitie volgt dat

$$\xi\underline{\chi}^2_\nu = \sum_{i=1}^{\nu} \xi\underline{u}_i^2 = \nu$$

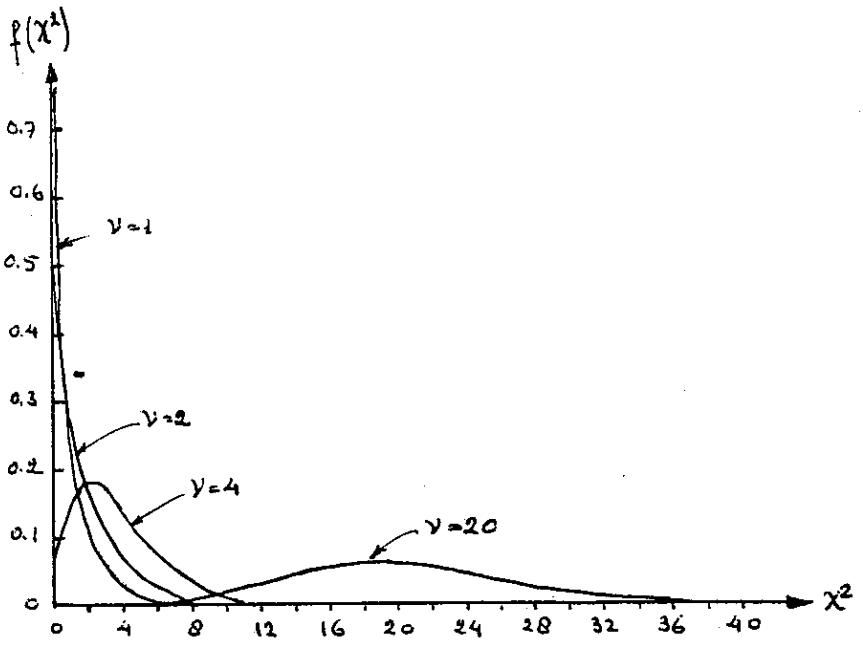
en

$$\text{var}\ \underline{\chi}^2_\nu = \sum_{i=1}^{\nu} \text{var}(\underline{u}_i^2) = 2\nu \ ,$$

want

$$\text{var}\ \underline{u}_i^2 = \xi\underline{u}^4 - (\xi\underline{u}^2)^2 = 3 - 1 = 2 \ .$$

Figuur 4.4.1.1. De χ^2_ν-verdeling.



Bron: K.A. Brownlee, Statistical Theory and Methodology in Science and Engineering, pag. 149. New York; Wiley, 1960.

Voor grote waarden van ν is de $\underline{\chi}^2_\nu$-verdeling praktisch normaal.
Een nog betere benadering is

$$\underline{\chi}_\nu \approx \sqrt{\nu} + \underline{u}\sqrt{\tfrac{1}{2}} \ .$$

Als $\underline{x}_i$ normaal verdeeld is, dan is

$$\frac{\underline{x}_i - \mu}{\sigma} \approx \underline{u}_i$$

zodat

$$\sum_{i=1}^{n} \left(\frac{\underline{x}_i - \mu}{\sigma}\right)^2 \approx \underline{\chi}^2_n \ .$$

Als we in deze formule μ vervangen door $\bar{\underline{x}}$ dan kan worden bewezen dat

$$\sum_{i=1}^{n} \left(\frac{\underline{x}_i - \bar{\underline{x}}}{\sigma}\right)^2 = \frac{\underline{KS}}{\sigma^2} \approx \underline{\chi}^2_\nu \ , \quad \nu = n-1 \ .$$

Dit bewijs valt buiten het bestek van dit college.
De $\underline{\chi}^2$-verdeling kent vele toepassingen. Een aantal daarvan worden later in hoofdstuk 7 besproken.

4.4.2. De $\underline{t}_\nu$-stochastiek is per definitie isomoor met het quotiënt van 2 onafhankelijke stochastieken. De teller bezit een u-verdeling, de noemer is isomoor met de verdeling van $\sqrt{\chi^2_\nu/\nu}$. Dus

$$\underline{t}_\nu \simeq \frac{\underline{u}}{\sqrt{\chi^2_\nu/\nu}} .$$

Voor een normale verdeling geldt dat $\bar{\underline{x}}$ en $\underline{s}$ onderling onafhankelijk zijn (dit geldt alleen voor de normale verdeling). Dus is

$$\frac{\bar{\underline{x}} - \mu}{\underline{s}/\sqrt{n}} = \frac{\bar{\underline{x}} - \mu}{\sigma/\sqrt{n}} : \frac{\underline{s}}{\sigma} \simeq \underline{u} : \sqrt{\frac{\underline{s}^2}{\sigma^2}} \simeq \underline{u} : \sqrt{\chi^2_\nu/\nu} \simeq \underline{t}_\nu .$$

Deze t-verdeling heeft, evenals de χ^2_ν-verdeling, maar één parameter en deze is ν. Het doet er voor de berekening van de toetsingsgrootheid dan ook niet toe of s bepaald is uit dezelfde serie als $\bar{x}$ of verkregen is door het samenvoegen van meerdere schattingen van σ^2.
Zoals in 4.3.2.5 al is opgemerkt kan de t_ν-verdeling voor $\nu > 30$ zonder bezwaar benaderd worden door de u-verdeling.

4.4.3. De $\underline{F}^{\nu_1}_{\nu_2}$-stochastiek tenslotte is isomoor met het quotiënt van 2 onafhankelijke χ^2_ν/ν-stochastieken met ν_1 en ν_2 vrijheidsgraden, dus

$$\underline{F}^{\nu_1}_{\nu_2} \simeq \frac{\underline{\chi}^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2} .$$

Behalve bij het vergelijken van varianties wordt de F-toets toegepast bij de variantie-analyse. Deze variantie-analyse is een methode om de nulhypothese H_0: $\mu_i = \mu$, $i = 1,\ldots,k$ $(k > 2)$, te toetsen onder de onderstelling $\sigma_i = \sigma$.
In hoofdstuk 9 komt de variantie-analyse nog ter sprake.

4.4.4. Onderlinge samenhang tussen de besproken verdelingen is gemakkelijk af te leiden uit hun definities. We hebben reeds gezien dat

$$\underline{t}_\infty \simeq \underline{u}$$

en

$$\underline{F}^{\nu_1}_{\nu_2} \simeq \frac{1}{\underline{F}_{\nu_1}^{\nu_2}} .$$

Uit

$$\underline{\chi}_1^2 \simeq \underline{u}^2$$

volgt

$$\underline{u} \simeq \underline{\chi}_1 .$$

Een volgende relatie is

$$\underline{t}_\nu^2 \simeq \left\{\frac{\underline{u}}{\sqrt{\chi_\nu^2/\nu}}\right\}^2 \simeq \frac{\underline{u}^2}{\chi_\nu^2/\nu} \simeq \frac{\chi_1^2}{\chi_\nu^2/\nu} \simeq \underline{F}_\nu^1 .$$

Omdat $\lim_{\nu\to\infty} \frac{\chi_\nu^2}{\nu} = 1$ geldt

$$\underline{F}_\infty^\nu \simeq \chi_\nu^2/\nu .$$

5. Het toetsen van normaliteit.

Zoals al eerder is opgemerkt wordt bij het toetsen van hypothesen vaak aangenomen dat de steekproef afkomstig is uit een normaal verdeelde populatie. Er bestaan nu verschillende methoden om deze onderstelling te toetsen. Hiervan zullen wij er in dit hoofdstuk twee behandelen. De eerste methode is een grafische methode. Deze levert geen objectief criterium in de vorm van een overschrijdingskans maar is vaak nuttig voor een globaal antwoord op de vraag of we althans in benadering met een normale verdeling te doen hebben. De tweede toets, die ontwikkeld is door Shapiro en Wilk geeft wel een objectief criterium.
In een hoofdstuk gewijd aan de χ^2-verdeling wordt een algemene toets behandeld. Hiermee kan men nagaan of een steekproef afkomstig is uit een willekeurige gespecificeerde verdeling.

5.1. De grafische methode.

5.1.1. Als de steekproef bestaat uit een groot aantal waarnamingen maken we eerst een klasse-indeling. We kunnen dan een cumulatieve frequentieverdeling berekenen (in %) en bij deze percentages de bijbehorende u-waarden opzoeken in tabel S.C. 1.1. Deze u-waarden zetten we vervolgens uit tegen de klassegrenzen. Als de steekproef afkomstig is uit een normale verdeling moeten de punten ongeveer op een rechte lijn liggen.
Voor deze grafiek maakt men vaak gebruik van zogenaamd (lineair) waarschijnlijkheidspapier.
De cumulatieve frequentieverdeling kan men ook maken voor de klassemiddens: men heeft dan één punt meer op de grafiek en de behandeling is meer symmetrisch.
Voor de cumulatieve frequentie, F'_k, behorende bij het k-de klassemidden kiezen we:

$$F'_k = \sum_{i=1}^{k-1} f_i + \tfrac{1}{2} f_k \ .$$

In de tabellen 5.1.1.1 en 5.1.1.2 en de figuur 5.1.1.1 worden deze grafische methoden geïllustreerd voor de gegevens van tabel 2.5.2.2.

Tabel 5.1.1.1. Toets op normaliteit voor de gegevens van tabel 2.5.2.2. Methode met klassegrenzen.

klasse	f	F	F in %	u	klassegrens
151-200	6	6	12	-1.18	200.5
201-250	9	15	30	-0.52	250.5
251-300	17	32	64	0.36	300.5
301-350	11	43	86	1.08	350.5
351-400	4	47	94	1.56	400.5
401-450	2	49	98	2.05	450.5
451-500	1	50	100		500.5

F = cumulatieve frequentie.

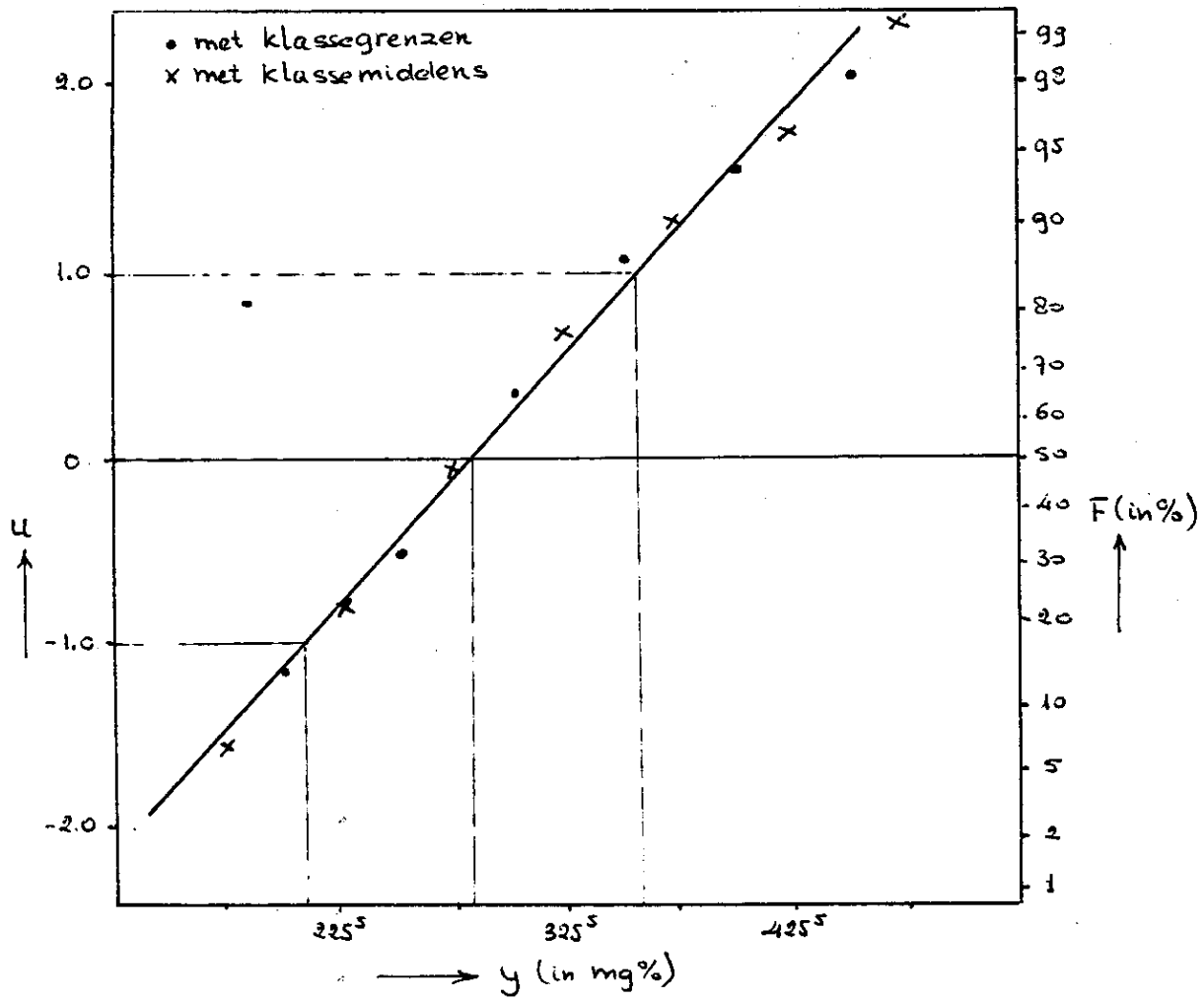
Tabel 5.1.1.2. Toets op normaliteit voor de gegevens van tabel 2.5.2.2. Methode met klassemiddens.

klasse	f	2F'	F' in %	u	klassemidden
151-200	6	6	6	-1.56	175.5
201-250	9	21	21	-0.81	225.5
251-300	17	47	47	-0.08	275.5
301-350	11	75	75	0.67	325.5
351-400	4	90	90	1.28	375.5
401-450	2	96	96	1.75	425.5
451-500	1	99	99	2.33	475.5
		100	100		

F' = cumulatieve frequentie.

Uit de getrokken rechte kan men een ruwe schatting afleiden voor $\bar{y}$ en s. Immers, $\bar{y}$ correspondeert met u = 0 en 2s is het verschil tussen de punten die corresponderen met u = 1 en u = -1. We vinden dan $\bar{y}$ = 284 en s = 75 wat aardig overeenkomt met de vroeger berekende waarden $\bar{y}$ = 284 en s = 69.

Figuur 5.1.1.1. De gegevens van de tabellen 5.1.1.1 en 5.1.1.2, uitgezet op lineair waarschijnlijkheidspapier.



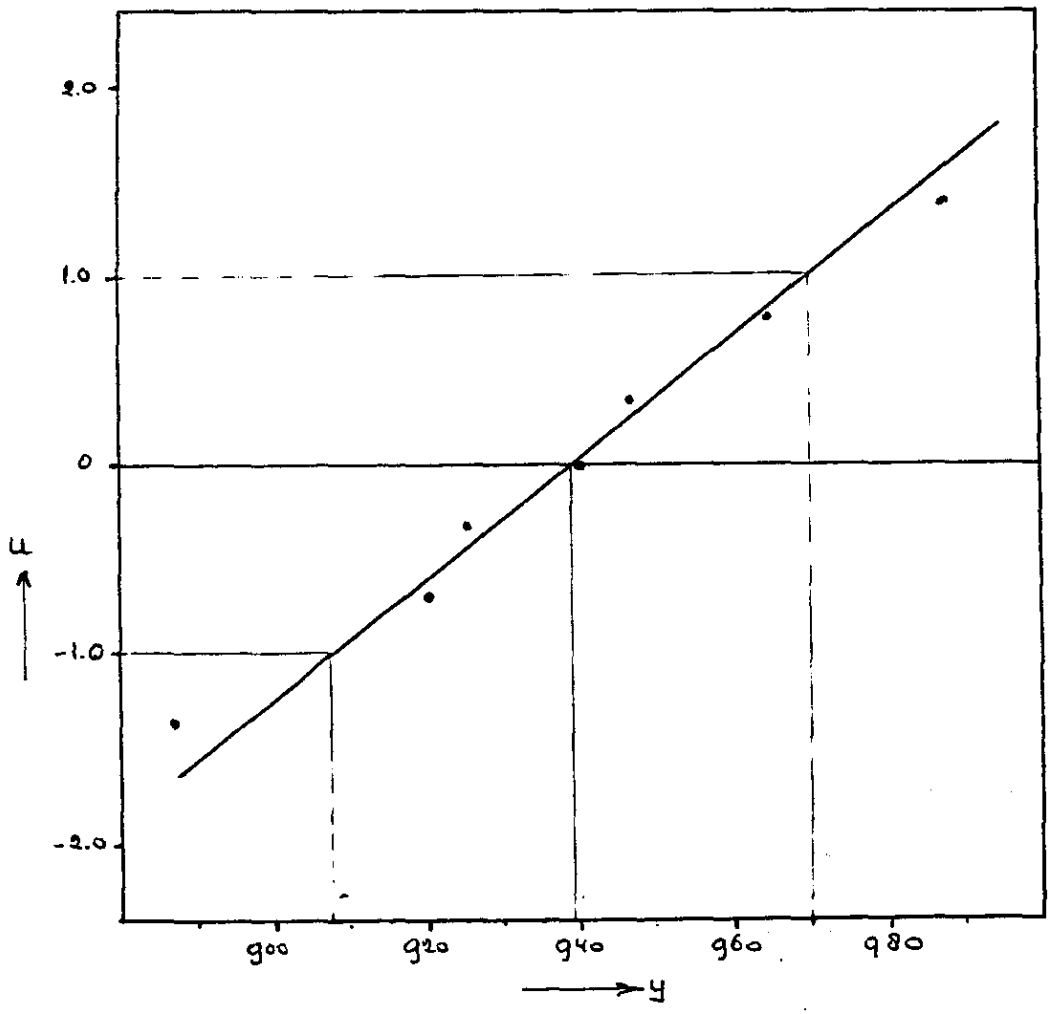
5.1.2. Als de steekproef bestaat uit een klein aantal waarnemingen zetten we $y_{(i)}$ uit tegen $\mathcal{E}(\underline{u}_{(i)})$. $\mathcal{E}(\underline{u}_{(i)})$ is getabelleerd voor verschillende steekproefgrootten, maar indien geen tabel beschikbaar is kan men $\mathcal{E}(\underline{u}_{(i)})$ heel goed benaderen met $u^*_{(i)}$ = de waarde van u waarvoor $\Phi(\underline{u}^*_{(i)}) = \frac{i - 3/8}{n + 1/4}$ (hierin is n de steekproefgrootte).

Als voorbeeld nemen we de waarnemingen van tabel 2.4.2.1. Onderstaande tabel geeft de berekeningen die nodig zijn voor het tekenen van de grafiek in figuur 5.1.2.1.

Tabel 5.1.2.1. Toets op normaliteit voor de gegevens van tabel 2.4.2.1.

y_i	$y_{(i)}$	$E(\underline{u}_{(i)})$	$\frac{i - 3/8}{n + 1/4}$	$u^*_{(i)}$
887	887	-1.35	0.086	-1.37
964	920	-0.76	0.224	-0.76
939	925	-0.35	0.362	-0.35
987	939	0	0.500	0
925	946	0.35	0.638	0.35
946	964	0.76	0.776	0.76
920	987	+1.35	0.914	1.37

Figuur 5.1.2.1. De gegevens van tabel 5.1.2.1 uitgezet op waarschijnlijkheidspapier.



Uit figuur 5.1.2.1 vinden we $\bar{y} = 938$ en $s = \frac{1}{2}(970 - 907) = 31\frac{1}{2}$ wat klopt met de berekende waarden ($\bar{y} = 938$, $s = 32$).

5.2. De toets van Shapiro en Wilk.

Shapiro en Wilk ontwikkelden voor steekproefgrootten van n = 3(1)20 de toetsingsgrootheid

$$\underline{W}_n = \frac{\left\{\sum_{i=1}^{n} a_{i,n}\, \underline{Y}_{(i)}\right\}^2}{\underline{KS}} .$$

De coëfficiënten $a_{i,n}$ zijn voor deze steekproefgrootten getabelleerd (tabel (S.C. 8.2). De kritieke waarden van $\underline{W}_n$ zijn gevonden met behulp van simulatie. De kritieke zône is linkseenzijdig.
Door transformatie vindt men

$$\underline{G} = \gamma + \delta\ {}^{10}\log \frac{\underline{W}_n - \varepsilon}{1 - \underline{W}_n} .$$

$\underline{G}$ heeft in goede benadering een standaardnormale verdeling. Ook nu is de kritieke zône linkseenzijdig.
Bij grotere steekproeven kan men de waarnemingen in K groepen verdelen ($3 \le n_k \le 20$, $k = 1,\ldots,K$) en de waarden $\underline{G}_k$, $k = 1,\ldots,K$, uit de verschillende reeksen combineren tot

$$\underline{G}_{tot} = \frac{\sum_{k=1}^{K} \underline{G}_k}{\sqrt{K}} ,$$

eveneens met een standaardnormale verdeling.
De waarden γ, δ en ε hangen af van de steekproefgrootte (zie tabel S.C. 8.2)
Passen we de toets van Shapiro en Wilk toe op de steekproef van tabel 2.4.2.1 dan vinden we

$$\sum_{i=1}^{7} a_{i,7}\, y_{(i)} = 78.6085 , \quad KS = 6235$$

zodat

$$W_7 = 0.9911 .$$

Uit tabel 5.2.2 blijkt dat de linkeroverschrijdingskans bij deze waarde groter is dan 0.50. Dit betekent dat de hypothese H_0: de steekproef is afkomstig uit een normale verdeling niet verworpen kan worden. De grafische toets (fig. 5.1.2.1) leidde eveneens tot deze conclusie.

6. De binomiale, de hypergeometrische en de Poisson verdeling.

6.1. Binomiale verdeling, normale en Poisson benadering.

In Wiskunde 31/49 hebben we reeds met de binomiale verdeling en met mogelijke benaderingen kennisgemaakt. We gebruiken hier de notatie

$$\underline{x} \sim BN(n,p) \ ,$$

waarbij n bekend is en p de parameter is die moet worden geschat of getoetst. Er geldt

$$\mathcal{E}\underline{x} = np \ , \quad \text{var } \underline{x} = np(1-p) = npq \ . \tag{6.1.1}$$

Voor de schatter $\hat{\underline{p}} = \underline{x}/n$ geldt

$$\mathcal{E}\hat{\underline{p}} = p \ , \quad \text{var } \hat{\underline{p}} = \frac{pq}{n} \ . \tag{6.1.2}$$

We veronderstellen verder $p \leq \frac{1}{2}$. Is $p > \frac{1}{2}$ dan beschouwen we $q = 1-p$ als de parameter en $n-x$ als de waarneming.
In het algemeen verdient het aanbeveling geen benadering te gebruiken als exacte tabellen beschikbaar zijn.
Benadering door de Poisson verdeling is bruikbaar als $p \leq 0{,}1$. Als $np \geq 5$ kan de normale benadering worden toegepast.

6.2. Toetsen van H_0: $p = p_0$, $p \geq p_0$, $p \leq p_0$, en een betrouwbaarheidsinterval voor p.

Exacte toetsen en betrouwbaarheidsintervallen zijn in Wiskunde 31/49 behandeld.
De normale benadering berust op

$$\underline{k} = \frac{\underline{x} - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\hat{\underline{p}} - p_0}{\sqrt{p_0q_0/n}} \ , \tag{6.2.1}$$

waarbij $\underline{k}$ bij benadering N(0,1) verdeeld is.

H_0: $p = p_0$	wordt verworpen wanneer $\lvert k \rvert \geq u_{\alpha/2}$
$p \geq p_0$	wordt verworpen wanneer $k \leq -u_{\alpha}$
$p \leq p_0$	wordt verworpen wanneer $k \geq u_{\bar{\alpha}}$.

Een betrouwbaarheidsinterval voor p vinden we door uit te gaan van de algemene definitie, die zegt dat een betrouwbaarheidsinterval bestaat uit alle parameterwaarden die met onbetrouwbaarheid α niet verworpen kunnen worden. Dat zijn dus die p-waarden waarvoor geldt:

$$-u_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{pq/n}} < u_{\alpha/2} \ . \qquad (6.2.2)$$

Op de grenzen is

$$u^2_{\alpha/2} = \frac{n(\hat{p} - p)^2}{pq} \ ,$$

of

$$p^2(1 + u^2_{\alpha/2}/n) - p(2\hat{p} + u^2_{\alpha/2}/n) + \hat{p}^2 = 0 \ ,$$

waaruit volgt

$$p_\ell, p_r = \frac{\hat{p} + u^2_{\alpha/2}/2n \mp u_{\alpha/2}\sqrt{\hat{p}\hat{q}/n + u^2_{\alpha/2}/4n^2}}{(1 + u^2_{\alpha/2}/n)} \ . \qquad (6.2.3)$$

Als we termen met $1/n^{3/2}$ verwaarlozen en $(1 + u^2/n)^{-1}$ benaderen door $(1 - u^2/n)$ krijgen we de eenvoudiger uitdrukking:

$$p_\ell, p_r = \hat{p} + u^2_{\alpha/2}(\tfrac{1}{2} - \hat{p})/n \mp u_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} \ . \qquad (6.2.4)$$

Als n groot is (bv. np > 50) kan dit verder worden vereenvoudigd tot

$$p_\ell, p_r = \hat{p} \mp u_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} \ . \qquad (6.2.5)$$

Deze laatste formule wordt bijvoorbeeld steeds toegepast bij enquête-onderzoek waar men doorgaans met steekproeven van enige honderdtallen werkt.

Een <u>voorbeeld</u> met n = 30, x = 9, $\hat{p}$ = 0,3, α = 0,05 geeft

(6.2.3): 0.166 < p < 0,479 ,
(6.2.4): 0.162 < p < 0.490 ,
(6.2.5): 0.136 < p < 0.464 .

Ter vergelijking het exacte betrouwbaarheidsinterval:

0.148 < p < 0.494 .

Dit interval is gevonden met behulp van tabellen die vermeld staan in S.C. pag. 96 onder de nummers 10 en 11.

Gezien de breedte van het interval zijn de verschillen tussen (6.2.3) en (6.2.4) onderling van geen betekenis, terwijl ook de verschillen met het exact bepaalde interval niet groot zijn.
Men kan (6.2.5) ook gebruiken om globaal de steekproefgrootte n te bepalen die vereist is om p met gegeven nauwkeurigheid Δp te bepalen (behoudens een onbetrouwbaarheid α). Dan moet namelijk gelden:

$$u_{\alpha/2}\sqrt{\frac{pq}{n}} = \Delta p \ ,$$

of

$$n = \frac{p(1-p)u_{\alpha/2}^2}{(\Delta p)^2} \ . \qquad (6.2.6)$$

Hierin moet voor p een voorlopige schatting worden ingevuld. Heeft men geen enkele informatie over de waarde van p, dan kan men eerst een kleine steekproef nemen om die informatie te krijgen. Een bovengrens voor n wordt gevonden door in (6.2.6) $p = \frac{1}{2}$ te stellen.

6.3. Toetsen voor H_0: $p_1 = p_2$, $p_1 \geq p_2$ en $p_1 \leq p_2$, bij twee binomiale verdelingen.

Een exacte toets voor deze hypothesen, met behulp van de hypergeometrische verdeling, wordt in de volgende paragraaf besproken.
Uitgaande van waarnemingen x_1 en x_2 bij steekproefgroottes van resp. n_1 en n_2 gaat de normale verdeling als volgt.
Voor het toetsen van boven vermelde hypothesen gaan we uit van $p_1 = p_2 = p$.
De schatting voor p is dan

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \ .$$

Bij benadering geldt dat

$$\underline{k} := \frac{(\hat{\underline{p}}_1 - \hat{\underline{p}}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \qquad (6.3.1)$$

standaard normaal verdeeld is. Vullen we in de noemer voor p_1 en p_2 de schatter $\hat{\underline{p}}$ in, dan wordt de toetsingsgrootheid

$$\underline{k} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} . \qquad (6.3.2)$$

Om een betrouwbaarheidsinterval voor $\delta := p_1 - p_2$ op te kunnen stellen hebben we een toets nodig voor de hypothese

$$H_0: p_1 - p_2 = \delta .$$

Nu gaan we er blijkbaar van uit dat $p_1 \neq p_2$ kan zijn en dus moeten p_1 en p_2 afzonderlijk worden geschat met resp.:

$$\hat{p}_1 = \frac{\bar{x}_1}{n_1} \quad \text{en} \quad \hat{p}_2 = \frac{x_2}{n_2} .$$

We gaan nu uit van

$$\underline{k} = \frac{\hat{p}_1 - \hat{p}_2 - \delta}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} . \qquad (6.3.3)$$

Door dit gelijk te stellen aan $\pm u_\alpha$ vinden we betrouwbaarheidsgrenzen voor δ. In de meeste gevallen waarin de toetsen worden toegepast zijn de verschillen tussen $\hat{p}_1$ en $\hat{p}_2$ slechts gering en dan is het verschil tussen de noemers van (6.3.2) en (6.3.3) praktisch onbelangrijk.

6.4. De hypergeometrische verdeling en haar benaderingen.

Uit een populatie bestaande uit N elementen waaronder M < N met het kenmerk A wordt een aselecte steekproef van n elementen getrokken zonder teruglegging. Asèlect betekent dat alle $\binom{N}{n}$ mogelijke steekproeven een gelijke kans bezitten te worden getrokken.
Het aantal, $\underline{x}$, elementen van A in de steekproef heeft dan de zg. hypergeometrische verdeling: $\underline{x} \sim HG(N,M,n)$

$$P(\underline{x} = x;\ N,M,n) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}} =$$

$$= \frac{n!\ M!(N-n)!(N-M)!}{N!\ x!(n-x)!(M-x)!(N-M-n+x)!} \qquad (6.4.1)$$

waarbij

$$\max(0,M+n-N) \leq x \leq \min(M,n)$$

$$\mathcal{E}\underline{x} = \frac{nM}{N}\ , \quad \sigma^2 = \frac{nM(N-n)(N-M)}{(N-1)N^2} = n\,\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}\ . \qquad (6.4.2)$$

Berekening kan geschieden met behulp van de tabel van log n! (S.C. 9.3).

Tabel 6.4.1. Rekenvoorbeeld. N = 52, M = 13, n = 4.

log n! + log M! + log(N-n)! + log(N-M)! - log N! = 50.6714						
x	log x!	log(n-x)!	log(M-x)!	log(N-M-n+x)!	log P	P
0	0.0000	1.3802	9.7943	40.0142	9.4827 - 10	0.3039
1	0.0000	0.7782	8.6803	41.5705	9.6424 - 10	0.4389
2	0.3010	0.3010	7.6012	43.1387	9.3295 - 10	0.2135
3	0.7782	0.0000	6.5598	44.7185	8.6149 - 10	0.0412
4	1.3802	0.0000	5.5598	46.3096	7.4218 - 10	0.0026
					Som =	1.0001

De HG verdeling is symmetrisch in M en n. De berekende kansen geven de kans op 0,1,2,3,4 azen in een aselecte trekking van 13 kaarten, of op 0,1,2,3,4 schoppenkaarten in een aselecte trekking van 4 kaarten uit een volledig kaartspel.

Tabel: G.J. Lieberman en D.B. Owen, Tables of the hypergeometric probability distribution, Stanford University Press, 1961.

Bij steekproefonderzoek treden vaak problemen op waarbij men in feite met de HG verdeling te maken heeft doch met één van de volgende benaderingen kan volstaan:

Tabel 6.4.2. Benaderingen voor de HG verdeling

	Voorwaarden	Benadering
1)	$n < 0.1\ N$	$\underline{x} \sim BN(n,\ p = \frac{M}{N})$
2)	$M < 0.1\ N$	$\underline{x} \sim BN(M,\ p = \frac{n}{N})$
3)	$n < 0.1\ N,\ M < 0.1\ N$	$\underline{x} \sim PS(\mu = \frac{nM}{N})$
4)	$\frac{nM}{N} > 5,\ \frac{M(N-n)}{N} > 5,$ $\frac{n(N-M)}{N} > 5,\ \frac{(N-n)(N-M)}{N} > 5$	$\underline{x} \sim N(\mu,\sigma),\ \begin{cases} \mu = \frac{nM}{N} \\ \sigma = \sqrt{\frac{nM(N-M)}{N^2} \cdot \frac{N-n}{N-1}} \end{cases}$

Het S.C. (pag. 17) stelt als voorwaarden voor 4) alleen

$$\frac{nM}{N} > 5 \quad \text{en} \quad \frac{n(N-M)}{N} > 5$$

maar dit is niet voldoende.

Belangrijk zijn vooral 1) en 4) bij enquête onderzoek en 3) bij keuring van partijen produkten in de industrie. In beide gevallen is de steekproef doorgaans klein t.o.v. de populatie ($n < 0.1\ N$); bij keuringen gaat het bovendien om een klein percentage defecte produkten ($M < 0.1\ N$).

Met behulp van de hypergeometrische verdeling kan ook een exacte toets worden gevonden voor de hypothese

$$H_0: P_1 = P_2$$

bij twee binomiale verdelingen. Er kan namelijk eenvoudig worden bewezen dat onder H_0 geldt:

$$P(\underline{x}_1 = x_1 \mid \underline{x}_1 + \underline{x}_2 = x) = \frac{\binom{n_1}{x_1}\binom{n_2}{x - x_1}}{\binom{n_1 + n_2}{x}}. \qquad (6.4.3)$$

Deze hypergeometrische verdeling kan als basis voor een toets dienen als de waarden van n_1, n_2, x_1 en x_2 te klein zijn voor een benadering met de normale verdeling.

6.5. De Poisson verdeling en de normale benadering.

In het Statistisch Compendium komen de volgende tabellen van de Poisson verdeling voor:

6.1. De kansen $P(\underline{x} = x)$.

6.2. De cumulatieve kansen $P(\underline{x} \leq x)$.

8.1. Waarden van μ voor gegeven waarden van c en $P(\underline{x} \leq c)$.

Een uitgebreidere tabel is:
E.C. Molina, Poisson's exponential binomial limit, Van Nostrand, 1942, Bibl. WSK, BN, 4201, bsa.

Zoals bekend is de formule voor de individuele kansen van de Poisson verdeling

$$P(\underline{x} = x) = \frac{e^{-\mu}\,\mu^{x}}{x!} ,$$

waarin de parameter μ zowel de verwachting als de variantie van $\underline{x}$ voorstelt:

$$\mathcal{E}\underline{x} = \mu \ , \quad \text{var } \underline{x} = \sigma^2 = \mu \ ,$$

zodat $\underline{x}$ zowel een zuivere schatter van μ als van σ^2 is.
Verder geldt dat, als

$$\underline{x}_1 \sim PS(\mu_1) \quad \text{en} \quad \underline{x}_2 \sim PS(\mu_2) \ ,$$

onderling onafhankelijk zijn,

$$\underline{x} := \underline{x}_1 + \underline{x}_2 \sim PS(\mu_1 + \mu_2) \ . \tag{6.5.1}$$

Bewijs. $P(\underline{x}_1 + \underline{x}_2 = x) = \sum_{x_1=0}^{x} P(\underline{x}_1 = x_1 \wedge \underline{x}_2 = x - x_1) =$

$$= \sum_{x_1=0}^{x} \frac{e^{-\mu_1}\,\mu_1^{x_1}}{x_1!}\,\frac{e^{-\mu_2}\,\mu_2^{(x-x_1)}}{(x-x_1)!} =$$

$$= \frac{e^{-(\mu_1+\mu_2)}\,(\mu_1+\mu_2)^{x}}{x!} \sum_{x_1=0}^{x} \frac{x!}{x_1!\,(x-x_1)!} \left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{x_1} \left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{x-x_1} .$$

Voor praktische doeleinden is $N(\mu, \sigma^2 = \mu)$ een redelijke benadering voor $PS(\mu)$ wanneer

$$\mu \geq 5$$

en een goede benadering wanneer

$$\mu \geq 10 .$$

Op analoge wijze kan worden bewezen dat, als

$$\underline{x}_1 \sim PS(\mu_1) \quad \text{en} \quad \underline{x}_2 \sim PS(\mu_2)$$

onderling onafhankelijk zijn,

$$P(\underline{x}_1 = x_1 \mid \underline{x}_1 + \underline{x}_2 = n) = \binom{n}{x_1} p^{x_1} (1-p)^{n-x_1} , \qquad (6.5.2)$$

met

$$p := \frac{\mu_1}{\mu_1 + \mu_2} .$$

Dus onder de voorwaarde dat $\underline{x}_1 + \underline{x}_2 = n$, is $\underline{x}_1$ binomiaal verdeeld (en uiteraard $\underline{x}_2$ eveneens).

6.6. Toetsen en betrouwbaarheidsintervallen bij één Poisson verdeling.

Op grond van één waarneming x kunnen we een van de volgende hypothesen toetsen:

$$H_0: \mu = \mu_0, \ \mu \leq \mu_0 \text{ of } \mu \geq \mu_0 .$$

Indien de exacte tabellen toereikend zijn verdient het aanbeveling deze te gebruiken. Tabel 6.1 zal bruikbaar zijn tot ongeveer $x \leq 15$. Voor grotere waarden van x, tot $x \leq 30$, kan tabel 8.1 worden toegepast. Dit gaat als volgt:

Zoek bij $c = x$ de waarde $\mu = m_r$ voor $P_A = \alpha$ resp. $\alpha/2$
en bij $c = x - 1$ de waarde $\mu = m_\ell$ voor $P_A = 1 - \alpha$ resp. $(1 - \alpha/2)$

$H_0: \mu \leq \mu_0$ wordt verworpen wanneer $\mu_0 < m_\ell(\alpha)$
$H_0: \mu \geq \mu_0$ wordt verworpen wanneer $\mu_0 > m_r(\alpha)$
$H_0: \mu = \mu_0$ wordt verworpen wanneer $\mu_0 < m_\ell(\alpha/2)$ of $> m_r(\alpha/2)$. (6.6.1)

Voorbeeld. $x = 20$, H_0: $\mu = 30$, $\alpha = 0{,}05$. In tabel 8.1 zien we dat

$$P(\underline{x} \leq 20 \mid m_r = 30{,}888) = 0{,}025 .$$

Dus volgens (6.6.1) wordt H_0 niet verworpen. Het is namelijk duidelijk dat

$$P(\underline{x} \leq 20 \mid \mu = 30) > 0{,}025 .$$

Bij gebruik van de normale benadering wordt berekend

$$k = \frac{x - \mu_0}{\sqrt{\mu_0}} .$$

Onder H_0: $\mu = \mu_0$ geldt nu bij benadering: $\underline{k} \sim N(0,1)$. De regels voor het toetsen luiden nu:

$$\boxed{\begin{array}{l} H_0\text{: } \mu = \mu_0 \text{ wordt verworpen als } |k| \geq u_{\alpha/2} \\ H_0\text{: } \mu \leq \mu_0 \text{ wordt verworpen als } k \geq u_{\alpha} \\ H_0\text{: } \mu \geq \mu_0 \text{ wordt verworpen als } k \leq -u_{\alpha} . \end{array}} \qquad (6.6.2)$$

In het bovenstaande voorbeeld is

$$k = \frac{20 - 30}{\sqrt{30}} = -1{,}83 ,$$

terwijl $u_{0,025} = 1{,}96$, zodat H_0 niet wordt verworpen.

6.7. Betrouwbaarheidsintervallen voor de parameter μ van een Poisson verdeling.

Gegeven: een realisatie x van een Poisson variabele.

$$\boxed{\begin{array}{l} \text{Is } x > 50, \text{ dan is een goede benadering} \\ x - u_{\alpha/2}\sqrt{x} < \mu < x + u_{\alpha/2}\sqrt{x} , \\ \text{met betrouwbaarheid } (1 - \alpha). \end{array}} \qquad (6.7.1)$$

Dit interval is daarop gebaseerd dat

$$\frac{\underline{x} - \mu}{\sqrt{\underline{x}}}$$

bij benadering normaal verdeeld is wanneer μ groot is.

Voor een globale snelle schatting van de grootte van een betrouwbaarheidsinterval is (6.7.1) bijzonder nuttig; bij $\alpha = 5\%$ neemt men dan ruwweg $u_{\alpha/2} = 2.0$ i.p.v. 1.96.
Een nauwkeuriger benadering zijn de grenzen

$$\left.\begin{matrix} m_\ell \\ m_r \end{matrix}\right\} = x + \tfrac{1}{2}u^2_{\alpha/2} \mp u_{\alpha/2}\sqrt{x + \tfrac{1}{4}u^2_{\alpha/2}} \ ,$$

$$m_\ell < \mu < m_r \text{ met betrouwbaarheid } (1-\alpha). \qquad (6.7.2)$$

Afleiding: Is μ de ware waarde van de parameter, dan geldt bij toepassing van de normale benadering en zonder continuïteitscorrectie

$$P(-u_{\alpha/2} < \frac{\underline{x} - \mu}{\sqrt{\mu}} < u_{\alpha/2}) = (1-\alpha) \ ,$$

waaruit volgt

$$P((\underline{x} - \mu)^2 - \mu u^2_{\alpha/2} < 0) = (1-\alpha) \ .$$

Wanneer echter $(\underline{x} - \mu)^2 - \mu u^2_{\alpha/2} < 0$ moet μ liggen tussen de wortels van de tweedegraads vergelijking

$$(\underline{x} - \mu)^2 - \mu u^2_{\alpha/2} = 0 \ .$$

Deze wortels zijn dus stochastische variabelen $\underline{m}_\ell$, $\underline{m}_r$ waarvoor geldt

$$P(\underline{m}_\ell < \mu < \underline{m}_r) = (1-\alpha) \ .$$

(6.7.2) is een eenmalige toepassing van de relatie.

Exacte betrouwbaarheidsintervallen voor $x \leq 20$ worden gegeven in tabel S.C. 6.3.
Voor $x \leq 30$ kan eveneens een exact betrouwbaarheidsinterval worden gevonden met behulp van S.C. 8.1 volgens (6.6.1).

Bv. $x = 25$. In tabel 8.1 vinden we bij $c = 25$ dat $\mu_\ell = 16{,}984$ voor $P_A = 0{,}975$ en $\mu_r = 36{,}905$ voor $P_A = 0{,}025$.
Een tweezijdig betrouwbaarheidsinterval met betrouwbaarheid 95% is dus

$$16{,}984 < \mu < 36{,}905 .$$

Ter vergelijking: met (6.72) wordt gevonden

$$16{,}93 < \mu < 36{,}91 .$$

6.8. Het vergelijken van twee Poisson verdelingen.

6.8.1. Het toetsen van de hypothesen: $\mu_1 = \mu_2$, $\mu_1 \geq \mu_2$, $\mu_1 \leq \mu_2$.

Gegeven: waarnemingen x_1, x_2.

Model: $\underline{x}_1 \sim PS(\mu_1)$, $\underline{x}_2 \sim PS(\mu_2)$.

De te toetsen hypothesen zijn:

$$\mu_1 = \mu_2,\ \mu_1 \leq \mu_2,\ \mu_1 \geq \mu_2$$

tegen de alternatieven

$$\mu_1 \neq \mu_2,\ \mu_1 > \mu_2,\ \mu_1 < \mu_2 .$$

Een belangrijk verschil met het geval van één Poisson verdeling is dat door een hypothese $\mu = \mu_0$ tevens de waarde van $\sigma = \sqrt{\mu_0}$ wordt vastgelegd, terwijl door de hypothese $\mu_1 = \mu_2$ de waarden van μ_1 en μ_2 zelf onbepaald blijven. We zijn derhalve genoodzaakt voor var $\underline{x}_1$ en var $\underline{x}_2$ hun puntschatters x_1, x_2 te gebruiken.
Toepassing van de normale benadering levert dan de volgende toets:

$$k = \frac{x_1 - x_2}{\sqrt{x_1 + x_2}} , \quad \underline{k} \sim N(0,1) .$$

H_0: $\mu_1 \leq \mu_2$ wordt verworpen wanneer $k > u_\alpha$,
H_0: $\mu_1 \geq \mu_2$ wordt verworpen wanneer $k < -u_\alpha$,
H_0: $\mu_1 = \mu_2$ wordt verworpen wanneer $|k| > u_{\alpha/2}$.

(6.8.1a)

Door berekening en door simulatie kan worden aangetoond dat deze toets nog goed bruikbaar is voor $\mu_1 = \mu_2 = 2$.
Een alternatieve toets is gebaseerd op stelling (6.5.2). Onder de nulhypothese $\mu_1 = \mu_2$ is

$$P(x_1 = x_1 \mid n = \underline{x}_1 + \underline{x}_2) = \binom{n}{x_1}(\tfrac{1}{2})^n .$$

De toets verloopt dan als volgt.

$k = x_1$ of x_2, $x_1 + x_2 = n$,

$$P(\underline{k} \le k) = \sum_{r=0}^{k} \binom{n}{r}(\tfrac{1}{2})^n .$$

H_0: $\mu_1 \le \mu_2$ wordt verworpen wanneer $P(\underline{x}_2 \le x_2 \mid n) < \alpha$,
H_0: $\mu_1 \ge \mu_2$ wordt verworpen wanneer $P(\underline{x}_1 \le x_1 \mid n) < \alpha$,
H_0: $\mu_1 = \mu_2$ wordt verworpen wanneer $P(\underline{x}_1 \le x_1 \mid n) < \alpha/2$
of $P(\underline{x}_2 \le x_2 \mid n) < \alpha/2$.

(6.8.1b)

De toets kan worden uitgevoerd met behulp van S.C. 7.1. Uit die tabel lezen we bijv.

voor $n = 10$, $k = 1$: $0.01 < P(\underline{k} \le 1) < 0.025$,

en

voor $n = 10$, $k = 2$: $0.05 < P(\underline{k} \le 2) < 0.125$.

Of een combinatie n,k significant is bij een gegeven α lezen we direkt uit deze tabel af.
De toetsen volgens (6.8.1b) zijn voorwaardelijke toetsen onder de voorwaarde $x_1 + x_2 = n$. Is echter voor een voorwaardelijke toets

$$P(\underline{k} \le k_n \mid n) < \alpha$$

dan is tevens voor de onvoorwaardelijke toets

$$P(\underline{k} \le k_n) = \sum_{n} P(n)P(\underline{k} \le k_n \mid n) < \alpha ,$$

en is een onvoorwaardelijke bovengrens voor $P(H_0)$. Dit geldt algemeen, en voorwaardelijke onbetrouwbaarheden worden vaker toegepast wanneer men er niet in slaagt een toets met een onvoorwaardelijke onbetrouwbaarheid te construeren.

6.8.2. Betrouwbaarheidsintervallen.

Onder H_0: $\mu_1 - \mu_2 = \delta$ en met toepassing van de normale benadering zal gelden:

$$\underline{k} = \frac{\underline{x}_1 - \underline{x}_2 - \delta}{\sqrt{\underline{x}_1 + \underline{x}_2}} \sim \underline{u} \sim N(0,1) \ .$$

Dit zou kunnen dienen om H_0: $\mu_1 - \mu_2 = \delta$ te toetsen, of om voor $\delta = \mu_1 - \mu_2$ een betrouwbaarheidsinterval te construeren; dit luidt

$$\left.\begin{matrix} d_\ell \\ d_r \end{matrix}\right\} = x_1 - x_2 \mp u_{\alpha/2}\sqrt{x_1 + x_2} \ ,$$

$$d_\ell < \mu_1 - \mu_2 < d_r \text{ met betrouwbaarheid } (1-\alpha). \qquad (6.8.2)$$

Uit een betrouwbaarheidsinterval voor p bij een BN-verdeling (zie 6.2) en stelling (6.5.2) kan ook een betrouwbaarheidsinterval worden gevonden voor de verhouding μ_1/μ_2. We zullen dit verder niet uitwerken.

6.9. Vergelijken van de toetsen (6.8.1a) en (6.8.1b).

De toets (6.8.1a) is gebaseerd op de normale benadering en kan daarom niet zonder meer bij kleine waarden van x_1 en x_2 worden toegepast; (6.8.1b) is daarentegen exact en is bij alle waarden van x_1 en x_2 toepasbaar. Het is interessant deze toetsen te vergelijken.
Als voorbeeld kiezen we $\alpha = 0.025$, $u_\alpha = 1.96$ en de éénzijdige hypothese $\mu_1 \geq \mu_2$ die alleen kan worden verworpen wanneer $x_2 > x_1$. Aan de hand van het criterium k volgens (6.8.1a) en met tabel S.C. 7.1 kunnen we dan een lijst opmaken van die combinaties van x_1 en x_2 die tot verwerping van de nulhypothese leiden, Tabel 6.9; en met behulp van de tabellen S.C. 6.1 en 6.2 kunnen we vervolgens voor gegeven waarden van μ_1 en μ_2 berekenen hoe groot de kans op verwerpen van H_0 in die situatie is.

Tabel 6.9. Vergelijking van de toetsen (6.8.1a) en (6.8.1b) bij lage waarden van μ_1 en μ_2.

Model: $\underline{x}_1 \sim PS(\mu_1)$, $\underline{x}_2 \sim PS(\mu_2)$, H_0: $\mu_1 \geq \mu_2$, $\alpha = 0.025$, $u_\alpha = 1.96$

$k = \frac{x_2 - x_1}{\sqrt{x_2 + x_1}}$ verwerpt H_0 voor				Tabel S.C. 7.1 verwerpt H_0 voor			
$x_1 = 0$	$x_2 \geq 4$	$x_1 = 6$	$x_2 \geq 15$	$x_1 = 0$	$x_2 \geq 6$	$x_1 = 6$	$x_2 \geq 17$
1	7	7	17	1	8	7	18
2	9	8	18	2	10	8	20
3	11	9	20	3	12	9	21
4	12	10	21	4	13	10	23
5	14	enz.		5	15	enz.	

Kansen $P(\not H_0)$ dat H_0 wordt verworpen

Toets		(6.8.1a)	(6.8.1b)
μ_1	μ_2	$P(\not H_0)$	
2.0	2.0	0.0207	0.0026
4.0	4.0	0.0223	0.0091
6.0	6.0	0.0226	0.0127
2.0	6.0	0.2725	0.1816
2.0	10.0	0.6787	0.5611
4.0	8.0	0.1928	0.1339
4.0	12.0	0.5365	0.3975

We zien uit deze tabel dat de toets (6.8.1a) ook bij lage waarden van μ_1 en μ_2 zeer goed voldoet. Onder H_0 is $P(\not H_0)$ slechts weinig kleiner dan de gestelde waarde $\alpha = 0.025$ en bovendien vrijwel onafhankelijk van de waarde van μ_1 en μ_2. Bij toepassing van S.C. 7.1 ligt $P(\not H_0)$ ver beneden α en is bovendien van de waarde van μ_1 en μ_2 afhankelijk. Ook het onderscheidingsvermogen van de eerste toets is beter zoals blijkt uit de hogere waarden van $P(\not H_0)$ voor de alternatieve hypothesen.

Ook kan het onderscheidingsvermogen van (6.8.1a) op eenvoudige wijze worden geschat door aan te nemen dat de stochastische variaties in $\sqrt{\underline{x}_1 + \underline{x}_2}$ mogen worden verwaarloosd en dat in goede benadering

$$\underline{k} = \frac{\underline{x}_2 - \underline{x}_1}{\sqrt{\underline{x}_2 + \underline{x}_1}} \approx \frac{\underline{x}_2 - \underline{x}_1}{\sqrt{\mu_2 + \mu_1}} \sim N\left(\frac{\mu_2 - \mu_1}{\sqrt{\mu_2 + \mu_1}}, 1\right) .$$

H_0 zal dan worden verworpen wanneer

$$\frac{\mu_2 - \mu_1}{\sqrt{\mu_2 + \mu_1}} + \underline{u} > u_\alpha .$$

Voor het voorbeeld van Tabel 6.9 geeft dit

$$\mu_1 = 2.0,\ \mu_2 = 6.0 \Rightarrow P(\not H_0) = P(\underline{u} > 0.55) = 0.2912\ (0.2725) ,$$

$$\mu_1 = 2.0,\ \mu_2 = 10.0 \Rightarrow P(\not H_0) = P(\underline{u} > -0.35) = 0.6368\ (0.6587) .$$

De tussen haakjes geplaatste waarden zijn de exact berekende uit Tabel 6.9; de verschillen zijn voor de praktijk onbelangrijk.

7. Toepassingen van de chi-kwadraatverdeling.

Een veel voorkomend statistisch probleem is het toetsen van de hypothese dat een aantal waargenomen frequenties overeenkomt met een overeenkomstig aantal kansen volgens een theoretisch model.

7.1. De χ^2-toets.

Een eenvoudig voorbeeld van het vergelijken van theoretische met waargenomen frequenties is het volgende. Met een dobbelsteen worden 120 worpen uitgevoerd met het volgende resultaat. In de rij aangegeven met o (observed) staan de waargenomen frequenties en in de rij aangegeven met e (expected) de bij een zuivere dobbelsteen verwachte frequenties.

Aantal ogen	1	2	3	4	5	6
o	12	21	27	22	20	18
e	20	20	20	20	20	20

Als maat voor de overeenstemming tussen dergelijke waargenomen en verwachte frequenties is het gebruikelijk een grootheid X^2 te berekenen, gedefinieerd door

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \qquad (7.1.1)$$

In ons geval is k = 6 en

$$X^2 = \frac{(12-20)^2}{20} + \frac{(21-20)^2}{20} + \frac{(27-20)^2}{20} + \frac{(22-20)^2}{20}$$

$$+ \frac{(20-20)^2}{20} + \frac{(18-20)^2}{20} = 6{,}10$$

Een waarde $X^2 = 0$ zou een volledige overeenstemming betekenen met de verwachting terwijl X^2 bij slechter wordende overeenstemming steeds hogere waarden aanneemt. Om te kunnen beslissen of de gevonden waarde $X^2 = 6{,}10$ aanleiding geeft om aan het model te twijfelen moeten we weten wat de kansverdeling is van $\underline{X}^2$ onder de aanname dat het model (kansen $\frac{1}{6}$ op elk van het aantal ogen) juist is. Hieruit kan dan de rechtséénzijdige overschrijdingskans worden berekend.
Men kan bewijzen dat $\underline{X}^2$ onder de nulhypothese bij benadering een chi-

kwadraat verdeling heeft met k - 1 vrijheidsgraden. Dat het aantal vrijheidsgraden één minder bedraagt dan k hangt samen met het feit dat er tussen de $\underline{o}_i$ één lineaire betrekking bestaat, namelijk : $\sum \underline{o}_i = 120$.
Een nodige voorwaarde voor een goede benadering is dat alle verwachte aantallen e_i minstens gelijk zijn aan 5.

In ons voorbeeld is $\nu = 6 - 1 = 5$. De bijbehorende rechter overschrijdingskans is > 0,25. Er is dus geen enkele aanleiding om de hypothese dat de dobbelsteen zuiver is te verwerpen.

7.2. Het verband met de multinomiale verdeling.

De multinomiale verdeling is een generalisatie van de binomiale verdeling. Bij de binomiale verdeling wordt verondersteld dat er n onafhankelijke experimenten plaatsvinden, waarbij er bij elk experiment twee mogelijke uitkomsten zijn. Deze uitkomsten worden dan vaak met "succes" en "mislukking" aangeduid, waarbij de kans op een succes p en op een mislukking (1 - p) is. Bij de multinomiale verdeling zijn er meerdere mogelijke uitkomsten, namelijk k inplaats van twee. Deze uitkomsten geven we aan met

$$A_1, A_2, \ldots, A_k$$

met als bijbehorende kansen

$$p_1, p_2, \ldots, p_k.$$

Hierbij moet dus gelden: $\sum_{i=1}^{k} p_i = 1$.
Het aantal keren dat, bij n onafhankelijke experimenten, de gebeurtenis A_i optreedt noemen we $\underline{x}_i$ $(i = 1, \ldots, k)$, dus $\sum_{i=1}^{k} \underline{x}_i = n$.

Gevraagd wordt nu de kans dat de eerste gebeurtenis precies x_1 keer, de tweede x_2 keer, ..., de k^{e} precies x_k keer optreedt, waarbij uiteraard weer $\sum x_i = n$.
Op soortgelijke wijze als voor de binomiale verdeling (vergelijk syllabus WSK 31/49) kan worden afgeleid dat deze kans gegeven wordt door

$$P(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \qquad (7.2.1)$$

Deze zelfde formule kan ook langs geheel andere weg worden verkregen. Stel dat $\underline{x}_1,\ldots,\underline{x}_k$ k onafhankelijke stochastische variabelen zijn, waarbij $\underline{x}_i$ een Poisson verdeling heeft met verwachting $\mu_i = np_i$. Dus

$$P(\underline{x}_i = x_i) = \frac{e^{-np_i}(np_i)^{x_i}}{x_i!} \tag{7.2.2}$$

Beschouw nu de verdeling van $(\underline{x}_1,\ldots,\underline{x}_n)$ onder de voorwaarde $\sum\underline{x}_i = n$.
De som $\sum\underline{x}_i$ is weer Poisson verdeeld met gemiddelde $\sum_i\mu_i = \sum_i np_i = n$.
Dus

$$P(\textstyle\sum\underline{x}_i = n) = \frac{e^{-n}n^n}{n!} \tag{7.2.3}$$

Uit (7.2.2) en (7.2.3) samen volgt dat de voorwaardelijke verdeling is

$$P(x_1,x_2,\ldots,x_k \mid \textstyle\sum\underline{x}_i = n) = \frac{\prod_{i=1}^{k} P(\underline{x}_i=x_i)}{P(\sum\underline{x}_i = n)} =$$

$$= \frac{n!}{x_1!x_2!\ldots x_k!}\, p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}, \tag{7.2.4}$$

precies de zelfde uitdrukking als (7.2.1). Wij merken op dat dit een generalisatie is van het geval behandeld in 6.5 voor k = 2 (vergelijk 6.5.2). Aangezien een Poisson-verdeelde variabele onder bepaalde voorwaarden bij benadering normaal is, zijn de gestandaardiseerde variabelen

$$\frac{\underline{x}_i - np_i}{\sqrt{np_i}}$$

bij benadering standaard-normaal verdeeld.
Volgens de in 4.4.1 gegeven definitie van de χ^2-verdeling zou men kunnen verwachten dat de uitdrukking

$$\sum_{i=1}^{k} \frac{(\underline{x}_i - np_i)^2}{np_i} \tag{7.2.5}$$

bij benadering een χ^2_k-verdeling zou hebben.
Ten gevolge van de afhankelijkheid opgelegd door de relatie $\sum \underline{x}_i = n$ wordt de verdeling echter beter benaderd door een χ^2-verdeling met $(k-1)$ vrijheidsgraden. Het verband met paragraaf 7.1 is nu duidelijk, want de X^2 van (7.1.1) is een realisatie van (7.2.5). Immers de verwachting e_i van de waarneming $\underline{x}_i$ (in plaats van $\underline{o}_i$) is np_i.
Men kan verder bewijzen dat als er meer (bijvoorbeeld $p > 1$) lineaire betrekkingen tussen de $\underline{x}_i$ bestaan, de verdeling van $\underline{X}^2$ benaderd wordt door een χ^2-verdeling met $k-p$ vrijheidsgraden.

7.3. Een toets voor k Poisson verdelingen.

In 6.8 werden voor twee Poisson grootheden de toetsen voor $H_0 : \mu_1 = \mu_2$, $\mu_1 \geq \mu_2$ en $\mu_1 \leq \mu_2$ besproken.
Een algemener probleem is het toetsen van

$$H_0: \frac{\mu_1}{a_1} = \frac{\mu_2}{a_2}\ ,\ \frac{\mu_1}{a_1} \geq \frac{\mu_2}{a_2}\ ,\ \frac{\mu_1}{a_1} \leq \frac{\mu_2}{a_2}\ .$$

Aan het eind van deze paragraaf wordt hiervan een voorbeeld gegeven (Voorbeeld 7.3.2).
Bij k verdelingen is de eenvoudigste hypothese:

$$H_0: \mu_i = \mu_j; \quad i,j = 1,\ldots,k$$

en algemener

$$H_0: \frac{\mu_i}{a_i} = c, \quad i = 1,\ldots,k$$

met a_i gegeven en c een onbekende constante.
De tweezijdige hypothesen voor twee verdelingen zijn speciale gevallen van de overeenkomstige hypothesen voor k verdelingen. In het laatste geval ($k > 2$) zijn geen duidelijke éénzijdige hypothesen aan te geven. Wordt de algemene nulhypothese verworpen, dan dient men nog apart na te gaan wat de oorzaak kan zijn. Het kan voorkomen dat één μ_i afwijkt van de andere, dat ze alle onderling verschillen, dat ze in twee of meer groepen uiteenvallen, of dat een ander model moet worden toegepast.

Voorbeeld 7.3.1. Stel dat de aantallen geregistreerde ongevallen in een afdeling van een fabriek als volgt zijn:

1972	1973	1e halfjaar 1974
13	10	2

We willen de hypothese toetsen:

$$H_0: \mu_1 = \mu_2 = 2\mu_3,$$

In het algemeen geldt:

$$\mu_i = c a_i,$$

$$\sum \mu_j = c \sum a_j.$$

Als de waarnemingen uit Poissonverdelingen afkomstig zijn, dan is onder de voorwaarde $\sum x_i = n$ de verdeling multinomiaal met kansen

$$p_i = \frac{\mu_i}{\sum \mu_j} = \frac{a_i}{\sum a_j}.$$

In dit voorbeeld:

$$a_1 = a_2 = 2;\ a_3 = 1,$$

want de nulhypothese kan worden geschreven als

$$H_0: \frac{\mu_1}{2} = \frac{\mu_2}{2} = \frac{\mu_3}{1}.$$

Dus

$$p_1 = p_2 = \frac{2}{5};\ p_3 = \frac{1}{5}.$$

Verder is $n = 25$, dus toepassing van (7.2.5) (of van (7.1.1)) levert op:

$$X^2 = \frac{(13 - \frac{2}{5}.25)^2}{\frac{2}{5}.25} + \frac{(10 - \frac{2}{5}.25)^2}{\frac{2}{5}.25} + \frac{(2 - \frac{1}{5}.25)^2}{\frac{1}{5}.25} = 2.7$$

Dit kan worden beschouwd (onder H_0) als een trekking uit een χ^2_2-verdeling. De bijbehorende overschrijdingskans (tabel S.C. 3.1) is > 25%. Dus geen reden om H_0 te verwerpen.

Toegepast op twee waarnemingen geeft bovenstaande methode, na herleiding, de toetsingsgrootheid

$$\frac{\left(\frac{\underline{x}_1}{a_1} - \frac{\underline{x}_2}{a_2}\right)^2}{\frac{\underline{x}_1 + \underline{x}_2}{a_1 a_2}} \sim \chi_1^2 \simeq \underline{u}^2 \qquad (7.3.1)$$

Ten eerste moet hierbij worden opgemerkt dat een χ^2-toets met $\nu = 1$ gelijkwaardig is aan een u-toets. Alleen gaat door het kwadrateren het teken verloren, waardoor een tweezijdige toets gebaseerd op u gelijkwaardig is met een rechtséénzijdige toets gebaseerd op χ_1^2. Vergelijking met (6.8.1a) laat zien dat daar een speciaal geval is beschreven van (7.3.1) met $a_1 = a_2$ (= 1).

In de tweede plaats kunnen we opmerken dat, evenals in 6.3 voor twee binomiale verdelingen, er een verschil optreedt tussen het toetsen van een hypothese en het construeren van een betrouwbaarheidsinterval.

Bij het toetsen van H_0 ligt het voor de hand, om, zoals hierboven in feite is gebeurd, door combinatie van beide waarnemingen een enkele schatting $\hat{c}$ van c te berekenen:

$$\left.\begin{aligned} \mathcal{E}\,\underline{x}_1 &= ca_1 \\ \mathcal{E}\,\underline{x}_2 &= ca_2 \end{aligned}\right\} \Rightarrow \hat{c} = \frac{x_1 + x_2}{a_1 + a_2} . \qquad (7.3.2)$$

Daaruit kan dan de variantie van het verschil

$$\underline{d} = \frac{\underline{x}_1}{a_1} - \frac{\underline{x}_2}{a_2}$$

worden bepaald. We vinden dan

$$\hat{\text{var}}\,\underline{d} = \frac{\hat{\mu}_1}{a_1^2} + \frac{\hat{\mu}_2}{a_2^2} = \hat{c}\left(\frac{1}{a_1} + \frac{1}{a_2}\right) = \frac{x_1 + x_2}{a_1 a_2} , \qquad (7.3.3)$$

(dit alles onder H_0).

Bij toepassing van de normale benadering is dan de toetsingsgrootheid

$$\underline{k} = \frac{\underline{d}}{\sqrt{\hat{\text{var}}\ \underline{d}}} = \frac{\underline{x}_1/a_1 - \underline{x}_2/a_2}{\left\{\frac{\underline{x}_1 + \underline{x}_2}{a_1 a_2}\right\}^{\frac{1}{2}}} = \frac{a_2\underline{x}_1 - a_1\underline{x}_2}{\{a_1 a_2(\underline{x}_1 + \underline{x}_2)\}^{\frac{1}{2}}} \qquad (7.3.4)$$

geheel in overeenstemming met (7.3.1).

Bij het construeren van een betrouwbaarheidsinterval voor het verschil $\frac{\mu_1}{a_1} - \frac{\mu_2}{a_2}$ daarentegen gaan we er in principe van uit dat $\frac{\mu_1}{a_1} \neq \frac{\mu_2}{a_2}$ en dan wordt var $\underline{d}$ geschat door

$$\hat{\text{var}}\ \underline{d} = \frac{x_1}{a_1^2} + \frac{x_2}{a_2^2} \neq \text{var}\ (\underline{d}\,|\,H_0).$$

Voorbeeld 7.3.2. Aan twee radio-actieve preparaten heeft men met een Geigerteller waargenomen:

bij A: 540 aanslagen in 1 minuut,
bij B: 390 aanslagen in ½ minuut, zodat

$$x_A = 540\ ,\quad a_A = T_A = 60\ ,\quad \hat{\mu}_A = 9\ ,$$
$$x_B = 390\ ,\quad a_B = T_B = 30\ ,\quad \hat{\mu}_B = 13\ .$$

De hypothese $\mu_A = \mu_B$ wordt dan getoetst met

$$u = \frac{9 - 13}{\sqrt{\frac{540 + 390}{30 \times 60}}} = -5.56$$

of

$$X_1^2 = \frac{(30 \times 540 - 60 \times 390)^2}{30 \times 60 \times (540 + 390)} = 31.0\ .$$

De grenzen van een betrouwbaarheidsinterval voor het verschil $\mu_A - \mu_B$ worden daarentegen gegeven door

$$\hat{\mu}_A - \hat{\mu}_B \mp u_{\alpha/2}\sqrt{\frac{x_A}{a_A^2} + \frac{x_B}{a_B^2}} = 9 - 13 \mp u_{\alpha/2}\sqrt{\frac{540}{60^2} + \frac{390}{30^2}} =$$

$$= -4 \mp u_{\alpha/2} \times 0{,}764 \ , \quad (1-\alpha) \ .$$

7.4. Afhankelijkheidstabellen.

Voorbeeld 7.4.1.

Tabel 7.4.1 geeft de resultaten van een wiskundetentamen waarbij twee verschillende klasse-indelingen zijn toegepast, en wel

a) naar studierichting,

b) naar niet deelgenomen, voldoende en onvoldoende.

Studierichting	ND (Niet deelgenomen)	V (Voldoende)	OV (Onvoldoende)	Totaal
E	51	81	51	183
N	31	45	17	93
T	25	61	39	125
W	49	58	52	159
totaal	156	245	159	560

Tabel 7.4.1.

Als we met p_{ij}, i = 1,2,3,4; j = 1,2,3, de kans aangeven dat een willekeurig gekozen student tot de i-de rij en de j-de kolom behoort, dan kan bij bekende p_{ij}'s de overeenkomst tussen de waargenomen en de verwachte aantallen weer met de χ^2-verdeling worden getoetst. Gewoonlijk (en ook in dit voorbeeld) zijn deze kansen echter onbekend en moeten ze uit de waarnemingen worden geschat. We zijn dan meestal geïnteresseerd in de hypothese dat de rij- en kolomindelingen onafhankelijk van elkaar zijn, zodat geldt:

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j} \ .$$

We schatten dan

$$\hat{p}_{1\cdot} = \hat{p}_E = \frac{183}{560} = 0.327$$

$$\hat{p}_{2\cdot} = \hat{p}_N = \frac{93}{560} = 0.166$$

$$\hat{p}_{3.} = \hat{p}_T = \frac{125}{560} = 0.223$$

$$\hat{p}_{4.} = \hat{p}_W = \frac{159}{560} = 0.284$$

$$\hat{p}_{.1} = \hat{p}_{ND} = \frac{156}{560} = 0.279$$

$$\hat{p}_{.2} = \hat{p}_V = \frac{245}{560} = 0.438$$

$$\hat{p}_{.3} = \hat{p}_{OV} = \frac{159}{560} = 0.284 \ .$$

Hieruit volgen dan schattingen voor de p_{ij}, bv.

$$\hat{p}_{31} = 0.223 \times 0.279 = 0.062 \ .$$

De bijdrage tot X^2 van de "cel" (3,1) ofwel (T,ND) is dan

$$\frac{(25 - 560 \times 0.062)^2}{560 \times 0.062} = \frac{(25 - \frac{125 \times 156}{560})^2}{\frac{125 \times 156}{560}} = \frac{(25 - 34.82)^2}{34.82} = 2.77 \ .$$

Opmerking. De verwachte frequenties worden in de praktijk rechtstreeks uit de randtotalen berekend, bv. $E_{31} = \frac{125 \times 156}{560}$ zonder eerst $\hat{p}_{31}$ te berekenen.

Alle $4 \times 3 = 12$ cellen geven een totale X^2 van 12.14.
Het bijbehorende aantal vrijheidsgraden is nu echter niet gelijk aan $12 - 1 = 11$.
Er zijn 4 parameters $\hat{p}_{i.}$ geschat, maar deze voldoen aan de relatie

$$\sum_i \hat{p}_{i.} = 1 \ ,$$

dus er zijn er slechts $(4 - 1) = 3$ te schatten, waarna de 4e vastligt.
Evenzo zijn we $(3 - 1) = 2$ parameters $\hat{p}_{.j}$ te schatten.
Tenslotte is er de relatie $\sum_{i,j} x_{ij} = N$, het totale aantal.
Het aantal vrijheidsgraden is in zo'n geval

$$12 - (4 - 1) - (3 - 1) - 1 = (4 - 1)(3 - 1) = 6 \ .$$

In het algemeen bij r rijen en k kolommen:

$$\nu = (r - 1)(k - 1) \ .$$

Opmerking. Beschouwen we de toets als voorwaardelijk, d.w.z. onder de voorwaarde van de gevonden randtotalen, dan leidt dit tot hetzelfde aantal vrijheidsgraden. Dit is nl. ook het aantal celfrequenties dat vrij kan worden gekozen bij deze vaste randtotalen.

De overschrijdingskans is in ons voorbeeld ongeveer 5% en de hypothese van onafhankelijkheid is aan twijfel onderhevig.
Nadere beschouwing leert dat de grootste bijdragen komen van de cellen (N,OV) en (T,ND), zoals blijkt uit Tabel 7.4.2.

		ND	V	OV
E	x_{ij}	51	81	51
	$\hat{\mathcal{E}}x_{ij}$	50.98	80.06	51.96
	ΔX^2	0.00	0.01	0.02
N	x_{2j}	31	45	17
	$\hat{\mathcal{E}}x_{2j}$	25.91	40.69	26.41
	ΔX^2	1.00	0.46	3.35
T	x_{3j}	25	61	39
	$\hat{\mathcal{E}}x_{3j}$	34.82	54.69	35.49
	ΔX^2	2.77	0.73	0.35
N	x_{4j}	49	58	52
	$\hat{\mathcal{E}}x_{4j}$	44.29	69.56	45.14
	ΔX^2	0.50	1.92	1.04

Tabel 7.4.2.

Men zou nu kunnen concluderen dat de N-studenten relatief weinig onvoldoendes scoorden en dat de T-studenten relatief weinig verzuimden deel te nemen. Daarbij is dan ook het teken van de verschillen $(\underline{x} - \mathcal{E}\underline{x})$ in aanmerking genomen.

Een speciaal geval doet zich voor als het aantal kolommen (of het aantal rijen) gelijk is aan 2. We hebben dan te maken met een zogenaamde k × 2 tabel. We kunnen dit opvatten als een toets voor k binomiale verdelingen.

Voorbeeld 7.4.2.

In tabel 7.4.3 zijn de resultaten weergegeven van 6 maal 100 worpen met 6 verschillende dobbelstenen. (x_i = aantal malen dat "6" wordt geworpen.)

				$p_i = 1/6$		$\hat{p}_i = \hat{p} = 130/600$	
Steen	n_i	x_i	$n_i - x_i$	$\mathcal{E}\underline{x}_i$	$\mathcal{E}(n_i - \underline{x}_i)$	$\hat{\mathcal{E}}\underline{x}_i$	$\hat{\mathcal{E}}(n_i - \underline{x}_i)$
A	100	13	87	16.7	83.3	21.7	78.3
B	100	24	76	16.7	83.3	21.7	78.3
C	100	16	84	16.7	83.3	21.7	78.3
D	100	19	81	16.7	83.3	21.7	78.3
E	100	27	73	16.7	83.3	21.7	78.3
F	100	31	69	16.7	83.3	21.7	78.3
	600	130	470	$X^2 = 27.7$		$X^2 = 13.8$	
	$X^2 = 10.8$			$\nu = 6$		$\nu = 5$	
	$\nu = 1$; $P = 0.001$			$P < 0.001$		$P \sim 0.02$	

Tabel 7.4.3.

Verschillende toetsen zijn hier mogelijk.

$$H_0: p_i = 1/6 \qquad (7.4.1)$$

$$X^2 = \frac{(13-16.7)^2}{16.7} + \frac{(24-16.7)^2}{16.7} + \ldots + \frac{(73-83.3)^2}{83.3} + \frac{(69-83.3)^2}{83.3} = 27.7.$$

In dit geval hebben we 6 vrijheidsgraden. Er zijn namelijk 2 × 6 = 12 waarnemingen en 6 lineaire betrekkingen:

$$x_i + (n_i - x_i) = n_i = 100 \quad (i = 1,\ldots,6).$$

Daar χ^2_6 (0.001) = 22.5 is de overschrijdingskans kleiner dan 0.001 en wordt H_0 verworpen.
Dit resultaat kan ook op een andere wijze verkregen worden. We beschouwen daartoe de stochastische grootheden

$$\frac{x_i - n_i p_i}{\sqrt{n_i p_i q_i}}, \quad i = 1,\ldots,6$$

welke onder zekere voorwaarden bij benadering standaardnormaal verdeeld zijn en bovendien onderling onafhankelijk zijn. De som van de kwadraten hiervan is dan χ^2-verdeeld met 6 vrijheidsgraden:

$$\sum_{i=1}^{6} \frac{(x_i - n_i p_i)^2}{n_i p_i q_i} \sim \chi^2_6$$

De berekening voor tabel 7.4.3 leidt dan tot

$$X^2 = \frac{(13 - 16.7)^2}{16.7 * 5/6} + \ldots + \frac{(31 - 16.7)^2}{16.7 * 5/6} = 27.7 ,$$

exact hetzelfde resultaat als hiervoor.
Men kan inderdaad gemakkelijk bewijzen, dat

$$\sum_{i=1}^{k} \frac{(x_i - n_i p_i)^2}{n_i p_i q_i} = \sum_{i=1}^{k} \left[\frac{(x_i - n_i p_i)^2}{n_i p_i} + \frac{\{(n_i - x_i) - n_i q_i\}^2}{n_i q_i}\right]$$

$$H_0: p_i = p; \hat{p} = \frac{130}{600} = 0.217 \qquad (7.4.2)$$

$$X^2 = \frac{(13 - 21.7)^2}{21.7} + \frac{(24 - 21.7)^2}{21.7} + \ldots + \frac{(69 - 78.3)^2}{78.3} = 13.8.$$

Er is nu één lineaire betrekking meer, namelijk $\sum x_i = 130$, zodat $\nu = 5$.
Uit de χ^2-tabel (S.C. 3.1) volgt dat de bijbehorende overschrijdingskans ligt tussen 0.025 en 0.01. Er is dus reden om aan de nulhypothese te twijfelen.
Tenslotte kan getoetst worden of de gemiddelde p gelijk is aan 1/6.

$$H_0: \bar{p} = 1/6 \qquad (7.4.3)$$

$$X^2 = \frac{(130 - 100)^2}{100} + \frac{(470 - 500)^2}{500} = 10.8.$$

Het aantal vrijheidsgraden bedraagt nu slechts 1, vanwege de relatie $x + (n - x) = n = 600$.

De kritieke waarde χ_1^2 (0.001) is juist 10.8, zodat H_0 wordt verworpen.

Het meest eenvoudige geval van de afhankelijkheidstabel is de 2 × 2 tabel.

Voorbeeld 7.4.3.

Van 100 studenten is bij het begin van de studie een serie tests afgenomen om de geschiktheid voor de studie te voorspellen en na een aantal jaren is op grond van een aantal studiescores de gebleken geschiktheid vastgesteld.
De resultaten zijn in tabel 7.4.4 samengevat.

		gebleken geschiktheid		
		-	+	
voorspelde geschiktheid	-	20	10	30
	+	10	60	70
		30	70	100

Tabel 7.4.4.

De nulhypothese betekent in dit geval dat de voorspelling geen enkele waarde heeft. De verwachte aantallen staan in tabel 7.4.5.

		gebleken		
		-	+	
voorspeld	-	9	21	30
	+	21	49	70
		30	70	100

Tabel 7.4.5.

$$X^2 = \frac{(20-9)^2}{9} + \frac{(21-10)^2}{21} + \frac{(21-10)^2}{21} + \frac{(60-49)^2}{49} = 27.4$$

De nulhypothese moet dus zeer duidelijk worden verworpen ($\nu = 1$).
De tellers van de termen in X^2 zijn allen gelijk, we kunnen in dit geval dus ook schrijven

$$X^2 = (11)^2 \left(\frac{1}{9} + \frac{1}{21} + \frac{1}{21} + \frac{1}{49}\right).$$

In deze paragraaf hebben we ons beziggehouden met het toetsen van onafhankelijkheid tussen twee variabelen (rij- en kolomindelingen) in een afhankelijkheidstabel. Meestal echter zijn we niet alleen geïnteresseerd in de vraag of er verband of samenhang bestaat tussen de variabelen maar ook in de sterkte van deze samenhang. Om nu deze samenhang te kunnen beschrijven zijn een groot aantal zogeheten associatiematen ontwikkeld. Daarbij kunnen twee typen worden onderscheiden:

1) Maten die gebaseerd zijn op de χ^2-verdeelde toetsingsgrootheid voor onafhankelijkheid; de waarde 0 correspondeert dan één-éénduidig met statistische onafhankelijkheid.
2) Maten met een voorspellende interpretatie; kennis van de waarde van de ene variabele geeft meer informatie over de waarde van de andere variabele naarmate het verband sterker is.

De belangrijkste eigenschappen waaraan associatiematen zouden moeten voldoen zijn:

1) De maat is 0 (dan en slechts dan) als geen associatie (samenhang) bestaat.
2) De maat is maximaal (dan en slechts dan) als de associatie maximaal is, waarbij het begrip "maximale associatie" nader moet worden gespecificeerd.
3) De maat moet, indien van toepassing, de richting van het verband aangeven.
4) De maat moet genormeerd zijn; het waardenbereik mag niet afhangen van
   i) het totaal aantal waarnemingen (N),
   ii) het aantal rijen (r) en kolommen (k).
5) De waarden moeten onderling vergelijkbaar zijn.
6) De maat moet interpreteerbaar zijn.

Wij zullen enkele van de meest gebruikte associatiematen bespreken.

7.4.1. Op χ^2 gebaseerde associatiematen.

Deze maten zijn afgeleid van de toetsingsgrootheid X^2 voor onafhankelijkheid. De waarde $X^2 = 0$ komt overeen met onafhankelijkheid en dat geldt ook voor de van χ^2 afgeleide associatiematen (vergelijk eis 1). Een maat die vol-

doet aan eis 4-i) is

$$\phi^2 := X^2 : N \ .$$

Een genormeerde maat kan worden verkregen door X^2 te delen door zijn maximale waarde. Bewezen kan worden dat deze maximale waarde gelijk is aan $N[\min(r,k) - 1]$. De zo verkregen associatiemaat, geïntroduceerd door Cramér luidt:

$$V := \sqrt{\frac{X^2}{N(\min(r,k) - 1)}}$$

Pearson heeft de zogeheten contingency coëfficiënt ingevoerd:

$$C := \sqrt{\frac{X^2/N}{X^2/N + 1}} = \sqrt{\frac{X^2}{X^2 + N}} \ .$$

Weliswaar geldt $C < 1$, maar de waarde 1 kan niet worden bereikt. Deze maat kan genormeerd worden als volgt:

$$C' := \frac{C}{C_{max}} = \sqrt{\frac{X^2}{X^2 + N}} \cdot \frac{\min(r,k)}{\min(r,k) - 1}$$

Cramérs V en de genormeerde contingency coëfficiënt C' voldoen aan de meeste van de genoemde eigenschappen. Een belangrijk bezwaar is echter dat beide maten afhankelijk zijn van de randtotalen, waardoor vergelijkbaarheid van verschillende tabellen (eis 5) niet goed mogelijk is. Interpretatie van deze maten is evenmin goed mogelijk; zij geven aan hoe groot de afwijking van onafhankelijkheid is maar niet hoe groot de overeenkomst is met een vorm van maximale samenhang.
Uit de waarnemingen van de afhankelijkheidstabel van voorbeeld 7.4.1. vinden we voor deze twee associatiematen

$$V = \sqrt{\frac{12,14}{560.2}} = 0,104 \quad , \quad C' = \sqrt{\frac{12,14}{12,14+560}} \cdot \frac{3}{2} = 0,178 \ .$$

Beide waarden wijzen op een geringe afwijking (in welke richting dan ook) van onafhankelijkheid.

7.4.2. Maten met een voorspellende interpretatie

We bespreken eerst de lambda-maat van Goodman en Kruskal aan de hand van voorbeeld 7.4.1. We vragen ons dan af in hoeverre de studierichting iets zegt over het tentamenresultaat. Voor een willekeurig gekozen student zullen we een voldoende voorspellen omdat de geschatte kans hierop het grootst is:

$$\frac{x_{.2}}{N} = \frac{245}{560} = 0{,}4375.$$

De geschatte kans dat deze voorspelling onjuist is, is

$$\hat{P}_1 = 1 - \frac{x_{.2}}{N} = 0{,}5625 .$$

Als de studierichting bekend is, b.v. E, dan is de beste voorspelling eveneens "voldoende". De geschatte kans dat deze voorspelling fout is, is

$$1 - \frac{x_{12}}{x_{1.}} = 1 - \frac{81}{183} = 0{,}5574.$$

Gemiddeld over alle studierichtingen is de geschatte kans dat de beste voorspelling fout is

$$\hat{P}_2 = \sum_{i=1}^{4} \left\{1 - \frac{\max_j(x_{ij})}{x_{i.}}\right\} \frac{x_{i.}}{N} =$$

$$= (1 - \tfrac{81}{183})\,\tfrac{183}{560} + (1 - \tfrac{45}{93})\tfrac{93}{560} + (1 - \tfrac{61}{125})\tfrac{125}{560} + (1 - \tfrac{58}{159})\tfrac{159}{560}$$

$$= 0.5625 .$$

Hierbij is $\frac{x_{i.}}{N}$ de geschatte kans dat een willekeurig gekozen student studierichting i heeft.

De lambda-maat λ_K, gedefinieerd door

$$\lambda_K := \frac{\hat{P}_1 - \hat{P}_2}{\hat{P}_1} ,$$

geeft dan een relatieve vermindering van de kans op een foute voorspelling als de studierichting bekend is.

Berekening geeft:

$$\lambda_K = \frac{0{,}5625 - 0{,}5625}{0{,}5625} = 0 .$$

In dit voorbeeld leidt kennis van de studierichting niet tot een betere voorspelling van het tentamenresultaat ("voldoende" is voor elke studierichting de beste voorspelling), maar de beide variabelen zijn wel afhankelijk.

De index K geeft aan dat de kolomvariabele wordt voorspeld. Het is uiteraard ook mogelijk, maar in dit voorbeeld niet interessant, om de studierichting te voorspellen en te zien of het tentamenresultaat de kans op een fout al of niet reduceert. Herschrijven van de uitdrukking λ_K geeft:

$$\lambda_K = \frac{\sum_{i=1}^{r} \max_j (x_{ij}) - \max_j (x_{.j})}{N - \max_j (x_{.j})} .$$

Analoog vinden we bij het voorspellen van de rijvariabele

$$\lambda_R = \frac{\sum_{j=1}^{k} \max_i (x_{ij}) - \max_i (x_{i.})}{N - \max_i (x_{i.})} .$$

Uit deze formules blijkt direkt de asymmetrie van deze maten. In het geval van onafhankelijkheid is $\lambda_K = 0$ (en ook $\lambda_R = 0$). Het omgekeerde is niet het geval zoals in het voorbeeld blijkt. De maximale waarde $\lambda_K = 1$ betekent dat de kolomvariabele perfekt voorspeld kan worden uit de rijvariabele. Ook de lambda-maten hebben het nadeel dat ze afhankelijk zijn van de randtotalen.

Als beide variabelen geordende klassen hebben is een associatiemaat gewenst die positief is als beide variabelen vaak samen een hoge of samen een lage waarde hebben, en die negatief is als de ene variabele een hoge waarde heeft terwijl de andere een lage waarde heeft. Een maat die hieraan voldoet is de gamma-maat van Goodman en Kruskal

$$\gamma := \frac{P - Q}{P + Q} .$$

Hierin is P het aantal concordante paren waarnemingen, d.w.z. het aantal paren waarvoor de ordening van beide variabelen dezelfde richting heeft. Q is het aantal discordante paren: de ordening voor de ene variabele is tegengesteld aan de ordening van de andere variabele. De paren waarvan beide elementen voor minstens één variabele in dezelfde klasse vallen worden buiten

beschouwing gelaten. Stel één element van een paar behoort tot cel (i,j), het andere element behoort tot cel (i',j'). Het paar is dan concordant als $(i-i')(j-j') > 0$ en het paar is discordant als $(i-i')(j-j') < 0$. Als $(i-i')(j-j') = 0$ dan is het paar noch concordant noch discordant. Ter toelichting beschouwen we voorbeeld 7.4.3. Uit de 100 studenten kunnen $\frac{1}{2} \cdot 100.99 = 4950$ paren worden gevormd. Een paar is concordant als één student tot cel (1,1) behoort en de andere tot (2,2). Er zijn dus $P = 20 \cdot 60 = 1200$ concordante paren. Een paar is discordant als één student tot cel 1,2) behoort en de andere tot (2,1). Het aantal discordante paren is dus $Q = 10 \cdot 10 = 100$. De overige $4950 - 1200 - 100 = 3650$ paren zijn noch concordant noch discordant (ga dat na). De associatiemaat voor deze tabel is

$$\gamma = \frac{1200 - 100}{1200 + 100} = \frac{11}{13} = 0{,}846 \ .$$

Deze associatiemaat kan worden geinterpreteerd als het verschil van twee voorwaardelijke kansen. Immers $P/(P+Q)$ is de kans op een concordant paar gegeven dat het paar óf concordant óf discordant is, $Q/(P+Q)$ is de kans op een discordant paar onder dezelfde voorwaarde. Er geldt dus $-1 \leq \gamma \leq 1$ en γ geeft de richting van de associatie aan. Ook is duidelijk dat $\gamma = 0$ als de variablelen onafhankelijk zijn (het omgekeerde is niet het geval). Verder is γ onafhankelijk van r en k. Wel is het zo dat tabellen met verschillende fracties paren die óf concordant óf discordant zijn niet zonder meer vergelijkbaar zijn. Tot slot geldt ook hier dat de associatiemaat afhankelijk is van de randtotalen.

7.5. De χ^2-toets voor aanpassing (goodness of fit).

De χ^2-toets kan ook worden gebruikt om te toetsen of een steekproef uit een bepaalde verdeling afkomstig is. In 7.1 is daarvan reeds een voorbeeld gegeven, namelijk 120 worpen met een dobbelsteen. In dat voorbeeld was de verdeling discreet en volledig gespecificeerd. In het nu volgende voorbeeld is dat laatste niet het geval.

Voorbeeld 7.5.1. Bij de kwaliteitscontrole van een productieproces worden om het uur steekproeven van 30 stuks genomen. Het aantal foute exemplaren in 60 steekproeven is weergegeven in de volgende tabel. We willen toetsen of het resultaat kan worden opgevat als een steekproef uit een Poisson verdeling.

x	o	e (PS(1))	ΔX^2
0	25	22.1	0.381
1	20	22.1	0.200
2	7	11.0	1.455
3	6	3.7 } 4.8	2.133
≥ 4	2	1.1 }	
	60	60	$X^2 = 4.17$

χ^2_2 (0.90) = 4,61

P > 10%

Hier moet μ worden geschat uit de waarnemingen

$$\hat{\mu} = \frac{\sum o_i x_i}{\sum o_i} = \frac{60}{60} = 1$$

Er zijn nu twee lineaire betrekkingen, namelijk

$$\sum o_i = 60 \text{ en } \sum o_i x_i = 60,$$

dus moeten twee vrijheidsgraden worden afgetrokken en blijven er twee over. De laatste twee klassen zijn samengenomen om een verwachte waarde te verkrijgen die niet te ver onder de grenswaarde 5 ligt.

Voorbeeld 7.5.2. Tenslotte behandelen we een voorbeeld van een steekproef uit een continue verdeling. In de volgende tabel staan 100 waarden voor de treksterkte van textiel (ponden/inch^2).

320	380	340	410	380	340	360	350	320	370
350	340	350	360	370	350	380	370	300	420
370	390	390	440	330	390	330	360	400	370
320	350	360	340	340	350	350	390	380	340
400	360	350	390	400	350	360	340	370	420
420	400	350	370	330	320	390	380	400	370
390	330	360	380	350	330	360	300	360	360
360	390	350	370	370	350	390	370	370	340
370	400	360	350	380	380	360	340	330	370
340	360	390	400	370	410	360	400	340	360

Gevraagd wordt te toetsen of deze steekproef uit een normale verdeling afkomstig is. We moeten daarom eerst μ en σ schatten:

$$\hat{\mu} = \bar{x} = 364.70$$

$$\hat{\sigma} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{99}} = 26.83 \ .$$

Daarna wordt de x-as in 10 intervallen ingedeeld, zoals in de volgende tabel is aangegeven, en per interval worden de verwachte frequenties berekend.

x_i	$\frac{x_i - 364.70}{26.83}$	$\Phi(\frac{x_i - 364.70}{26.83})$	$\mathcal{E}f_i$	f_i	ΔX^2
$(-\infty$; 325)	$-\infty$; -1.48	0 ; 0.0694	6.94	6	0.13
(325 ; 335)	-1.48 ; -1.11	0.0694 ; 0.1335	6.41	6	0.03
(335 ; 345)	-1.11 ; -0.73	0.1335 ; 0.2327	9.92	11	0.12
(345 ; 355)	-0.73 ; -0.36	0.2327 ; 0.3594	12.67	14	0.14
(355 ; 365)	-0.36 ; 0.01	0.3594 ; 0.5040	14.46	16	0.16
(365 ; 375)	0.01 ; 0.38	0.5040 ; 0.6480	14.40	15	0.02
(375 ; 385)	0.38 ; 0.76	0.6480 ; 0.7764	12.84	8	1.82
(385 ; 395)	0.76 ; 1.13	0.7764 ; 0.8708	9.44	10	0.03
(395 ; 405)	1.13 ; 1.50	0.8708 ; 0.9332	6.24	8	0.50
(405 ; ∞	1.50 ; ∞	0.9332 ; 1.0000	6.68	6	0.07
			100.00	100	3.02

Het resultaat is $X^2 = 3.02$.
Het aantal vrijheidsgraden is 7, want van de 10 moeten er 3 worden afgetrokken: 1 voor de betrekking $\sum f_i = 100$ en één voor elk van de geschatte parameters.
De kritieke waarde ($\alpha = 0.05$) bij $\nu = 7$ is 14.07 en de nulhypothese wordt dus niet verworpen.

Opmerking. Volledig juist is het aantal vrijheidsgraden niet, omdat de χ^2-benadering alleen geldt als de parameters worden geschat uit de klassemiddens met bijbehorende waargenomen freqenties. In werkelijkheid wordt de verdeling van X^2 begrensd door de χ^2_{k-p}-verdeling en de χ^2_k-verdeling (k = aantal klassen; p = aantal lineaire betrekkingen).

8. Regressie analyse.

8.1. Correlatie.

Stel $\underline{x}$ en $\underline{y}$ hebben een simultane verdeling. In WSK 49 (pag. 38) zagen we:

$$\text{var}(\underline{x}+\underline{y}) = \text{var}\,\underline{x} + \text{var}\,\underline{y}$$

als $\underline{x}$ en $\underline{y}$ onafhankelijk zijn. Wat wordt dit indien $\underline{x}$ en $\underline{y}$ wel stochastisch afhankelijk zijn? Wel,

$$\text{var}(\underline{x}+\underline{y}) = \mathcal{E}[\underline{x}+\underline{y}-\mu_x-\mu_y]^2 = \mathcal{E}[(\underline{x}-\mu_x)+(\underline{y}-\mu_y)]^2 =$$

$$= \mathcal{E}(\underline{x}-\mu_x)^2 + 2\,\mathcal{E}[(\underline{x}-\mu_x)(\underline{y}-\mu_y)] + \mathcal{E}(\underline{y}-\mu_y)^2 ,$$

nl. $\mathcal{E}$ is een lineaire operator. We definiëren nu de covariantie van $\underline{x}$ en $\underline{y}$.

8.1.1. Definitie.

$$\text{cov}(\underline{x},\underline{y}) = \sigma_{xy} := \mathcal{E}(\underline{x}-\mu_x)(\underline{y}-\mu_y) ,$$

dan ontstaat

8.1.2. $$\text{var}(\underline{x}+\underline{y}) = \text{var}\,\underline{x} + \text{var}\,\underline{y} + 2\,\text{cov}(\underline{x},\underline{y}) .$$

Analoog aan $\text{var}\,\underline{x} = \mathcal{E}\underline{x}^2 - (\mathcal{E}\underline{x})^2$ geldt: $\text{cov}(\underline{x},\underline{y}) = \mathcal{E}(\underline{xy}) - \mathcal{E}\underline{x}\,\mathcal{E}\underline{y}$. We zien dat $\text{cov}(\underline{x},\underline{x}) = \text{var}\,\underline{x}$.

8.1.3. Zijn $\underline{x}$ en $\underline{y}$ stochastisch onafhankelijk, dan geldt (zie WSK 49, pag. 38):

$$\mathcal{E}(\underline{xy}) = \mathcal{E}\underline{x}\,\mathcal{E}\underline{y} \quad \text{oftewel} \quad \text{cov}(\underline{x},\underline{y}) = 0 .$$

In het bijzonder is dus $\text{cov}(\underline{x},\underline{y}) = 0$ als $\underline{x}$ en/of $\underline{y}$ constant zijn.
Nu is $\text{cov}(\underline{x},\underline{y})$ gevoelig voor schaaltransformaties (nl. $\text{cov}(a\underline{x},b\underline{y}) = ab\,\text{cov}(\underline{x},\underline{y})$) en is dus geen goede maat voor afhankelijkheid. Daarom wordt nu een andere maat ingevoerd:

8.1.4. Definitie. De (populatie) correlatiecoëfficiënt

$$\rho(\underline{x},\underline{y}) := \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{\text{cov}(\underline{x},\underline{y})}{\sqrt{\text{var}\,\underline{x}\;\text{var}\,\underline{y}}} .$$

In feite is dit de covariantie tussen de gestandaardiseerde variabelen

$$\underline{x}^* = (\underline{x} - \mu_x)/\sigma_x \quad \text{en} \quad \underline{y}^* = (\underline{y} - \mu_y)/\sigma_y \ ,$$

dus (ga na)

$$\rho(\underline{x},\underline{y}) = \text{cov}(\underline{x}^*,\underline{y}^*) \ .$$

8.1.5. Stelling. $\rho(\underline{x},\underline{y})$ is, op het teken na, invariant onder lineaire transformaties van $\underline{x}$ en $\underline{y}$.

Bewijs. Stel $\underline{u} = a\underline{x} + b$, $\underline{v} = c\underline{y} + d$ en $ac \neq 0$. Nu is

$$\sigma_u = |a|\sigma_x \quad \text{en} \quad \sigma_v = |c|\sigma_y$$

(zie WSK 49, pag. 34) en

$$\text{cov}(\underline{u},\underline{v}) = \mathcal{E}[(a\underline{x} + b - a\mu_x - b)(c\underline{y} + d - c\mu_y - d)] =$$

$$= \mathcal{E}[ac(\underline{x} - \mu_x)(\underline{y} - \mu_y)] = ac\ \text{cov}(\underline{x},\underline{y}) \ ,$$

dus

$$\rho(\underline{u},\underline{v}) = \frac{\sigma_{uv}}{\sigma_u\sigma_v} = \frac{ac\ \sigma_{xy}}{|a||c|\sigma_x\sigma_y} = \frac{ac}{|a||c|}\rho(\underline{x},\underline{y}) = \pm\rho(\underline{x},\underline{y}) \ . \qquad \square$$

Opmerking. De correlatiecoëfficiënt tussen $\underline{x}$ en een constante c is niet gedefinieerd, nl. dan is $\text{cov}(\underline{x},c) = 0$ en $\sigma_c = 0$.

8.1.6. Stelling. $\rho^2(\underline{x},\underline{y}) \leq 1$.

Bewijs. Voor alle a geldt:

$$\text{var}(a\underline{x} + \underline{y}) = a^2\ \text{var}\ \underline{x} + 2a\ \text{cov}(\underline{x},\underline{y}) + \text{var}\ \underline{y} \geq 0 \ .$$

Dus de discriminant van de vierkantsvergelijking in a is niet positief:

$$4\ \text{cov}^2(\underline{x},\underline{y}) - 4\ \text{var}\ \underline{x}\ \text{var}\ \underline{y} \leq 0$$

oftewel

$$\rho^2 = \frac{\text{cov}^2(\underline{x},\underline{y})}{\text{var}\ \underline{x}\ \text{var}\ \underline{y}} \leq 1 \ . \qquad \square$$

Opmerking. $\rho^2 = 1$ dan en slechts dan als $a\underline{x} + \underline{y} = c$ (constant), dus bij volledige lineaire afhankelijkheid. Deze ρ is een maat voor lineaire afhanke-

lijkheid en het ware beter geweest als ρ de lineaire correlatiecoëfficiënt genoemd werd.

8.1.7. In 8.1.3 zagen we dat, indien $\underline{x}$ en $\underline{y}$ stochastisch onafhankelijk zijn, $\mathrm{cov}(\underline{x},\underline{y}) = 0$ en dus ook $\rho(\underline{x},\underline{y}) = 0$. Als $\rho = 0$ noemen we $\underline{x}$ en $\underline{y}$ ongecorreleerd.

Dus: stochastisch onafhankelijke variabelen zijn ongecorreleerd, echter ongecorreleerd betekent nog niet stochastisch onafhankelijk, zoals blijkt uit het volgende voorbeeld:

Stel $\underline{x}$ neemt de waarde $-1,0,1$ aan, elk met kans $1/3$. Stel $\underline{y} = \underline{x}^2$. Nu is $\mathcal{E}\underline{x} = 0$ en $\mathcal{E}(\underline{xy}) = \mathcal{E}(\underline{x}^3) = 0$, dus $\rho = 0$, oftewel $\underline{x}$ en $\underline{y}$ zijn ongecorreleerd, terwijl $\underline{y} = \underline{x}^2$, dus sterk afhankelijk.

Opmerking 1. Zijn $\underline{x}$ en $\underline{y}$ simultaan normaal verdeeld dan betekent ongecorreleerd wel stochastisch onafhankelijk (zie Multivariate Analyse).

Opmerking 2. Ongecorreleerd betekent dus niet dat $\underline{x}$ en $\underline{y}$ stochastisch onafhankelijk zijn, wel echter dat $\underline{x}$ en $\underline{y}$ lineair onafhankelijk zijn.

8.1.8. Vaak is $\rho(\underline{x},\underline{y})$ niet bekend. Deze moeten we dan schatten uit een steekproef van n paren waarnemingen $(x_1,y_1),\ldots,(x_n,y_n)$.

Als voor de hand liggende schatting gebruiken we de (steekproef) correlatiecoëfficiënt

$$r(\underline{x},\underline{y}) := \frac{s_{xy}}{s_x s_y}$$

waarin

$$s_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})/(n-1) ,$$

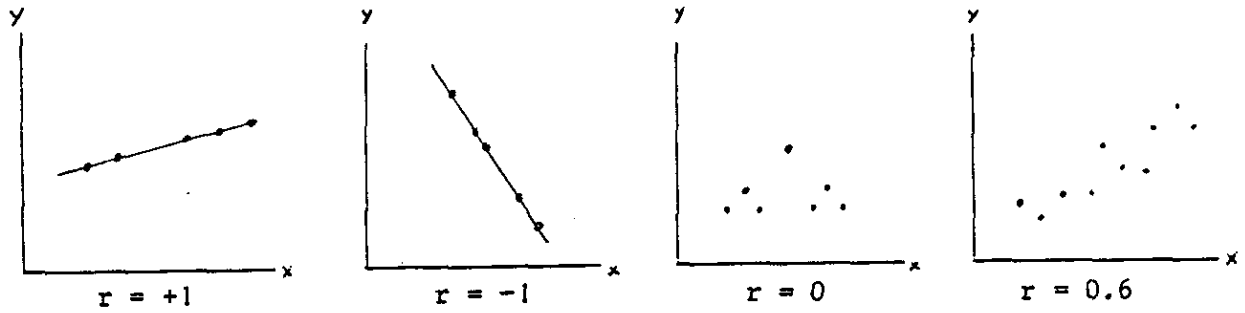
$$s_x^2 = \sum_i (x_i - \bar{x})^2/(n-1) \quad \text{en} \quad s_y^2 = \sum_i (y_i - \bar{y})^2/(n-1) .$$

Voor berekeningen is de volgende formule beter:

8.1.9.
$$r = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{[n\Sigma x_i^2 - (\Sigma x_i)^2][n\Sigma y_i^2 - (\Sigma y_i)^2]}} .$$

Voor r gelden analoge eigenschappen als voor ρ, o.a. is r^2 ongevoelig voor lineaire transformaties en geldt ook $r^2 \leq 1$. We kunnen dus de waarnemingen eerst coderen, de gecodeerde waarnemingen geven dezelfde ρ.

In de volgende figuur zijn enkele "puntenwolken" getekend met bijbehorende r.



8.1.10. Om de nulhypothese H_0: $\rho = 0$ te toetsen tegen H_1: $\rho \neq 0$, gebruikt men als toetsingsgrootheid:

$$\frac{\underline{r}\sqrt{n-2}}{\sqrt{1-\underline{r}^2}} \approx \underline{t}_{n-2} \ ,$$

d.w.z. onder H_0 heeft deze grootheid een Studentverdeling met $n-2$ vrijheidsgraden. Ook kan men direkt S.C. 2.2 hanteren. Men moet wel bedenken dat beide uitgaan van de veronderstelling dat $\underline{x}$ en $\underline{y}$ simultaan normaal verdeeld zijn (zie ook 8.2.12).

8.1.11. Wat vertelt ons nu deze correlatiecoëfficiënt? Men moet hiermee zeer kritisch zijn. Enkele aanhalingen:

"Thurstone, one of the chief architects of modern factor analysis was obliged to call the correlation coefficient a symbol of complete ignorance."
Een opmerking in het boek van H.W. Alexander, pag. 295: "The usefulness of the correlation coefficient is severely limited by the difficulty of interpreting it."

We zagen reeds dat ρ een maat is voor <u>lineaire</u> afhankelijkheid, d.w.z. ρ bijna +1 of -1 geeft een sterk <u>lineair</u> verband aan tussen $\underline{x}$ en $\underline{y}$ en ρ bijna 0 geeft aan dat er geen lineair verband bestaat.
Het is onjuist om ρ te koppelen aan het begrip "causaal verband". Men maakt vaak de fout door te onderstellen dat, indien $\underline{x}$ en $\underline{y}$ hoog gecorreleerd zijn, $\underline{x}$ en $\underline{y}$ ook sterk <u>causaal</u> gerelateerd zouden zijn. Zo ontstaat de bekende <u>nonsens</u> correlatie: de prijs $\underline{x}$ van de rumbonen en het salaris $\underline{y}$ van de dominees in de loop der jaren is hoog gecorreleerd. Wordt nu $\underline{x}$ door $\underline{y}$ veroorzaakt?

Vaak wordt een hoge correlatie veroorzaakt door het feit dat zowel $\underline{x}$ als $\underline{y}$ beide weer hoog gecorreleerd zijn met een derde variabele $\underline{z}$ (bv. de tijd, economische index).
Ook een zeer gebruikelijke (o.a. in de Psychologie) doch onjuiste procedure bij een probleem waarin tal van factoren een rol spelen is het berekenen van een groot aantal correlaties, de significante correlaties eruit te lichten en hiervoor een verklaring op te bouwen. Overigens een methodische denkfout die ook elders in de Statistiek gemaakt wordt.
Er zijn nog meer coëfficiënten van correlatie, o.a. rang-, partiële, multiple correlatie. Deze behandelen we hier niet.

8.2. Regressie analyse.

Het algemene regressieprobleem.

Vaak komen er problemen voor waarbij een stochastische variabele $\underline{y}$ wordt waargenomen als functie van een aantal variabelen $x_1,\ldots,x_k$. Het algemene model is dan $\mathcal{E}(\underline{y} \mid x) = f(x_1,\ldots,x_k;\ \beta_1,\ldots,\beta_p)$, waarbij $\beta_1,\ldots,\beta_p$ onbekende modelparameters zijn.
$\underline{y}$ kan bijv. de opbrengst zijn in een chemisch proces, $x_1,\ldots,x_k$ de temperatuur, druk, concentratie e.d., waar de opbrengst van afhangt.
Men schrijft ook wel

8.2.1. $$\underline{y} = f(x_1,\ldots,x_k;\ \beta_1,\ldots,\beta_p) + \underline{e} \qquad (\underline{e} \text{ van error};\ \mathcal{E}\underline{e} = 0).$$

Dit heet een regressievergelijking, y de afhankelijke, te verklaren variabele en de $x_1,\ldots,x_k$ de onafhankelijke of verklarende variabelen.
De regressie analyse houdt zich bezig met het probleem uit een steekproef, dat zijn n waarnemingen van $\underline{y}$ met n instellingen of keuzen van $x_1,\ldots,x_k$, de parameters te schatten met bijbehorende nauwkeurigheid en eventuele toetsen uit te voeren. Het doel van de regressie-analyse kan o.a. zijn dat men y als functie van de x_i wil kennen om te gebruiken bij interpoleren of kalibreren.
De vorm van de functie f is soms gegeven op grond van theoretische beschouwingen, doch in vele gevallen gaat het alleen om een doelmatige beschrijving van de samenhang tussen te verklaren en verklarende variabelen en kiest men voor f een vorm die praktisch hanteerbaar en aanvaardbaar is.
Is f een lineaire functie van de parameters, dan spreken we van meervoudige lineaire regressie als $k > 1$.

We beperken ons echter tot <u>enkelvoudige</u> lineaire regressie (k = 1) en de regressievergelijking luidt dan

8.2.2. $$\mathcal{E}(\underline{y} \mid x) = \alpha + \beta x$$

oftewel

$$\underline{y} = \alpha + \beta x + \underline{e}$$

oftewel voor de waarnemingen

$$\underline{y}_i = \alpha + \beta x_i + \underline{e}_i \ , \quad i = 1,\ldots,n \ .$$

De onderstellingen behorende bij dit model zijn meestal:

1) De variabele x is <u>niet</u> stochastisch, d.w.z. de waarden x_i zijn exact bekend of althans toevallige fouten in x zijn verwaarloosbaar. Deze aanname is essentieel voor het vervolg!
2) $\text{cov}(\underline{e}_i,\underline{e}_j) = 0$ voor $i \neq j$, d.w.z. de toevallige fouten zijn onderling ongecorreleerd. $\mathcal{E}\underline{e}_i = 0$ en $\text{var}\ \underline{e}_i = \sigma^2$ voor $i = 1,\ldots,n$. D.w.z. de variantie is onafhankelijk van x_i. Deze aanname kan men eventueel laten vallen.
3) Om toetsen te kunnen uitvoeren neemt men vaak aan dat $\underline{e}_i \sim N(0,\sigma^2)$, oftewel dat de fout normaal verdeeld is.

8.2.3. <u>Schatting van de parameters</u>.

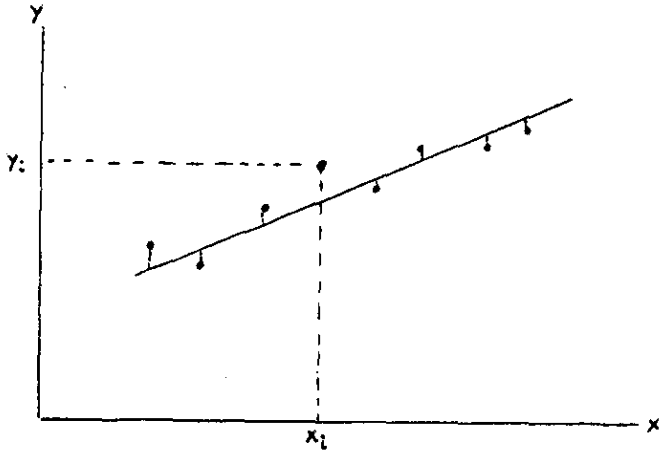
We proberen nu uit een "steekproef" de modelparameters α, β (de <u>regressiecoëfficiënten</u>) en σ^2 te schatten. Deze schatters zijn resp. $\underline{a}$, $\underline{b}$ en $\underline{s}^2$.

Het biedt voordeel het model enigszins anders te schrijven (zie ook 8.2.7), nl.

8.2.4. $$\underline{Y}_i = \beta_0 + \beta_1(x_i - \bar{x}) + \underline{e}_i \ .$$

Dus $\beta_1 = \beta$ en $\alpha = \beta_0 - \beta_1\bar{x}$. $\bar{x} = \Sigma x_i/n$.

Fig. 8.2.5.



$\underline{b}_0$ en $\underline{b}_1$ zijn de schatters voor resp. β_0 en β_1. Om deze b_0 en b_1 te bepalen zoeken we een lijn

$$\hat{y} = b_0 + b_1(x - \bar{x})$$

die "zo goed mogelijk" bij de waarnemingen past. Men zou op het oog een lijn door de puntenwolk kunnen trekken, doch deze is natuurlijk erg onnauwkeurig. Een gebruikelijke methode is die der kleinste kwadraten (least squares method). Hierbij worden b_0 en b_1 zó bepaald dat

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i [y_i - b_0 - b_1(x_i - \bar{x})]^2 = \min_{\beta_0,\beta_1} \sum_i [y_i - \beta_0 - \beta_1(x_i - \bar{x})]^2 .$$

In woorden: de som van de kwadraten der residuen $(y_i - \hat{y}_i)$ is minimaal.
Het is verhelderend dit meetkundig te interpreteren:
Noteer: $y = (y_1,\ldots,y_n)'$; $x_0 = (1,\ldots,1)'$; $x - \bar{x} = (x_1-\bar{x},\ldots,x_n-\bar{x})'$, $\hat{y} = (\hat{y}_1,\ldots,\hat{y}_n)'$, alle vectoren in de $\mathbf{R}^n$. Dan is $\hat{y} = b_0x_0 + b_1(x - \bar{x})$, d.w.z. $\hat{y}$ is een lineaire combinatie van x_0 en $x - \bar{x}$.
$\sum_i (y_i - \hat{y}_i)^2 = |y - \hat{y}|^2$, de (lengte)^2 van de verschilvector $y - \hat{y}$, de vector der residuen.

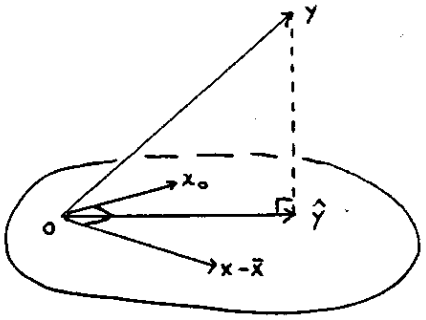


Fig. 8.2.6.

Het probleem is dus nu geworden: zoek in het vlak, opgespannen door de vectoren x_0 en $x - \bar{x}$, een vector $\hat{y}$ zodat $|y - \hat{y}|$ minimaal is. Dan is $\hat{y}$ juist de loodrechte projectie van y op dit vlak. Dus $y - \hat{y} \perp x_0$ en $\perp (x - \bar{x})$.
Zo ontstaan de normaalvergelijkingen

8.2.6. $$(y - \hat{y}, x_0) = 0$$
$$(y - \hat{y}, x - \bar{x}) = 0$$

oftewel

$$(y - b_0x_0 - b_1(x-\bar{x}) \, , x_0) = 0 \qquad (y,x_0) - b_0(x_0,x_0) - b_1(x-\bar{x},x_0) = 0$$

$$(y - b_0x_0 - b_1(x-\bar{x}) \, , x-\bar{x}) = 0 \qquad (y,x-\bar{x}) - b_0(x_0,x-\bar{x}) - b_1(x-\bar{x},x-\bar{x}) = 0 \, .$$

Nu is

$$(x_0,x_0) = n \; ; \quad (y,x_0) = \Sigma y_i \; ; \quad (x_0,x-\bar{x}) = \sum_i (x_i - \bar{x}) = 0 \, ,$$

zodat de normaalvergelijkingen worden:

8.2.7.
$$\Sigma y_i - nb_0 = 0 \qquad \text{oftewel} \quad b_0 = \Sigma y_i/n = \bar{y}_. ,$$

$$\Sigma y_i(x_i - \bar{x}) - b_1 \Sigma (x_i - \bar{x})^2 = 0 \qquad b_1 = \frac{\Sigma y_i(x_i - \bar{x})}{\Sigma (x_i - \bar{x})^2} \, .$$

b_1 wordt berekend met

$$b_1 = \frac{n \Sigma x_i y_i - \Sigma x_i \, \Sigma y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2} \, .$$

Opmerking. Dat we ons model in de vorm 8.2.4 in plaats van 8.2.2 schreven heeft tot gevolg gehad dat de normaalvergelijkingen eenvoudiger zijn geworden, dan is nl. $(x_0,x-\bar{x}) = 0$.
Natuurlijk is $a = b_0 - b_1\bar{x}$.
b_0 en b_1 had men ook kunnen vinden door de afgeleiden van
$\Sigma[y_i - \beta_0 - \beta_1(x_i - \bar{x})]^2$ naar β_0 en β_1 nul te stellen.
De som der residuen $\Sigma(y_i - \hat{y}_i) = 0$, nl $= (x_0,y-\hat{y}) = 0$ daar $(y - \hat{y}) \perp x_0$.

8.2.8. $\underline{b}_0$ en $\underline{b}_1$, en dus ook $\underline{a}$, zijn zuivere schatters voor β_0, β_1, α resp.

Bewijs.

$$\mathcal{E}\underline{b}_0 = \mathcal{E}\bar{\underline{y}} = \mathcal{E}(\beta_0 + \Sigma\underline{e}_i/n) = \beta_0$$

daar $\mathcal{E}\underline{e}_i = 0$ verondersteld is.

$$\mathcal{E}\underline{b}_1 = \mathcal{E} \frac{\Sigma[\beta_0 + \beta_1(x_i - \bar{x}) + \underline{e}_i](x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2} = \mathcal{E}[\beta_1 + \frac{\Sigma\underline{e}_i(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}] = \beta_1 \, ,$$

$$\mathcal{E}\underline{a} = \mathcal{E}(\underline{b}_0 - \underline{b}_1\bar{x}) = \beta_0 - \beta_1\bar{x} = \alpha \, . \qquad \square$$

Hieruit volgt tevens dat $\hat{\underline{y}}$ een zuivere schatter is voor $\mathcal{E}\underline{y}$, nl.

$$\mathcal{E}\hat{\underline{y}} = \mathcal{E}[\underline{b}_0 + \underline{b}_1(x-\bar{x})] = \beta_0 + \beta_1(x-\bar{x}) = \mathcal{E}\underline{y} \ .$$

8.2.9. Voor de <u>variantie</u> van deze schatters geldt:

$$\text{var } \underline{b}_0 = \text{var } \bar{\underline{y}} = \sigma^2/n \ ,$$

daar var $\underline{y}_i$ = var $\underline{e}_i = \sigma^2$ verondersteld is en $\text{cov}(\underline{e}_i,\underline{e}_j) = 0$.

$$\text{var } \underline{b}_1 = \frac{\Sigma(x_i-\bar{x})^2\sigma^2}{[\Sigma(x_i-\bar{x})^2]^2} = \frac{\sigma^2}{\Sigma(x_i-\bar{x})^2} \ ;$$

$\text{cov}(\underline{b}_0,\underline{b}_1) = 0$ (probeer dit zelf eens), d.w.z. $\underline{b}_0$ en $\underline{b}_1$ zijn ongecorreleerd, hetgeen <u>niet</u> geldt voor $\underline{a}$ en $\underline{b}$.

$$\text{var } \underline{a} = \text{var}(\underline{b}_0 - \underline{b}_1\bar{x}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i-\bar{x})^2}\right) \geq \text{var } \underline{b}_0 \ ;$$

$$\text{var } \hat{\underline{y}} = \text{var}[\underline{b}_0 + \underline{b}_1(x-\bar{x})] = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right) \ .$$

We zien dat var $\hat{\underline{y}}_i$ in het algemeen veel kleiner is dan var $\underline{y}_i$.
Hieruit zien we dat $\hat{y}$ het nauwkeurigst is in het zwaartepunt, nl. dan is $(x-\bar{x})^2 = 0$.
De berekende regressielijn gaat door het zwaartepunt $(\bar{x},\bar{y})$ van de puntenwolk, nl.

$$\hat{y} = b_0 + b_1(x-\bar{x}) = \bar{y} + b_1(x-\bar{x}) \ .$$

De regressiecoëfficiënt b_1 en de correlatiecoëfficiënt $r(\underline{x},\underline{y})$ hangen nauw samen, nl.

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{s_y}{s_x} r \quad \text{of ook wel} \quad \frac{\hat{y}-\bar{y}}{s_y} = r\,\frac{x-\bar{x}}{s_x} \ .$$

8.2.10. <u>Een schatter voor σ^2</u>.

Hiervoor moeten we gebruik maken van de som der kwadraten van de residuen

$$KS_r = \sum_i (y_i - \hat{y}_i)^2 \ ,$$

de zogenaamde restkwadratensom. Men kan nu bewijzen dat

$$\mathcal{E}\underline{KS}_r = (n-2)\sigma^2 \ .$$

Een zuivere schatter voor σ^2 krijgen we dus door de restkwadratensom te delen door het bijbehorende aantal vrijheidsgraden $\nu = n-2$. Dit aantal is gelijk aan het aantal waarnemingen n verminderd met het aantal aangepaste parameters 2. Dus

$$\underline{\hat{\sigma}}^2 = \underline{s}^2 := \frac{\underline{KS}_r}{n-2} = \frac{|\underline{y} - \underline{\hat{y}}|^2}{n-2} \ .$$

Onder de aanname dat de $\underline{e}_i$ simultaan normaal verdeeld zijn (en de overige aannamen uit 8.2.2) geldt

$$\nu\underline{s}^2/\sigma^2 \simeq \chi^2_\nu \ , \quad \nu = n-2 \ .$$

De restkwadratensom KS_r en de residuen r_i zijn belangrijke grootheden.
Aan de residuen kan men mede beoordelen in hoeverre een gekozen model redelijk past. Eén abnormaal hoge r_i kan reden zijn aan de juistheid van de betreffende waarneming te twijfelen; een systematisch verloop in de r_i's geeft een aanwijzing dat het gekozen model niet deugt en geeft reden naar andere modellen om te zien.

8.2.11. Betrouwbaarheidsintervallen en toetsen.

In WSK 49 (pag. 40) is bewezen dat een lineaire combinatie van onafhankelijke normaal verdeelde variabelen weer normaal verdeeld is.
Volgens 8.2.8 is

$$\underline{b}_0 = \beta_0 + \Sigma\underline{e}_i/n \quad \text{en} \quad \underline{b}_1 = \beta_1 + \frac{\Sigma\underline{e}_i(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2} \ .$$

Onder de aanname dat de $\underline{e}_i$ simultaan normaal verdeeld zijn en ongecorreleerd (dus ook onafhankelijk nu) is dus

$$\frac{\underline{b}_j - \beta_j}{\sigma(\underline{b}_j)} \sim N(0,1) \ , \quad j = 0,1 \ .$$

Meestal is σ^2 onbekend en vervangen door $\underline{s}^2$. Dan wordt $\sigma(\underline{b}_j)$ geschat door $s(\underline{b}_j)$. Dan geldt:

$$\frac{\underline{b}_j - \beta_j}{\underline{s}(\underline{b}_j)} \simeq \underline{t}_{n-2} ,$$

een Student-variabele met $n-2$ vrijheidsgraden. Analoog is

$$\frac{\hat{\underline{y}}_i - \mathcal{E}\hat{\underline{y}}_i}{\underline{s}(\hat{\underline{y}}_i)} \simeq \underline{t}_{n-2} , \quad i = 1,\ldots,n .$$

De bijbehorende betrouwbaarheidsintervallen worden dan, met betrouwbaarheid $1-\alpha$

$$\underline{b}_j - t_\nu(\tfrac{1}{2}\alpha)\underline{s}(\underline{b}_j) < \beta_j < \underline{b}_j + t_\nu(\tfrac{1}{2}\alpha)\underline{s}(\underline{b}_j) , \quad j = 0,1 , \ \nu = n-2$$

en

$$\hat{\underline{y}}_i - t_\nu(\tfrac{1}{2}\alpha)\underline{s}(\hat{\underline{y}}_i) < \mathcal{E}\underline{y}_i < \hat{\underline{y}}_i + t_\nu(\tfrac{1}{2}\alpha)\underline{s}(\hat{\underline{y}}_i) , \quad i = 1,\ldots,n .$$

Met behulp van S.C. 3.2 kan men, uitgaande van s^2, een betrouwbaarheidsinterval voor σ opstellen.

8.2.12. In 8.2.9 is de samenhang tussen de regressie- en de correlatiecoëfficiënt even genoemd. Laten we dit eens nader uitwerken:
Noteer $\bar{y} = (\bar{y}_\cdot,\ldots,\bar{y}_\cdot)'$, een vector met componenten $\bar{y}_\cdot$, dus $\bar{y} = \lambda x_0$ met $\lambda = \bar{y}$. Dus, daar nu $(y-\hat{y}) \perp (\bar{y}-\hat{y})$, zie fig. 8.2.6, geldt:

$$|y-\bar{y}|^2 = |\hat{y}-\bar{y}|^2 + |y-\hat{y}|^2 ,$$

ofwel in woorden:
KS_{tot} = KS tengevolge van regressie + KS_r, met bijbehorend aantal vrijheidsgraden,
$n-1 = 1 + (n-2)$.

Nu volgt uit 8.2.9:

$$|\hat{y}-\bar{y}|^2 = r^2 s_y^2 \frac{|x-\bar{x}|^2}{s_x^2} = r^2|y-\bar{y}|^2 ,$$

zodat

$$|y-\hat{y}|^2 = (1-r^2)|y-\bar{y}|^2 .$$

Is $r = 0$, dan is $KS_{tot} = KS_r$, is $r = 1$, dan is KS_{tot} = KS regressie.
De toetsingsgrootheid voor de hypothese H_0: $\beta = 0$ tegen H_1: $\beta \neq 0$ is

$$\frac{\underline{KS}\ \text{regressie}}{\underline{s}^2} = \frac{\underline{r}^2}{(1-\underline{r}^2)/(n-2)} \simeq \underline{F}^1_{n-2} \simeq \underline{t}^2_{n-2} \quad (\text{onder } H_0, \text{ vgl. } 8.1.10).$$

8.3. Een rekenvoorbeeld.

Gegeven de leeftijd x en bloeddruk y van 12 personen:

x:	56	42	72	36	63	47	55	49	38	42	68	60
y:	147	125	160	118	149	128	150	145	115	140	152	155

We gaan de regressielijn $\hat{y} = a + bx$ ($= b_0 + b(x - \bar{x})$) bepalen.
Is er een computerprogramma voor lineaire regressie aanwezig, of beschikt men over een electronische tafelrekenmachine, dan kan men deze gegevens zo invoeren. Anders verdient het aanbeveling de waarnemingen eerst te coderen, bijvoorbeeld: stel u := x - 50 en v := y - 140. Zo ontstaat:

u:	6	- 8	22	-14	13	- 3	5	- 1	-12	- 8	18	10
v:	7	-15	20	-22	9	-12	10	5	-25	0	12	15

Men kan nu natuurlijk de afgeleide formules voor de regressiecoëfficiënt e.d. gaan gebruiken. Het is echter handig de volgende "matrices" in te voeren. Hierin zijn x_0, x, y, $x - \bar{x}$, $y - \bar{y}$ de reeds eerder ingevoerde n-dimensionale vectoren.

	x_0	x	y
x_0	n	Σx	Σy
x	.	Σx^2	Σxy
y	.	.	Σy^2

A

	$x - \bar{x}$	$y - \bar{y}$
$x - \bar{x}$	$\Sigma(x - \bar{x})^2$	$\Sigma(x - \bar{x})(y - \bar{y})$
$y - \bar{y}$	.	$\Sigma(y - \bar{y})^2$

B

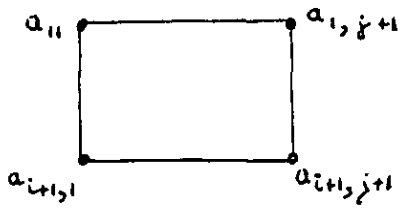
KS_r

C

Toelichting.

1) A is de matrix der ruwe produktsommen. De elementen zijn de inwendige produkten van de bijbehorende randvectoren. A is symmetrisch.

2) B is de matrix der gecorrigeerde produktsommen. De elementen zijn weer de inwendige produkten van de bijbehorende randvectoren. B is symmetrisch. Deze elementen berekent men uit die van A als volgt:



$$b_{ij} = a_{i+1,j+1} - a_{i+1,1}\, a_{1,j+1} / a_{11} .$$

Zo is dus

$$b_{12} = b_{21} = \Sigma xy - \frac{\Sigma x\ \Sigma y}{n} ,$$

etc.

3) Matrix C ontstaat op dezelfde manier uit matrix B. Deze heeft slechts 1 element en is juist KS_r.
Ga zelf na dat zo inderdaad $KS_r = \Sigma(y - \hat{y})^2$ ontstaat!

Toegepast op ons getallenvoorbeeld levert dit:

	x_0	u	v
x_0	12	28	4
u	.	1616	1774
v	.	.	2502

	$u - \bar{u}$	$v - \bar{v}$
$u - \bar{u}$	1550,67	1764,67
$v - \bar{v}$	.	2500,67

492,47

Stel nu

$$\hat{v} = c_0 + c_1(u - \bar{u}) .$$

Dan is

$$s^2 = KS_r/(n-2) = 49{,}247 ,$$

$$c_0 = \frac{a_{13}}{a_{11}} = \frac{4}{12} = 0{,}33 ,$$

$$c_1 = \frac{b_{12}}{b_{11}} = \frac{1764{,}67}{1550{,}67} = 1.138 ,$$

$$\bar{u} = \frac{a_{12}}{a_{11}} = \frac{28}{12} = 2{,}33 ,$$

$$s^2(c_1) = \frac{s^2}{b_{11}} = \frac{49{,}247}{1550{,}67} = 0{,}0318 ,$$

$$s(c_1) = 0{,}178 .$$

Oftewel:

$$\hat{v} = 0{,}33 + 1{,}138(u - 2{,}33) \ ,$$

$$\hat{y} - 140 = 0{,}33 + 1{,}138(x - 52{,}33) \ ,$$

$$\hat{y} = 80{,}78 + 1{,}138x$$

of m.b.v. afrondingsregels beter

$$\hat{y} = 81 + 1{,}14x \ ;$$

$$\hat{\text{var}}\ \underline{a} = s^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x - \bar{x})^2}\right] = 49{,}247\left[\frac{1}{12} + \frac{(52{,}33)^2}{1550{,}67}\right] = 91{,}07 \ .$$

Het 95%-betrouwbaarheidsinterval voor β wordt $(t_{10}(\frac{1}{2}\alpha) = 2{,}23)$

$$1{,}138 - 2{,}23(0{,}178) < \beta < 1{,}138 + 2{,}23(0{,}178) \ ,$$

$$0{,}74 < \beta < 1{,}53 \ .$$

De gebruikelijke vuistregel: bloeddruk = 100 + leeftijd, is dus niet zo gek.

8.4. Meervoudige lineaire regressie.

Eerst enkele definities.

Laat $\underline{y}$ een kolomvector zijn met n stochastische componenten $\underline{y}_i$ $(i = 1,\ldots,n)$. De verwachting $\mathbb{E}\underline{y}$ is de vector met componenten $\mathbb{E}\underline{y}_i$. De covariante-matrix van $\underline{y}$ (VAR($\underline{y}$)) is de matrix met elementen $\text{cov}(\underline{y}_i,\underline{y}_j)$. Er geldt:

$$\text{VAR}(\underline{y}) = \mathbb{E}\{(\underline{y} - \mathbb{E}\underline{y})(y - \mathbb{E}y)^T\}.$$

Ga na dat geldt:

$$\text{VAR}(A\underline{y}) = A.\text{VAR}(\underline{y}).\ A^T\ ,$$

waarbij A een niet-stochastische matrix is met n kolommen.

Het spoor van een vierkante matrix A (sp(A)) is gedefinieerd door

$$\text{sp}(A) = \sum_i A_{ii}\ .$$

Ga na dat geldt:

$$\text{sp}(A + B) = \text{sp}(A) + \text{sp}(B)\ ,$$

$$\text{sp}(AB) = \text{sp}(BA)\ .$$

De lengte of de Euclidische norm van een vector v is gedefinieerd door

$$\|v\| = \left(\sum_i v_i^2\right)^{\frac{1}{2}}$$

Er geldt:

$$\|v\|^2 = v^T v\ .$$

Bij meervoudige lineaire regressie luidt de regressievergelijking:

$$(1) \qquad \underline{y}_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \underline{e}_i\ ,\quad (i = 1,\ldots,n)\ ,$$

waarbij de x_{ij}'s de exact bekende zogenaamde instelwaarden, $\beta_1,\ldots,\beta_p$ de parameters met onbekende waarden en de $\underline{e}_i$'s de fouttermen zijn.

Als $x_{i1} = 1$ voor alle i, dan wordt β_1 het algemeen effect genoemd.

In vectornotatie kan (!) worden geschreven als

$$\underline{y} = X\beta + \underline{e} \ .$$

(X wordt wel de design-matrix genoemd.)

Aangenomen wordt dat

$$\mathbb{E}\,\underline{e} = 0 \text{ en } \mathrm{VAR}(\underline{e}) = \sigma^2 I \ .$$

De kleinste kwadraten schatter voor β is die waarde van β die $\|y - X\beta\|^2$ minimaliseert. Omdat $X\beta$ een lineaire combinatie is van de kolommen van X (ga na!) komt het minimaliseren van $\|y - X\beta\|^2$ neer op het loodrecht projecteren van y op de ruimte $<X>$ opgespannen door de kolommen van X (zie figuur 1).

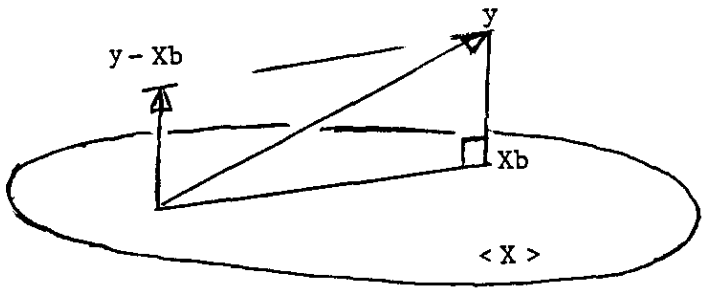


fig.1

Dus: $y - Xb \perp X$.

Oftewel: $(y - Xb)^T X = 0$ (de normaalvergelijkingen).

Als X^TX regulier is volgt daaruit:

$$b = (X^TX)^{-1} X^T y \ .$$

Er geldt:

$$\mathbb{E}\,\underline{b} = \mathbb{E}\,((X^TX)^{-1} X^T \underline{y}) = (X^TX)^{-1} X^T \mathbb{E}\,\underline{y} = (X^TX)^{-1} X^TX\beta = \beta \ ,$$

($\underline{b}$ is een zuivere schatter voor β).

Tevens:

$$\mathrm{VAR}(\underline{b}) = (X^TX)^{-1} X^T \mathrm{VAR}(\underline{y})\, X(X^TX)^{-1} = (X^TX)^{-1} \sigma^2 .$$

Hoe kan σ^2 geschat worden?
Er geldt:

$$KS_r = \|y - Xb\|^2 = \|y - X(X^TX)^{-1}X^Ty\|^2 = \|Py\|^2 ,$$

waarbij $P = I - X(X^TX)^{-1}X^T$.
Er geldt $P = P^T$, $PP = P$ en $\mathbb{E}\,P\underline{y} = 0$ (ga na!), en dus:

$$\mathbb{E}\,\underline{KS}_r = \mathbb{E}\,((P\underline{y})^TP\underline{y}) = \mathbb{E}\,\mathrm{sp}((P\underline{y})^TP\underline{y}) =$$

$$= \mathrm{sp}\;\mathbb{E}\,((P\underline{y})^TP\underline{y}) = \mathrm{sp}\;\mathbb{E}\,(P\underline{y}(P\underline{y})^T) =$$

$$= \mathrm{sp}\;\mathrm{VAR}(P\underline{y}) = \sigma^2 \cdot \mathrm{sp}(PP^T) = \sigma^2\,\mathrm{sp}(P) =$$

$$= \sigma^2\,\mathrm{sp}(I - X(X^TX)^{-1}X^T) = \sigma^2(n - \mathrm{sp}(X(X^TX)^{-1}X^T)) =$$

$$\sigma^2(n - \mathrm{sp}((X^TX)^{-1}X^TX)) = \sigma^2(n - p) .$$

Een zuivere schatter voor σ^2 is dus:

$$\underline{\hat{\sigma}}^2 = \underline{s}^2 = \underline{KS}_r/(n - p) .$$

Betrouwbaarheidsintervallen voor de componenten van β worden geconstrueerd zoals elders beschreven voor de parameters van het enkelvoudige regressiemodel, waarbij we moeten bedenken dat

$$\mathrm{var}(\underline{b}_j) = ((X^TX)^{-1})_{jj}\sigma^2 .$$

Voor de constructie van betrouwbaarheidsintervallen voor lineaire combinaties van componenten hebben we nodig dat

$$\mathrm{var}(a^T\underline{b}) = a^T\mathrm{VAR}(\underline{b})\,a = a^T(X^TX)^{-1}a\cdot\sigma^2 .$$

Voorbeeld: Aan een kg lakverf wordt x gram van een stof toegevoegd, die het drogen moet bevorderen. De fabrikant vermeldt dat met een toevoeging van 5g. per kg. verf de droogtijd 6 uur bedraagt. Om dat na te gaan zetten we een experiment zodanig op, dat we met een gerust hart de boven beschreven theorie kunnen toepassen .

We nemen als model voor de waargenomen droogtijd aan:

$$\underline{y} = \beta_1 + \beta_2 x + \beta_3 x^2 + \underline{e}$$

en we veronderstellen dat meetfouten onderling onafhankelijk en identiek normaal verdeeld zijn. Het experiment resulteerde in de waarnemingen:

y	12.0	10.5	10.0	8.0	7.0	8.0	7.5	8.5	9.0
x	0	1	2	3	4	5	6	7	8

De design-matrix wordt gegeven door:

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 0 & 1 & 4 & 9 & 16 & 25 & 36 & 49 & 64 \end{pmatrix}$$

zodat

$$X^TX = \begin{pmatrix} 9 & 36 & 204 \\ 36 & 204 & 1296 \\ 204 & 1296 & 8772 \end{pmatrix}, \quad (X^TX)^{-1} = \begin{pmatrix} .6606 & -.3091 & .03030 \\ -.3091 & .2245 & -.02597 \\ .03030 & -.02597 & .003247 \end{pmatrix},$$

$y^TX = (80.5,\ 299.0,\ 1697.0)$, $b^T = y^TX(X^TX)^{-1} = (12.176,\ -1.8281,\ .18428)$

$KS_r = \|y - \hat{y}\|^2 = \|y - Xb\|^2 = 1.738$

$s^2 = KS_r/6 = .2897$ en $s = \sqrt{s^2} = .5382$.

Opmerking: Een ogenschijnlijk handige formule voor de berekening van KS_r wordt gegeven door

$$KS_r = \|Py\|^2 = y^TPy = y^Ty - y^TXb \ .$$

Echter, deze manier van berekening is alleen betrouwbaar als bij de tussenresultaten extra decimale cijfers meegenomen worden.

In termen van ons model komt de claim van de fabrikant op:

$$\beta_1 + 5\beta_2 + 25\beta_3 = 6 \ .$$

Onder deze hypothese is $\underline{b}_1 + 5\underline{b}_2 + 25\underline{b}_3$ normaal verdeeld met gemiddelde 6 en variantie

$$(1 \quad 5 \quad 25)(X^TX)^{-1} \ (1 \quad 5 \quad 25)^T\sigma^2 = .8084\,\sigma^2 .$$

Een 95%-betrouwbaarheidsinterval voor $\beta_1 + 5\beta_2 + 25\beta_3$ wordt gegeven door:

$$b_1 + 5b_2 + 25b_3 \pm t_6(.025)* \ .8991*s =$$

$$7.64 \pm 2.45*.8991*.5382 = 7.64 \pm 1.19 .$$

Dit interval bevat de door de fabrikant gegeven waarde niet. De waarnemingen spreken zijn claim tegen!

9. Variantie-analyse.

Dit onderwerp is te groot om in zijn geheel te behandelen en we zullen ons daarom beperken tot 1 factor met herhalingen.
We gaan de theorie van 4.3 voor twee series waarnemingen uitbreiden tot die voor k series waarnemingen.

9.1. De series bestaan uit een ongelijk aantal waarnemingen.

Laten we dit aan een voorbeeld uitwerken:
Een laboratorium bepaalt volgens 4 methoden het zwavelgehalte van steenkool.
De gecodeerde waarnemingen en tevens de nodige rekengrootheden zijn gegeven in tabel 9.1.1.
Voorbeeld:

methode	waarnemingen	n_i	$x_{i.}$	$\sum_j x_{ij}^2$	$x_{i.}^2/n_i$	KS_i
I	13 15 20	3	48	794	768,00	26,00
II	20 19 18 27 24	5	108	2390	2332,80	57,20
III	14 18 13	3	45	689	675,00	14,00
IV	18 20 22 14 9 10 10	7	103	1685	1515,57	169,43
		$n_. = 18$	$x_{..} = 304$	5558	5291,37	266,63

Tabel 9.1.1. De waarnemingen geven we weer als x_{ij} met $i = 1,2,3,4$;
$j = 1,\ldots,n_i$.
Verder is $x_{i.} = \sum_j x_{ij}$; $x_{i.}^2 = (x_{i.})^2$; $KS_i = \sum_j x_{ij}^2 - x_{i.}^2/n_i$;
$\nu_i = n_i - 1$; $x_{..} = \sum_{i,j} x_{ij}$; $n_. = \sum_i n_i$.

In principe zijn de 4 modellen (4.3.1) t/m (4.3.4) weer alle mogelijk.

9.1.1. De nulhypothese H_0: $\sigma_i^2 = \sigma^2$, $i = 1,\ldots,k$.

Deze kan getoetst worden met Bartlett's toets die gebaseerd is op de benadering

$$\frac{1}{c}\left[\nu \; {}^{e}\log \underline{s}^2 - \Sigma(\nu_i \; {}^{e}\log \underline{s}_i^2)\right] \approx \chi_{k-1}^2$$

met

$$\nu = \Sigma\nu_i \; ; \quad \underline{s}^2 = \Sigma\nu_i \, \underline{s}_i^2/\nu \quad \text{en} \quad c = 1 + \frac{1}{3(k-1)} \left(\Sigma \; 1/\nu_i - 1/\nu\right) .$$

Met de waarnemingen uit ons voorbeeld ontstaat tabel 9.1.2. Bovendien is

$$s^2 = \Sigma KS_i/\nu = 266{,}63/14 = 19{,}0 \; ; \quad \nu\,{}^{10}\log s^2 = 17{,}906 \; ;$$

$$c = 1{,}15 \; ; \quad {}^{e}\log 10 = 2{,}30 \; .$$

ν_i	$s_i^2 = KS_i/\nu_i$	${}^{10}\log s_i^2$	$\nu_i\,{}^{10}\log s_i^2$
2	13,0	1,114	2,228
4	14,3	1,155	4,620
2	7,0	0,845	1,690
6	28,2	1,450	8,700
$\nu = 14$			17,238

Tabel 9.1.2.

Zodat

$$\chi_3^2 = \frac{2{,}30}{1{,}15}\,(17{,}906 - 17{,}238) = 1{,}36 \; .$$

Nu is $P(\chi_3^2 > 2.37) = 0{,}50$, zodat we kunnen concluderen: er is geen enkele reden om een verschil in de σ_i^2 aan te nemen.
Hierbij dient te worden opgemerkt dat in de praktijk deze toets vaak achterwege wordt gelaten en de hypothese 9.1.1 als onderstelling wordt aanvaard, en wel:

1) omdat de series zijn verkregen onder omstandigheden die slechts in geringe mate verschillen en men daarom geen verschillen in de σ_i verwacht;
2) omdat het aan de series zelf direkt te zien is dat Bartlett's toets geen significante uitkomst zal geven; alleen zeer in het oog lopende verschillen in de spreidingen kunnen tot verwerpen van de hypothese 9.1.1 leiden;
3) omdat de volgende toets op de hypothese H_0: $\mu_i = \mu$, $i = 1,\ldots,k$, niet ernstig wordt beïnvloed door enigszins verschillende waarden van de σ_i's.

9.1.2. <u>De nulhypothese</u> H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$.

Dit is de belangrijkste hypothese, onder de onderstelling $\sigma_i = \sigma$ $(i = 1,\ldots,k)$.
Deze toets staat bekend als de zg. <u>variantie-analyse</u>.
We splitsen de waarnemingen als volgt:

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

oftewel de totale afwijking is gesplitst in twee componenten, nl.
$x_{ij} - \bar{x}_{i.}$ = de afwijking <u>binnen series</u> en $\bar{x}_{i.} - \bar{x}_{..}$ = de afwijking <u>tussen series</u>.
Daar deze componenten orthogonaal zijn, d.w.z. $\sum_{i,j} (x_{ij} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{x}_{..}) = 0$, geldt

$$\sum_{i,j} (x_{ij} - \bar{x}_{..})^2 = \sum_{i,j} (x_{ij} - \bar{x}_{i.})^2 + \sum_{i,j} (\bar{x}_{i.} - \bar{x}_{..})^2 ,$$

of in woorden:
de totale KS_T = KS_b binnen series + KS_t tussen series, en
$\nu_T = n_. - 1 = \nu_b + \nu_t$ met $\nu_b = \Sigma\nu_i = n_. - k$ en $\nu_t = k - 1$, de bijbehorende splitsing in het aantal vrijheidsgraden.

<u>Opmerking</u>. $\bar{x}_{..} = x_{..}/n_.$; $\bar{x}_{i.} = (x_{i.})/n_i$.

De statistische theorie leert nu dat onder het model (4.3.1) $\underline{x}_{ij} = \mu + \underline{u}_{ij}\sigma$, de zg. <u>gemiddelde kwadraten</u> (GK := KS/ν)

$$\underline{GK}_b = \underline{KS}_b/\nu_b \quad \text{en} \quad \underline{GK}_t = \underline{KS}_t/\nu_t$$

onderling onafhankelijke zuivere schatters zijn van σ^2, en beide een $\sigma^2\chi^2_\nu/\nu$-verdeling bezitten (met $\nu = \nu_b$ of ν_t).
Onder het alternatieve model (4.3.2): $\underline{x}_{ij} = \mu_i + \underline{u}_{ij}\sigma$ geldt echter

$$\mathcal{E}\underline{GK}_b = \sigma^2$$

doch

9.1.3. $$\mathcal{E}\underline{GK}_t = \frac{\Sigma n_i(\mu_i - \bar{\mu})^2}{k-1} + \sigma^2 \text{ , dus groter dan } \sigma^2 .$$

Hierin is $\bar{\mu} := \Sigma n_i\mu_i/\Sigma n_i$.
Het model (4.3.1) wordt daarom getoetst tegen het alternatief (4.3.2) door als toetsingsgrootheid te nemen:

$$\underline{GK}_t/\underline{GK}_b = \underline{F}^{\nu_t}_{\nu_b} \quad \text{(een onder } H_0 \text{ F-variabele).}$$

Dit is een éénzijdige toets: alleen als $GK_t > GK_b$ is kan model (4.3.1) worden verworpen.
De KS berekenen we als volgt:

$$KS_b = \sum_{i,j} x_{ij}^2 - \sum_i x_{i.}^2/n_i \ ,$$

$$KS_t = \sum_i x_{i.}^2/n_i - x_{..}^2/n_. \ .$$

Deze zijn deels reeds berekend in tabel 9.1.1, zodat

$$KS_b = 5558 - 5291{,}37 = 266{,}63 \ ,$$

$$KS_t = 5291{,}39 - (304)^2/18 = 157{,}15 \ .$$

Zo ontstaat de volgende variantie-analyse:

Variatiebron	KS	ν	GK	$\mathcal{E}_{GK}$	F
tussen series	157,15	3	52,38		2,75
binnen series	266,63	14	19,04	σ^2	
$F_{14}^3 = 52{,}38/19{,}04 = 2{,}75 < 3{,}34$ (zie S.C. 4.1 met $\alpha = 0.05$)					

Tabel 9.1.3. Variantie-analyse behorende bij de waarnemingen uit tabel 9.1.1.

Conclusie: er is geen reden het model (4.3.1) of de H_0: $\mu_i = \mu$, te verwerpen.

9.2. De series bestaan uit een gelijk aantal waarnemingen.

De berekeningen worden nu eenvoudiger.
We werken dit weer uit aan de hand van een voorbeeld:
Acht laboratoria doen elk 3 waarnemingen. De gecodeerde waarnemingen en tevens de nodige rekengrootheden zijn gegeven in tabel 9.2.1.

Lab	x_{ij}			$x_{i.}$	$x_{i.}^2$	$\sum_j x_{ij}^2$
A	26	26	20	72	5184	1752
B	20	20	20	60	3600	1200
C	37	36	38	111	12321	4109
D	16	22	23	61	3721	1269
E	20	19	18	57	3249	1085
F	27	31	29	87	7569	2531
G	27	24	23	74	5476	1834
H	25	26	26	77	5929	1977
				$x_{..} = 599$	47049	15757

Tabel 9.2.1.

Analoog vinden we voor de KS (nu met $n_i = 3$):

$$KS_b = 15757 - 47049/3 = 74{,}00$$

$$KS_t = 47049/3 - (599)^2/24 = 732{,}96 \ .$$

Als we definiëren

$$\sigma_L^2 := \sum_i \frac{(\mu_i - \bar{\mu}_.)^2}{k - 1}$$

met $\bar{\mu}_. = \Sigma\mu_i/k$, $n = n_i$, ontstaat de volgende variantie-analyse:

Variatiebron	KS	ν	GK	$\mathcal{E}(\underline{GK})$	F
tussen laboratoria	732,96	7	104,71	$3\sigma_L^2 + \sigma^2$	22,7
binnen laboratoria	74,00	16	4,62	σ^2	
$F_{16}^{7} = 104{,}71/4{,}62 = 22{,}7 >> 2{,}66$ (zie S.C. 4.1 met $\alpha = 0.05$; deze F-waarde is zelfs significant bij $\alpha = 0.005$)					

Tabel 9.2.2. Variantie-analyse behorende bij de waarnemingen uit tabel 9.2.1.

Conclusie: de verschillen tussen de laboratoria zijn groter dan kan worden verklaard als gevolg van de toevallige fouten binnen de laboratoria.

Het is uit tabel 9.2.1 direkt duidelijk dat dit vooral te wijten is aan lab. C dat veel hogere uitkomsten geeft dan de overige laboratoria.
Voor dit geval vereenvoudigt 9.1.3 tot

$$\mathcal{E}_{\underline{GK}_t} = n \frac{\Sigma(\mu_i - \bar{\mu}_.)^2}{k - 1} + \sigma^2 = n\sigma_L^2 + \sigma^2 \ .$$

9.3. Tweeweg-variantieanalyse.

Tot nu toe hebben we slechts 1 factor met herhalingen bestudeerd. Thans beschouwen we proeven met 2 kwalitatieve factoren, eventueel met herhalingen. Hierbij beperken we ons echter tot "gekruiste" factoren, d.w.z. elk niveau van de ene factor komt voor bij elk niveau van de tweede factor. Bovendien nemen we aan dat het aantal herhalingen "per cel" constant is. Zo'n proefschema wordt een orthogonaal proefschema genoemd.

Stel we hebben een proef met 2 factoren: A en B op resp. a en b vaste niveaus en met n herhalingen per cel. Met vaste niveaus bedoelt men: bij herhaling van de proef komen A en B weer op dezelfde niveaus voor. Zonder herhaling betekent dus n = 1.

We hebben nu het volgende model:

$$\underset{\sim}{x}_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \underset{\sim}{e}_{ijk} , \qquad \begin{matrix} i = 1,\ldots,a \\ j = 1\ldots,b \\ k = 1,\ldots,n \end{matrix}$$

Hierin is: $\underset{\sim}{x}_{ijk}$ de k^e herhaling van factor A op niveau i en factor B op niveau i.

μ het algemeen gemiddelde.

α_i het hoofdeffect A, β_j het hoofdeffect B en $(\alpha\beta)_{ij}$ de interactie tussen A en B

$\underset{\sim}{e}_{ijk}$ de stochastische fout per cel; var $\underset{\sim}{e}_{ijk} = \sigma_0^2$ (onafhankelijk i,j)

μ, α_i, β_j en σ_0^2 zijn de parameters van het model. Om deze te kunnen schatten leggen we de betrekkingen op:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0 .$$

Een praktijk voorbeeld:
4 merken autobanden (factor A, a = 4) wil een consumentenbond vergelijken wat betreft slijtage. Daar dit ook erg afhangt van het automerk (schokbrekers e.d.), voert men een tweede factor in, B = automerk b.v. drie merken, (factor B, b = 3). Om een interaktie te kunnen beschouwen (samenhang merk band, auto) herhaalt men deze proef.
In werkelijkheid zal men deze proef veel "grootser" moeten opzetten, daar slijtage ook samenhangt met: wegdek, rijstijl, snelheid, bandenspanning e.d.

Voor deze, meer-factoren-schemas, wordt vewezen naar het college "Proef-opzetten".
De variantie-analyse geeft nu antwoord op de volgende vragen:
Is er een (significant) verschil tussen de bandenmerken wat betreft slijtage?
Is er een (significant) verschil in slijtage bij de verschillende automerken?
Reageren alle bandenmerken hetzelfde bij verschillende automerken m.a.w. is er een "significante" interaktie bandmerk-automerk?

Kleinste kwadraten-schattingen van de parameters zijn:

$$\hat{\mu} = \bar{x}_{...} \; ; \; \hat{\alpha}_i = \bar{x}_{i..} - \bar{x}_{...} \; ; \; \hat{\beta}_j = \bar{x}_{.j.} - \bar{x}_{...} \; ;$$

$$(\hat{\alpha\beta})_{ij} = \bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...}$$

Deze voldoen aan de relaties die aan de corresponderende parameters zijn opgelegd. Hierin is $\bar{x}_{i..} = x_{i../bn}$; $\bar{x}_{ij./n}$ etc.

Er geldt nu:

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

oftewel

$$x_{ijk} - \hat{\mu} = \hat{\alpha}_i + \hat{\beta}_j + (\hat{\alpha\beta})_{ij} + (x_{ijk} - \bar{x}_{ij.})$$

Omdat de dubbele producten 0 geven is

$$\sum_{ijk} (x_{ijk} - \bar{x}_{...})^2 = nb \sum_i \hat{\alpha}_i^2 + na \sum_j \hat{\beta}_j^2 + n \sum_{ij} (\hat{\alpha\beta})_{ij}^2 + \sum_{ijk} (x_{ijk} - \bar{x}_{ij.})^2$$

of uitgedrukt in kwadratensommen:

$$KS_{tot} = KS_A + KS_B + KS_{AB} + KS_r \; .$$

KS_r = restkwadratensom. De splitsing van het aantal bijbehorende vrijheidsgraden is:

$$abn - 1 = (a - 1) + (b - 1) + (a - 1)(b - 1) + ab(n - 1) \; .$$

We definiëren, alhoewel het geen echte varianties zijn:

$$\sigma_A^2 := \sum_i^2 \alpha_i^2/(a-1) \; ; \; \sigma_B^2 := \sum_j \beta_j^2/(b-1) \; ; \; \sigma_{AB}^2 := \sum_{ij} (\alpha\beta)_{ij}^2/(a-1)(b-1) \; .$$

Om te kunnen toetsen moeten we aannamen doen omtrent de verdeling der $\underline{e}_{ijk}$. Hierbij onderstellen we meestal dat de $\underline{e}_{ijk}$ alle onderling onafhankelijk zijn en normaal verdeeld met gemiddelde 0 en variantie σ_0^2 (de restvariantie). We kunnen de volgende variantie-analysetabel opstellen:

variatiebron	KS'	ν	GK	EGK
factor A	KS_A	a-1	GK_A	$\sigma_0^2 + nb\sigma_A^2$
factor B	KS_B	b-1	GK_B	$\sigma_0^2 + na\sigma_B^2$
interaktie AB	KS_{AB}	(a-1)(b-1)	GK_{AB}	$\sigma_0^2 + n\sigma_{AB}^2$
rest	KS_t	ab(n-1)	GK_r	σ_0^2
Totaal	KS_{tot}	abn-1		

Hierin is $GK_A = KSA/(a-1)$ etc., de gemiddelde kwadratensom van factor A. EGK is de verwachting van de corresponderende gemid.K.S.

Men kan nu bewijzen dat, vanwege de orthogonaliteit van het proefschema en de aanname betreffende de verdeling der $\underline{e}_{ijk}$, de GK alle onderling onafhankelijke $\sigma^2\underline{\chi}^2_\nu/\nu$ - verdeelde grootheden zijn. Deze zijn x^2-stochasten onder de bijbehorende H_0, zodat quotiënten een F-verdeling geven.

Zo wordt de nulhypothese $H_0 : \alpha_1 = \alpha_2 = \ldots = 0$ oftewel $\sigma_A^2 = 0$ (in woorden: er is geen verschil tussen de niveaus van factor A) getoetst met

$F^{\nu_1}_{\nu_2} = \dfrac{GK_A}{GK_r}$ met $\nu_1 = a - 1$ en $\nu_2 = ab(n-1)$. Analoog worden de nulhypothesen

$\sigma_B^2 = 0$ en $\sigma_{AB}^2 = 0$ getoetst met resp. $F = GK_B/GK_r$ en $F = GK_{AB}/GK_r$.

Dit zijn éénzijdige toetsen.

Een zuivere schatting voor σ_0^2, de restvariantie, is dus GK_r.

De kwadratensommen KS berekent men als volgt:

$$KS_A = \sum_{ijk} (\bar{x}_{i..} - \bar{x}_{...})^2 = \sum_{i} x_{i..}^2/bn - x_{...}^2/abn = S_A - S_0 ,$$

$$KS_B = \sum_{ijk} (\bar{x}_{.j.} - \bar{x}_{...})^2 = \sum_{j} x_{.j.}^2/an - x_{...}^2/abn = S_B - S_0 ,$$

$$KS_{AB} = \sum_{ijk} (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2 = \sum_{ij} x_{ij.}^2/n - S_A - S_B + S_0 ,$$

$$KS_r = \sum_{ijk} (x_{ijk} - \bar{x}_{ij.})^2 = \sum_{ijk} x_{ijk}^2 - \sum_{ij} x_{ij.}^2/n .$$

Controle: $KS_{tot} = \sum_{ijk} x_{ijk}^2 - S_0$.

Een rekenvoorbeeld:
Faktor A op a = 4 niveaus, factor B op b = 3 niveaus en n = 2 (men zegt: 2 herhalingen. Dit is verwarrend daar men 1 x"herhaalt").
De waarnemingen zijn:

	j→						$x_{i..}$	$x_{i..}^2$	
	2	2	10	8	6	7	35	1225	
	12	9	12	15	9	10	67	4489	
i↓	12	12	12	13	15	15	79	6241	
	5	5	10	11	8	2	41	1681	
$x_{.j.}$	59		91		72		222	13636	$= \sum x_{i..}^2$
$x_{.j.}^2$	3481		8261		5184		→ 16946		

$S_0 = x_{...}^2/abn = 222^2/24 = 2053{,}50$,

$KS_A = 13636/6 - S_0 = 219{,}17$,

$KS_B = 16946/8 - S_0 = 64{,}75$,

$KS_{AB} = (4^2+21^1+24^2+\ldots+10^2)/2 - S_A - S_B + S_0 = 2391 - S_A + S_0 = 53{,}58$,

$KS_{tot} = 2^2 + 2^2 + 12^2 + 9^2 + \ldots + 8^2 + 2^2 - S_0 = 2422 - S_0 = 368{,}50$.

$KS_r = 2422 - 2391 = 31{,}00$.

De variantie-analyse wordt nu:

Variatiebron	KS	ν	GK	EGK
factor A	219,17	3	73,1	$\sigma_0^2 + 6\sigma_A^2$
factor B	64,75	2	32,4	$\sigma_0^2 + 8\sigma_B^2$
interactie AB	53,58	6	8,9	$\sigma_0^2 + 2\sigma_{AB}^2$
rest	31,00	12	2,6	σ_0^2
Totaal	368,50	23		

De toets $H_0 : \sigma_A^2 = 0$ geeft: $F_{12}^3 = 28{,}3 \gg 3{,}49$ dus significant A-effect, ($\alpha = 0.05$)

$H_0 : \sigma_B^2 = 0$ geeft: $F_{12}^2 = 12{,}5 > 3{,}89$, significant B-effect

$H_0 : \sigma_{AB}^2 = 0$ geeft: $F_{12}^6 = 3{,}5 > 3$, significante interactie.

Een schatting voor de restvariatie σ_0^2 is: $S_0^2 = 2{,}6$ met $\nu = 12$.

Opmerking: als men per cel geen herhalingen doet, dus n = 1, kan men uiteraard geen rest KS etc. berekenen. Verwacht men totaal geen interactie, dan kan men deze als "rest" definiëren. We hebben dan als KS: KS_A, KS_B en KS_r met resp. a, b en (a-1)(b-1) vrijheidsgraden.

10. Steekproeven

10.0. Inleiding

In de praktijk van de statistiek hebben we meestal te maken met onvolledig bekende modellen. Soms is van een te bestuderen populatie (stochast) de verdeling bekend op 1 of meer parameters na. We hebben dan waarnemingen nodig om deze onbekende parameters te schatten. Een belangrijk deel in de toegepaste statistiek is gewijd aan het toetsen van hypothesen, dat zijn statistische onderstellingen betreffende een populatie, b.v. de fractie defecten in de populatie is hoogstens 5%, de gemiddelde levensduur van een partij lampen is minstens 1000 branduren, e.d.

Ook hierbij zullen we steeds waarnemingen moeten verrichten oftewel een steekproef moeten nemen om de toetsen te kunnen uitvoeren. Onderzoekt men de gehele populatie, dan heeft men exact de waarde van de onbekende parameter gevonden. Maar, afgezien van het feit dat dit vaak te tijdrovend en/of te duur is, is het ook vaak onmogelijk. O.a. bijv. bij het bepalen van levensduur treksterkte, vermoeidheid, e.d. Na de waarneming is het element kapot. Men zal zich dus veelal tevreden moeten stellen met slechts een deel, een steekproef, van de populatie in het onderzoek te betrekken. Dit gaat natuurlijk ten koste van de nauwkeurigheid waarmee we een uitspraak betreffende de populatie willen doen. Niet alleen bij het toetsen heeft men steekproeven nodig. Ook in andere situaties o.a. bij:

1) Kwaliteitscontrole en procesbeheersing, waarbij met kleine tussenpozen uit de lopende productie een kleine steekproef wordt genomen teneinde tijdig vast te stellen wanneer het proces ontregeld is en moet worden bijgeregeld.
2) Partijkeuringen in de industrie
3) Modelanalyse en opiniepeilingen Ook hierbij is het ondoenlijk, te kostbaar en te tijdrovend om de gehele populatie te ondervragen.
   Wil men iets te weten komen over het gebruik van een wasmiddel door huisvrouwen, dan zal men slechts een deel van alle huisvrouwen benaderen.

Uit het voorgaande is duidelijk dat het belangrijk is iets te weten over "Steekproeftechnieken". Het is nl. niet zo eenvoudig als men wellicht zou denken. Men moet nl. m.b.v. een steekproef conclusies trekken (dus extrapoleren naar) over de gehele populatie. Deze moet dus "representatief" zijn. Hier duiken al veel moeilijkheden op.

Om te beginnen moet men de te bestuderen populatie (= universum) goed definiëren. Als men iets te weten wil komen over de besteding van het gezinsinkomen, moet men eerst goed definiëren wat een gezin is. Ook de vrijgezellen, samenwonenden e.d.? Nadat het universum goed gedefinieerd is, komt direct de vraag op: Hoe moeten we een steekproef trekken, hoe groot zal deze moeten zijn?

Het is niet eenvoudig hier een éénduidig antwoord op te geven. De steekproefkosten, de gewenste nauwkeurigheid, de grootte van de populatie en eventuele beperkte mogelijkheden, zullen daarbij zeker een rol spelen. Daar de steekproeftheorie zeer uitgebreid is, zullen we ons slechts tot enkele technieken beperken.
Voor diepgaande studie wordt verwezen naar:

[1] Moors, Muilwijk: Steekproeven, Agon Elsevier A'dam (1975), zeer elementair met veel informatie.

[2] SOM : A manual of sampling techniques, Heinemann Educ. Books, London 1973.
Elementair

[3] Cochran : Sampling techniques, John Wiley, London 1963.
Bevat meer wisk. bewijzen.

1.1. Enkelvoudige steekproeven. (simple random sampling)

In het vervolg bestuderen we eindige populaties ter grootte N. Voor zo'n populatie willen we de stochastische grootheid $\underline{X}$ bestuderen. Deze neemt de waarden aan $X_1, X_2, \ldots, X_N$, waaronder dezelfde kunnen zijn.

Nu is bekend dat i.h.a. voor een discrete stochast $\underline{x}$ met waarden $x_1, \ldots, x_n$ gedefinieerd zijn: het populatiegemiddelde $\mu = \mathcal{E}\underline{x} = \sum_1^n p_i x_i$ en de populatievariantie

$$\sigma^2 = \mathcal{E}(\underline{x} - \mu)^2 = \sum_1^n p_i (x_i - \mu)^2 .$$

Voor een eindige populatie is $p_i = f_i/N$ waarbij f_i het aantal keren is dat waarde x_i optreedt. Telt men deze f_i-waarden ook mee in de populatie, dan bereikt men hiermee dat $P(\underline{X} = X_i) = 1/N$ voor alle $i = 1, \ldots, N$.

Dan wordt $\mu = \frac{1}{N}\sum_{1}^{N} X_i = \overline{X}$. Analoog $\text{var}^{*}\underline{X} = \sigma_*^2 = \frac{1}{N}\sum_{i}(X_i - \mu)^2$.

In het vervolg gebruiken we ook de definitie $\text{var}\,\underline{X} = \sigma^2 = \sum(X_i - \mu)^2/(N-1)$. Dit blijkt bij trekking zonder teruglegging eenvoudigere resultaten te geven

Er geldt de relatie $\frac{\sigma_*^2}{N-1} = \frac{\sigma^2}{N}$.

Voor grote N zijn deze vrijwel gelijk.

We zullen ons voornamelijk bezighouden met het schatten van het populatiegemiddelde μ ($\overline{X}$ in literatuur).

Gegeven een eindige populatie ter grootte N: $X_1,\ldots,X_N$. Hieruit nemen we een enkelvoudige, aselecte steekproef ter grootte n: $x_1,\ldots,x_n$.

$f = \frac{n}{N}$ heet de steekproeffractie.

"Enkelvoudig" slaat op het feit dat de trekking direct plaats vindt uit de gehele populatie zonder verdere indeling vooraf.

"Aselect", omdat bij het trekken geen selectie wordt toegepast, m.a.w. de kans om in de steekproef opgenomen te worden is voor elk element even groot.

Definitie: Een enkelvoudige steekproef ter grootte n uit een populatie met verdelingsfunctie F is een n-dim.stochastische vector $\underline{x} = (\underline{x}_1,\ldots,\underline{x}_n)$ van n stochasten, alle met dezelfde verdelingsfunctie F.

Gebeurt trekking met teruglegging, dan zijn alle $\underline{x}_i$ 0.0 (aselecte steekproef).

Gebeurt trekking zonder teruglegging, wat meestal het geval is, dan zijn deze $\underline{x}_i$ niet meer 0.0.

Een goede, zuivere schatter voor μ is $\overline{\underline{x}} = \sum_{1}^{n} \underline{x}_i/n$, het steekproefgemiddelde.

Wat is de variantie van deze schatter?

i) met teruglegging

Trekken we met terugleggen, dan zijn de $\underline{x}_i$ alle onderling onafhankelijk en geldt:

$$\text{var}\,\overline{\underline{x}} = \frac{1}{n^2}\sum \text{var}\,\underline{x}_i = \sigma_*^2/n$$

ii) Zonder teruglegging

Nu blijkt

$$\text{var}\,\overline{\underline{x}} = \frac{\sigma_*^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N}\right) = \frac{\sigma^2}{n}(1-f)$$

De factor $\frac{N-n}{N-1}$ of ook 1 - f wordt <u>eindigheidscorrectie</u> genoemd.
Deze factor zagen we ook al optreden bij de variantie van de hypergeometrische verdeling (trekken zonder terugleggen)

$$\sigma^2 = n \frac{M}{N}\left(1 - \frac{M}{N}\right) \frac{N-n}{N-1} \quad .$$

Voor grote N is deze correctie te verwaarlozen.
Nu bewijs:

$$\text{var } \bar{\underline{x}} = \frac{1}{n^2} \sum_i \text{var } \underline{x}_i = \frac{1}{n^2} [n\sigma_*^2 + 2 \binom{n}{i} \text{cov}(\underline{x}_i, \underline{x}_j)] =$$

$$= \sigma_*^2/n + \frac{n(n-1)}{n^2} \text{cov}(\underline{x}_i, \underline{x}_i) \qquad i \neq j \ .$$

Stel even n = N, dus de steekproef is de gehele populatie. Dan is

$$\text{var}(\underline{x}_1 + \ldots + \underline{x}_N) = 0 = N\sigma_*^2 + N(N-1) \text{cov}(x_i, \underline{x}_j) \qquad i \neq j \ ,$$

oftewel $\text{cov}(\underline{x}_i, \underline{x}_j) = \frac{-\sigma_*^2}{(N-1)}$. Dit invullen geeft:

$$\text{var } \bar{\underline{x}} = \sigma_*^2/n + \frac{n-1}{n}\left(\frac{-\sigma_*^2}{N-1}\right) = \frac{\sigma_*^2}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma_*}{n}\left(\frac{N-n}{N-1}\right)$$

<u>Zuivere schatter voor de populatie variantie</u> σ_*^2 of σ^2

We hebben in Wis 49 gezien dat de steekproefvariantie (in steekproef met teruglegging)

$$\underline{s}^2 = \frac{\sum (\underline{x}_i - \bar{\underline{x}})^2}{n-1} \text{ een zuivere schatter is voor } \sigma_*^2.$$

Wat, indien we trekken zonder terugleggen?
Nu is

$$\sum(\underline{x}_i - \mu)^2 = \sum(\underline{x}_i - \bar{\underline{x}})^2 + \sum(\bar{\underline{x}} - \mu)^2$$

nl. dubbelproducten geven 0.

$$(n-1)\,\underline{s}^2 = \sum(\underline{x}_i - \bar{\underline{x}})^2 = \sum(\underline{x}_i - \mu)^2 - (\bar{\underline{x}} - \mu)^2$$

oftewel

$$(n-1)\,\mathcal{E}\,\underline{s}^2 = n\sigma_*^2 - n\,\frac{\sigma_*^2}{n}\left(\frac{N-n}{N-1}\right)$$

$$= \sigma_*^2(n-1)\,\frac{N}{N-1}\,.$$

Dus $\mathcal{E}\,\underline{s}^2 = \sigma^2$, dus $\underline{s}^2$ is zuiver voor σ^2, niet voor σ_*^2

Schatter voor de populatie parameter p.

De populatiewaarden zijn weer $X_1,\ldots,X_N$ maar met $X_i = 0$ of 1 alnaargelang het "mislukking" resp. een "succes" betreft. Dus $p = X./N$.
Een zuivere schatter voor p is $\hat{\underline{p}} = \sum \underline{x}_i/n = \bar{\underline{x}}$.
Uit Wis 49 weten we dat de verdeling van $\sum \underline{x}_i$ is:

met teruglegging: $\sum \underline{x}_i \sim BN(n,p)$,

zonder teruglegging: $\sum \underline{x}_i \sim HG(N,M,n)$

dus var $\hat{\underline{p}} = pq/n$ (met teruglegging) en $= \frac{pq}{n}\,\frac{N-n}{N-1}$ (zonder teruglegging)

met $p = \frac{M}{N}$, $q = \frac{N-M}{N}$.

Zuivere schatter voor var $\hat{\underline{p}}$

Met teruglegging: $s_p^2 = \hat{p}\,\hat{q}/(n-1)$; zonder teruglegging $s_p^2 = \frac{\hat{p}\,\hat{q}}{n-1}(1-f)$

Bewijs: Nu is

$$\underline{s}^2 = \frac{1}{n-1}\,\Sigma(\underline{x}_i - \bar{\underline{x}})^2 = \frac{1}{n-1}(\Sigma \underline{x}_i^2 - n\,\bar{\underline{x}}^2) \;;\; \frac{1}{n-1}(n\hat{p} - n\hat{p}^2) \quad \text{nl.}$$

$$\underline{x}_i = \underline{x}_i^2 \;;\; \underline{s}^2 = \frac{n\hat{p}\hat{q}}{n-1}$$

Dus

$$\mathcal{E}\,\underline{s}_p^2 = \mathcal{E}\,\underline{s}^2/n = \frac{\sigma_*^2}{n} = \frac{pq}{n} = \text{var } \hat{p} \text{ (met teruglegging).}$$

Zonder teruglegging: $\mathcal{E}\,\underline{s}_p^2 = \mathcal{E}\,\underline{s}^2(1-f)/n = \frac{1}{n}\,\sigma^2(1-f) = \frac{N}{N-1}\,\frac{1}{n}\,\sigma_*^2(1-f) =$

$$= \frac{pq}{n}\,\frac{N-n}{N-1} = \text{var } \hat{p}\,.$$

Steekproefomvang en nauwkeurigheid

Stel we willen μ schatten met een onbetrouwbaarheidsdrempel α, en het bijbehorende betreffende interval mag niet breder zijn dan Δ. Oftewel, met normale aanpassing, moet

$$u(\tfrac{1}{2}\alpha)\sigma(\bar{x}) < \tfrac{1}{2}\Delta$$

met teruglegging:

$$u(\tfrac{1}{2}\alpha)\sigma/\sqrt{n} < \tfrac{1}{2}\Delta \text{ dus } n > 4u^2(\tfrac{1}{2}\alpha)\sigma^2/\Delta^2 \quad ,$$

zonder teruglegging:

$$u(\tfrac{1}{2}\alpha)\,\frac{\sigma}{\sqrt{n}}\,\sqrt{\frac{N-n}{N-1}} < \tfrac{1}{2}\Delta \quad \text{dus } n > \frac{4u^2(\tfrac{1}{2}\alpha)\sigma^2}{\Delta^2(1-f) + 4u^2(\tfrac{1}{2}\alpha)\sigma^2/N} \quad .$$

Voor $N \gg n$ worden beide grenzen ongeveer gelijk. Echter in deze benadering zit nog steeds de onbekende σ^2. Om deze te bepalen houdt men vooraf een klein onderzoek of schat deze uit vorige steekproeven. Zijn dus α en Δ gegeven en σ^2 uit het verleden geschat, dan kan men de gewenste steekproefgrootte n bepalen.Deze benadering laat de kosten van het steekproef nemen buiten beschouwing.

Wil men de steekproefgrootte weten om p te schatten bij gegeven α en Δ dan krijgen we

$$n > 4u^2(\tfrac{1}{2}\alpha)\,pq/\Delta^2 \quad ;$$

met p en q geschat geeft dit

$$n > 4u^2(\tfrac{1}{2}\alpha)\,\hat{p}\hat{q}/\Delta^2 \quad .$$

Nu is max $\hat{p}\hat{q} = \frac{1}{4}$ dus kunnen we benaderen $n > u^2(\tfrac{1}{2}\alpha)/\Delta^2$.

Willen we bijvoorbeeld via een enquête het percentage ±5% weten van de huisvrouwen die een zeker wasmiddel gebruiken, met $\alpha = 0{,}10$, dan nemen we een steekproef van $n > (1.645)^2/\frac{1}{10^2} \approx 300$.

Tot nu toe hebben we steeds enkelvoudige steekproeven beschouwd. In het vervolg zullen we ons bezighouden met meervoudige steekproeven.

10.2. Gelede of gestratificeerde steekproeven (stratified sampling)

Bij gelede steekproeven wordt de populatie ter grootte N eerst verdeeld in k deelpopulaties (strata) met resp. $N_1, N_2, \ldots, N_k$ elementen, dus

$\sum_{1}^{n} N_i = N.$

Dit splitsen heet stratificeren.

Uit elk stratum wordt een steekproef getrokken en we beperken ons tot het geval van een enkelvoudige aselecte steekproef uit elk stratum. Zo'n stelsel van steekproeven wordt een gelede of gestratificeerde aselecte steekproef genoemd (stratified random sampling). Dit is dus een meervoudige steekproef.

Gelede steekproeven worden toegepast op grond van verschillende motieven:

a) Men wil afzonderlijke gegevens over deelpopulaties weten.
b) De te bestuderen populatie is al in strata ingedeeld (provincies, gemeenten, rijks- en provinciale wegen, fabrieken, scholen, etc.)
c) De omvang in aantal elementen per stratum is nogal verschillend. Bedrijven bijv. deelt men daarom meestal in naar grootte.
d) De variantie kan gereduceerd worden door stratificatie in deelpopulaties die homogener zijn dan de populatie als geheel.

Stel we hebben een populatie $X_1, \ldots, X_N$ met $\mathcal{E}\underline{X} = \mu$ en var $X = \sigma^2$. Deze verdelen we in k strata ter grootte $N_1, \ldots, N_k$ met dus $\Sigma N_i = N$. In i^e stratum zitten de elementen $X_{i1}, \ldots, X_{iN_i}$.
Voor het i^e stratum is

$$\mu_i = \sum_{j}^{N_i} X_{ij}/N_i \text{ en } \sigma_i^2 = \sum_{j}^{N_i} \frac{(X_{ij} - \mu_i)^2}{N_i - 1}.$$

Noteren we $N_i/N = w_i$ (stratumgewicht) dan is evident $\mu = \sum_{ij} x_{ij}/N =$

$= \frac{1}{N} \sum_{i} \sum_{j} X_{ij} = \mu = \sum_{i} N_i\mu_i/N = \sum_{i} w_i\mu_i$, het gewogen stratumgemiddelde; $\sum w_i = 1$.

Nu trekken we uit i^e stratum ($i = 1, \ldots, k$) een aselecte steekproef zonder teruglegging, ter grootte n_i : $x_{i1}, \ldots, x_{in_i}$ met $\sum n_i = n$.

$\frac{n_i}{N_i} = f_i$ is de steekproeffractie voor stratum i.

Een goede zuivere schatter voor μ is:

$$\bar{\bar{x}} = \frac{\sum_i N_i \bar{x}_{i.}}{N} = \sum_i w_i \bar{x}_{i.} \quad ,$$

waarbij $\bar{x}_{i.} = \sum x_{ij}/n_i$. Pas op, dit is niet gelijk aan het steekproefgemiddelde $\bar{x}_{..} = \sum_{ij} x_{ij}/n$; $\bar{\bar{x}}$ is het gewogen steekproefgemiddelde.

De vraag blijft: hoe moeten we onze steekproef ter grootte n (gegeven ondersteld) verdelen over de k strata? Dit is het probleem van de allocatie, het toewijzen der n_i .

Onder Neyman-allocatie (bij gegeven steekproefgrootte n) verstaat men een zodanige verdeling van n over de k steekproeven, dat de variantie van $\bar{\bar{x}}$ zo klein mogelijk is. Nu is, met teruglegging,

$$\text{var } \bar{\bar{x}} = \sum w_i^2 \text{ var } \bar{x}_{i.} = w_i^2 \frac{\sigma_i^2}{n_i} (1 - f_i) \quad . \qquad (*)$$

We gaan deze minimaliseren onder de nevenvoorwaarde $\sum n_i = n$.
Dit doen we met de multiplicatorenmethode van Lagrange:
We gaan minimaliseren:

$$f(n_i) = \sum w_i^2 \frac{\sigma_i^2}{n_i} (1 - f_i) + \lambda(\sum n_i - n) \quad ,$$

$$\frac{\partial f}{\partial n_i} = \frac{-w_i^2 \sigma_i^2}{n_i^2} + \lambda = 0 \quad ; \quad n_i\sqrt{\lambda} = w_i\sigma_i \quad ; \quad n\sqrt{\lambda} = \sum w_i\sigma_i \quad ,$$

dus

$$n_i = \frac{w_i\sigma_i}{\sum w_i\sigma_i} n \quad .$$

We zien hieruit dat, hoe "gewichtiger" stratum i, en/of hoe grotere spreiding, des te groter n_i . Deze σ_i zullen we weer uit eerdere, kleine steekproeven geschat moeten hebben.

Opmerking: Zijn de σ_i voor alle strata gelijk, dan $n_i = \frac{N_i}{N} n$ dus
$f_i = \frac{n_i}{N_i} = \frac{n}{N}$ is constant. Dit heet een proportionele steekproef
nl. $\frac{N_i}{n} = \frac{N_i}{N}$.
In dat geval is $\bar{x} = \sum_i w_i \bar{x}_{i.} = \sum_i w_i \sum_j x_{ij}/w_i n = \sum_{ij} x_{ij}/n = \bar{x}_{..}$
juist het steekproefgemiddelde, met var $\bar{x} = \frac{}{n} (1 - \frac{n}{N})$ volgt uit (*) .

Men kan ook de kosten van het steekproefnemen erbij betrekken. Deze zijn veelal afhankelijk van het aantal waarnemingen dat men in de steekproef opneemt. De kostenfunctie in haar eenvoudigste vorm wordt dan

$$\text{kosten } C = c_0 + n_i c_i \text{ , waarbij}$$

c_0 vaste kosten; c_i variabele kosten per waarneming in stratum i.

Met behulp van de methode van Lagrange gaan we weer minimaliseren

$$\text{var } \bar{x} + \lambda(\sum n_i c_i + c_0 - C) .$$

Dit geeft volkomen analoog aan zoëven: $n_i = \frac{w_i \sigma_i / \sqrt{c_i}}{\Sigma w_i \sigma_i / \sqrt{c_i}} n$; $i = 1,\ldots,k$.

Ga na!
Dit heet een optimale locatie. Voor c_i = constant krijgen we uiteraard het vorige terug.

10.3. Trossteekproeven (cluster sampling)

Bij trossteekproeven worden de waarnemingen in groepen bijeen genomen, in "clusters" dus. Deze methode is i.h.b. van belang wanneer de clusters ongeveer gelijk gemiddelde hebben en ongeveer dezelfde interne spreiding. Hieruit trekken we dan een steekproef van n clusters die we in z'n geheel waarnemen.

Voorbeeld. In een zuivelfabriek worden per dag 500 dozen van elke 24 pakjes boter geproduceerd. Ten behoeven van de keuringsdienst voor waren worden 25 dozen getrokken, waarvan alle pakjes worden gecontroleerd. Men behoeft nu slechts een gering aantal dozen te openen en het trekken van steekproeven uit de getrokken dozen kan achterwege blijven.

Stel we hebben N trossen (clusters) van elk M elementen. De elementen in de i^e tros zijn weer $X_{i1},\ldots,X_{iM}$;Totaal dus NM elementen;μ_i is het gemiddelde van tros i. Het totale gemiddelde is weer μ. We nemen een enkel-

voudige aselecte steekproef zonder teruglegging ter grootte n uit de N trossen. Dan is

$$\bar{\bar{x}} = \frac{1}{n} \sum_i \bar{x}_{i.} / M = \sum_i \bar{x}_{i.} / n$$

een zuivere schatter voor μ.

Opmerking: deze is juist = $\bar{\bar{x}}_{..}$. De steekproef bevat nM elementen!

Er geldt:

$$\text{var } \bar{\bar{x}} = \frac{1 - n/N}{n} \frac{NM - 1}{M^2(N-1)} \sigma^2[1 + (M-1)\rho] .$$

Hierin is

$$\sigma^2 = \frac{\sum(X_{ij} - \mu)^2}{NM - 1} \quad ; \text{ NM = populatie-omvang.}$$

$$\rho = \frac{\sum(X_{ij} - \mu)(X_{ik} - \mu)}{\sum(X_{ij} - \mu)^2} \quad , \text{ de binnentrossen correlatiecoëf.}$$

Voor de afleiding zie Cochran [3] § 9.4 .

Is M = 1 dan ontstaat de bekende formule var $\bar{\bar{x}} = \frac{\sigma^2}{n}(1 - f)$.

Stellen we $\frac{NM - 1}{N(M-1)} \sim 1$ dan wordt var $\bar{\bar{x}} = \frac{1 - f}{nM} \sigma^2[1 + (M-1)\rho]$.

Hadden we geen trossen genomen, doch direct een steekrpoef ter grootte nM,

dan zou var $\bar{\bar{x}} = \frac{\sigma^2}{nM}(1 - f)$. Dit scheelt de factor $1 + (M-1)\rho$.

Alleen dus voor $\rho < 0$ wordt de variantie bij trossteekproeven kleiner.

10.4. Getrapte steekproeven (multi-stage sampling)

In de vorige paragraaf werd een steekproef genomen uit een populatie trossen (clusters), en de getrokken clusters werden geheel waargenomen. Nu beschouwen we het geval waar slechts een deel van elk getrokken cluster wordt waargenomen. De trekking van de steekproef verloopt dus in twee trappen (of meer). Men noemt deze dan ook getrapte steekproeven. In de eerste trap wor-

den samengestelde eenheden getrokken die we nu liever primaire eenheden noemen in plaats van clusters. In de tweede trap worden uit alle getrokken primaire eenheden secundaire eenheden getrokken, dat zijn dus eenheden waaruit de primaire eenheid is samengesteld.

Voorbeeld: Ten behoeve van een onderzoek naar de gebitten van schoolkinderen, wordt eerst een steekproef genomen uit alle scholen (= primaire eenheden) en in de tweede trap trekt men per getrokken school uit alle klassen (= secundaire eenheden).Men kan nu de hele klas onderzoeken of in een 3e trap een steekproef nemen van leerlingen uit de getrokken klassen. Een trossteekproef is dus een speciaal geval van een getrapte steekproef (two-stage) nl. waarbij alle secundaire eenheden in de steekproef worden opgenomen.

De motieven voor het trekken van getrapte steekproeven zijn dan ook, evenals bij trossteekproeven, de moeilijkheid of onmogelijkheid de onderzoeks elementen rechtstreeks te trekken, en de reductie van onderzoekssteekproefkosten.

Een getrapte steekproef brengt doorgaans hogere kosten met zich mee dan een trossteekproef van dezelfde omvang, omdat uit de getrokken primaire eenheden opnieuw steekproeven moeten worden getrokken. Getrapte steekproeven liggen, wat betreft nauwkeurigheid, tussen enkelvoudige en trossteekproeven in.

We beschouwen het geval van N evengrote clusters (primaire eenheden), elk met M secudaire eenheden. Uit de clusters nemen we een aselecte steekproef ter grootte n zonder teruglegging, en uit elke cluster weer m secundaire eenheden.

Nu is $\bar{\underline{x}}_{..} = \sum_{ij} \underline{x}_{ij}/mn$, een zuivere schatter voor μ,

met

$$\text{var } \bar{\underline{x}}_{..} = (1 - n/N) \frac{\sigma_1^2}{n} + (1 - \frac{m}{M}) \frac{\sigma_2^2}{nm} \qquad (1)$$

waarbij

$$\sigma_1^2 = \sum_i \frac{(\mu_i - \mu)^2}{N - 1} ,$$

de "variantie" tussen gemiddelden van primaire eenheden ,

$$\sigma_2^2 = \sum_{ij} \frac{(X_{ij} - \mu_i)^2}{N(M-1)} = \sum_i \sigma_i^2/N \quad ,$$

de variantie tussen elementen binnen primaire eenheden.
(Voor bewijs zie [3] Cochran 10.3).

Is m = M, dan hebben we een trossteekproef. De 2e term geeft bijdrage 0. Is n = N en m_i = m dan hebben we het geval van een gestratificeerde steekproef met evenredige allocatie. Nu is de 1e term in de var $\bar{x}_{..}$ nul. De variantie voor een gestratificeerde steekproef wordt nu:

$$\sum w_i^2 \frac{\sigma_i^2}{n_i}(1 - f_i) = \sum_i \frac{M^2}{M^2N^2} \frac{\sigma_i^2}{m}(1 - \frac{m}{M}) = (1 - \frac{m}{M}) \frac{1}{nm} \sum_i \frac{\sigma_i^2}{N} \quad ,$$

juist de zoëven verkregen variantie. (1)

Vrijwel altijd is, bij constante steekproefgrootte nm, de vairantie minimaal als n zo groot mogelijk is, dus m zo klein mogelijk.

De variantie is niet de enige factor die van belang is. Even belangrijk zijn de kosten die voor de steekproef zijn gemaakt. Het is ook beter de variantie te minimaliseren bij gegeven toelaatbare kosten. We nemen aan dat elke getrokken primiare eenheid een bedrag c_1 kost en de waarnemingskosten per secudaire eenheid c_2. De totale kosten C bedragen zijn
$C = nc_1 + nmc_2$.
We gaan weer, volgens de methode van Lagrange, minimaliseren:

$$g(n,m) = \text{var } \bar{x}_{..} + \lambda(nc_1 + nmc_2 - C)$$

Differentiëren naar n en m geeft:

$$\frac{\partial g}{\partial n} = -\sigma_1^2/n^2 - \frac{(1 - \frac{m}{M})}{n^2 m} \sigma_2^2 + \lambda(c_1 + mc_2) = 0 \quad ,$$

$$\frac{\partial g}{\partial m} = -\frac{1}{nm^2} \sigma_2^1 + n\lambda c_2 = 0 \quad .$$

Dit geeft via $n^2\lambda$:

$$n^2\lambda = \frac{m\sigma_1^2 + (1 - \frac{m}{M})\sigma_2^2}{mc_1 + m^2c_2} = \frac{\sigma_2^2}{m^2c_2} \quad ,$$

oftewel

$$m_{opt} = \sqrt{\frac{c_1/c_2}{\sigma_1^2/\sigma_2^2 - 1/M}} \quad .$$

Hebben we C van te voren vastgelegd, dan wordt

$$n_{opt} = \frac{C}{c_1 + c_2 m_{opt}} \quad .$$

Uit de formule voor m_{opt} kan men de volgende, voor de hand liggende conclusies trekken:

Wanneer c_1 veel groter is dan c_2, dan moet men m groot, dus n klein nemen. Eventueel kan m = M worden, hetgeen betekent dat men een trossteekproef moet gebruiken als de voorbereidingskosten c_1 hoog zijn en de waarnemingskosten laag. Zijn echter de voorbereidingskostenlaag t.o.v. de waarnemingskosten, dan kiest men m klein, dus n groot; eventueel kan n = N worden, met andere woorden een gestratificeerde steekproef. Evenzo leidt een grote σ_1/σ_2 tot een kleine m, dus grote n. Dit is dus het geval als de primaire eenheden onderling sterk uiteenlopen.

11. Verdelingsvrije methoden.

11.1. Inleiding.

De meest gebruikelijke statistische methoden voor het analyseren van continue stochastische grootheden gaan uit van de veronderstelling dat wij met normale verdelingen te maken hebben. Als deze veronderstelling juist is hebben deze methoden (zoals t-toets, regressie- en variantie-analyse) bepaalde optimale eigenschappen, bijv. maximaal onderscheidingsvermogen. Als aan de veronderstelling van normaliteit niet is voldaan kunnen deze methoden verre van optimaal zijn en bovendien kan de nominale onbetrouwbaarheid (bijv. van een betrouwbaarheidsinterval) ernstig afwijken van de werkelijke waarde.
In die gevallen waarin de verdeling onbekend is en een benadering met de normale niet voldoende met een beroep op de centrale limietstelling te rechtvaardigen is, zijn methoden vereist die toepasbaar zijn wat de vorm van de verdeling ook is. Dit zijn de zogenaamde verdelingsvrije of parametervrije methoden.
Enkele van de meest toegepaste methoden worden in dit hoofdstuk besproken.

11.2. De tekentoets.

De meest eenvoudige verdelingsvrije methode is de tekentoets, die gebruikt wordt om een verschil tussen gemiddelden (eigenlijk tussen medianen) te toetsen bij waarnemingen die in paren zijn uitgevoerd. Dit is dezelfde probleemstelling als in § 4.3.3, maar daar werd van de normale verdeling uitgegaan. Zoals de naam aangeeft wordt bij de hier te bespreken methode alleen gelet op de tekens van de verschillen tussen de gepaarde waarnemingen.
Als per paar $\underline{x}_i$, $\underline{y}_i$ $(i = 1,\ldots,n)$ de waarnemingen $\underline{x}_i$ en $\underline{y}_i$ uit dezelfde continue verdeling komen, dan geldt

$$P(\underline{x}_i > \underline{y}_i) = P(\underline{x}_i < \underline{y}_i) = \tfrac{1}{2} ,$$

ofwel

$$P(\underline{x}_i - \underline{y}_i < 0) = P(\underline{x}_i - \underline{y}_i > 0) = \tfrac{1}{2} ,$$

of

$$P(\underline{d}_i < 0) = P(\underline{d}_i > 0) = \tfrac{1}{2} ,$$

als

$$\underline{d}_i := \underline{x}_i - \underline{y}_i \quad (i = 1,\ldots,n) .$$

Bij een continue simultane verdeling van $(\underline{x}_i,\underline{y}_i)$ is de kans dat $\underline{x}_i = \underline{y}_i$ gelijk aan nul.
In de praktijk komen vaak, bijv. door afronding, toch paren voor met een verschil $d_i = 0$. We laten deze paren voor de toets buiten beschouwing en we gaan uit van de verschillen $d_i \neq 0$. Onder deze voorwaarde toetsen we de hypothese

$$H_0: P(\underline{x}_i > \underline{y}_i) = P(\underline{x}_i < \underline{y}_i) = \tfrac{1}{2} \quad (i = 1,\ldots,n) .$$

Als we het aantal positieve verschillen $\underline{d}_i$ aangeven met $\underline{P}$ en het aantal negatieve met $\underline{N}$, dan is zowel $\underline{P}$ als $\underline{N}$ onder de nulhypothese binomiaal verdeeld met $p = \frac{1}{2}$.
De tekentoets komt daarom neer op een binomiale toets voor de hypothese

$$H_0: p = \tfrac{1}{2} .$$

Kritieke waarden bij deze speciale waarde van p zijn te vinden in tabel 7.1 van het S.C. voor n = 1(1)100.
Hierbij valt op te merken dat de tekentoets alleen bruikbaar is als n niet te klein is. Bij een tweezijdige toets met onbetrouwbaarheidsdrempel $\alpha = 0,05$ kan voor het eerst significantie optreden bij n = 6. Alle verschillen moeten dan hetzelfde teken hebben. Bij n = 6 is nl. $2(\frac{1}{2})^n = 0,03125 < 0,05$, terwijl dit voor n = 5 nog niet het geval is.

Voorbeeld 12.2. Twee soorten varkensvoer zijn met elkaar vergeleken door 18 paren biggen (per paar afkomstig uit hetzelfde toom) er mee te voeren. De toenamen in gewicht (in kg) over een bepaalde tijd staan de tabel 11.2.

Tabel 11.2.1 Toenamen in gewicht (in kg) van biggen gevoerd met twee soorten voer.

Paar	Voer A	Voer B	Teken van verschil
1	25	19	+
2	30	32	-
3	28	21	+
4	34	34	0
5	23	19	+
6	25	25	0
7	27	25	+
8	35	31	+
9	30	31	-
10	28	26	+
11	32	30	+
12	29	25	+
13	30	29	+
14	30	31	-
15	31	25	+
16	29	25	+
17	23	20	+
18	26	25	+

Er zijn twee gevallen met $d = 0$, dus we hebben $n = 16$ met $P = 13$, $N = 3$. Toetsen we tweezijdig met $\alpha = 0{,}05$ dan is de kritieke waarde voor de kleinste van de waarden P en N volgens S.C. tabel 7.1 gelijk aan 3. Dus H_0 moet worden verworpen.

Als we geen speciale tabel ter beschikking hebben kan voor $n \geq 10$ de normale benadering worden gebruikt. In ons geval vinden we

$$u = \frac{P - \frac{1}{2}n}{\sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}}} = \frac{13 - 8}{2} = 2{,}50 \ .$$

De hierbij behorende tweezijdige overschrijdingskans is

$$2 \times 0{,}0062 = 0{,}0124 < 0{,}05 \ .$$

De conclusie is dus hetzelfde.

Opmerking. De tekentoets kan ook worden toegepast als we niet uitgaan van een continue verdeling. Stel bijvoorbeeld dat aan 18 mensen wordt gevraagd of ze van twee soorten frisdrank merk A of merk B lekkerder vinden en dat de antwoorden als volgt zijn:

A lekkerder	13
B lekkerder	3
Geen voorkeur	2
	18

Dit zijn dezelfde aantallen als in voorbeeld 12.2 en de conclusie is dus dat de hypothese dat er evenveel mensen zijn die A lekkerder vinden dan B als omgekeerd moet worden verworpen bij $\alpha = 0{,}05$.
Is in een geval als dit het aantal met "geen voorkeur" erg groot dan moet dit in de conclusie worden vermeld. Er kan dan bij een gering percentage van de proefpersonen dat verschil proeft nog wel een voorkeur voor één der produkten bestaan maar de praktische betekenis is dan misschien gering.

11.3. De twee-steekproeventoets van Wilcoxon.

Voor het vergelijken van twee gemiddelden op grond van twee steekproeven is de op normaliteit gebaseerde toets in § 4.3.2 besproken. Bij de toets die we nu gaan behandelen luidt de nulhypothese dat we twee steekproeven

$$\underline{x}_1,\ldots,\underline{x}_{n_1}$$

en

$$\underline{y}_1,\ldots,\underline{y}_{n_2}$$

hebben, afkomstig uit dezelfde continue verdeling. De alternatieve hypothese is dat de verdelingen t.o.v. elkaar verschoven zijn, zodat de verwachtingen verschillen. Men kan bewijzen dat, evenals bij de tekentoets, de toets gevoelig is voor afwijkingen van $P(\underline{x} > \underline{y})$ van $\frac{1}{2}$.
We zullen de methode uiteenzetten aan de hand van een voorbeeld.
In de paragrafen 4.2 en 4.3 werden de waarnemingen gegeven van twee methoden om het zwavelgehalte van steenkool te bepalen. Deze waarnemingen waren (na codering)

methode M_1	18	20	22	14	9	10	10	$(x_1,\ldots,x_7)$
methode M_2	20	19	18	27	24			$(y_1,\ldots,y_5)$

Bij de toets van Wilcoxon worden de $n_1 + n_2$ (hier 12) waarnemingen in opklimmende volgorde gerangschikt, waarna de rangnummers 1 t/m 12 worden toegekend. Bij steekproeven uit continue verdelingen kunnen, evenmin als bij de tekentoets gelijke waarnemingen voorkomen. Als dit, bijv. door afronding zoals hier toch het geval is, dan wordt aan een groepje van twee of meer waarnemingen het gemiddelde rangnummer toegekend. Zo zouden in ons voorbeeld aan de waarnemingen x_6 en x_7, die beide gelijk zijn aan 10, de rangnummers 2 en 3 moeten worden gegeven. Wij kennen nu aan beide het rangnummer $2\frac{1}{2}$ toe.
De rangnummers zijn:

methode M_1	$5\frac{1}{2}$	$8\frac{1}{2}$	10	4	1	$2\frac{1}{2}$	$2\frac{1}{2}$
methode M_2	$8\frac{1}{2}$	7	$5\frac{1}{2}$	12	11		

Als beide steekproeven uit dezelfde verdeling komen, dan is elke verdeling van de 12 rangnummers over beide steekproeven even waarschijnlijk. Als de verdelingen verschoven zijn t.o.v. elkaar, dan zullen in de ene steekproef hogere en in de andere lagere rangnummers voorkomen.
Wij berekenen nu de som T_1 van de rangnummers van de eerste steekproef:

$$T_1 = 34$$

en eveneens van de tweede steekproef:

$$T_2 = 44 \ .$$

Uiteraard geldt

$$T_1 + T_2 = 1 + 2 + \ldots + 12 = 78$$

en in het algemeen

$$T_1 + T_2 = \tfrac{1}{2}(n_1 + n_2)(n_1 + n_2 + 1) \ .$$

Als toetsingsgrootheid gebruikt men die T_i die hoort bij de <u>kleinste</u> n_i, $i = 1,2$. De linker kritieke waarden voor $\alpha = 0{,}05$ en $\alpha = 0{,}01$ vindt men in tabel 7.2 van het S.C. De rechter kritieke waarden volgen uit de relatie: linker k.w. + rechter k.w. = $n_i(n_1 + n_2 + 1)$. In ons voorbeeld is, bij een onbetrouwbaarheid $\alpha = 0{,}05$ en steekproefgroottes 7 en 5, de linker kritieke waarde 20. De rechter kritieke waarde is dus
$5(7 + 5 + 1) - 20 = 45$.
De gevonden waarde $T_2 = 44$ ligt hier onder, dus H_0 wordt niet verworpen bij $\alpha = 0{,}05$.

<u>Opmerking</u>. De t-toets toegepast op dezelfde waarnemingen leverde een significant resultaat bij $\alpha = 0{,}05$ (4.3.2.4).

We gaan hier nog iets nader in op de verdelingen van $\underline{T}_1$ en van $\underline{T}_2$ onder H_0, in het geval van continue verdelingen. Daar $\underline{T}_1 + \underline{T}_2$ constant is, kunnen we ons tot de verdeling van b.v. $\underline{T}_1$ beperken. De waarde T_1 die de stochastische grootheid $\underline{T}_1$ aanneemt wordt bepaald door de rangschikking van de n_1 x-waarden en de n_2 y-waarden. De minimale waarde van T_1 krijgen we als alle x-en kleiner zijn dan alle y's. We hebben dan:

11.3.1. $$\min T_1 = 1 + 2 + \ldots + n_1 = \tfrac{1}{2}n_1(n_1 + 1).$$

Evenzo vinden we

11.3.2. $$\max T_1 = (n_2 + 1) + (n_2 + 2) + \ldots + (n_2 + n_1) = \tfrac{1}{2}n_1(n_1 + 2n_2 + 1) .$$

Bij iedere rangschikking die T_1 oplevert hoort precies één rangschikking die de waarde $n_1(n_1 + n_2 + 1) - T_1$ geeft. Als we namelijk de rij van achteren naar voren lezen verandert een rangnummer r_i in $(n_1 + n_2 + 1 - r_i)$ en als we sommeren over de rangnummers van de x-waarden krijgen we juist de hiervoor genoemde relatie. Onder H_0 is de verdeling van $\underline{T}_1$ dus symmetrisch om $\tfrac{1}{2}n_1(n_1 + n_2 + 1)$ en dus moet ook gelden:

11.3.3. $$\mathcal{E}\,\underline{T}_1 = \tfrac{1}{2}n_1(n_1 + n_2 + 1) ,$$

Dit is ook als volgt in te zien. Als $\underline{r}_i$ het rangnummer van $\underline{x}_i$ is, dan geldt onder H_0:

$$\mathcal{E}\,\underline{r}_i = \tfrac{1}{2}(n_1 + n_2 + 1) ,$$

het gemiddelde van alle rangnummers. Dus

$$\mathcal{E}\,\underline{T}_1 = \mathcal{E}\left(\sum_1^{n_1} \underline{r}_i\right) = \tfrac{1}{2}n_1(n_1 + n_2 + 1) .$$

Onder H_0 zijn alle $\binom{n_1 + n_2}{n_1}$ verschillende rangschikkingen van n_1 x-en en n_2-y's even waarschijnlijk. Als n_1 en n_2 niet te groot zijn kunnen kritieke waarden voor betrekkelijk kleine α nog wel gevonden worden door het tellen van de aantallen rangschikkingen die het kritieke gebied vormen. Neem b.v. $n_1 = n_2 = 4$ en $\alpha = 0{,}05$ (tweezijdig). Er zijn $\binom{8}{4} = 70$ verschillende con-

figuraties. Het kritieke gebied bestaat uit twee delen en in elk deel zitten hoogstens 0.025 × 70 = 1,75 gevallen. Zodra we dus 1 of meer hebben kunnen we stoppen. Dus alleen de minimale waarde

$$T_1 = 1 + 2 + 3 + 4 = 10$$

en de maximale waarde

$$T_1 = 5 + 6 + 7 + 8 = 26 = 4(4 + 4 + 1) - 10 .$$

vormen het kritieke gebied. De bijbehorende onbetrouwbaarheid van de toets is 2/70 ; 0.029. De linker kritieke waarde klopt uiteraard met de in tabel S.C.7.2 opgegeven waarde.

Als $n_2 > 20$, zodat de tabel niet meer kan worden gebruikt en $n_1 > 5$, dan kan de normale benadering worden toegepast. Hiervoor hebben we de variantie van $\underline{T}_1$ nodig, die als volgt wordt afgeleid:

We noemen de rangnummers van $\underline{x}_1,\ldots,\underline{x}_{n_1}$ weer $\underline{r}_1,\ldots,\underline{r}_{n_1}$. Voor iedere $\underline{r}_i$ geldt:

$$\underline{r}_i = \tfrac{1}{2}(n_1 + n_2 + 1) ,$$

$$\operatorname{var} \underline{r}_i = \frac{1}{12}[(n_1 + n_2)^2 - 1] ,$$

want de verdeling van $\underline{r}_i$ is diskreet homogeen op $\{1,2,\ldots,(n_1 + n_2)\}$.
Noemen we de rangnummers van $\underline{y}_1,\ldots,\underline{y}_{n_2}$, respectievelijk $\underline{r}_{n_1+1},\ldots,\underline{r}_{n_1+n_2}$, dan hebben alle rangnummers $\underline{r}_1,\ldots,\underline{r}_{n_1+n_2}$ de bovengenoemde verdeling.
We definiëren

$$\sigma^2 := \operatorname{var} \underline{r}_i \quad (i = 1,\ldots,n_1 + n_2)$$

$$\rho\sigma^2 := \operatorname{cov}(\underline{r}_i,\underline{r}_j) \ (i \neq j) .$$

Dan geldt:

$$0 = \operatorname{var}\left(\sum_{1}^{n_1+n_2} \underline{r}_i \right) = (n_1 + n_2)\sigma^2 + (n_1 + n_2)(n_1 + n_2 - 1)\rho\sigma^2 .$$

Dus $\rho = -1/(n_1+n_2 - 1)$.

Hieruit volgt:

$$\text{var } \underline{T}_1 = \text{var}\left(\sum_1^{n_1} \underline{r}_i \right) = n_1\sigma^2 + n_1(n_1 - 1)\rho\sigma^2 =$$

$$= n_1 \left(1 - \frac{n_1 - 1}{n_1 + n_2 - 1}\right) \sigma^2 =$$

$$= \frac{n_1 n_2}{n_1 + n_2 - 1} \quad \frac{1}{12}(n_1 + n_2 + 1)(n_1 + n_2 + 1) =$$

11.3.4. $$= \frac{1}{12}\, n_1 n_2 (n_1 + n_2 + 1) \ .$$

Benaderde kritieke waarden voor $\underline{T}_1$ kunnen dus met (11.3.3) en (11.3.4) als volgt worden gevonden:

11.3.5. $$(\ell, r) = \tfrac{1}{2} n_1(n_1 + n_2 + 1) \mp u_{\alpha/2}\left[\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)\right]^{\frac{1}{2}} \ .$$

Vergelijking van een aantal exacte kritieke waarden met de benaderde waarden volgens (12.3.5) geeft het volgende beeld:

		$\alpha = 0{,}05$		$\alpha = 0{,}01$	
n_1	n_2	exact	ben.	exact	ben.
5	20	35	36,15	28	27,08
10	20	110	110,45	97	96,45
15	20	210	211,20	193	192,72
20	20	337	337,54	315	314,77

Tabel 11.3.1. Linker kritieke waarden van de toetsingsgrootheid T_1 van Wilcoxon bij tweezijdige onbetrouwbaarheid α. Exacte waarden en normaal benaderde waarden.

11.3.a De rangtekentoets of symmetrietoets van Wilcoxon

In 12.2 is de tekentoets behandeld om bij paren waarnemingen de volgende nulhypothese te toetsen:

$$H_0: P(\underline{x}_i > \underline{y}_i) = P(\underline{x}_i < \underline{y}_i) = \tfrac{1}{2} \qquad (i = 1,\ldots,n) .$$

Hierbij wordt alleen op het teken van de verschillen en niet op hun grootte gelet. Als de verschillen zelf zijn waargenomen en dus niet alleen het teken bekend is bestaat er een verdelingsvrije toets die in het algemeen een groter onderscheidingsvermogen heeft dan de tekentoets.

We zullen deze methode illustreren aan voorbeeld 12.2. De 16 verschillen $\neq 0$ worden van een rangnummer voorzien dat betrekking heeft op de absolute waarden:

Verschil d :	+6	−2	+7	+4	+2	+2	−1	+2
rangnr.van |d|	$14\frac{1}{2}$	$6\frac{1}{2}$	16	$11\frac{1}{2}$	$6\frac{1}{2}$	$11\frac{1}{2}$	$2\frac{1}{2}$	$6\frac{1}{2}$
verschil d	+2	+4	+1	−1	+6	+4	+3	+1
rangnr.van |d|	$6\frac{1}{2}$	$11\frac{1}{2}$	$2\frac{1}{2}$	$2\frac{1}{2}$	$14\frac{1}{2}$	$11\frac{1}{2}$	9	$2\frac{1}{2}$

Nu kunnen we de rangnummers van de positieve verschillen en ook die van de negatieve verschillen optellen. Het resultaat is

$$R_+ = 14\tfrac{1}{2} + 16 + \ldots + 2\tfrac{1}{2} = 124\tfrac{1}{2}$$
$$R_- = 6\tfrac{1}{2} + 2\tfrac{1}{2} + 2\tfrac{1}{2} = 11\tfrac{1}{2} .$$

Uiteraard geldt:

$$R_+ + R_- = \tfrac{1}{2}n(n+1) = \tfrac{1}{2} \cdot 16 \cdot 17 = 136 .$$

Onder de nulhypothese is van elk rangnummer de kans dat het bij een positief verschil hoort gelijk aan $\frac{1}{2}$. In principe kan zo de kansverdeling van $\underline{R}_+$ worden uitgerekend. Bovendien is het duidelijk dat $\underline{R}_+$ en $\underline{R}_-$ dezelfde verdeling hebben. Voor het geval alle rangnummers verschillend zijn is deze verdeling tot

n = 50 berekend en kritieke waarden zijn te vinden in tabel S.C. 7.1a. Deze tabel kan als benadering worden gebruikt als er wel gelijke verschillen zijn, zodat we met gemiddelde rangnummers moeten werken, zoals in dit voorbeeld. We zien dat de linker kritieke waarde bij n = 16 en α = 0,01 (tweezijdig) gelijk is aan 19. De gevonden waarde voor R_- is kleiner en H_0 kan dus worden verworpen met α = 0,01. Berekening van de rechter kritieke waarde voor R_+ levert op: 136 - 19 = 117. Deze wordt overschreden en de conclusie is (uiteraard) dezelfde.

Voor grote n kan de normale benadering worden gebruikt. Verwachting en variantie van $\underline{R}_+$ en $\underline{R}_-$ zijn respectievelijk.

11.3a.1. $$\mathcal{E}\underline{R}_+ = \mathcal{E}\underline{R}_- = \tfrac{1}{4}n(n+1)$$

11.3a.2. $$\text{var}\,\underline{R}_+ = \text{var}\,\underline{R}_- = \frac{1}{24}n(n+1)(2n+1) \ .$$

We bewijzen dat op de volgende manier. De rangnummers van de absolute verschillen $|d_1|,\ldots,|d_n|$ noemen we $\underline{r}_1,\ldots,\underline{r}_n$. We definiëren nu

$$\underline{p}_i := \underline{r}_i \text{ als } \underline{d}_i > 0 \ ,$$

$$\underline{p}_i := 0 \text{ als } \underline{d}_i < 0 \ .$$

Er geldt dus

$$\underline{R}_+ = \sum_{i=1}^{n} \underline{p}_i \ .$$

Aangezien $P(\underline{d}_i > 0) = P(\underline{d}_i < 0) = \frac{1}{2}$ hebben we

$$\mathcal{E}\underline{p}_i = \tfrac{1}{2}\mathcal{E}\underline{r}_i = \tfrac{1}{4}(n+1)$$

en

$$\mathcal{E}\underline{R}_+ = \sum_{i=1}^{n} \underline{r}_i = \tfrac{1}{4}n(n+1) \ .$$

Ook geldt

$$\mathcal{E}\underline{p}_i^2 = \tfrac{1}{2}\,\mathcal{E}\underline{r}_i^2 \quad ,$$

dus

$$\operatorname{var}\underline{p}_i = \tfrac{1}{2}\,\mathcal{E}\underline{r}_i^2 - \tfrac{1}{4}(\mathcal{E}\underline{r}_i)^2 = \tfrac{1}{2}\operatorname{var}\underline{r}_i + \tfrac{1}{4}(\mathcal{E}\underline{r}_i)^2 =$$

$$= \frac{1}{24}(n^2 - 1) + \frac{1}{16}(n+1)^2 =$$

$$= \frac{1}{48}(n+1)(5n+1) \quad . \qquad \text{(vlg. §11.3)}$$

Verder is

$$\operatorname{cov}(\underline{p}_i,\underline{p}_j) = \mathcal{E}(\underline{p}_i\underline{p}_j) - (\mathcal{E}\underline{p}_i\,\mathcal{E}\underline{p}_j) = \tfrac{1}{4}\operatorname{cov}(\underline{r}_i,\underline{r}_j) =$$

$$= -\tfrac{1}{4}\cdot\frac{1}{12}(n^2-1)\cdot\frac{1}{n-1} = -\frac{1}{48}(n+1) \quad .$$

Tenslotte is dus

$$\operatorname{var}\underline{R}_+ = \operatorname{var}\sum_1^n \underline{p}_i = n\operatorname{var}\underline{p}_i + n(n-1)\operatorname{cov}(\underline{p}_i,\underline{p}_j) =$$

$$= \frac{1}{48}n(n+1)(5n+1) - \frac{1}{48}n(n-1)(n+1) =$$

$$= \frac{1}{24}n(n+1)(2n+1).$$

11.4. De rangcorrelatietoets van Spearman.

In paragraaf 8.1 werd, uitgaande van een tweedimensionale normale verdeling, de toets voor H_0: $\rho = 0$ behandeld. Een verdelingsvrij analogon hiervan is de toets van Spearman. Hiertoe worden de gepaarde waarnemingen (x_i, y_i) gerangnummerd, d.w.z. de x-waarnemingen en de y-waarnemingen *afzonderlijk*. Tussen deze rangnummers wordt dan de (steekproef) correlatiecoëfficiënt uitgerekend. Het is eenvoudig na te gaan dat dit oplevert:

$$r_s = 1 - \frac{6 \Sigma d_i^2}{n^3 - n},$$

waarin d_i het verschil tussen de rangnummers in het i-de paar voorstelt. Tabel S.C. 7.4 geeft kritieke waarden van

$$S := \Sigma d_i^2 .$$

Tabel S.C. 7.4 berust voor het grootste deel op een benadering voor de verdeling van S. Tot en met n = 16 zijn de waarden exact. Deze waarden geven we in tabel 11.4.1.

Tabel 11.4.1. Kritieke waarden voor de tweezijdige rangcorrelatietoets van Spearman.

	onbetrouwbaarheid (tweezijdig)							
n	0.01	0.02	0.05	0.10	0.10	0.05	0.02	0.01
4	-	-	-	0	20	-	-	-
5	-	0	0	2	38	40	40	-
6	0	2	4	6	64	66	68	70
7	4	6	12	16	96	100	106	108
8	10	14	22	30	138	146	154	158
9	20	26	36	48	192	204	214	220
10	34	42	58	72	258	272	288	296
11	54	64	84	102	338	356	376	386
12	78	92	118	142	430	454	480	494
13	108	128	160	188	540	568	600	620
14	146	170	210	244	666	700	740	764
15	194	222	268	310	810	852	898	926
16	248	284	338	388	972	1022	1076	1112

De verdeling van r_S en van S onder de nulhypothese berust weer op het feit dat als H_0 waar is alle n! rangschikkingen van de x-waarde, gegeven de volgorde van de y-waarden, even waarschijnlijk zijn.

Voorbeeld 11.4.1. Stel dat 10 kinderen gerangschikt zijn naar hun bekwaamheid in wiskunde en in muziek. De rangnummers zijn als volgt

Wiskunde	:	7	4	3	10	6	2	9	8	1	5
Muziek	:	5	7	3	10	1	9	6	2	8	4

We berekenen hieruit

$$S = \Sigma d_i^2 = 2^2 + 3^2 + \ldots + 1^2 = 182$$

$$r_S = 1 - \frac{6 \times 182}{1000 - 10} = -0{,}103 \ .$$

We zien in de tabel dat S = 182 verre van significant is.

Voorbeeld 11.4.2 (evenals voorbeeld 1 ontleend aan Kendall: Rank Correlation Methods). Twee juryleden rangschikken een aantal deelneemsters aan een schoonheidswedstrijd:

Jurylid A :	1	2	3	4	5	6	7	8	9
Jurylid B :	2	5	1	3	4	7	6	9	8

$S = 20$

$$r_S = 1 - \frac{6 \times 20}{729 - 9} = 0{,}83 \ .$$

Het ligt hier voor de hand om éénzijdig te toetsen, maar de hypothese dat er geen correlatie is kan hier zelfs worden verworpen als tweezijdig wordt getoetst met $\alpha = 0{,}01$.

Opmerking. In beide voorbeelden werden de gegevens al in de vorm van rangnummers gepresenteerd. Maar de methode is even goed toepasbaar als we met waarnemingen uit een continue (2-dimensionale) verdeling te maken hebben die we eerst moeten rangschikken.

11.5. De methode der m rangschikkingen.

Deze methode (ontwikkeld door Friedman) wordt toegepast in het geval dat we meer dan 2 rangschikkingen (of groepen rangnummers van continu verdeelde waarnemingen) hebben.
Als in voorbeeld 11.4.2 nog een 3e jurylid meedoet zijn de rangschikkingen bijv. als volgt:

Voorbeeld 11.5.1.

Jurylid A :	1	2	3	4	5	6	7	8	9
Jurylid B :	2	5	1	3	4	7	6	9	8
Jurylid C :	5	4	1	7	2	8	3	6	9
Totalen :	8	11	5	14	11	21	16	23	26

We zouden per paar juryleden de rangcorrelatie kunnen berekenen, maar we willen graag één getal hebben voor de mate van overeenstemming van de 3 juryleden als groep.
Dit gaat als volgt: De rangnummers worden per object opgeteld. Als er m rangschikkingen zijn (hier: $m = 3$) en n objecten (hier: $n = 9$) dan is het gemiddelde van de kolomtotalen $\frac{1}{2}m(n+1) = 15$. De afwijkingen t.o.v. dit gemiddelde zijn:

$$-7 \quad -4 \quad -10 \quad -1 \quad -4 \quad 6 \quad 1 \quad 8 \quad 11$$

De som van de kwadraten van deze afwijkingen is

$$S = 404 \ .$$

Deze som is maximaal

$$\frac{1}{12} m^2(n^3 - n) \ .$$

Dit is het geval als alle rangschikkingen precies overeenstemmen.
Dus

$$W := \frac{12S}{m^2(n^3 - n)}$$

is een maat voor de overeenstemming met $0 \leq W \leq 1$.
In ons voorbeeld is

$$W = \frac{12 \times 404}{9(729 - 9)} = 0,75 \ .$$

In dit voorbeeld hebben we tot nu toe de nadruk gelegd op de beoordelaars. Vrijwel altijd gaat het echter in de eerste plaats om de beoordeelde objecten. De nulhypothese is dat er geen waarneembare verschillen tussen de objecten zijn. Als die er wel zijn zullen de beoordelaars globaal overeenstemmen en wordt S en daarmee W groot. De toets is daarom eenzijdig.
Tabel 7.5 van het S.C. geeft kritieke waarden voor S als $n \leq 7$.
Daarboven gebruiken we de χ^2-benadering:

$$\frac{12S}{mn(n + 1)} \approx \chi^2_{n-1} \ .$$

In ons voorbeeld:

$$\frac{12S}{mn(n + 1)} = \frac{12 \times 404}{3 \times 9 \times 10} = 17,96 \ .$$

Dit is groter dan de kritieke waarde van χ^2_8 voor $\alpha = 0,05$, dus er is een duidelijke aanwijzing dat de nulhypothese (geen verschillen tussen de deelneemsters) moet worden verworpen.

11.6. De k-steekproeventoets van Kruskal-Wallis

In 11.3 werd de toets van Wilcoxon besproken voor het probleem van twee steekproeven. Een voor de hand liggende generalisatie is het probleem van k-steekproeven ($k > 2$). De nulhypothese is dat k-steekproeven

$$\underline{x}_{11},\cdots,\underline{x}_{1n_1} \quad ,$$

$$x_{21},\cdots,\underline{x}_{2n_2} \quad ,$$

$$x_{k_1},\cdots,\underline{x}_{kn_k} \quad ,$$

allen uit dezelfde continue verdeling F komen. De alternatieve hypothese is dat de verdelingen verschillende geniddelden hebben. Bij de toets van Kruskal-Wallis worden alle $n := \sum_1^k n_i$ waarnemingen in opklimmende volgorde gerangschikt en van een rangnummer voorzien. Het rangnummer van $\underline{x}_{ij}$ wordt aangeduid met $\underline{r}_{ij}$. Per steekproef wordt het gemiddelde rangnummer berekend:

$$\bar{\underline{r}}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} \underline{r}_{ij} \ .$$

Aangezien de som van alle rangnummers gelijk is aan $\frac{1}{2}n(n+1)$ is het over-all gemiddelde van de rangnummers

$$\bar{r}_{..} = \tfrac{1}{2}(n+1).$$

Als H_0 juist is zullen de gemiddelde rangnummers per steekproef niet al te veel afwijken van het over-all gemiddelde. De toetsingsgrootheid

11.6.1.
$$\underline{K} = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\bar{\underline{r}}_{i.} - \tfrac{1}{2}(n+1)\right)^2$$

is een maat voor deze afwijkingen. Uit (12.3.4) volgt dat de variantie van $n_i\bar{\underline{r}}_i$ gelijk is aan

$$\mathrm{var}(n_i\bar{\underline{r}}_i) = \mathrm{var}\left(\sum_1^{n_i} \underline{r}_{ij}\right) = \frac{1}{12} n_i(n - n_i)(n+1) \ .$$

Dit is in te zien door te bedenken dat $\sum_1^{n_i} \underline{r}_{ij}$ de toetsingsgrootheid van Wilcoxon is als we de i^e steekproef zouden toetsen tegen de overige $(k-1)$ bij elkaar genomen.

Dus

11.6.2. $$\mathcal{E}_{\underline{K}} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{1}{12}(n - n_i)(n + 1) = (k - 1) .$$

De exacte verdeling van $\underline{K}$ is zeer lastig te berekenen, behalve voor kleine k en kleine n_i ($k \leq 3$ en $n_i \leq 5$) .
Als $k > 3$ en alle $n_i \geq 4$ kan de verdeling van $\underline{K}$ worden benaderd door de χ^2-verdeling met $(k - 1)$ vrijheidsgraden. Het is niet moeilijk te bewijzen dat in plaats van formule (12.6.1) ook de volgende kan worden gebruikt:

11.6.3. $$\underline{K} = \frac{12}{n(n + 1)} \sum_{i=1}^{k} \frac{\underline{r}_{i.}^2}{n_i} - 3(n + 1)$$

waarin $\underline{r}_{i.} = \sum_{j=1}^{n_i} \underline{r}_{ij}$.

Voorbeeld 11.6.1

In de volgende tabel staan de groeicijfers van 25 ratten na 12 weken op 4 verschillende diëten.

Dieet	Groeicijfers	Rangnummers
A	257,205,206,164,190,214,228,203	22,10,11, 1, 4,14,18, 8
B	201,231,197,185	6,20, 5, 2
C	248,265,187,220,212,215,281	21,23, 3,16,13,15,25
D	202,276,207,204,230,227	7,24,12, 9,19,17

Hier is $r_{1.} = 88$, $r_{2.} = 33$, $r_{3.} = 116$, $r_{4.} = 88$

$$K = \frac{12}{25 \cdot 26}\left(\frac{88^2}{8} + \frac{33^2}{4} + \frac{116^2}{7} + \frac{88^2}{6}\right) - 3.26 = 4{,}21 .$$

In tabel S.C 3.1 zien we dat $\chi_3^2(0{,}10) = 6{,}25$ en $\chi_3^2(0{,}25) = 4{,}11$. De overschrijdingskans ligt dus dicht bij 25% en er is geen reden om de nulhypothese te verwerpen.