EXAMEN STATISTISCH ANALIST

1952 – 1964

ALGEMEEN GEDEELTE

Georganiseerd door de Vereniging voor Statistiek

UITGEGEVEN DOOR HET MATHEMATISCH CENTRUM AMSTERDAM

SP 85

Voorwoord

Deze verzameling vraagstukken bevat de opgaven (met de antwoorden) van de eerste twaalf Examens Statistisch Analist, Algemeen Gedeelte, die in de periode van 1952 tot en met 1964 onder auspiciën van de Vereniging voor Statistiek werden gehouden.

De opgaven werden samengesteld door leden van de Examencommissie van de Vereniging voor Statistiek. Zij werden reeds verspreid gepubliceerd in de jaargangen van Statistica Neerlandica.

Onder redactie van drs. R. DOORNBOS, werkzaam bij de Unilever en lid van de Examencommissie, en op verzoek van de Vereniging voor Statistiek, publiceerde het Mathematisch Centrum in 1961 de opgaven met antwoorden van de eerste acht examens. Daarbij werd, o.a. ten behoeve van een uniforme notatie, de oorspronkelijke formulering van de opgaven op een aantal plaatsen enigszins gewijzigd. Ook in de antwoorden werden in vergelijking met de in Statistica Neerlandica gepubliceerde standaardoplossingen vrij veel veranderingen aangebracht.

De huidige verzameling vraagstukken bestaat uit een ongewijzigde herdruk van deze in 1961 verschenen collectie, aangevuld met de opgaven met oplossingen van de examens die in de jaren 1961 tot en met 1964 zijn gehouden. Deze laatsten werden langs fotografische weg uit Statistica Neerlandica overgenomen.

Amsterdam, 1965.

VRAAGSTUKKEN

Eerste examen, december 1952

1. Ter vergelijking van twee chemische, physische of biologische meetmethoden A en B worden n (niet noodzakelijk verschillende) objecten volgens beide methoden gemeten. De waarnemingsresultaten zijn $a_1, a_2, \ldots, a_n$ resp. $b_1, b_2, \ldots, b_n$, waarbij de metingen a_i en b_i $(i = 1, \ldots, n)$ aan hetzelfde object verricht zijn. Het gaat er om na te gaan, of misschien één van beide meetmethoden een systematisch hogere meetuitkomst geeft dan de andere.
   Welke statistische methoden komen in dit geval in aanmerking?
   Geef een korte discussie der verschillende mogelijkheden en der voorwaarden, waaronder ieder daarvan toepasbaar is. (Antwoord op blz.49.)

2. Een gemeentelijke belastingdienst wil, onafhankelijk van de belastinggegevens, nagaan hoeveel honden er in de gemeente zijn. Men trekt daartoe op aselecte wijze ("at random") 400 adressen uit het woningregister van het bevolkingsregister. De adressen laat men door enquêteurs bezoeken, die vragen of men geen, dan wel een of meer honden bezit. Het resultaat van het onderzoek luidt:

geen honden	247	gevallen
een of meer honden	44	"
geen gehoor	27	"
antwoord geweigerd	82	"
totaal	400	gevallen

   Men redeneert nu: van de 291 gevallen, waarin antwoord werd ontvangen, zijn er 44 "met hond". Er zijn in de gemeente dus $\frac{44}{291}$ maal zoveel honden als adressen.

Welke bezwaren hebt U tegen dit onderzoek en tegen de redenering? (Antwoord op blz.50.)

3. Van een Poisson-verdeling is gegeven

$$\frac{P(\underline{x}=2)}{P(\underline{x}=3)} = 0,9.$$

Welke waarde heeft het gemiddelde $E\underline{x}$? (Antwoord op blz.51.)

4. Aan de Nederlandse woningtelling 1947 zijn de volgende gegevens ontleend.

Aantal personen per huishouding	Alleenwonende huishoudingen		Samenwonende huishoudingen	
	Totaal aantal	Met onvoldoende slaapruimte	Totaal aantal	Met onvoldoende slaapruimte
	in honderdtallen			
1	771	-	1468	3
2	3149	13	2387	25
3	3547	54	1661	57
4	3562	125	899	59
5	2413	130	430	44
6	1480	118	203	30
7	891	93	98	19
8	546	71	49	11
9 en meer	809	143	55	15
Totaal	17168	747	7250	263

Het blijkt dus, dat van de alleenwonende huishoudingen 4,35% onvoldoende slaapruimte heeft en van de samenwonende huishoudingen 3,63%. Kan men hieruit concluderen, dat de toestand op het gebied van de huisvesting voor de samenwonende huishoudingen beter is dan voor de

alleenwonende huishoudingen? (Antwoord op blz.51.)

5. Twee van elkaar onafhankelijke stochastische variabelen $\underline{x}$ en $\underline{y}$ zijn gedefinieerd door de onderstaande verdelingen.

$\underline{x}$ =	1	2	3		$\underline{y}$ =	1	2
P =	0,3	0	0,7		P =	0,4	0,6

Hoe luidt de verdeling van $2\underline{x}$ resp. $\underline{x}+\underline{y}$ resp. $\underline{x}-\underline{y}$? (Antwoord op blz.51.)

6. In een bedrijf wordt een poedervormig product door een pakmachine in grote aantallen per uur in zakjes verpakt. Om een inzicht te krijgen in de variaties, die bij dit doseringsproces in de gewichten van de gevulde zakjes voorkomen, werd een groot aantal steekproeven uit de productie genomen. Hiertoe werden telkens vier opeenvolgende zakjes uit de productiestroom afgezonderd en als steekproef beschouwd.
Tien van deze steekproeven (aselect gekozen) geven het volgende resultaat:

Gewicht der onderzochte zakjes
(verminderd met het nominale gewicht van 40 g):

volgnummer v.d. steekproef									
1	2	3	4	5	6	7	8	9	10
+1.0	+3.7	+2.0	-1.3	-1.1	-0.2	+1.9	+5.5	+0.9	+1.0
+2.6	+2.3	-0.7	+3.1	+1.6	+0.7	+1.0	+1.4	+1.8	+4.5
-1.3	+2.9	+0.9	-0.6	-2.8	+3.3	+2.0	+0.4	+1.2	+0.5
-1.1	+0.8	-1.1	+3.9	+2.3	-1.7	-3.7	-1.1	-1.1	+2.0

De variaties, die tussen vier opeenvolgende exemplaren optreden, zijn een maatstaf voor de nauwkeurigheid van het doseringsproces.
Vermoed wordt echter, dat er nog een andere oorzaak is,

die maakt, dat de spreiding in de gewichten van de inhoud der zakjes groter is, dan op grond van de spreiding der gewichten binnen de steekproeven zou worden verwacht.

a. Onderzoek aan de hand van de 40 gegeven meetresultaten of dit vermoeden gegrond is en of dus, van statistisch standpunt gezien, geadviseerd mag worden een nader onderzoek in te stellen.
Welke onderstellingen liggen ten grondslag aan de door U toegepaste methode voor het analyseren der waarnemingsuitkomsten?

De producten worden door de afnemer op het gewicht gecontroleerd. Deze afnemer eist, dat de inhoud van 5 pakjes - op aselecte wijze uit de partij genomen - tezamen gewogen, tenminste 200 gram bedraagt. De leverancier vermoedt, dat hij met deze afnemer ernstige moeilijkheden zal krijgen, indien hij in meer dan één op de tien gevallen niet aan deze eis voldoet. Deze moeilijkheden wenst hij te vermijden. De hiervóór gegeven meetresultaten zijn in alle opzichten representatief voor de resultaten van dit doseringsproces.

b. Op welk gewicht moet de leverancier - hangende het onder a bedoelde onderzoek - zijn pakmachine afstellen, zodat hij geen moeilijkheden met zijn afnemer krijgt? (Antw.blz.52.

7. In Bibelonië was de productie van gas en electriciteit als volgt:

Jaar	Gas (in 10^6 m^3)	Electriciteit (in 10^6 kWh)
1920	78	11
1930	107	31
1935	114	44
1940	142	57
1949	174	79
1950	183	85

Teken een grafiek, waarin op duidelijke wijze de ontwikkeling van gas- en electriciteitsproductie met elkaar kunnen worden vergeleken.
De tekening kan in potlood worden uitgevoerd, doch dient overigens met de nodige zorg te worden samengesteld. (Antwoord op blz.53.)

Tweede examen, december 1953

1. Wat is de kans dat althans de grootste van drie waarden, gevonden bij drie onafhankelijke trekkingen uit een normaal universum (met gemiddelde nul en standaarddeviatie één), groter is dan twee? (Antwoord op blz.54.)
2. Uit een zeker universum trekt men een steekproef I van 15 elementen. Aan ieder van de elementen dezer steekproef wordt een waarneming verricht. Vervolgens trekt men een andere steekproef van 20 elementen uit een universum, dat mogelijk ten aanzien van de onderzochte eigenschap op dezelfde wijze is samengesteld. Ook aan de elementen van deze steekproef worden de waarnemingen verricht.

De resultaten zijn in onderstaand overzicht gegeven.

Steekproef I			Steekproef II			
0,44	-1,20	-2,55	1,10	0,16	1,62	-0,67
0,00	2,08	0,30	1,24	1,29	0,36	2,55
0,86	-0,50	1,70	1,00	1,39	-0,34	0,89
1,23	-0,28	1,23	1,97	-1,64	0,83	1,09
-0,05	1,06	-0,60	1,19	1,53	-0,36	1,60

Onderzoek met behulp van de toets van Wilcoxon of beide steekproeven ten aanzien van de onderzochte eigenschap beschouwd kunnen worden als onafhankelijke steekproeven uit hetzelfde universum.

3. Bij een onderzoek in 371 gezinnen naar het aantal verdienende en niet-verdienende gezinsleden verkreeg men de volgende gegevens:

		Aantal verdienende gezinsleden								
		0	1	2	3	4	5	6	7	8
Aantal niet-verdienende gezinsleden	0	-	17	34	8	3	2	1	-	-
	1	25	46	36	30	5	1	1	2	-
	2	8	38	30	8	3	-	-	-	-
	3	1	23	7	3	6	1	-	-	1
	4	-	15	2	5	1	-	-	-	-
	5	-	2	1	-	1	-	-	-	-
	6	-	-	1	-	-	-	-	-	-
	7	-	2	1	-	-	-	-	-	-

1. Bereken het gemiddelde aantal verdienende gezinsleden en het gemiddelde aantal niet-verdienende gezinsleden

per gezin.

2. Bereken het percentage gezinnen dat meer verdienende dan niet-verdienende leden heeft.
3. Bereken de gemiddelde afwijking (t.o.v. de mediaan) van het aantal verdienende gezinsleden.
4. Bereken de correlatie-coefficient voor bovenstaande verdeling. (Antwoord op blz.56.)

4. Bij het keuren van een partij gaat men door tot men 3 afgekeurde exemplaren heeft aangetroffen. Het derde afgekeurde exemplaar blijkt het 50ste onderzochte exemplaar te zijn. Is 6% een goede schatting van het afkeuringspercentage der partij?
   Licht Uw antwoord toe. (Antwoord op blz.58.)

5. Een dame beweert, dat zij bij het drinken van een kopje thee kan proeven, of de melk dan wel de thee er het eerst in gedaan is. Om na te gaan of dit juist is, worden haar achtereenvolgens 20 kopjes thee voorgezet om te proeven. Daarbij worden 2 proefopzetten overwogen:

   1. Bij ieder kopje met een zuivere munt werpen, om te beslissen of de melk dan wel de thee er het eerst ingedaan zal worden.
   2. In 10 kopjes eerst de melk en in 10 kopjes eerst de thee doen en deze vervolgens in een aselecte (gelote) volgorde aan de dame voorzetten.

   De dame wordt in geen van beide gevallen over de proefopzet ingelicht.

   Beantwoord de volgende vragen:

   a. Hoe luidt de te toetsen hypothese?
   b. Welke toets zoudt U in dit geval toepassen?
   c. Welke van de twee proefopzetten zoudt U gebruiken en waarom? (Antwoord op blz.58.)

6. Men beschikt over 10 steekproeven, waarvan de elementen in 2 groepen worden verdeeld, naar gelang zij het kenmerk "A" al dan niet blijken te bezitten:

Nummer v/d steekproef	Aantal elementen	waarvan	
		"A"	"niet-A"
1	57	45	12
2	35	27	8
3	31	24	7
4	29	19	10
5	43	32	11
6	32	26	6
7	112	88	24
8	32	22	10
9	34	28	6
10	32	25	7

1. Toets voor elk van de 10 steekproeven de hypothese dat de steekproef afkomstig is uit een universum waarin 75% der elementen het kenmerk "A" bezit:
   a) met behulp van de normale verdeling;
   b) met behulp van de χ^2-toets.
   Welk verband bestaat er tussen de onder a) en b) gemaakte berekeningen?
2. Vat de 10 steekproeven samen tot 1 steekproef en ga na met behulp van de χ^2-toets of het percentage "A-elementen" daarin significant van 75% afwijkt.
3. Bepaal de som van de 10 waarden van χ^2 welke onder 1b) zijn gevonden en toets de significantie van deze grootheid.
4. Maak een berekening, als onder 3 gevraagd, waarbij thans echter niet wordt verondersteld, dat in het uni-

versum het percentage "A-elementen" 75% is, doch waarbij dit percentage wordt geschat op grond van het gezamenlijk resultaat der 10 steekproeven.

5. Bespreek de betekenis van de onder 2, 3 en 4 gemaakte berekeningen.

Indien bijv. (eventueel bij andere cijfers voor de uitkomsten der verschillende steekproeven) significante waarden voor χ^2 worden gevonden, welke conclusies kan men daaruit dan trekken?
Zou het bijv. kunnen voorkomen dat onder 2 een niet-significante waarde van χ^2 wordt gevonden, terwijl de onder 3 gevonden waarde van χ^2 wel significant is?

6. Wanneer de gestelde hypothese juist is, zijn de onder 1b. berekende waarden van χ^2 uiteraard verdeeld volgens een χ^2-verdeling.
   Stel nu dat men beschikt over een groter aantal steekproeven - bijv. 100 - en dus over meer waarden van χ^2.
   Op welke wijze zou men kunnen controleren of deze 100 waarden inderdaad kunnen worden opgevat als een steekproef uit een χ^2-verdeling?

Aanwijzing: Vorm van de 100 χ^2-waarden een frequentieverdeling, en kies de klassegrenzen zodanig dat de theoretische frequentie in elke klasse onmiddellijk uit de χ^2-tabel kan worden afgeleid. (Antwoord op blz.60.)

7. De emigratie van Nederland naar het buitenland in 1949 omvatte 58185 personen. Deze waren als volgt naar leeftijd en nationaliteit (Nederlanders en vreemdelingen) verdeeld (Bron: Statistisch Zakboek 1950):

Leeftijd	Aantal emigranten	Waarvan Nederlanders
0 - 14 jaar	15197	13906
15 - 19 jaar	3419	2923
20 - 29 jaar	15908	13368
30 - 49 jaar	19649	17848
50 - 64 jaar	3211	2722
65 jaar en ouder	801	585
Totaal	58185	51352

Teken een histogram, waarin de verdeling naar leeftijd en nationaliteit duidelijk tot uitdrukking wordt gebracht. Aan de uitvoering van de tekening - die in potlood mag geschieden - dient de nodige zorg te worden besteed. (Antwoord op blz.62.)

Derde examen, mei 1955

1. Welk percentage ongeveer van de gezinnen met 4 kinderen zal 3 jongens en 1 meisje bezitten? (Antwoord op blz.63.)
2. Een stochastische grootheid $\underline{x}$ kan de waarden 1 en 2 aannemen, met waarschijnlijkheden resp. $\frac{1}{4}$ en $\frac{3}{4}$.
   Een van $\underline{x}$ onafhankelijke stochastische grootheid $\underline{y}$ kan de waarden 1, 2 en 3 aannemen, met waarschijnlijkheden resp. $\frac{1}{6}$, $\frac{1}{3}$ en $\frac{1}{2}$. Wat is de waarschijnlijkheidsverdeling van $\underline{x}^2+3\underline{y}$? (Antwoord op blz.63.)
3. In stad A vindt men in een steekproef van 900 personen 425 gebruikers van een bepaald artikel. In stad B vindt men 825 gebruikers op 1600 personen. De vraag of er een significant verschil bestaat tussen het percentage verbruikers in A en B werd als volgt opgelost:

Neem aan dat er geen significant verschil bestaat tussen A en B. De beste schatting van het percentage verbruikers is dan $100.\frac{425+825}{900+1600}\ \% = 50\%$. Indien de hypothese juist is, kan men verwachten (95% kans) dat de steekproef in A een resultaat oplevert binnen de grenzen: $\frac{1}{2} \pm 1{,}96\sqrt{\frac{\frac{1}{2}\ x\ \frac{1}{2}}{900}}$. Dit is inderdaad het geval.
Eveneens kan men verwachten dat de steekproef in B een resultaat oplevert binnen de grenzen: $\frac{1}{2} \pm 1{,}96\sqrt{\frac{\frac{1}{2}\ x\ \frac{1}{2}}{1600}}$.
Aangezien ook dit het geval is, is er geen reden de hypothese te verwerpen.
Zet Uw eventuele bezwaren tegen deze oplossing uiteen. (Antw.op blz.64.)

4. Van een in serie vervaardigd product wordt ieder exemplaar gekeurd op grove fouten. De keuring is zo streng, dat er voor ieder ondeugdelijk exemplaar slechts een kans 0,001 bestaat, om doorgelaten te worden. Bereken, onderstellende dat de keuringsresultaten onderling onafhankelijk zijn, hoe groot bij benadering de kans is, dat er van 100 ondeugdelijke exemplaren één of meer worden doorgelaten.
Wat kunt U zeggen van de kans, dat er onder 100 willekeurige exemplaren, die de keuring zijn gepasseerd, één of meer ondeugdelijke exemplaren voorkomen? (Antw.op blz. 64.)

5. In een stad wil men gegevens verzamelen omtrent de aldaar wonende gezinshoofden. Daartoe bezoekt men alle gezinshoofden die op enig huisnummer 7 wonen (er bestaan geen nummers als 7a, 7b, iedere woning heeft een eigen nummer).
Gevraagd: Welke bezwaren zouden er kunnen bestaan ten aanzien van de representativiteit van deze steekproef?
(Antwoord op blz.65.)

6. Uit een partij geplukt, ongesorteerd fruit wordt een aselect monster van 12 vruchten genomen. Van iedere vrucht worden lengte (L) en breedte (B) in mm gemeten. Het resultaat is als volgt:

L	B
95	78
89	82
83	71
81	71
108	94
66	53
68	61
83	68
85	77
90	70
106	86
95	80
1049	891

Gevraagd:

a) Bereken een schatting van de variantie van L, van B en van L - B.

b) Bereken de lineaire regressievergelijking $L = cB + d$.

c) Bereken de correlatiecoefficient tussen L en B.

d) Bereken de residuele variantie van L.

e) De gehele partij fruit wordt in een aantal breedte-klassen gesorteerd. Is de correlatiecoefficient, berekend uit L en B van de vruchten in een bepaalde breedte-klasse, een goede schatting van de correlatiecoefficient in de gehele partij?
Motiveer Uw antwoord.

f) Maak in potlood een behoorlijk ontwerp van een spreidingsdiagram, dat U zoudt kunnen gebruiken in een

rapport over het verband tussen lengte en breedte van dit soort fruit. (Antwoord op blz.65.)

7. Uit 5 fiches, waarop de cijfers 1, 2, 3, 4 resp. 5 staan vermeld, worden er aselect achtereenvolgens twee getrokken, zonder teruglegging van het eerste. Noemt men de getallen, die op de twee getrokken fiches staan resp. $\underline{x}$ en $\underline{y}$, dan bezitten $\underline{x}$ en $\underline{y}$ een simultane waarschijnlijkheidsverdeling. Bereken de correlatiecoëfficiënt $\varrho(\underline{x},\underline{y})$. (Antwoord op blz.66.)

8. In een fabrieksafdeling werken 10 arbeiders. Men kiest steekproeven uit de productie van telkens een andere arbeider. De arbeiders worden in lotingsvolgorde gekozen. Voor de loting maakt men gebruik van nummerplaatjes. Iedere arbeider heeft zo'n plaatje. Men schudt de plaatjes en trekt ze blindelings. Bij de met het getrokken plaatje corresponderende arbeider wordt een steekproef genomen. Het getrokken plaatje wordt niet teruggelegd. Zijn alle 10 arbeiders aan de beurt geweest, dan begint men aan een volgende ronde (met een nieuwe loting).
   Gevraagd: Hoe groot is de kans dat de tweede helft van een ronde tezamen met de eerste helft van de daarop volgende ronde ook weer een ronde vormt, waarin alle arbeiders aan de beurt komen? (Antwoord op blz.68.)

9. In een magazijn ligt een groot aantal platte schijven, waarvan de dikte normaal verdeeld is met een gemiddelde van 12 cm en een standaarddeviatie van 2 cm. Het magazijn is binnenwerks 320 cm hoog. Men heeft de gewoonte 25 schijven op elkaar te stapelen. Bij welk gedeelte van de stapels zal het niet gelukken deze compleet te maken, aangenomen dat de schijven vóór het stapelen grondig door elkaar liggen? (Antwoord op blz.68.)

10. De volgende gegevens zijn ontleend aan de resultaten der volkstellingen van 1920 en 1930.

Provincies	% Rooms Katholieken	
	1920	1930
Groningen	5,7	5,4
Friesland	7,0	7,0
Drenthe	6,2	6,1
Overijssel	27,6	28,4
Gelderland	36,1	36,6
Utrecht	31,7	31,0
Noordholland	27,2	27,2
Zuidholland	24,0	23,8
Zeeland	25,7	25,1
Noordbrabant	89,1	88,6
Limburg	94,6	93,4
Het Rijk	35,6	36,4

Voor de provincies zijn de percentages in 1930 vrijwel over de gehele linie lager dan in 1920, voor het Rijk als geheel daarentegen is het percentage over 1930 hoger dan voor 1920. Hoe verklaart U dit? (Antwoord op blz.68.)

11. Het volgende stuk stelt voor het rapport van de statistische analyse van een experiment door een (slecht) statisticus.
Gevraagd wordt:
a) Welke critiek hebt U op de opzet van het experiment?
b) T.a.v. welke punten zijn fouten gemaakt in de zin van ten onrechte of verkeerd toegepaste statistische methoden?
c) Welke controlemaatregelen van statistische aard zijn

nagelaten?

d) Welke andere punten zoudt U als statisticus in Uw rapport nog hebben opgenomen?

Naar rekenfouten behoeft niet te worden gezocht.

Verslag van de statistische verwerking van een experiment over de grondstofwisseling van zuigelingen.

§ 1. Beschrijving van het experiment

Een aantal pasgeboren kinderen werden volgens drie verschillende systemen (I, II en III) gevoed. Bij ieder van deze kinderen werd, voor zover mogelijk op verschillende dagen, de grondstofwisseling (= basaal metabolisme, verder afgekort als B.M.) bepaald; tegelijkertijd werd het gewicht vastgesteld. Daar voor de bepaling van het B.M. het kind volledig rustig moet zijn werd, zo dit het geval was, van de gelegenheid gebruik gemaakt om meer dan één bepaling te verrichten.
De experimenten vielen in twee reeksen uiteen, nl. die verricht in de eerste week na de geboorte en die verricht in de derde week na de geboorte.

§ 2. Uitkomsten

In de tabellen 1 en 2 (zie bijlage) zijn de uitkomsten der waarnemingen opgenomen. Indien bij eenzelfde kind in één week meerdere experimenten (dus op verschillende dagen) verricht werden, zijn deze naast elkaar opgenomen. Indien tijdens hetzelfde experiment, dus op één dag meerdere bepalingen van het B.M. verricht werden, dan staan de uitkomsten onder elkaar opgegeven.

§ 3. Statistische verwerking

A. Verband tussen B.M. en gewicht

In de eerste week is er een sterk significante correlatie tussen gewicht en B.M. (r=0,87; 29 graden van vrij-

heid).
In de derde week was de correlatie tussen gewicht en B.M. niet significant (r=0,16; 23 graden van vrijheid).

B. Verschil tussen jongens en meisjes

In de eerste week was blijkens de t-toets het verschil tussen jongens en meisjes noch ten aanzien van gewicht, noch ten aanzien van B.M. significant. (t=0,02; 9 graden van vrijheid, P > 0,90 voor gewicht en t=0,09; 29 graden van vrijheid, P > 0,90 voor B.M.).
In de derde week was het verschil ten aanzien van het gewicht wel significant, t.a.v. het B.M. niet significant (gewicht: t=9,97; 9 graden van vrijheid, P < 0,001; B.M. t=0,77; 23 graden van vrijheid, P=0,05).

C. Verschil tussen de drie methoden van voeding

Teneinde na te gaan of de verschillende methoden van voeding verschil in gewicht resp. in B.M. gegeven hebben, werden op de uitkomsten van de derde week een tweetal variantie-analyses toegepast. De uitkomsten hiervan waren als volgt:

Gewicht

	Kwadraat som	Graden van vrijheid	Gem. kwadraat-som	F
Totaal	3173,6	10		
Tussen methoden van voeding	2088,1	2	1044,0	7,69
Rest	1085,5	8	135,7	

B.M.

Totaal	1058,96	24		
Tussen methoden van voeding	310,72	2	155,36	4,57
Rest	748,24	22	34,01	

Zowel t.a.v. gewicht als t.a.v. B.M. bestaat er derhalve een significant verschil tussen de drie methoden van voeding. ($P < 0,05$)

Bijlage

TABEL I. Uitkomsten 1e week

Methode van voeding	geslacht	nr kind	1e meting 1)G	BM	2e meting G	BM	3e meting G	BM
I	jongen	1	3400	111	3380	97		
				113		99		
						95		
I	jongen	2	3420	129	3430	124		
						125		
II	meisje	3	3120	103	3090	107	3110	108
						105		111
						108		109
II	meisje	4	3080	97				
				97				
				100				
III	jongen 2)	5	1720	54				
				57				
				53				
				58				
III	meisje	6	3210	92	3180	87		
				88		88		
				93		84		
III	jongen	7	3470	118				
				119				
				117				

1) G = gewicht in grammen. 2) 7-maands kind.

TABEL II. Uitkomsten 3e week

Methode van voeding	geslacht	nr kind	1e meting 1)G	BM	2e meting G	BM
I	jongen	1	3470	120	3490	123
						121
I	jongen	2	3510	111		
				114		
II	meisje	3	3180	117	3200	114
				118		115
II	meisje	4	3140	101	3170	104
				100		102
				102		
III	jongen	5	geen waarnemingen			
III	meisje	6	3260	112	3270	115
				113		115
				111		114
III	jongen	7	3580	102	3600	111
				105		109
						107

1) G = gewicht in grammen (Antwoord op blz.69.)

Vierde examen, oktober 1956

1. Van 16 aselect gekozen verbruikers van een bepaald artikel heeft men nagegaan het verbruik voor en na een reclamecampagne en geconstateerd, dat in 12 gevallen het verbruik is gestegen en in 4 gevallen gedaald. Kan men hieruit concluderen, dat de reclamecampagne succes heeft gehad:
   a) indien men a priori aanneemt, dat de reclamecampagne zeker niet een ongunstige invloed op het verbruik

zal hebben gehad;

b) indien men een ongunstige invloed niet a priori uitsluit?

Men kan ervan uitgaan, dat er, behalve eventueel de reclamecampagne, geen andere invloeden zijn geweest welke het verbruik systematisch hebben beïnvloed. (Antw.op blz.71.)

2. Het resultaat van een worp met een zuivere munt (kruis = 0, munt = 1) telt men op bij het resultaat van een worp met een zuivere dobbelsteen.
   a) Bereken gemiddelde en variantie van deze som.
   b) Wat is de kans, dat deze som kleiner dan 4 is? (Antw.op blz. 72.)

3. Een schooltandarts onderzoekt om het halfjaar, van het moment af, dat een grote groep kinderen op school komt, het gebit van deze kinderen, en tekent telkens aan bij hoeveel kinderen hij voor het eerst één of meer aangetaste melkkiezen vindt.
   Hij vindt de volgende uitkomsten:

halfjaar	
1e	139
2e	75
3e	94
4e	67
5e	34
6e	14
7e en latere	12

   a) Geef deze cijferreeks als histogram weer.
      Wij stellen nu, dat aangenomen mag worden, dat de groep kinderen met aangetaste kiezen een steekproef is uit de populatie van kinderen, die aangetaste melkkiezen krijgen, zodat wij hier een frequentieverdeling van deze kinderen naar de beginleeftijd van aantasting hebben.

b) Geef deze frequentieverdeling op normaal waarschijnlijkheidspapier weer.

c) Geeft een vergelijking tussen het histogram en de andere grafiek aanwijzingen, waaruit verklaard kan worden, dat het aantal gevallen in het tweede halfjaar lager ligt dan in het eerste en derde halfjaar, hoewel verder de frequentie met de leeftijd afneemt?

N.B. Op de netheid, volledigheid en duidelijkheid van het tekenwerk zal worden gelet. (Antwoord op blz.73.)

4. Een automobilist rijdt over de volgende afstanden met de opgegeven snelheden

Afstand in km	Snelheid in km/uur
20	60
100	80
35	90
20	100

Hoeveel km heeft hij gemiddeld per uur gereden? (Antwoord op blz.75.)

5. De volgende waarden stellen een steekproef uit een normale verdeling voor:

6,53
2,83
5,58
4,73
4,41
3,66
5,92
4,48
6,46
5,70

Bereken een schatting van het achtste deciel van deze verdeling. (Antwoord op blz.75.)

6. In een weverij verwerkt men garen, dat op spoelen gewikkeld ontvangen wordt. Het nominale gewicht van het garen per spoel is 300 gram. Men weegt een zeer groot aantal spoelen met garen en vindt een gemiddeld gewicht van 405,2 gram en een standaardafwijking van 12,8 gram. Na het weven weegt men de lege spoelen. Daarbij vindt men een gemiddeld gewicht van 99,7 gram en een standaardafwijking van 11,9 gram.
Men neemt aan, dat het gewicht van het garen op de spoelen normaal verdeeld is en onafhankelijk van het gewicht van de spoel, waar het op gewikkeld is.
Bereken:
   a) welk percentage der spoelen volgens bovenstaande gegevens minder garen bevat dan het nominale gewicht aangeeft;
   b) hoe groot de kans is, dat het gemiddelde garengewicht van een steekproef van 5 spoelen met garen kleiner zal zijn dan het nominale gewicht. (Antwoord op blz.75,)

7. Bij een grootscheepse inentingscampagne tegen een gevaarlijke ziekte, waarbij van een nieuw serum gebruik gemaakt wordt, blijken er na afloop 1730 personen, die ingeënt zijn, de ziekte toch opgelopen te hebben. Het aantal der niet-ingeënten, die de ziekte hebben gekregen, bedraagt 752. Het totale aantal in het onderzoek betrokken personen bedraagt vele tienduizenden. Uit deze resultaten wordt van bepaalde zijde de conclusie getrokken, dat men zich beter niet kan laten inenten.
Is deze uitspraak juist?
Licht Uw antwoord toe. (Antwoord op blz.76.)

8. De lucht in een bepaalde ruimte bevat gemiddeld $2{,}5 \times 10^{19}$ moleculen per cm^3.
   Welke afwijkingen van het gemiddelde aantal moleculen kan men in een volumenelement van 4 cm^3 van deze ruimte verwachten? (Antwoord op blz.76.)

9. Een steekproef bestaat uit 11 waarnemingen.
   Kunt U aangeven hoe de waarden van het gemiddelde, de mediaan en de standaardafwijking veranderen, indien achtereenvolgens:
   a) alle waarnemingen met 10 worden verhoogd;
   b) daarna alle waarnemingen met 2 worden vermenigvuldigd;
   c) daarna de op 2 na grootste waarneming met 11 wordt verhoogd.

   Indien U niet over voldoende gegevens beschikt om alle vragen te kunnen beantwoorden, wordt gevraagd aan te geven welk additioneel gegeven het vraagstuk oplosbaar zou maken. (Antwoord op blz.77.)

10. Een bedrijf heeft een aantal vertegenwoordigers op provisiebasis werken. De provisieregeling is vrij ingewikkeld; zo is bijv. de hoogte van de provisie afhankelijk van het verkochte artikel en van de categorie klant. Omdat de provisie-afrekening administratief veel tijd in beslag nam, heeft men een systeem van afrekening ingevoerd, dat voor orders beneden *f* 1000 (het grootste deel) gebaseerd wordt op steekproeven.
    De methode, die voor iedere vertegenwoordiger apart wordt toegepast, is als volgt:
    a) Berekening en afrekening vinden eens per maand plaats.
    b) Het aantal orders beneden *f* 1000 van de beschouwde vertegenwoordiger in de afgelopen maand (N) wordt vastgesteld.

c) Achteraf wordt door loting een werkdag van die maand als peildatum (i) getrokken.

d) Van alle op de peildatum verkregen orders (beneden *f* 1000) wordt de provisie berekend. Het totaalbedrag aan provisie voor die dag zij P_i.

e) Het gemiddelde provisiebedrag per order, $\bar{p}_i$, wordt nu berekend uit de formule

$$\bar{p}_i = P_i/n_i$$

waarin n_i het aantal orders beneden *f* 1000 op de peildatum is.

f) De totale provisie van de beschouwde vertegenwoordiger wordt berekend als: $N.\bar{p}_i$ + de provisie voor orders boven *f* 1000.

Toen men deze regeling enige tijd had toegepast, bleek dat de vertegenwoordigers systematisch méér provisie ontvingen dan bij volledige berekening (zonder steekproeven). Geef hiervoor een mogelijke verklaring. (Antwoord op blz.78.)

11. Bij een onderzoek naar de materiële welstand van twee bevolkingsgroepen (A en B te noemen) werd van 10 personen uit iedere groep (op aselecte wijze getrokken) het jaarinkomen vastgesteld. De inkomens (afgerond op honderdtallen guldens) bedroegen:

Groep A	Groep B
f 3.100	*f* 4.600
" 3.400	" 600
" 10.900	" 400
" 1.400	" 1.900
" 2.100	" 500
" 6.900	" 2.400
" 3.500	" 2.600
" 4.800	" 4.700
" 6.000	" 5.100
" 6.200	" 2.700

Beantwoord naar aanleiding van deze gegevens de volgende vragen:

a) Onderzoek of de inkomens van één van beide groepen systematisch hoger zijn dan die van de andere groep (onbetrouwbaarheidsdrempel 0,05) en wel:
   a.1. onder de veronderstelling, dat de inkomens normaal verdeeld zijn en in beide groepen dezelfde variantie bezitten;
   a.2. zonder deze veronderstelling te maken.

b) Geef aan en motiveer, welke van beide toetsen U in de praktijk in dit geval zoudt toepassen.

c) Als alle inkomens uit groep B met een gelijk bedrag x (positief of negatief) worden verhoogd, hoe groot moet dan x zijn, opdat bij toetsing van het aldus gewijzigde cijfermateriaal volgens de methode van vraag a.1. juist een tweezijdige overschrijdingskans van 0,01 ontstaat. (Antwoord op blz.79.)

Vijfde examen, oktober 1957

1. Bij een onderzoek naar de afmetingen van de dikte van bepaalde onderdelen uit één produktie-serie werden de volgende resultaten verkregen:

Dikte	Frequentie
10,9 - 11,4	31
11,5 - 12,0	168
12,1 - 12,6	339
12,7 - 13,2	381
13,3 - 13,8	274
13,9 - 14,4	130
14,5 - 15,0	41
15,1 - 15,6	6
Totaal	1370

Gevraagd wordt:
een normale verdeling aan te passen en de bij de klassen behorende theoretische frequenties te berekenen. (Antwoord op blz.83.)

2. Bij een onderzoek werd de relatie tussen twee grootheden x en y onderzocht. In het verslag van dit onderzoek werden slechts de beide regressievergelijkingen als uitkomsten vermeld. Deze luidden:

$$y^* = 2,1x - 7,2$$
$$x^* = 0,7y + 0,3.$$

Gevraagd wordt:

a. Te concluderen of hier met x, y, x^*, y^* al dan niet de afwijkingen van het gemiddelde zijn bedoeld.
b. Volgens welke formule berekent men uit een dergelijk stel regressievergelijkingen de correlatie-coëfficient?
c. Voer de berekening uit en geef Uw commentaar op de uitkomst. (Antwoord op blz.85.)

3. Bij een strenge keuring van rollen weefseldoek (ieder lang 40 m) op uiterlijk is het de bedoeling de rollen met hoogstens 1 fout (zgn. rollen met A-kwaliteit) te scheiden van de overige rollen, die dus ieder twee of meer fouten bevatten (B-kwaliteit).
De keuringsresultaten van 200 rollen waren als volgt:

aantal fouten per rol	aantal rollen
0	70
1	80
2	32
3	15
4	3

Een afnemer heeft belangstelling voor rollen van 60m

lengte, maar alleen van A-kwaliteit.
Indien aangenomen mag worden dat het karakter van de verdeling van de fouten over het doek bewaard blijft, wordt gevraagd:

a. De verdeling van het aantal fouten per rol over de rollen van 60m lengte te schatten.
b. Te berekenen hoe het percentage meters A-kwaliteit wordt gewijzigd.
c. Te berekenen bij welke rollengte het percentage meters A-kwaliteit ca 90 is. (Antwoord op blz.85.)

4. Een verzekeringsagent werkt een lijst adressen in alfabetische volgorde af. Onder de eerste 300 bezoeken zijn er 41, waarbij hij succes heeft.
Bepaal een minimum-percentage aan succesvolle bezoeken voor de overige adressen van de lijst, met onbetrouwbaarheidsdrempel 0,05. Op welke veronderstelling(en) berust de door U gebruikte methode? (Antwoord op blz.86.)

5. Bij een doorlichtingsonderzoek van de bevolking maakte men gebruik van 10 ploegen, die ieder eenzelfde zeer groot aantal personen uit de bevolking onderzochten. Het aantal gevallen, dat voor nader onderzoek in aanmerking kwam, staat voor iedere ploeg in onderstaande tabel vermeld.

Ploeg	Aantal verdachte gevallen
A	137
B	94
C	103
D	62
E	21
F	138
G	110
H	60
I	43
J	42

a. Uit de getallen is wel te zien, dat de verschillende ploegen tot een significant verschillend resultaat kwamen. Hoe kunt U statistisch bewijzen, dat deze uitkomsten inderdaad significant verschillen? (Geef slechts de methode: geen berekeningen).

b. Welke mogelijke verklaring of verklaringen zou U voor de grote verschillen kunnen geven?

c. Stel, men weet op grond van vorige onderzoekingen, dat het percentage verdachte gevallen onder de 10 groepen van de bevolking resp. bedroeg: 0,25% - 0,19% - 0,17% - 0,08% - 0,05% - 0,25% - 0,18% - 0,09% - 0,08% - 0,06%.

Onderzoek of de nieuwe uitkomsten al dan niet in strijd zijn met de vroegere bevindingen. (Antwoord op blz.86.)

6. Een bedrijf exporteert 3 artikelen, A, B en C. De waarde van de export in 1950 en 1956 is gegeven in de volgende tabel:

	waarde van de export (mln guldens)	
Artikel	1950	1956
A	51	85
B	25	40
C	12	16

a. Bereken een indexcijfer voor de waarde van de totale export in 1956 (1950=100).

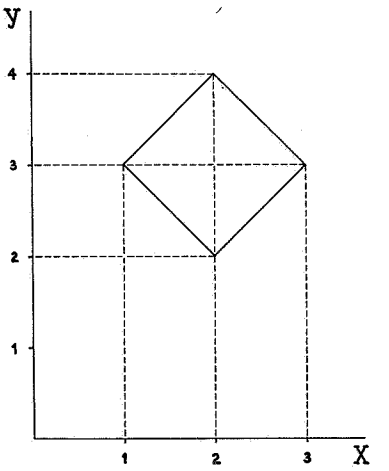
b. Splits het gevonden indexcijfer in een prijscomponent en een hoeveelheidscomponent, indien gegeven is dat de exportprijzen van de artikelen A, B en C in 1956 resp. 10%, 20% en 25% hoger waren dan in 1950. (Antwoord op blz.88.)

7. Een schietschijf bestaat uit 3 concentrische cirkels. Het treffen van de binnenste cirkel levert een score

van 3 punten, van de andere cirkels is de score resp. 2 en 1. De kansen dat men de schietschijf niet raakt, resp. de buitenste, middelste en binnenste cirkel raakt vormen een meetkundige reeks met reden $\frac{1}{2}$.
Gevraagd wordt:
a. de verwachting van het kwadraat van de score.
b. de standaardafwijking van de gemiddelde score bij 25 schoten. (Antwoord op blz.88.)

8. De hoogste waterstand van het jaar, die op een bepaald punt van de kust waargenomen wordt, bedraagt voor een 20-tal op elkaar volgende jaren (in m boven Nieuw Amsterdams Peil): 2,75; 2,93; 2,85; 2,40; 3,05; 2,90; 3,41; 2,60; 2,99; 2,40; 3,25; 3,01; 2,87; 3,07; 2,71; 3,85; 2,87; 3,30; 3,10; 2,98.
Ga na of deze waterstanden een verloop met de tijd vertonen. (Antwoord op blz.89.)

9. Twee stochastische grootheden $\underline{x}$ en $\underline{y}$ bezitten een verdeling van het volgende type. Over het vierkant, dat in de figuur (blz. 92) is aangegeven, is de verdeling homogeen, terwijl de kans om daarbuiten terecht te komen, gelijk aan 0 is.



Gevraagd wordt:

a. Hoe groot is de correlatiecoëfficiënt van x en y?

b. Zijn x en y stochastisch onafhankelijk? (Antwoord op blz.90.)

Zesde examen, november 1958

1. In een bepaalde afdeling van een bedrijf komt gemiddeld per jaar één ernstig ongeval voor. In de eerste 3 jaar volgende op de vervanging van de oude machines door nieuwe machines van een ander type komen in totaal 10 dergelijke ongevallen voor, nl. 3 in het eerste jaar, 4 in het tweede en 3 in het derde jaar. De personeelssterkte van de afdeling is in deze 3 jaar niet noemenswaard veranderd ten opzichte van die in de voorafgaande periode.
   Kan men uit deze gegevens de conclusie trekken, dat het werken met de nieuwe machines gevaarlijker is dan met de oude?
   Is het bij de beantwoording van deze vraag nodig de grootte van de personeelsbezetting te kennen? (Antwoord op blz. 91.)

2. Gegeven is de volgende tabel:

   Gezinnen naar inkomen en uitgaven voor voeding in guldens per jaar

Uitgaven voor voeding	Inkomen							
	4600-5199	5200-5799	5800-6399	6400-6999	7000-7599	7600-8199	8200-8799	8800 en meer
	Aantal gezinnen							
1600-1999	-	1	1	1	-	-	-	-
2000-2399	6	1	2	1	1	-	-	-
2400-2799	1	5	9	4	2	-	1	-
2800-3199	-	3	4	8	6	3	-	1
3200-3599	-	-	2	3	5	3	2	1
3600-3999	-	-	-	2	1	3	3	-
4000 en meer	-	-	-	-	1	-	2	1

Gevraagd wordt de correlatiecoëfficient tussen inkomen en uitgaven voor voeding te berekenen (de bedoeling is niet alleen de uitkomst maar ook de berekeningswijze te geven). (Antwoord op blz.92.)

3. In een bepaald land wil men overgaan tot een andere wijze van honoreren van tandartsen voor hun ziekenfondspraktijk. Teneinde hiervoor de nodige gegevens te verkrijgen wordt 10% van alle tandartsen door middel van een aselect proces uitgekozen. De zo gekozen tandartsen tekenen gedurende één jaar op hoeveel tijd zij besteden aan alle patiënten, die wegens klachten op hun spreekuur komen, benevens het aantal van dergelijke patiënten. Per tandarts i wordt hieruit de gemiddelde duur d_i van een dergelijk consult berekend.
Men kan nu op twee wijzen het algemeen gemiddelde van de duur van een consult berekenen:

   I. Door de waarden van d_i te middelen over alle tandartsen.
   II. Door de totale tijd door de tandartsen aan dergelijke consulten besteed te sommeren en deze som te delen door de som van het aantal patiënten, die de tandartsen bezocht hebben.

   Vragen:

   a. Zullen de uitkomsten berekend volgens I steeds gelijk zijn aan die berekend volgens II?
   b. Men is voornemens de tandartsen te honoreren per consult.
   Men moet dus de gemiddelde duur van één consult kennen om een zodanige hoogte van het honorarium per consult te kunnen vaststellen dat bij een redelijk geachte werktijd per dag een redelijk inkomen wordt verkregen.

Zoudt U de gemiddelde duur per consult volgens methode I of volgens methode II berekenen?
Licht Uw antwoord toe. (Antwoord op blz.93.)

4. Op een chemisch laboratorium wordt een bepaald soort routinebepaling op een monster steeds in duplo (tweevoud) uitgevoerd. Het is bekend, dat de werkelijke gehalten van monster tot monster kunnen verschillen; voorts wordt bij elke bepaling uitgegaan van eenzelfde hoeveelheid materiaal. Indien de duplo-uitkomsten meer dan 0,05 verschillen, worden beide bepalingen overgedaan; de eerst gevonden waarden worden dan weggeworpen.
Na een zekere periode, waarin zeer vele van dergelijke bepalingen verricht zijn, wordt uit alle duplobepalingen, die geaccepteerd waren, de standaardafwijking berekend uit de gemiddelde spreidingsbreedte (door vermenigvuldiging met 0,886).
Geef een gemotiveerd antwoord op de volgende vragen:
a. Men kan als eis stellen dat duplobepalingen overgedaan moeten worden indien de absolute waarde van hun onderling verschil een bepaalde waarde overschrijdt (in ons geval dus 0,05), maar ook dat de bepalingen overgedaan moeten worden indien het verschil een bepaald percentage van het gemiddelde der twee uitkomsten overschrijdt (indien dit percentage 10 bedraagt zullen uitkomsten 29 en 31, met een verschil van $100 \times \frac{31-29}{\frac{1}{2}(29+31)} = 6{,}7\%$ niet en uitkomsten 9 en 11, met een verschil van $100 \times \frac{11-9}{\frac{1}{2}(11+9)} = 20\%$ wel overgedaan moeten worden).
Waarvan zoudt U de keuze tussen deze beide methoden laten afhangen?

b. Geeft de door het laboratorium berekende standaardafwijking een juist beeld van de werkelijke nauwkeurigheid?
c. Geef aan of U de bovenbeschreven wijze van handelen (overdoen van beide bepalingen bij een verschil van meer dan 0,05) redelijk acht indien:
   1. als standaardafwijking wordt gevonden 0,012;
   2. als standaardafwijking wordt gevonden 0,028.
d. Geeft de toegepaste methode (overdoen van beide bepalingen bij een verschil van meer dan 0,05) zuivere schattingen van de gemiddelden?

Voor ieder der vier vragen dient afzonderlijk aangegeven te worden, welke van de onderstaande veronderstellingen U bij Uw beantwoording gebruikt:

a. Onafhankelijkheid der waarnemingen.
b. Normaliteit.
c. Gelijke spreiding.
d. Mogelijkheid van blunders bij het waarnemen. (Antwoord op blz.95.)

5. Een bepaald soort artikel is aan breuk tijdens transport onderhevig. Als een exemplaar met een breeksterkte $\underline{b}$ tijdens het transport een kracht $\underline{k}$ ondergaat die groter is dan b, breekt het exemplaar. Is $\underline{k}$ kleiner dan $\underline{b}$, dan breekt het exemplaar niet.
   Over een lange tijdsperiode genomen bleek gemiddeld 3,0% breuk te zijn opgetreden. Over deze zelfde periode bleken de breeksterkten volgens genomen steekproeven vrijwel normaal verdeeld te zijn met gemiddelde μ_b=200 en standaardafwijking σ_b=40.
   De leiding van het bedrijf vond het percentage breuk te hoog en besloot door gewijzigde samenstelling de breeksterkte op te voeren. Uit nieuwe proeven bleek de gemiddelde breeksterkte nu μ_b=231,5 te zijn geworden. Vorm van de verdeling en spreiding bleven ongewijzigd. Als gemiddeld breukcijfer werd 0,6% gevonden.

Van de grootste tijdens transport optredende breekkrachten $\underline{k}$ mag worden aangenomen dat ze normaal verdeeld zijn met gemiddelde μ_k en standaardafwijking σ_k en dat ze onafhankelijk van $\underline{b}$ zijn.
Gevraagd wordt:

a. Bereken μ_k en σ_k.
b. Hoe groot moet bij gelijkblijvende spreiding de gemiddelde breeksterkte (μ_b) worden om een breuk van gemiddeld 1,1% te kunnen verwachten? (Antwoord op blz.97.)

6. Uit de statistieken van Amerikaanse verzekeringsmaatschappijen bleek, dat in 1955 52000 mannelijke autobestuurders betrokken waren bij ernstige ongelukken en bijna 2000000 bij minder ernstige, terwijl voor de vrouwelijke bestuurders deze aantallen 4000 en 341000 bedroegen.
De voorzitter van de Bond van Voetgangers in Engeland trok hieruit de conclusie, dat vrouwelijke bestuurders voorzichtiger rijden dan mannelijke, omdat onder de ongevallen, waarbij zij betrokken zijn, een kleiner percentage ernstige voorkomt dan bij de mannelijke bestuurders.
Bent U het met deze conclusie eens? Geef commentaar. (Antwoord op blz.99.)

7. Twee stochastische grootheden $\underline{x}$ en $\underline{y}$ zijn onafhankelijk verdeeld. Voor de standaardafwijkingen geldt:

$$\sigma_y = 2\sigma_x.$$

Een grootheid z wordt gedefinieerd door:

$$\underline{z} = 2\underline{x}+\underline{y}.$$

Bereken de correlatiecoefficient $\rho(\underline{x},\underline{z})$. (Antwoord op blz.100.)

8. Voor een proef met 3 verschillende bewerkingsmethoden A, B en C van een kostbaar halffabrikaat zijn 17 exem-

plaren beschikbaar. Deze exemplaren wenst men aselect ("at random") over de 3 methoden A, B en C te verdelen, waarbij 7 exemplaren met methode A, 5 met methode B en 5 met methode C behandeld moeten worden. De exemplaren zijn genummerd met de nummers 1,2,...,17.
Voer de verdeling in 3 groepen uit met behulp van de onderstaande tabel van aselecte getallen.
Geef duidelijk aan, hoe U dit doet en wat het resultaat is.

Tabel van aselecte getallen (0,1,...,9)

7160	6043	0767	0230	6082
3637	4556	6654	8972	9607
7965	7435	8397	9741	6297
2297	6491	7961	0243	6897
6708	0600	2765	1911	0813
2268	3554	7976	4102	0414
4159	6804	3838	4255	9664
7044	3067	6720	7416	4748
6592	1846	2269	9136	7107
0676	9782	8061	2715	2932

(Antwoord op blz.100.)

9. In Sigmanië hebben de dobbelstenen de vorm van een regelmatig viervlak (tetraëder), waarvan de vlakken resp. van 1, 2, 3 en 4 ogen zijn voorzien. Hiermee wordt op de volgende manier gespeeld: De spelers werpen om de beurt met drie stenen. Wie de hoogste uitkomst gooit wint. Als uitkomst geldt de som van het aantal onderliggende ogen, met dien verstande dat, als meerdere stenen in één worp een gelijk aantal ogen opleveren, slechts één van deze aantallen meetelt voor de som. Gooit men dus 2/4/2, dan is de uitkomst 6.
Beantwoord de volgende vragen:
a. Hoeveel verschillende uitkomsten zijn er en wat zijn

de bijbehorende kansen?

b. Als men 1/2/4 heeft gegooid en men mag één der stenen naar keuze opnieuw gooien om zijn worp te verbeteren is het dan verstandig dit te doen?

c. Als de vlakken der stenen resp. van 0, 1, 2 en 3 ogen waren voorzien, zou dan de frequentieverdeling der uitkomsten anders van vorm zijn? (Antwoord op blz. 102.)

10. Uit een Poisson-verdeling met verwachting μ neemt men een steekproef van n waarnemingen. De steekproef heeft gemiddelde $\underline{m}$ en standaardafwijking $\underline{s}$.
Hoe groot is:

a. de verwachting van $\underline{m}$;

b. de verwachting van $\underline{s}^2$;

c. de variantie van $\underline{m}$? (Antwoord op blz. 105.)

11. Van tweehonderd kinderen van elf jaar werd het gewicht bepaald. De gewichten werden in hectogrammen nauwkeurig opgetekend. De volgende tabel geeft de frequentieverdeling van de uitkomsten.

Gewicht in kg	Aantal kinderen
33,0 - 34,9	5
35,0 - 36,9	30
37,0 - 37,9	57
38,0 - 38,9	53
39,0 - 39,9	30
40,0 - 40,9	20
41,0 en meer	5

a. Teken een histogram dat de frequentieverdeling weergeeft.

b. Teken op normaal waarschijnlijkheidspapier de cumulatieve frequentieverdeling.

c. Aangenomen mag worden dat de verdeling van de gewichten van elfjarige kinderen normaal is. Geef in grafiek b aan de lijn die deze verdeling zo goed mogelijk weergeeft. Het is de bedoeling, dat U de lijn niet berekent, doch zo goed mogelijk op het oog trekt.

d. Neem aan, dat de lijn die U in c getrokken heeft, de werkelijke (populatie-) frequentieverdeling is. Leid met behulp van deze lijn, voor de in de tabel gegeven gewichtsklassen, de overeenkomstige klassefrequenties voor de populatie af en teken deze met een stippellijn in het onder a bedoelde histogram.

Op netheid van de tekening en op volledigheid, duidelijkheid en kortheid van het bijschrift zal worden gelet. (Antwoord op blz.105.)

Zevende examen, oktober 1959

1. Van de eerste honderd telefoonnummers, die voorkomen op blz. 17 van gids no. 109 van de Plaatselijke Telefoondienst te Rotterdam is het aantal keren geteld, dat in de laatste positie een 0, 1, 2 ... 9 voorkomt. Hetzelfde is gedaan met de voorlaatste positie. De waargenomen aantallen zijn in de velden van onderstaande tabel vermeld.

Positie \ Cijfer	0	1	2	3	4	5	6	7	8	9
Laatst	24	8	9	9	9	9	11	8	7	6
Voorlaatst	9	7	7	12	12	10	12	11	15	5

Toets met behulp van de χ^2-methode, met een onbetrouwbaarheid van 0,05, de volgende hypothesen:

a. De laatste cijfers van opeenvolgende telefoonnummers in bovengenoemde gids zijn onderling onafhankelijke aselecte trekkingen uit een verdeling met gelijke

kansen 1/10 op de cijfers 0,1,2,...,9.

b. De onder a genoemde hypothese, maar nu voor de op één na laatste cijfers.

c. De laatste cijfers en de op één na laatste cijfers zijn onderling onafhankelijke trekkingen uit dezelfde verdeling met onbekende kansen $p_0, p_1, \ldots, p_9$ op de cijfers 0,1,...,9.

d. Zijn de antwoorden die U in de gevallen a, b en c vindt met elkaar in tegenspraak? (Antwoord op blz.108.)

2. Men heeft van een bepaalde stof 3 verschillende oplossingen gemaakt van sterktes van ongeveer 10%, 15% en 20%. Iedere oplossing wordt over 7 flesjes verdeeld, die aan 7 verschillende laboratoria worden toegezonden met het verzoek door middel van een nauwkeurig omschreven chemische analyse het gehalte van de stof in kwestie te bepalen. De 3 flesjes die ieder laboratorium kreeg toegezonden, waren, naar opklimmende sterkte, gemerkt met A, B en C.
De volgende resultaten zijn verkregen:

Laboratorium	Monster		
	A	B	C
1	10,23	15,17	21,10
2	10,29	15,52	21,18
3	10,22	15,20	21,07
4	10,33	15,31	21,12
5	10,42	15,40	21,82
6	10,15	15,03	21,05
7	10,20	15,19	21,01

Het is bekend, dat de spreiding van de waarnemingsuitkomsten bij toenemend gehalte niet constant is. De

spreiding bij hetzelfde gehalte op verschillende laboratoria mag als constant worden beschouwd.
Gevraagd wordt na te gaan of er systematische verschillen tussen de laboratoria bestaan in die zin dat het ene laboratorium systematisch hogere waarden vindt dan het andere. (Antwoord op blz.109.)

3. Op een werkstaat komen de volgende gegevens voor over het inkomen in een bepaald jaar (x) en de uitgaven voor voeding in datzelfde jaar (y) van 16 gezinnen. De bedragen zijn opgegeven in honderden guldens per jaar.

Inkomen (x)	Uitgaven aan voeding (y)
55	25
58	30
56	24
61	31
31	63
61	29
56	24
50	21
57	28
47	24
58	27
50	23
54	26
64	28
59	26
52	23

a. Stel bovenstaande gegevens voor in een spreidingsdiagram.

b. Bereken een schatting van de toeneming van de uitgaven voor voeding corresponderende met een vermeerdering van het inkomen van 100 gulden.

c. Bereken de correlatiecoefficient tussen x en y. (Antwoord op blz.110.)

4. In een fabriek zijn drie machines die alle hetzelfde produkt vervaardigen. Door kleine verschillen in constructie en door ongelijke slijtagetoestand leveren deze machines een ongelijk percentage foutieve exemplaren op. Uit controles is gebleken dat bij machine A gemiddeld 3,6% foutieve exemplaren optreedt, bij machine B 1,1% en bij machine C 1,6%. De ongesorteerde produkten van alle machines gaan door elkaar naar het magazijn. Ze zijn dan niet meer van elkaar te onderscheiden.
Uit dit magazijn wordt 1 exemplaar gepakt en onderzocht. Dit blijkt foutief te zijn. Hoe groot is de kans dat dit van machine A afkomstig is, als gegeven is dat op het moment van trekking in het magazijn aanwezig waren: 3017 exemplaren van machine A, 2655 exemplaren van machine B en 1614 exemplaren van machine C? (Antwoord op blz.111.)

5. Teneinde de verontreiniging van de lucht in een lokaal te meten, wordt een glazen plaatje gedurende een van te voren vastgestelde tijdsduur horizontaal neergelegd. Daarna wordt met een microscoop het aantal in het gezichtsveld zichtbare deeltjes geteld en wel op elf aselect gekozen plaatsen, die geheel buiten elkaar liggen. Het gezichtsveld van de microscoop heeft een oppervlakte van 3,03 mm^2. De volgende aantallen deeltjes werden gevonden:

117, 103, 89, 85, 97, 104, 112, 92, 94, 99, 108.

Op grond van voorafgaande onderzoekingen is bekend, dat de verdeling van de aantallen deeltjes per gezichtsveld,

en in het algemeen per eenheid van oppervlakte, een Poisson-verdeling is.

a. Gevraagd wordt de beste schatting te geven van het gemiddelde aantal deeltjes per cm^2 en van de variantie van dit gemiddelde.

b. Indien U het gegeven, dat de aantallen deeltjes per gezichtsveld volgens een Poisson-verdeling verdeeld is zou wantrouwen, op welke wijze zoudt U dan een indruk kunnen krijgen over het al dan niet juist zijn van deze veronderstelling. De desbetreffende berekening behoeft niet te worden uitgevoerd. (Antwoord op blz.112.)

6. Voor drie artikelen A, B en C zijn de volgende gegevens beschikbaar over het gemiddelde verbruik per hoofd der bevolking en de gemiddelde prijzen. Het artikel C kwam voor het eerst in periode 1 aan de markt. Met dit feit dient rekening te worden gehouden bij de samenstelling van de hierna gevraagde indexcijfers.

Periode	A		B		C	
	Hoeveelheid in kg	Prijs in gld/kg	Hoeveelheid in kg	Prijs in gld/kg	Hoeveelheid in kg	Prijs in gld/kg
0	10	0,50	5	0,40	-	-
1	10	0,60	5	0,50	5	1,00
2	10	0,55	4	0,55	6	0,75

Gevraagd wordt:

a. een waarde-index zowel als een prijsindex te berekenen, waarbij de periode 0 gelijk 100 wordt gesteld.

b. een hoeveelheidsindex te berekenen (periode 0 is 100). (Antwoord op blz.113.)

7. In 793 gemeenten is in 1958 vermakelijkheidsbelasting geheven, in de overige gemeenten niet. Teneinde een in-

druk te verkrijgen van de totale opbrengst werden de 793 bedragen opgeteld, na weglating van gedeelten van guldens: dus werd bijv. *f* 8123,87 geteld als *f* 8123. De opgegeven bedragen werden dus niet afgerond, maar afgekapt.
Als aangenomen mag worden, dat voor elke gemeente elk der honderd mogelijke uitkomsten in centen -,00; -,01; ...; -,99 even waarschijnlijk is en als gegeven is, dat de uitkomst der optelling *f* 26074281,- bedroeg, wordt gevraagd aan te geven tussen welke grenzen de totale opbrengst van de gemeentelijke vermakelijkheidsbelasting met een betrouwbaarheid van 0,95 zal liggen. (Antwoord op blz.114.)

8. Bij een vermoeidheidsproef wordt een stuk metaal afwisselend belast en ontlast. Het aantal wisselingen in de belasting totdat breuk optreedt is voor een gegeven soort metaal een stochastische grootheid met een logarithmisch-normale verdeling. Men neemt een aselecte steekproef van 200 stukken van een gegeven soort metaal en onderwerpt deze alle aan een vermoeidheidsproef. De uitkomsten staan opgegeven in de volgende tabel.

Aantal wisselingen	Aantal stukken
10 - < 50	2
50 - < 100	3
100 - < 500	14
500 - < 1000	13
1000 - < 5000	44
5000 - < 10000	24
10000 - < 50000	52
50000 - < 100000	16
100000 - < 500000	23
500000 - < 1000000	4
1000000 en hoger	5

Teken een histogram van de verdeling met logarithmische indeling van de abscis. (Antwoord op blz.116.)

9. Er bestaat een kansverdeling die bekend staat als de verdeling van CAUCHY. Deze verdeling bezit o.a. de volgende eigenschappen:

   1. De formule luidt:

   $$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \qquad -\infty < x < \infty .$$

   2. De formule voor de verdelingsfunctie (cumulatieve verdeling) luidt:

   $$F(x) = \frac{1}{\pi}\int_{-\infty}^{x}\frac{dt}{1+t^2} = 0{,}5 + \frac{1}{\pi}\,\text{arctg}x.$$

   3. De verdeling is symmetrisch.
   4. Het gemiddelde $\mu = 0$.
   5. De variantie is oneindig.
   6. Het eerste kwartiel ligt bij $x=-1$, het derde kwartiel bij $x=1$.

   Gevraagd:

   a. Welk gedeelte van de oppervlakte onder de verdeling van CAUCHY ligt tussen de mediaan en het derde kwartiel?

   b. Welke der bovengenoemde eigenschappen zijn nodig voor de beantwoording van vraag a? (Antwoord op blz.117.)

10. De personeelsverenigingen De Kaaymannen van de N.V. Kaay en Zonen wil een Sinterklaasfeest voor de gezinnen van ca 150 personeelsleden organiseren. De organisatoren moeten, om dit plan te concretiseren, over een aantal gegevens beschikken. Gedacht is nl. aan afzonderlijke bijeenkomsten voor de kinderen tussen 4 en 12 jaar, voor de jongens van 12 tot en met 17 jaar en voor de volwassenen.

De leeftijdsafgrenzing is niet alleen nodig met het oog op de samenstelling van het feestprogramma, maar ook in verband met de keuze van de kleine verrassingen die men wil uitreiken. Ook zal men een globaal idee moeten hebben van de aantallen te verwachten deelnemers aan iedere bijeenkomst.
Om een deel van de gewenste informatie te krijgen, zou men de gegevens van de personeelsafdeling kunnen gebruiken. Om allerlei redenen wil en kan men dit niet doen. De voorkeur wordt gegeven aan een eenvoudige schriftelijke enquête onder het personeel. Wel is het van belang dat van zoveel mogelijk personeelsleden een antwoord wordt verkregen.
Gevraagd wordt een beknopte beschrijving te geven van een mogelijke opzet van een dergelijke enquête, een vragenformulier te ontwerpen (dit laatste op een afzonderlijk vel) en zo nodig een motivering van de keuze der vragen en van de formulering te geven. (Antwoord op blz.117.)

Achtste examen, oktober 1960

1. Een bepaalde hoeveelheid van een voedingsmiddel wordt gesuspendeerd in steriel water en met vloeibare voedingsbodem gemengd. Na stollen hiervan wordt bij 20°C gedurende drie dagen gebroed. Daarna wordt het aantal bacteriekolonies geteld. Deze proef is met 15 verschillende monsters uitgevoerd. Hierbij werden de volgende aantallen kolonies geteld:

31	149	5
200	90	270
1	430	62
267	184	24
417	61	53

Daarna werd dezelfde proef uitgevoerd met 15 monsters waaraan een conserveringsmiddel was toegevoegd. De resultaten (één proef mislukte) waren:

96	6	9
68	0	18
239	12	58
16	176	65
26	3	

Toets met een onbetrouwbaarheid van 5% of de toevoeging van dit middel het aantal bacteriekolonies doet afnemen. (Antwoord op blz. 119.)

2. De statistiek van de woningbouw in Nederland bevat gegevens over de grootte van de woningvoorraad op 1 januari van elk jaar, de toeneming van de woningvoorraad gedurende het jaar ten gevolge van a. afbraak, vernietiging door brand, enz., b. onbewoonbaarverklaring, en c. verbouwing en wijziging van bestemming.

   Bovenbedoelde gegevens zijn beschikbaar voor de vier jaren 1956 tot en met 1959.

   Gevraagd wordt ten behoeve van een artikel over de woningmarkt een tabel te ontwerpen waarin de bovengenoemde cijfers voor Nederland over de jaren 1956-1959 op zodanige wijze moeten worden vermeld dat daaruit het verband tussen de woningvoorraad aan het begin en aan het einde van elk jaar en de oorzaken van de toeneming en de vermindering duidelijk blijken.
   Welke is de eenheid waarin U de cijfers zoudt uitdrukken?
   (Te Uwer oriëntering diene dat bij de in 1956 gehouden Algemene Woningtelling het aantal woningen in Nederland ruim 2,5 miljoen bedroeg, terwijl de netto-toeneming per jaar ongeveer 70.000 woningen bedraagt.) (Antwoord op blz. 121.)

3. Een biscuitfabriek fabriceert rollen inhoudende 40 biscuits. De biscuits wegen gemiddels 3 g, met een standaardafwijking van 0,2 g (scheve verdeling). De verpakking weegt gemiddeld 15 g met een standaardafwijking van 0,5 g (normaal verdeeld).

   Gevraagd:

   a. Hoe groot is de kans dat een rol minder dan 131 g weegt?
   b. Als men een steekproef van 5 rollen neemt, wat is dan de verwachting van de spreidingsbreedte (range) der gewichten?
   c. Bereken de correlatiecoëfficiënt tussen het brutogewicht van de rollen en het gewicht van de verpakking. (Antwoord op blz. 121.)

4. Men heeft twee analisten een aantal bepalingen laten verrichten van het verzepingsgetal van kokosolie. De eerste analist vond achtereenvolgens voor monsters, afkomstig uit eenzelfde fles:

   253,8 255,4 256,2 256,1 255,2 255,4

   De tweede analist vond voor monsters uit dezelfde fles:

   253,2 258,5 256,4 255,7 254,2.

   Toets, met een onbetrouwbaarheid van 0,05 de hypothese dat deze beide analisten even nauwkeurig werken. (Antwoord op blz. 123.)

5. Er bestaat een ziekte, die slechts kan optreden, wanneer een persoon daarmede erfelijk belast is. Men mag aannemen, dat in een welomschreven bevolkingsgroep er voor alle erfelijk belaste personen een gelijke kans bestaat om de ziekte te krijgen; niet iedere erfelijk belaste persoon wordt dus ziek. Indien de ziekte echter optreedt, geschiedt dit direct na de geboorte; de ziekte beïnvloedt de sterftekansen niet.

Het is bekend, dat de ziekte via de vrouwelijke lijn overgeërfd wordt, en wel zo, dat als de moeder erfelijk belast is, alle kinderen (jongens zowel als meisjes) eveneens erfelijk belast zijn. (Erfelijk belaste vaders dragen de belasting niet op hun kinderen over.)

Men onderzoekt van een bepaalde bevolkingsgroep een aantal twee-kinder gezinnen, waarin tenminste één der kinderen de ziekte vertoont en wil hieruit afleiden hoe groot de kans, p, is, dat bij een erfelijk belast kind de ziekte tot uiting komt. Men vond 135 twee-kinder gezinnen, waarin tenminste één kind de ziekte had en wel 108 gezinnen, waarin slechts één van beide kinderen de ziekte had en 27 gezinnen, waarin beide kinderen de ziekte hadden.

Gevraagd wordt de kans, p, om de ziekte te krijgen, indien de erfelijke belasting aanwezig is, te schatten. (Antwoord op blz. 124.)

6. Voor een frequentieverdeling (n = 100) heeft men de volgende grootheden berekend:

$$\Sigma x = -414$$
$$\Sigma x^2 = 1030.$$

Ga na of deze uitkomst juist kan zijn. (Antwoord op blz. 124.)

7. Uit een urn met 1 rode, 1 witte en 1 zwarte bal worden, met teruglegging 4 aselecte trekkingen gedaan. We noemen x het aantal rode, y het aantal witte en z het aantal zwarte ballen dat in deze vier trekkingen in totaal getrokken wordt.

Gevraagd wordt te berekenen:

a) De kansverdeling van $w = 2x - y + 2z$

b) De variantie van w. (Antwoord op blz. 125.)

8. Een explosieve stof wordt in de vorm van kleine blokjes vervaardigd. Teneinde van deze stof de gevoeligheid voor schokken na te gaan, laat men bij diverse valhoog-

ten een metalen kubus met een van tevoren vastgesteld gewicht op telkens een nieuw blokje vallen en registreert het aantal blokjes, dat explodeert. Elk blokje wordt slechts éénmaal beproefd.

Men vindt de volgende uitkomsten:

Valhoogte (cm)	Aantal onderzochte blokjes	Aantal geëxplodeerde blokjes
50	50	3
75	100	21
100	100	35
120	100	48
130	100	61
140	100	70
150	100	76
175	100	85
200	100	92
225	50	48
250	50	49

Uit deze waarnemingen kan een schatting gemaakt worden van de fractie der blokjes, die juist exploderen bij valhoogten binnen een bepaald interval, dus bijv. de fractie welke juist explodeert in het interval tussen 120 en 130 cm (ondergrens niet, bovengrens wel inbegrepen).

Gevraagd wordt voor elk der intervallen tussen de in de eerste kolom opgegeven valhoogten, benevens voor het interval 0-50 cm en het interval: groter dan 250 cm bovenbedoelde fracties te berekenen en vervolgens hiervan een histogram (niet cumulatief) te tekenen. (Antwoord op blz. 125.)

9. De volgende indexcijfers zijn beschikbaar over de produktie van bouwmaterialen in Nederland (1953 = 100).

	1956	1957	1958	1959
Jan. - Maart	93	106	100	99
April - Juni	127	135	122	133
Juli - Sept.	127	132	125	133
Okt. - Dec.	114	117	111	122

a. Maak een grafische voorstelling van bovenstaande tijdreeks

b. Hoe stelt U vast of U in dit geval voor de berekening van de seizoenbeweging de additieve of de multiplicatieve methode zult gebruiken?

c. Voer de berekening van de seizoenbeweging uit.

d. Voor het eerste kwartaal 1960 blijkt de index gelijk te zijn aan 108. Bereken het voor seizoenbeweging gecorrigeerde indexcijfer voor het eerste kwartaal 1960. (Antwoord op blz.128.)

ANTWOORDEN

Eerste examen, december 1952

1. De te toetsen hypothese luidt, dat er geen systematisch verschil is tussen de meetmethoden A en B. De volgende toetsen kunnen worden toegepast, indien de daarbij vermelde onderstellingen vervuld zijn.

   1. De tekentoets kan worden toegepast op de n verschillen $a_i - b_i$ $(i=1,\ldots,n)$, indien de 2n waarnemingen onderling onafhankelijk zijn.
   2. Indien de verschillen bovendien normaal verdeeld zijn met gelijke spreidingen, kan men in plaats van de tekentoets de toets van Student voor de hypothese dat de mathematische verwachting van deze verschillen gelijk aan nul is, toepassen. Een voldoende voorwaarde hiervoor is, dat de oorspronkelijke waarnemingen zelf normaal en onafhankelijk verdeeld zijn, alle met dezelfde spreiding.
   3. Indien bekend is, dat de n gemeten objecten, voor zoverre het de gemeten maat betreft, gelijk zijn, kan de toets van Wilcoxon voor het vergelijken van twee steekproeven worden toegepast. Voorwaarde: de waarnemingen moeten onafhankelijk zijn.
   4. Indien bovendien de verdeling der waarnemingen normaal is (met gelijke spreidingen der meetfouten) kan de toets van Student voor twee steekproeven worden toegepast.

   Er zijn nog meer mogelijkheden, die echter niet onder de stof voor het examen vallen. Het onderscheidingsvermogen der toetsen neemt, grofweg, toe in de volgorde, waarin zij zijn opgenoemd, in overeenstemming met het feit, dat de voorwaarden voor toepassing in die richting strenger worden. Tenslotte kan nog worden opgemerkt, dat de toet-

sen 1 en 2 ruimte laten voor een algemenere conclusie dan 3 en 4, daar men bij toepassing van de eerstgenoemde toetsen verschillende waarden der te meten maat tegelijk in het onderzoek kan betrekken.

2. Tegen de redenering zijn de volgende bezwaren aan te voeren

   1. Een voor de hand liggende reden om antwoord te weigeren is het bezit van een hond die niet voor de belastingen is opgegeven. Onder de 82 weigeraars kan men dus een onevenredig groot aantal hondenbezitters verwachten.
   2. Uit het grote aantal weigeraars blijkt dat vele ondervraagden gevreesd hebben in moeilijkheden te kunnen geraken door antwoord te geven. Maar dan zal ook een aantal ondervraagden een onjuist antwoord gegeven hebben, om zich aan moeilijkheden te onttrekken. Onder de 247 die zeiden geen hond te bezitten, zullen dus een aantal clandestiene hondenbezitters zijn.
   3. De 27 gevallen, waar geen gehoor werd gegeven vormen een onzekere factor. Het is bijv. denkbaar dat uithuizige personen zelden honden bezitten.
   4. Een aantal van de 44 personen uit de tweede groep heeft meer dan een hond gehad.

   Tegen het onderzoek valt in te brengen dat men de bovengenoemde bezwaren had kunnen voorzien en dus een betere enquêtetechniek had moeten toepassen of er niet aan had moeten beginnen. Een belangrijk voordeel van een onafhankelijke enquête is de mogelijkheid om de ondervraagden eerst gerust te stellen, waarin de enquêteurs blijkbaar slechts ten dele zijn geslaagd. De leiding van het enquêtewerk had zeker geen genoegen mogen nemen met de 82 gevallen waar een antwoord werd geweigerd.

3. Het gemiddelde E $\underline{x}$ is de bekende parameter van de Poisson-verdeling, die we hieronder voorstellen door μ.

$$P(\underline{x}=2) = e^{-\mu}\mu^2/2!$$

$$P(\underline{x}=3) = e^{-\mu}\mu^3/3!$$

Het quotient is dus gelijk aan

$$\frac{3!}{2!\mu} = 0,9$$

Hieruit volgt $\mu = 3\frac{1}{3}$.

4. Vergelijkt men, voor ieder aantal personen per huishouding afzonderlijk, het percentage alleenwonende huishoudingen met onvoldoende slaapruimte, met het overeenkomstige percentage bij samenwonende huishoudingen, dan blijkt het eerstgenoemde percentage zonder uitzondering lager te liggen dan het laatstgenoemde. Men kan dus zeker niet concluderen dat de toestand op het gebied van de huisvesting voor de samenwonende huishoudingen beter is dan voor de alleenwonende. Dat nochtans van de alleenwonende huishoudingen een groter percentage onvoldoende slaapruimte heeft dan van de samenwonende, komt doordat de samenwonende huishoudingen overwegend uit kleinere gezinnen bestaan.

5. a) De verdeling van 2 $\underline{x}$ is:

$2\underline{x}$ =	2	4	6
P =	0,3	0	0,7

want $P(\underline{x}=x) = P(2\underline{x}=2x)$.

b) Wegens de stochastische onafhankelijkheid van x en y mag de productregel worden toegepast. De simultane verdeling is dus als volgt:

y \ x	1	3	
1	0,12	0,28	0,4
2	0,18	0,42	0,6
	0,3	0,7	1

De waarden van $(x+y)$ en $(x-y)$ voor de verschillende combinaties van x en y zijn:

y \ x	1	3
1	2	4
2	3	5

$x+y$

y \ x	1	3
1	0	2
2	-1	1

$x-y$

De gevraagde verdelingen zijn dus

$x+y$ =	2	3	4	5
P =	0,12	0,18	0,28	0,42

$x-y$ =	-1	0	1	2
P =	0,18	0,12	0,42	0,28

6. a. Deze vraag kan met behulp van de methode der variantie-analyse worden opgelost. Op de bekende wijze wordt het onderstaande schema verkregen:

Variantiebron	Kwadraatsom	Vrijheidsgraden	Gem.kwadraten
tussen steekproeven	24,75	9	2,75
binnen steekproeven	131,64	30	4,39
Totaal	156,39	39	

Er is op grond van het feit, dat de gemiddelde kwadraatsom "tussen" kleiner is dan die "binnen", geen reden om aan te nemen dat er een spreiding tussen de steekproeven aanwezig is.
De veronderstellingen der variantie-analyse zijn: gelijke spreiding, onafhankelijkheid en normale verdeling

van de grootheden in de universa waaruit deze 10 steekproeven genomen zijn.

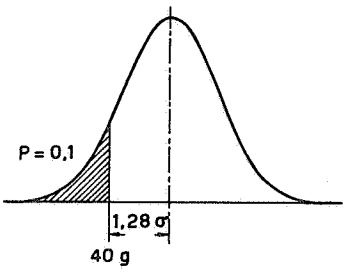
b. Gezien de uitkomst van vraag a. kunnen we de processpreiding σ schatten uit de totale kwadraatsom:

$$S^2 = \frac{156,39}{39} = 4,01 \quad \text{en} \quad S = 2,01.$$

De geschatte standaardspreiding van het gemiddelde van 5 stuks is dus

$$\frac{S}{\sqrt{n}} = \frac{2,01}{\sqrt{5}} = 0,90 \text{ gram.}$$

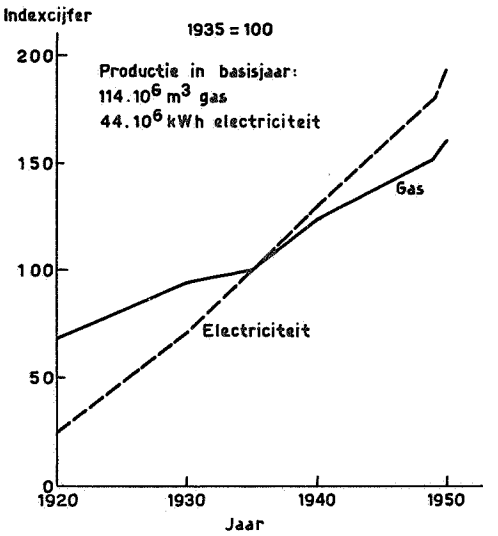
Om te voorkomen dat meer dan 1 op de 10 keer een bedrag kleiner dan 200 gram wordt gevonden, moet het productieproces dus op 40 + 1,28 x 0,90 = = 41,15 gr. worden afgesteld.



7. In dit tekenvraagstuk wordt gevraagd de ontwikkeling van gas- en electriciteitsproductie te vergelijken. Er zijn weer verschillende oplossingen mogelijk. De hierbij gevoegde grafiek wordt als de beste beschouwd. Andere oplossingen bijv. indexcijfers met 1920 als basisjaar, logarithmische schaal al of niet met indexcijfers zijn ook goed.

Veel gemaakte fouten:

Verschillende schalen voor de beide grootheden,
Verschillend nulpunt voor beide grootheden,
Niet aangeven van de absolute waarden in het basisjaar bij gebruik van indexcijfers,
Niet vermelden van schalen,
Onvolledig hoofd der grafiek,
Te veel getallen bij de assen.



Tweede examen, december 1953

1. Het bovenstaande kan ook aldus worden uitgedrukt: gevraagd wordt naar de kans P dat niet bij alle drie trekkingen een waarde kleiner dan twee wordt gevonden. De kans P is derhalve gelijk aan één minus de kans dat bij alle drie trekkingen waarden kleiner dan twee worden

gevonden.
Uit de tabel van de normale verdeling blijkt dat de kans dat bij één trekking een waarde kleiner dan twee wordt gevonden gelijk is aan 0,9772.
De kans dat bij drie trekkingen uitsluitend waarden kleiner dan twee worden gevonden is dus $0{,}9772^3$ = = 0,933.
De gevraagde kans P is dus:

$$P = 1-0{,}933 = 0{,}067.$$

2. Voor de toepassing van de toets van Wilcoxon moet de statistische grootheid V bepaald worden, die aangeeft bij hoeveel van de m.n combinaties van één element uit de eerste steekproef met één element uit de tweede steekproef de waargenomen waarde van het element uit de eerste steekproef groter (kleiner) is dan die van het element uit de tweede steekproef.
(m = uitgebreidheid van de eerste, n = uitgebreidheid van de tweede steekproef).
De bepaling van V geschiedt het eenvoudigst door de gezamenlijke waarnemingsuitkomsten van beide steekproeven in één reeks van hoog naar laag (resp. van laag naar hoog) te rangschikken en voor elk element van de eerste steekproef te bepalen hoeveel elementen van de tweede steekproef er op volgen en deze aantallen vervolgens te sommeren.
Indien ieder der beide steekproeven meer dan tien elementen bevat, zoals in dit vraagstuk, dan is $V-\frac{1}{2}mn$, onder de hypothese dat beide steekproeven uit hetzelfde universum afkomstig zijn, met voldoende benadering normaal verdeeld met gemiddelde nul en standaardafwijking $\sqrt{\frac{1}{12}\,mn(m+n+1)}$.
Bij dit probleem is m=15 en n=20, terwijl voor V de waarde 105 (resp. 195) gevonden wordt.

Derhalve is $V-\frac{1}{2}mn = 105-150 = -45$ (resp. 195-150=45)
terwijl voor de standaardafwijking $\sqrt{\frac{1}{12} \times 15 \times 20 \times 36} =$
$= \sqrt{900} = 30$ gevonden wordt.
Aangezien $(V-\frac{1}{2}mn)$ gelijk is aan 1,5 maal de standaardafwijking wordt de hypothese dat beide steekproeven t.a.v. het onderzochte kenmerk uit hetzelfde universum afkomstig zijn, niet verworpen.
Aangezien bij het gegeven vraagstuk tweezijdig getoetst moet worden is de overschrijdingskans P=0,13.
Daar de grootheid V slechts discrete waarden kan aannemen, is het gewenst een correctie voor continuïteit aan te brengen en niet $(V-\frac{1}{2}mn)$ doch $(V-\frac{1}{2}mn)-\frac{1}{2}$ te toetsen.
In dit geval heeft de continuïteitscorrectie geen invloed op de conclusie, terwijl de overschrijdingskans P verandert in 0,14.

3. Bereken eerst de som van rijen en kolommen:

		Verdienende gezinsleden									
		0	1	2	3	4	5	6	7	8	
Niet verdienende gezinsleden	0	-	17	34	8	3	2	1	-	-	65
	1	25	46	36	30	5	1	1	2	-	146
	2	8	38	30	8	3	-	-	-	-	87
	3	1	23	7	3	6	1	-	-	1	42
	4	-	15	2	5	1	-	-	-	-	23
	5	-	2	1	-	1	-	-	-	-	4
	6	-	-	1	-	-	-	-	-	-	1
	7	-	2	1	-	-	-	-	-	-	3
		34	143	112	54	19	4	2	2	1	371

1a. Gemiddeld aantal verdienende gezinsleden

0	x	34	= -
1	x	143	= 143
2	x	112	= 224
3	x	54	= 162
4	x	19	= 76
5	x	4	= 20
6	x	2	= 12
7	x	2	= 14
8	x	1	= 8
			659

$\frac{659}{371} = 1,8$

1b. Gemiddeld aantal niet-verdienende gezinsleden

0	x	65	= -
1	x	146	= 146
2	x	87	= 174
3	x	42	= 126
4	x	23	= 92
5	x	4	= 20
6	x	1	= 6
7	x	3	= 21
			585

$\frac{585}{371} = 1,6$

2. Dit betreft de gezinnen boven de diagonaal van links boven naar rechts onder in de tabel = 159. Dit is 43% van het totaal aantal gezinnen ($100 \times \frac{159}{371}$).

3. De mediaan is 2
Gemiddelde afwijking

$$\frac{34\,|0-2| + 143\,|1-2| + \ldots}{371} = \frac{339}{371} = 0,9.$$

4. Bereken eerst de som van de producten van de variabelen en de frequenties, dus 1 x 1 x 46 + 1 x 2 x 36 + .. + + 2 x 1 x 38 + .. + 7 x 1 x 2 + 7 x 2 x 1 = 1002. Deel dit door het aantal gezinnen en verminder deze uitkomst met het product van de gemiddelden.
$\frac{1002}{371} - \frac{659}{371} \times \frac{585}{371} = -0,10$. Dit is de covariantie.
Bereken van de variantie:
0 x 65 + 1 x 146 + 4 x 87 + 9 x 42 + 16 x 23 + 25 x 4 + + 36 x 1 + 49 x 3 = 1523
variantie = $\frac{1523}{371} - \left(\frac{585}{371}\right)^2 = 4,105 - 2,486 = 1,72$

Analoog: 0 x 34 + 1 x 143 + 4 x 112 enz. = 1715

$\frac{1715}{371} - \left(\frac{659}{371}\right)^2 = 4{,}623 - 3{,}157 = 1{,}47.$

De correlatiecoëfficient is: $r = \frac{-0{,}10}{\sqrt{1{,}72 \times 1{,}47}} = -0{,}06.$

4. Bij de gebruikelijke wijze van steekproef nemen is N (grootte steekproef) een vast getal en x (aantal defecten) een stochastische variabele.

   $100\,\frac{x}{N}$ is dan een zuivere schatting van % defecten in de partij. In het geval van dit vraagstuk is N (aantal defecten) een vast getal en x (grootte v/d steekproef) een stochastische variabele. Het geval ligt dus geheel anders en men kan niet zonder meer concluderen dat $100\,\frac{N}{x}$ een zuivere schatting is.
   Intuïtief kan men als volgt redeneren: In het "gewone" geval (vaste N) zal de laatste waarneming in het algemeen niet juist een defect opleveren. Men zal daarom N groter dan 50 moeten nemen om te kunnen verwachten dat er in één steekproef van N gemiddeld 3 defecten zullen voorkomen. 6% is daarom te hoog.
   Een exact bewijs werd niet verlangd.

5. a. De getoetste hypothese luidt, dat de bewering van de dame in kwestie onjuist is en dat zij in het geheel geen verschil proeft tussen de twee soorten kopjes thee.

   b. en c. Hier zijn twee juiste antwoorden mogelijk:

   1 b. Tekentoets
   c. Eerstgenoemde proefopzet.

   Toelichting: Uit de onder a geformuleerde hypothese volgt, in geval de eerste proefopzet gebruikt wordt, dat de kans op een goed antwoord steeds gelijk aan $\frac{1}{2}$ is, terwijl de opeenvolgende 20 proeven bij die proefopzet onderling onafhankelijk zijn.

Indien de dame, om welke reden dan ook, de neiging heeft, om vaker het ene antwoord te geven dan het andere, is de tekentoets strikt genomen bij de tweede proefopzet niet zonder meer toepasbaar, maar bij de eerste wel. Bij de tweede proefopzet zal dan nl., als de getoetste hypothese juist is, de kans op een juist antwoord in de ene helft van de gevallen groter dan $\frac{1}{2}$ zijn en in de andere helft van de gevallen evenveel kleiner. Indien men echter iedere keer loot is de kans op een juist antwoord steeds gelijk aan $\frac{1}{2}$, wat de dame ook zegt; zelfs is dit zo, als zij altijd het ene of altijd het andere antwoord geeft.
Men kan bewijzen (maar dat behoeven de candidaten niet te weten), dat toepassing van de tekentoets bij het tweede proefschema in het geval, dat de dame een voorkeur voor één der beide antwoorden heeft, niet leidt tot een grotere, maar tot een kleinere onbetrouwbaarheid dan de tekentoets. Als gevolg hiervan wordt ook het onderscheidingsvermogen kleiner, althans in de buurt van de getoetste hypothese. Een correctie om dit ongunstige effect op te heffen, zou alleen geconstrueerd kunnen worden, als de mate van voorkeur van de dame voor één der twee antwoorden bekend was en dit is natuurlijk in de regel niet zo.

2 b. Methode der 2 x 2-tabel, waarbij het ene tweetal kenmerken is, of de melk of de thee er in werkelijkheid het eerst in gedaan is en de tweede splitsing, wat de dame zegt dat er het eerst ingedaan is. Men kan dan bijv. de χ^2-toets met continuïteitscorrectie gebruiken.

c. In dit geval kan men zowel de eerste proefopzet als de tweede gebruiken.

Toelichting: Uit de onder a geformuleerde hypothese volgt, dat wat de dame zegt onafhankelijk is van wat

er werkelijk gebeurd is. Dit is echter juist de hypothese, die met een 2 x 2-tabel getoetst wordt en het aantal kopjes thee van beide soorten, dat hierbij gebruikt wordt is daarbij van geen belang. Er valt, bij gebruik van deze toets, iets te zeggen voor de tweede proefopzet tegenover de eerste, daar de eerste de mogelijkheid openlaat, dat alle of een overwegend aantal der kopjes thee van één der beide soorten zijn, waardoor de dame in de war zou kunnen raken. Dit is bij de tweede opzet niet mogelijk.

Opmerking: De keuze tussen de mogelijke toetsingen met bijbehorende proefopzet hangt van hun onderscheidingsvermogen af. Hierover werd van de candidaten geen kennis verwacht.

6. 1a. Bereken voor elke steekproef

$$\frac{0{,}75 - \text{waargenomen verhouding}}{\sqrt{\frac{0{,}25 \times 0{,}75}{N}}}$$

b. De χ^2-toets komt in dit geval hierop neer dat men het kwadraat van de onder a gegeven uitdrukking berekent (1 vrijheidsgraad). De χ^2 waarden zijn voor de 10 steekproeven resp.

1 : 0,47	6 : 0,67
2 : 0,09	7 : 0,76
3 : 0,10	8 : 0,67
4 : 1,39	9 : 0,98
5 : 0,00	10 : 0,17

Geen dezer waarden is significant.
(N.B. De conclusie verandert niet indien men een continuïteitscorrectie aanbrengt, daar de overschrijdingskansen hierdoor iets groter worden.)

2. Men krijgt dan een steekproef van 437 stuks, waarvan

336 A-elementen.
χ^2=0,96 (1 vrijheidsgraad). Niet significant.

3. Deze som is verdeeld volgens χ^2 met 10 vrijheidsgraden. Men vindt χ^2=5,30: niet significant.

4. Het percentage wordt nu 77% i.p.v. 75%.
Men krijgt nu een χ^2 met 9 vrijheidsgraden i.p.v. 10. De waarde is niet significant.

5. Indien een onder 1 gevonden waarde significant is, kan men concluderen dat de universumwaarde niet 75% is (of dat de steekproef niet aselect is, welke mogelijkheid wij verder buiten beschouwing zullen laten).
Een enkele significante uitkomst bij 1 zegt nog niet veel, daar men bij herhaalde toepassing van een toets, nu en dan grotere afwijkingen moet verwachten. Men zou zich in dit geval dan ook eerst moeten overtuigen dat ook 3 een significante uitkomst geeft. Omgekeerd, indien weliswaar geen enkele uitkomst onder 1 wordt gevonden die significant is, doch de grens in een aantal gevallen dicht wordt genaderd, kan 3 een significante uitkomst leveren. Daarbij kan 2 een niet-significante waarde geven, nl. indien de "bijna significante" afwijkingen aan weerskanten van 75% liggen, zodat ze elkaar opheffen. De spreiding tussen de steekproeven is dan te groot. Indien de spreiding tussen de steekproeven uitzonderlijk klein is, maar het niveau enigszins afwijkt van 75%, kan met 2 een significante waarde worden gevonden, terwijl 3 juist een niet-significante uitkomst geeft.
Geeft 3 een significante uitkomst, dan is of de gemeenschappelijke universumwaarde niet 75% of er is geen gemeenschappelijke universumwaarde. Dit laatste

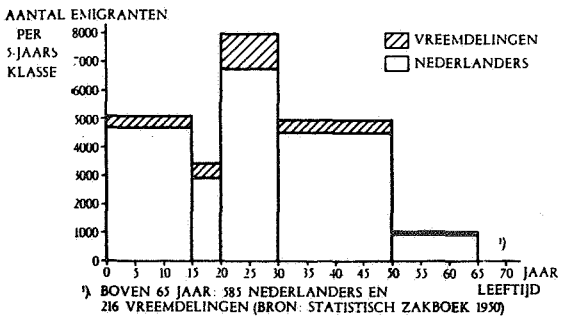
kan men nader onderzoeken met de onder 4 genoemde toets.

6. Men kieze als klassegrenzen waarden uit de χ^2 tabel (voor 1 vrijheidsgraad). Zo is bijv. de mathematische verwachting van het aantal waarden dat zal liggen tussen 0,0158 en 0,0642: 10% van het aantal waargenomen waarden van χ^2. Heeft men bijv. 100 waarden van χ^2 dan kan men de volgende tabel opstellen:

Waarden van χ^2	Frequenties	
	Theoretisch	waargenomen
< 0,0158	10	x
0,0158-0,0642	10	x
> 2,706	10	x
	100	100

Het verschil tussen de theoretische en de waargenomen frequenties kan men onderzoeken met de χ^2 toets (waarbij het aantal vrijheidsgraden = het aantal klassen verminderd met 1).

7. Aantal uit Nederland geëmigreerde personen in 1949, naar leeftijd en nationaliteit



Kritieke punten

1. Volledige titel en eenheidsomschrijving
2. Lineaire schaal voor leeftijd
3. Verticale schaal herleiden tot eenheidsklassebreedte
4. Moet histogram zijn
5. Vreemdelingen boven en Nederlanders onder is beter dan omgekeerd
6. Contrast in arcering niet te sterk (optische vervorming!)
7. Bronvermelding
8. Netheid van uitvoering.

De onderstreepte punten zijn het belangrijkst.

Derde examen, mei 1955

1. De verdeling over jongens en meisjes zal ongeveer worden gegeven door de termen van de ontwikkeling van $(\frac{1}{2}+\frac{1}{2})^4$.
   Dus 3 jongens en 1 meisje in $\frac{4}{16}$ ofwel 25% van de gezinnen met 4 kinderen.

2. Verdeling van x : 1 met kans $\frac{1}{4}$, 2 met kans $\frac{3}{4}$.
   dus verdeling van x^2 : 1 met kans $\frac{1}{4}$, 4 met kans $\frac{3}{4}$.

   Verdeling van y : 1 met kans $\frac{1}{6}$, 2 met kans $\frac{1}{3}$,
   3 met kans $\frac{1}{2}$.

   dus verdeling van $3y$: 3 met kans $\frac{1}{6}$, 6 met kans $\frac{1}{3}$,
   9 met kans $\frac{1}{2}$.

   Verdeling van $x^2 + 3y$:

$x^2 + 3y$:	1 + 3	1 + 6	1 + 9	4 + 3	4 + 6	4 + 9
$x^2 + 3y$:	4	7	10	7	10	13
kans:	$\frac{1}{4} \times \frac{1}{6}$	$\frac{1}{4} \times \frac{1}{3}$	$\frac{1}{4} \times \frac{1}{2}$	$\frac{3}{4} \times \frac{1}{6}$	$\frac{3}{4} \times \frac{1}{3}$	$\frac{3}{4} \times \frac{1}{2}$

Dus:

$\underline{x}^2 + 3\underline{y}$: 4 7 10 13

kans : $\frac{1}{24}$ $\frac{5}{24}$ $\frac{9}{24}$ $\frac{9}{24}$.

3. Door de beide steekproeven afzonderlijk te beschouwen toetst men niet met een overschrijdingskans van 5% zoals wordt gesuggereerd. Men dient de afwijking van nul van het verschil der beide steekproeven te toetsen. Het blijkt dan dat het verschil significant is bij onbetrouwbaarheidsdrempel 0,05.
Verschil: 0,5156 - 0,4722 = 0,0434.
Standaardafwijking van het verschil

$$\approx \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{900} + \frac{\frac{1}{2} \times \frac{1}{2}}{1600}} = 0,0208 .$$

Het verschil is groter dan 2 x de standaardafwijking, namelijk 2,08 keer.

Opmerking: Hetzelfde antwoord wordt gevonden door de gebruikers en niet-gebruikers in A en B te rangschikken in een 2 x 2-tabel en hierop een χ^2-toets toe te passen. De toetsingsgrootheid, een χ^2 met één vrijheidsgraad, is het kwadraat van de waarde 2,08 gevonden volgens de hierboven beschreven methode.

4. a. De kans dat er van 100 ondeugdelijke exemplaren één of meer zullen worden doorgelaten is:

1 - (de kans dat er geen ondeugdelijk exemplaar wordt doorgelaten) = $1 - \{1-0,001\}^{100} = 1 - 0,9048 =$ = 0,0952.

b. Van de kans dat er onder 100 willekeurige exemplaren, die de keuring zijn gepasseerd, één of meer ondeugdelijke exemplaren voorkomen is niets te zeggen, aangezien over de kansverdeling van het aantal ondeugde-

lijke exemplaren onder die 100 niets gegeven is.

5. Indien men het systeem van het vraagstuk volgt bij het trekken van de steekproef, zullen alle straten - met uitzondering van de straten met minder dan 7 nummers - één vertegenwoordiger in de steekproef hebben. Korte en lange straten zijn dus gelijkgeschakeld: er treedt geen vertegenwoordiging in evenredigheid met het aantal gezinnen op. Indien nu - wat niet ondenkbaar is - een samenhang bestaat tussen de lengte van de straat en een bepaald persoonlijk kenmerk bijv. het welstandsniveau van de bewoners, dan zal de steekproef niet representatief zijn.

6. a) $\text{var } L = \frac{1}{11}(\Sigma L^2 - \frac{(\Sigma L)^2}{12}) = \frac{1814,92}{11} = 164,99$

$\text{var } B = \frac{1}{11}(\Sigma B^2 - \frac{(\Sigma B)^2}{12}) = \frac{1348,25}{11} = 122,57$

$\text{var}(L-B) = \frac{1}{11}\left[\Sigma(L-B)^2 - \frac{\{\Sigma(L-B)\}^2}{12}\right] = \frac{229,67}{11} = 20,88.$

b) $c = \dfrac{\Sigma LB - \frac{\Sigma L \Sigma B}{12}}{\Sigma B^2 - \frac{(\Sigma B)^2}{12}}$

$\Sigma LB = \frac{1}{2}\left[\Sigma L^2 + \Sigma B^2 - \Sigma(L-B)^2\right] = 79355$

Dus $c = \frac{1466,75}{1348,25} = 1,088$

$d = \overline{L} - c\overline{B} = 6,63.$

Dus $L = 1,088B + 6,63$.

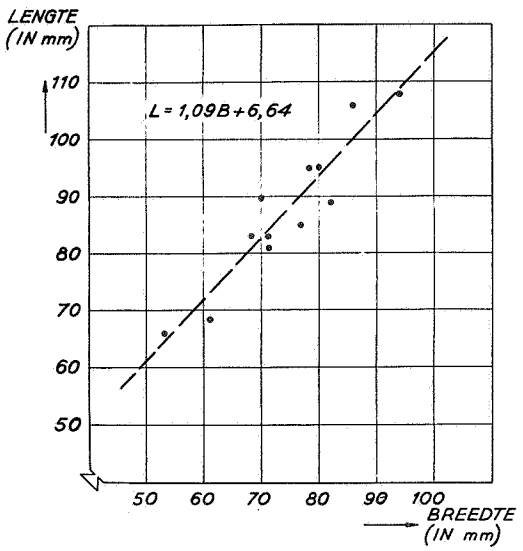
c) $r = \dfrac{\text{cov}(L,B)}{\sqrt{\text{var } L.\text{var } B}} = \dfrac{1466,75}{11\sqrt{164,99 \times 122,57}} = \dfrac{1466,75}{1564,3} =$

$= 0,938.$

d) Residuele variantie = $(1-r^2)$var L = 19,80.

e) De correlatiecoëfficiënt berekend uit L en B van de vruchten in een bepaalde breedte-klasse is geen goede schatting van de correlatiecoëfficiënt in de gehele partij, aangezien men dan niet meer te maken heeft met een aselecte steekproef. Om dit te verduidelijken kan men het volgende stellen: Zou een bepaalde klasse zeer smal worden, dan zou de regressielijn een geheel andere richting krijgen dan die welke uit het gehele materiaal wordt verkregen. De correlatiecoëfficiënt wordt in een kleine klasse veel kleiner dan in de gehele steekproef.

f)



Verband tussen lengte en breedte van 12 monsters ongesorteerd fruit.

7. Geef het eerste getrokken getal aan met $\underline{x}$ en het tweede met $\underline{y}$, dan is dus steeds $\underline{x} \neq \underline{y}$. Voor $\underline{x}$ zijn er 5 mogelijke waarden en voor $\underline{y}$, bij gegeven $\underline{x}$, 4 mogelijke

waarden, dus er zijn 20 mogelijke combinaties, die alle gelijke kans bezitten, daar de trekkingen aselect zijn.
In volgend schema zijn deze 20 mogelijkheden aangegeven, met de bij ieder daarvan behorende waarde van het product $\underline{xy}$.

Tabel van waarden x.y

$\underline{y}$ \ $\underline{x}$	1	2	3	4	5
1	X	2	3	4	5
2	2	X	6	8	10
3	3	6	X	12	15
4	4	8	12	X	20
5	5	10	15	20	X

Ieder der vakjes heeft een kans $\frac{1}{20}$.
Derhalve is de verwachting van het product $\underline{xy}$ gelijk aan de som van deze waarden gedeeld door 20, dus

$$E\underline{xy} = \frac{1}{20}\{(2+3+4+5) + (2+6+8+10) + \ldots\} = 8,5.$$

Voor $\underline{x}$ en $\underline{y}$ afzonderlijk geldt, dat ieder der waarden $1,\ldots,5$ een kans $\frac{1}{5}$ bezit, dus

$$E\underline{x} = E\underline{y} = \frac{1}{5}(1+\ldots+5) = 3$$

en

$$\sigma^2\{\underline{x}\} = \sigma^2\{\underline{y}\} = \frac{1}{5}\{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2\} = 2.$$

Dus is

$$\mathrm{cov}(\underline{x},\underline{y}) = E\underline{xy} - E\underline{x}E\underline{y} = 8,5-9 = -0,5,$$

en

$$\rho(\underline{x},\underline{y}) = \frac{\mathrm{cov}(\underline{x},\underline{y})}{\sigma\{\underline{x}\}\,\sigma\{\underline{y}\}} = -\frac{0,5}{2} = -0,25.$$

8. De gevraagde kans is gelijk aan de kans dat in de eerste helft van de tweede steekproef vijf bepaalde nummers voorkomen (nl. die welke ook in de eerste helft van de eerste steekproef voorkwamen). Deze kans is:

$$\frac{5}{10} \times \frac{4}{9} \times \frac{3}{8} \times \frac{2}{7} \times \frac{1}{6} = \frac{1}{252} .$$

9. De gemiddelde hoogte van stapels van 25 schijven zal zijn

25 x 12 cm = 300 cm.

Standaardafwijking = $\sqrt{\Sigma(\text{Standaardafwijking schijfdikte})^2}$ = $\sqrt{\Sigma 4}$ = $\sqrt{25x4}$ = $\sqrt{100}$ = 10.

De kans dat een stapel zou worden gevormd welke hoger is dan 320 cm is gelijk aan de kans dat bij een normale verdeling een afwijking, in positieve zin, groter dan 2σ (20=2x10) optreedt. Deze kans is gelijk aan 0,0228. Bij 2,28% van de stapels zal het dus niet gelukken deze compleet te maken.

10. De gemiddelden voor het Rijk zijn gelijk aan gewogen gemiddelden der percentages voor de provincies, met als gewichten de bevolking der provincie. Deze gewichten zijn voor 1920 en voor 1930 niet gelijk. In het Zuiden van het land is de bevolking tussen 1920 en 1930 sterker toegenomen dan in de andere gebieden tezamen, waardoor de gewichten voor de zuidelijke provincies in 1930 groter waren dan in 1920. Aangezien in deze provincies de percentages voor het aandeel Rooms Katholieken zeer hoog zijn, heeft de verhoging van de gewichten aldaar geleid tot een hoger Rijksgemiddelde - dit ondanks de omstandigheid dat ook in de zuidelijke provincies de percentages voor de Rooms Katholieken in 1930 lager waren dan in 1920.

11. a. Kritiek ten opzichte van het experiment

1. De jongens en meisjes zijn niet gelijkelijk over de drie voedingsgroepen verdeeld. Het verschil in uitwerking van de voeding is dus "confounded" met het verschil in sexe. Aangezien in het begin (in de eerste week) de invloed van de voeding nog maar kort geduurd heeft en dus vermoedelijk niet groot is geweest, geven de gewichtsmetingen in de eerste week reeds een aanwijzing dat althans het begingewicht van jongens en meisjes verschillend was.
2. Het 7-maands kind wijkt duidelijk af van de andere kinderen en waar hiervoor geen vergelijkingsmateriaal is in de vorm van andere 7-maands kinderen hoort het niet in dit experiment thuis.
3. Het heeft geen zin bij één kind op drie verschillende dagen waarnemingen te doen en bij een ander kind maar op één dag. Bovendien had bij elk kind de waarneming op dezelfde dag na de geboorte moeten plaatsvinden anders zijn de waarnemingen niet zonder meer vergelijkbaar.

b. Foutief toegepaste statistische methoden

1. De herhaalde waarnemingen van het B.M. op één tijdstip geven alleen aanwijzing over de nauwkeurigheid waarmee deze grootheid bepaald is of kan worden en over de variabiliteit ervan op korte termijn. Zij mogen niet als onafhankelijke waarnemingen beschouwd worden bij de vergelijking van groepen kinderen, zoals blijkens de aantallen graden van vrijheid bij het toepassen der t-toets (onder B) en der variantieanalyse (onder C) gedaan is. Ook de meervoudige waarnemingen bij één kind in dezelfde week zijn niet onafhankelijk.

Het beste wat de statisticus had kunnen doen t.a.v. deze meervoudige waarnemingen is vermoedelijk uitsluitend de eerste waarneming te gebruiken.

2. De statisticus had moeten opmerken, dat het 7-maands kind niet in de serie thuis hoort en deze waarneming moeten weglaten. Voordat hij de correlatiecoëfficiënt (par. 3A) berekende had hij tenminste een grafiek (puntendiagram) van het verband tussen B.M. en gewicht moeten tekenen. Hij zou dan gezien hebben, dat de gevonden correlatie vrijwel uitsluitend bepaald wordt door dit 7-maands kind. Daar er hier een buitenstatistische reden voor de afwijkende waarneming is, is weglaten ervan geoorloofd, ja noodzakelijk. Bij de berekening van de correlatiecoëfficiënt zijn blijkens het opgegeven aantal graden van vrijheid de meervoudige waarnemingen als onafhankelijk beschouwd.
   De afwijkende waarden van het 7-maands kind veroorzaken bij de t-toets een abnormaal hoge standaardafwijking waardoor geen significantie gevonden wordt; een serie met een dergelijke afwijkende waarde is bovendien zeker geen steekproef uit een normale verdeling, zodat de t-toets zo niet toegepast mocht worden.
3. Bij de berekening van de variantieanalyse (§ 3C) had de statisticus moeten opmerken dat sexe en methodé van voeding "confounded" zijn zodat de uitkomst der analyse niets bewijst over een eventuele invloed van de methode van voeding.
4. De statisticus had zich, als hij dan al een variantieanalyse wilde uitvoeren, moeten afvragen of hij deze niet beter had kunnen toepassen op de verandering in het B.M. resp. het gewicht van de eerste op de derde week. Dit zou de invloed van de verschillen in begingewicht min of meer hebben geëlimineerd, zodat hij vermoedelijk een lagere restvariantie en dus een

scherper onderscheidingsvermogen zou hebben gekregen.

c. Niet genomen controlemaatregelen van statistische aard

1. Voor de berekening van een correlatiecoëfficiënt is het wenselijk steeds een stippendiagram van het te onderzoeken verband te tekenen (reeds genoemd onder b 2).
2. Er is geen onderzoek naar normaliteit gedaan; dit was met het geringe aantal waarnemingen ook niet mogelijk; de statisticus had echter moeten vermelden hetzij dat uit andere onderzoekingen gebleken was, dat normaliteit aangenomen mag worden, of anders parametervrije toetsen moeten toepassen.

d. Overige in een statistisch rapport vermelde punten

1. De statisticus had zeker een duidelijke omschrijving van de doelstelling van het experiment moeten opnemen.

Vierde examen, oktober 1956

1. Men passe de tekentoets toe, resp. éénzijdig (vraag a) en tweezijdig (vraag b). Uit een tabel van de kritieke waarden van de tekentoets leest men af dat, bij een onbetrouwbaarheidsdrempel 0,05, in het eerste geval de afwijking (8-4) significant is (kritieke waarde 4), in het tweede geval niet significant (kritieke waarde 3). Opmerking: de normale benadering, met continuïteitscorrectie, geeft voor dit geval hetzelfde resultaat. De formule

$$\frac{x-\frac{1}{2}n-\frac{1}{2}}{\frac{1}{2}\sqrt{n}} \quad \text{geeft} \quad \frac{12-8-\frac{1}{2}}{\frac{1}{2}\sqrt{16}} = 1{,}75.$$

Dit is dus significant bij éénzijdige, doch niet bij tweezijdige toetsing.

2. We hebben hier twee onafhankelijke stochastische grootheden $\underline{x}$ en $\underline{y}$; de waarden daarvan, met bijbehorende kansen, zijn

$$x \begin{cases} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{cases} \qquad y \begin{cases} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{cases}$$

Voor $(\underline{x}+\underline{y})$ heeft men

$$(x+y) \begin{cases} 1 & 2 & 3 & 4 & 5 & 6 & 7. \\ \frac{1}{12} & \frac{2}{12} & \frac{2}{12} & \frac{2}{12} & \frac{2}{12} & \frac{2}{12} & \frac{1}{12}. \end{cases}$$

$$E(\underline{x}+\underline{y}) = \frac{1}{12} \times 1 + \frac{2}{12} \times 2 + \frac{2}{12} \times 3 + \ldots = 4.$$

Men kan dit resultaat ook vinden door toepassing van de formule

$$E(\underline{x}+\underline{y}) = E\underline{x} + E\underline{y} = \tfrac{1}{2} + 3\tfrac{1}{2} = 4.$$

$$\text{Var. } (\underline{x}+\underline{y}) = \frac{1}{12}(1-4)^2 + \frac{2}{12}(2-4)^2 + \ldots = \frac{38}{12}.$$

Deze uitkomst kan men ook als volgt verkrijgen:

$$\text{Var. } (\underline{x}+\underline{y}) = \text{var. } \underline{x} + \text{var. } \underline{y} = \frac{1}{4} + \frac{35}{12} = \frac{38}{12}.$$

De kans dat $\underline{x}+\underline{y}$ kleiner dan 4 is, is de som van de kansen behorende bij de waarden 1, 2 en 3. Deze kans is dus $\frac{1}{12} + \frac{2}{12} + \frac{2}{12} = \frac{5}{12}$.

3. a) Het gevraagde histogram is in figuur 1 getekend.

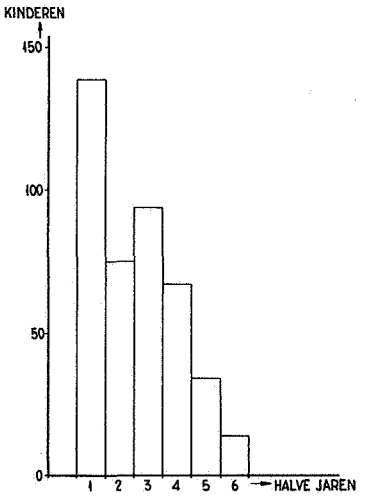


Fig. 1. Frequentiehistogram van het aantal kinderen, bij welke bij controle aangetaste melkkiezen werden gevonden, naar het halve jaar na het op school komen, waarin deze voor het eerst ontdekt werden. Na het zesde jaar werden nog bij 12 kinderen, die tevoren geen aangetaste melkkiezen hadden, aangetaste melkkiezen ontdekt.

b) In figuur 2 is de frequentieverdeling op normaal waarschijnlijkheidspapier getekend.

c) In de figuur op waarschijnlijkheidspapier liggen de punten behoorlijk rechtlijnig georienteerd. Dit wijst er op, dat de verdeling van de begin-leeftijd, waarop bij kinderen, die aangetaste melkkiezen krijgen, de aantasting van de melkkiezen optreedt, normaal is. Voorts volgt uit deze figuur, door extrapolatie, dat bij het op school komen reeds circa 15% aangetaste melkkiezen heeft. In het eerste halfjaar worden dus de kinderen ont-

dekt, die bij het op school komen aangetaste kiezen hadden, plus zij, die in dat halfjaar aangetaste kiezen krijgen. Hierdoor is het aantal, dat in het eerste halfjaar gevonden wordt, groter, dan het in het tweede halfjaar gevondene. In het eerste halfjaar worden er dus schijnbaar te veel kinderen met aantastingen gevonden.
Volgens de geschatte cumulatieve frequentieverdeling zijn de verwachte aantallen in het 2e en in het 3e halfjaar aan elkaar gelijk. Dat het gevonden aantal in het 2e halfjaar kleiner is moet door toevallige oorzaken worden verklaard.

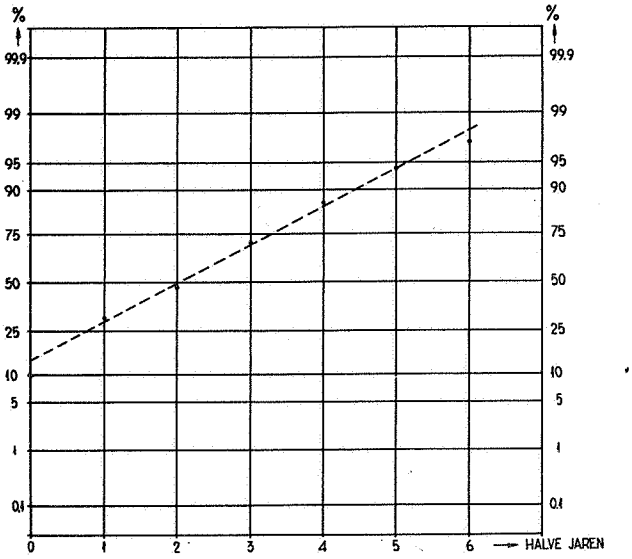


Fig. 2. Cumulatieve frequentieverdeling van het aantal kinderen, bij welke bij controle aangetaste melkkiezen werden gevonden, naar het halve jaar na het op school komen, waarin deze voor het eerst ontdekt werden.

4. De automobilist rijdt

60 km/u	gedurende	$\frac{20}{60}$ = 0,333	uur
80 "	"	$\frac{100}{80}$ = 1,250	"
90 "	"	$\frac{35}{90}$ = 0,389	"
100 "	"	$\frac{20}{100}$ = 0,200	"
			+
Totale rijtijd		2,172	uur

Totale afstand = 175 km,
dus zijn gemiddelde snelheid is $\frac{175}{2,17}$ = 80,6 km/uur.

5. $m = \frac{1}{10}\Sigma x_i = 5,03$,
$s^2 = \frac{1}{9}\Sigma(x_i-m)^2 = 1,48$, dus $s = \sqrt{1,48} = 1,22$.

Het 8e deciel van een normale verdeling is het punt $\mu + 0,84\sigma$. Dit punt kan dus geschat worden door

$$m + 0,84s = 5,03 + 0,84.1,22 = 6,05.$$

6. a) Geven we het spoelgewicht aan met $\underline{x}$, het garengewicht met $\underline{y}$ en het totale gewicht van spoel + garen met $\underline{z}$, dan is

$$\underline{z} = \underline{x} + \underline{y}$$
$$\mu_z = \mu_x + \mu_y$$
$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$$

Gegeven is: μ_z=405,2, σ_z=12,8 dus σ_z^2=163,84,
μ_x= 99,7 σ_x=11,9 dus σ_x^2=141,61,
Dus $\mu_y = \mu_z - \mu_x$=305,5 en $\sigma_y^2 = \sigma_z^2 - \sigma_x^2$=22,23; σ_y=4,71.

Het percentage spoelen met minder dan 300 g garen wordt gevonden door de overschrijdingskans van het getal

$$\frac{305,5 - 300}{4,71} = 1,17$$

op te zoeken in de tabel van de normale verdeling. Dit geeft 12%.

b) Is $\underline{m}$ het gemiddelde garengewicht van 5 aselect genomen spoelen, dan is $\underline{m}$ normaal verdeeld met hetzelfde gemiddelde als $\underline{y}$, maar met een 5 maal zo kleine variantie. Deze variantie is dus $\frac{22,23}{5}=4,45$ en de spreiding is dus 2,11. De gevraagde kans is nu de overschrijdingskans van

$$\frac{305,5 - 300}{2,11} = 2,61$$

en deze is 0,0046.

7. De gegevens zijn onvoldoende om uit te maken of de uitspraak juist is of niet. Daartoe zou nl. ook opgegeven moeten worden, hoeveel wel ingeënte en hoeveel niet ingeënte personen er bij het onderzoek betrokken zijn geweest. Bovendien doet zich dan nog de vraag voor, op welke indicatie tot inenting (of tot het nalaten daarvan) is overgegaan, m.a.w. de mogelijkheid bestaat nog dat de beide groepen niet gelijkwaardig zijn wat hun vatbaarheid (afgezien van inenting) voor de beschouwde ziekte betreft.

8. Stel het totaal aantal moleculen in de beschouwde ruimte N. Verdeel deze ruimte in gedachten in n gelijke delen (n groot). Het gemiddelde aantal moleculen in één elementje bedraagt $\frac{N}{n}$. Het werkelijke aantal op een bepaald ogenblik is binomiaal verdeeld, met gemiddelde $\frac{N}{n}$ en standaardafwijking $\sqrt{N \times \frac{1}{n} \times (1 - \frac{1}{n})}$. Hierin is $1 - \frac{1}{n}$ ongeveer 1, dus de standaardafwijking is $\sqrt{\frac{N}{n}}$. Indien het volumenelement 4 cm^3 is, heeft men $\frac{N}{n} = 4 \times 2,5 \times 10^{19}$ en $\sqrt{\frac{N}{n}} = 2 \times 5 \times 10^9 = 10^{10}$. Men kan dus afwijkingen van (tweemaal de standaardafwijking)

2×10^{10} verwachten.
(Aequivalent hiermee: het aantal moleculen in een volumenelement bezit een Poisson-verdeling; voor 4 cm^3 is daarvan $\mu = \sigma^2 = 10^{10}$; wegens de grote μ is de verdeling bij zeer goede benadering normaal, enz.).

9. a) Het gemiddelde en de mediaan worden 10 hoger, de standaardafwijking verandert niet.
   b) Het gemiddelde wordt 2 x zo groot, dus indien de oorspronkelijke waarde $\bar{x}$ was, wordt de nieuwe waarde $2\bar{x} + 20$. Hetzelfde geldt voor de mediaan. Als de oorspronkelijke standaardafwijking s is, dan wordt de nieuwe 2s.
   c) Het gemiddelde neemt met 1 toe en wordt dus $2\bar{x} + 21$. De mediaan behoudt de onder b) verkregen waarde. Wat met de standaardafwijking gebeurt valt zonder nadere gegevens niet te zeggen.

Noemt men de oorspronkelijke waarnemingen $x_i (i=1,\ldots,11)$, dan zijn de nieuwe waarden $2x_i + 20$ $(i=1,\ldots,8,10,11)$, resp. $2x_9 + 31$.
We bepalen de kwadraatsom der afwijkingen van het gemiddelde:

$$\sum_{i\neq 9} (2x_i+20-2\bar{x}-21)^2 + (2x_9+31-2\bar{x}-21)^2 =$$

$$= \sum_{i\neq 9} (2x_i-2\bar{x}-1)^2 + (2x_9-2\bar{x}+10)^2 =$$

$$= \sum_{i\neq 9} (2x_i-2\bar{x})^2 - \sum_{i\neq 9} 2(2x_i-2\bar{x}) + 10 + (2x_9-2\bar{x})^2 + 20(2x_9-2\bar{x}) + 100 =$$

$$= \sum_{i=1}^{11} (2x_i-2\bar{x})^2 - 2 \underbrace{\sum_{i=1}^{11} (2x_i-2\bar{x})}_{=0} + 22(2x_9-2\bar{x}) + 110.$$

Het verschil met de onder b) bereikte waarde van de kwadraatsom der afwijkingen van het gemiddelde - en daarmee de variantie - zou dus bekend zijn indien we $(x_9-\bar{x})$ zouden kennen.

10. Als dagen met weinig orders tevens dagen zijn met hoge gemiddelde provisie per order (hetgeen zeer goed mogelijk is, omdat grote orders meer tijd in beslag zullen nemen dan kleine), dan zal het gemiddelde, genomen over alle dagen, van de gemiddelde dagprovisies hoger zijn dan het gemiddelde van alle individuele orderprovisies, m.a.w. men zal op de beschreven wijze een hoger gemiddeld provisiebedrag per order berekenen dan bij exacte berekening het geval is. Men kan dit gemakkelijk inzien bij extreme gevallen; als er bijv. 20 dagen met één zeer grote order zijn en 5 dagen met ieder vele kleine orders. Er is dan 20/25 kans op een zeer hoog gemiddeld orderbedrag uit te komen, hoewel de grote orders toch slechts een kleine fractie van het gehele orderpakket vormen.

Bewijs:
De verwachting van de door middel van een steekproef berekende gemiddelde provisie per order bedraagt:

$$E(\underline{p}_i) = \frac{1}{k}\sum_i \bar{p}_i \ ,$$

waarin k het aantal werkdagen in een maand is, waarover wordt gesommeerd.
De zuivere gemiddelde orderprovisie over dezelfde maand bedraagt:

$$\bar{\bar{p}} = \frac{1}{N}\sum_i p_i = \frac{1}{N}\sum_i n_i\bar{p}_i \ .$$

Stel nu:

$$E(\bar{p}_i) = \frac{1}{k} \sum_i \bar{p}_i = \mu_1 ,$$

$$\frac{1}{k} \sum_i n_i = \mu_2 ,$$

$$\bar{p}_i - \mu_1 = x ,$$

$$n_i - \mu_2 = y .$$

Dan is:

$$\bar{\bar{p}} = \frac{1}{N} \sum_i (\mu_1+x)(\mu_2+y) =$$

$$= \frac{1}{N} \sum_i (\mu_1 \mu_2 + \mu_2 x + \mu_1 y + xy).$$

Men bedenke dat:

$$N = \sum_i n_i = k \mu_2 ,$$

$$\Sigma x = \Sigma y = 0.$$

Hieruit volgt dat:

$$\bar{\bar{p}} = \mu_1 + \frac{1}{N} \sum_i xy.$$

Als Σxy negatief is, m.a.w. als over de beschouwde maand het gemiddelde orderbedrag $\bar{p}_i$ negatief gecorreleerd is met het aantal orders n_i, dan is de zuivere provisie $\bar{\bar{p}}$ kleiner dan de berekende μ_1.
(Dit bewijs werd niet van de candidaten gevergd.)

11. a.1. In het geval dat de verdeling der inkomens als normaal mag worden beschouwd, komt voor toetsing van verschillen tussen de groepen de toets van STUDENT het meest in aanmerking, subs. de F-toets, hetgeen hier hetzelfde is. In de hierna volgende berekeningen worden alle bedragen uitgedrukt in eenheden van *f* 100.-.

	A	B	A^2	B^2
	31	46	961	2116
	34	6	1156	36
	109	4	11881	16
	14	19	196	361
	21	5	441	25
	69	24	4761	576
	35	26	1225	676
	48	47	2304	2209
	60	51	3600	2601
	62	27	3844	729
$\Sigma x =$	483	255	30369	9345
$(\Sigma x)^2/10 =$			23328,9	6502,5
$\Sigma(x-\bar{x})^2 =$			7040,1	2842,5

$s_A^2 = 782,2$

$s_B^2 = 315,8$

Als schatting van de variantie binnen de groepen, s_0^2, dient

$$s_0^2 = \frac{7040,1 + 2842,5}{10-1 + 10-1} = \frac{9882,6}{18} = 549,0,$$

$$s_0 = 23,4.$$

De standaardafwijking van het verschil tussen de gemiddelde inkomens van A en B bedraagt

$$s = 23,4\sqrt{\frac{1}{10} + \frac{1}{10}} = 10,46$$

$$t = \frac{48,3 - 25,5}{10,46} = \frac{22,8}{10,46} = 2,18.$$

Volgens de tabel van de STUDENT-verdeling behoort hierbij een tweezijdige overschrijdingskans van iets minder dan 0,05 bij 18 vrijheidsgraden. Het verschil tussen de gemiddelde inkomens der beide groepen is dus significant

bij de gestelde onbetrouwbaarheidsdrempel, hetgeen het vermoeden rechtvaardigt dat groep A gemiddeld een hoger inkomen heeft dan groep B.
Men kan tot dezelfde uitkomst komen door toepassing van de rekentechniek der variantie-analyse. In dat geval wordt het gemiddelde kwadraat tussen de groepen, Q_t, uitgerekend:

$$\begin{array}{rrr} \Sigma x_A = 483 & 483^2 = 233289 & \\ \Sigma x_B = 255 & 255^2 = \ 65025 & \\ \hline \Sigma x \ = 738 & \underline{298314} & + \\ & 29831,4 & : 10 \\ 738^2/20 = & 27232,2 & \\ & \overline{2599,2} & \end{array}$$

$$Q_t = 2599,2 : 1 = 2599,2$$

$$s_o^2 = \ 549,0. \text{ (zie boven)}$$

$$F = \frac{Q_t}{s_o^2} = \frac{2599,2}{549,0} = 4,73 = t^2 \text{ bij 1 en 18 vrij-}$$

heidsgraden.
Ook nu wordt uiteraard een overschrijdingskans gevonden van iets minder dan 0,05. (Daar $F = t^2$ is, komt een rechtséénzijdige overschrijdingskans van F overeen met een tweezijdige van de corresponderende waarde van t.)

a.2. Indien geen normaliteit der inkomensverdelingen mag worden verondersteld, komt slechts een verdelingsvrije toets in aanmerking. In dit geval is het de toets van WILCOXON. Voor berekening van de toetsingsgrootheid W wordt in navolging van "Handleiding voor de toets van WILCOXON" door ir DORALIEN WABEKE en CONSTANCE VAN EEDEN (Math.Centrum, Asd. 1955) de volgende opstelling gemaakt. (Zie tabel op blz.82.)
Bij m=n=10 wordt voor W=46 een tweezijdige overschrij-

dingskans van 0,044 gevonden, hetgeen wederom tot verwerping van de getoetste hypothese leidt. (Opmerking: In plaats van W wordt ook wel een toetsingsgrootheid U = $\frac{1}{2}$W gebruikt.)

b. Hoewel de uitkomsten der beide gebruikte toetsen in dit geval vrijwel gelijk zijn, dient men zich toch te realiseren, welke van beide in dit geval de juiste is. Dit hangt er van af in hoeverre de inkomens normaal verdeeld geacht mogen worden. Zelfs zonder nadere gegevens daarover valt gemakkelijk in te zien dat inkomens als regel verre van normaal verdeeld zijn; tegenover veel kleine inkomens staan weinig grote inkomens. Bij vraagstukken met inkomens is het daarom nooit geoorloofd

Rangnr.:	A	B	Bijdrage tot W
0		6,4,5	0
1	14		
2		19	2
3	21		
4		24,26,27	12
5	31		
6			
7	34		
8			
9	35		
10		46,47	20
11	48		
12		51	
13	60		
14			
15	62		
16			
17	69		
18			
19	109		
20			
	m=10	n=10	W=46

zonder nadere gegevens normaliteit te veronderstellen. Vandaar dat hier de toets van WILCOXON de juiste is.

c. Bij 18 vrijheidsgraden en een tweezijdige overschrijdingskans van 0,01 behoort blijkens de STUDENT-tabel een waarde t = 2,88.
Bij de gestelde verandering van de inkomens van groep B blijft de spreiding binnen de groepen onaangetast. Slechts het verschil tussen de gemiddelden van A en B ($\overline{A}$ en $\overline{B}$ te noemen) verandert. Dus

$$t = \frac{48,3 - \overline{B}}{10,46} = 2,88.$$

$\overline{B} = 48,3 - 2,88 \times 10,46 = 48,3 - 30,1 = 18,2.$
$\overline{B}$ was 25,5 zodat dus alle waarden van B 25,5 - 18,2 = = 7,3 kleiner zouden moeten worden om $\overline{B}$ = 18,2 en dus t = 2,88 te vinden.

Opmerking: t kan ook = -2,88 worden gesteld. In dat geval wordt $\overline{B}$ = 78,4 verkregen, hetgeen 52,9 meer is dan de oorspronkelijke waarde van $\overline{B}$. Door verhoging van alle inkomens in deze groep met 52,9 kan men dus hetzelfde effect bereiken.

Vijfde examen, oktober 1957

1. Uit onderstaande bererekeningsstaat blijkt voldoende duidelijk, op welke wijze een normale verdeling kan worden aangepast en de theoretische frequenties kunnen worden berekend.
Weglating van de correctie van SHEPPARD op de variantie (vermindering met 1/12e van het kwadraat van de klassebreedte) is niet fout. Juister is het echter, deze correctie wel toe te passen bij dit grote aantal waarnemingen.

BEREKENINGSSTAAT

Dikte d	Klas-semid-dens x	f	fx	fx^2	Klas-se-gren-zen x	u	$\int_{-\infty}^{u} f(u)du$	p^*	f^*
					-5,5	-4,09	0,0000		
9,7-10,2	-5	0						0,0004	1
					-4,5	-3,34	0,0004		
10,3-10,8	-4	0						0,0045	6
					-3,5	-2,58	0,0049		
10,9-11,4	-3	31	-93	279				0,0287	39
					-2,5	-1,83	0,0336		
11,5-12,0	-2	168	-336	672				0,1065	146
					-1,5	-1,08	0,1401		
12,1-12,6	-1	339	-339	339				0,2306	316
					-0,5	-0,33	0,3707		
12,7-13,2	0	381						0,2921	400
					0,5	0,42	0,6628		
13,3-13,8	1	274	274	274				0,2182	299
					1,5	1,18	0,8810		
13,9-14,4	2	130	260	520				0,0922	126
					2,5	1,93	0,9732		
14,5-15,0	3	41	123	369				0,0231	32
					3,5	2,68	0,9963		
15,1-15,6	4	6	24	96				0,0034	5
					4,5	3,43	0,9997		
15,7-16,2	5	0						0,0003	0
					5,5	4,18	1,0000		
Totaal		1370	-87	2549				1,0000	1370

$87^2/1370 = 6$

2543

$\bar{x} = -87/1370 = -0,06$

$s'^2 = 2543/1369 = 1,86$

Correctie van Sheppard:

$s^2 = 1,86 - 0,08 = 1,78$

$s = \sqrt{1,78} = 1,33$

x = (d-12,95)/0,6

f = absolute frequentie (waargenomen)

u = $(x-\bar{x})/s$

f(u) = normale verdeling

p^* = relatieve frequentie (berekend)

f^* = absolute frequentie (berekend)

2. a. Indien y^*, x^*y en x alle afwijkingen van hun gemiddelden voorstelden, zouden de vergelijkingen geen constante termen bevatten. Dit is dus niet het geval. In principe kan echter in ieder der beide vergelijkingen nog één der beide variabelen een afwijking van het gemiddelde voorstellen. Dit is echter ongebruikelijk.

   b. $|r| = \sqrt{bb'}$, waarin b en b' de beide regressiecoëfficiënten zijn.

   c. $\sqrt{bb'} = \sqrt{2,1.0,7} = \sqrt{1,47} = 1,21 > 1$.
   Dit kan niet juist zijn, dus één van beide (of beide) regressievergelijkingen is (zijn) onjuist. Uit de gegevens valt niet op te maken, welke vergelijking fout is.

3. De verdeling van het aantal fouten in rollen van 40 m lengte is een POISSON-verdeling. De parameter λ wordt geschat door het gemiddelde aantal fouten:

$$\frac{80 + 2.32 + 3.15 + 4.3}{200} = \frac{201}{200} \approx 1.$$

Uit een tabel van de POISSON-verdeling volgen de percentages:

Aantal fouten	%	Op 200 rollen	
0	37	74	De overeenstemming met de gevonden aantallen is goed.
1	37	74	
2	18	36	
3	6	12	
4	1	2	

a. In rollen van 60 meter is $\lambda = \frac{60}{40} \cdot 1 = 1,5$.

Is $\underline{x}$ het aantal fouten, dan is dus

$$P[\underline{x}=x] = e^{-1,5} \frac{(1,5)^x}{x!} .$$

b. De percentages volgen weer uit een POISSON-tabel.
Voor lengte 40 m ($\lambda = 1$) is percentage rollen van A-kwal. = 74.
Voor lengte 60 m ($\lambda = 1{,}5$) is percentage rollen van A kwal. = 56.
Het percentage meters van A-kwaliteit is gelijk aan het percentage rollen van die kwaliteit.

c. Een percentage 90 wordt bereikt voor $\lambda = 0{,}53$, hetgeen behoort bij een lengte $L = 40.\ \lambda$ meter = $= 40.\ 0{,}53 \approx 21$ meter.

4. De beste schatting voor de kans op succes bij ieder der bezoeken is $\frac{41}{300} = 0{,}137$. De standaardafwijking van deze schatting is ongeveer gelijk aan

$$\sqrt{\frac{41}{300} \cdot \frac{259}{300} \,/\, 300} = 0{,}0198 \ .$$

Een onderste betrouwbaarheidsgrens voor deze kans, met onbetrouwbaarheidsdrempel 0,05, is dus

$$0{,}137 - 1{,}65 \ . \ 0{,}0198 = 0{,}104 \ .$$

In een lange reeks van verdere bezoeken, met dezelfde kans op succes, zal dus het percentage successen ongeveer gelijk aan deze kans, dus minstens 0,104 zijn.

Hierbij zijn twee veronderstellingen gemaakt:

1e de eerste 300 bezoeken op de alfabetische lijst kunnen - voor zoverre het al of niet behalen van succes betreft - als een aselecte steekproef uit de gehele lijst worden beschouwd (m.a.w. de kans op succes is onafhankelijk van de plaats van de naam in het alfabet);

2e de totale lijst is zeer lang.

5. a. Noemen wij x_1 het aantal verdachte gevallen bij de 1ste ploeg, dan is de verwachting, gegeven het totale

aantal Σx_i, voor ieder der ploegen gelijk aan $\bar{x}$. Dus is:

$$\chi^2_{k-1} = \frac{\Sigma(x_i-\bar{x})^2}{\bar{x}} \quad (k-1 \text{ vrijheidsgraden}).$$

Deze waarde van χ^2 kan op de bekende wijze worden getoetst.

$\Sigma x_i = 810$; $\bar{x} = 81$; $\Sigma(x_i-\bar{x})^2 = 15246$;

$\chi^2_9 = \frac{15246}{81} = 188$; $E\chi^2_9 = 9$, dus 188 is veel te groot. (Opzoeken in tabel is overbodig.) De numerieke berekening werd niet van de candidaten gevraagd.

b. Een of meer der volgende verklaringen lijken redelijk:
   1. Geografische verschillen, als de ploegen op verschillende plaatsen werken.
   2. Milieu- en beroepsverschillen tussen de onderzochte groepen.
   3. Verschillen in de diagnose-methode, die bij de verschillende ploegen gebruikt werd.

c. Onder de hypothese dat de kansen voor de 10 groepen nog steeds gelijk aan de gegeven percentages zijn, zal het verwachte aantal verdachte gevallen der groepen evenredig zijn met deze percentages. De beste schattingen van de verwachte aantallen zijn die schattingen, waarvan de som gelijk is aan de som van de werkelijke aantallen verdachte gevallen (=810). De berekening is in onderstaande tabel uitgevoerd. Ook nu kan weer de waarde van χ^2_9 worden berekend. Deze berekening is eveneens in de tabel opgenomen. Daarbij geldt:

$$f^* = \frac{810}{1,4} p^* .$$

Ploeg	p^*	f	f^*	$f-f^*$	$(f-f^*)^2/f^*$
A	0,25	137	144,6	-7,6	0,40
B	0,19	94	109,9	-15,9	2,30
C	0,17	103	98,4	4,6	0,21
D	0,08	62	46,3	15,7	5,32
E	0,05	21	28,9	-7,9	2,16
F	0,25	138	144,6	-6,6	0,30
G	0,18	110	104,1	5,9	0,33
H	0,09	60	52,1	7,9	1,20
I	0,08	43	46,3	3,3	0,24
J	0,06	42	34,7	7,3	1,54
Totaal	1,40	810	809,9	0,1	14,00 = $= \chi^2_9$

De bijbehorende overschrijdingskans is groter dan 0,10, dus er is geen reden om de waarnemingen in strijd te achten met de vroegere bevindingen.

6. a. $100 \times \frac{85 + 40 + 16}{51 + 25 + 12} = 160,2.$

b. Prijsindex: $\frac{51 \times 110 + 25 \times 120 + 12 \times 125}{51 + 25 + 12} = 114,9.$

Hoeveelheidsindex: $100 \times \frac{160,2}{114,8} = 139,4.$

Opmerking. In plaats van volgens bovenstaande berekening (formule van LASPEYRES) kan men het prijsindexcijfer ook volgens de formule van PAASCHE berekenen (dit is echter niet

$$\frac{85 \times 110 + 40 \times 120 + 16 \times 125}{85 + 40 + 16}\ ).$$

7. De kansen op de scores 0, 1, 2 en 3 zijn resp. a, $\frac{1}{2}$a, $\frac{1}{4}$a en $\frac{1}{8}$a. Aangezien $a + \frac{1}{2}a + \frac{1}{4}a + \frac{1}{8}a = 1$, is $a = \frac{8}{15}$.

a. De verwachting van het kwadraat van de score is

$$\frac{8}{15} \times 0 + \frac{4}{15} \times 1 + \frac{2}{15} \times 4 + \frac{1}{15} \times 9 = \frac{21}{15} .$$

b. De verwachting van de score is

$$\frac{8}{15} \times 0 + \frac{4}{15} \times 1 + \frac{2}{15} \times 2 + \frac{1}{15} \times 3 = \frac{11}{15} .$$

De variantie van de score is $\frac{21}{15} - (\frac{11}{15})^2 = \frac{194}{225}$.

De standaardafwijking van de gemiddelde score bij 25 schoten bedraagt dus

$$\sqrt{\frac{1}{25} \times \frac{194}{225}} = 0,186.$$

Opmerking: "gemiddelde score bij 25 schoten" zou men ook kunnen opvatten als "verwachting van de totale score bij 25 schoten". Deze verwachting is gelijk aan $25.\frac{194}{225} = \frac{194}{5}$, een constant getal. De standaardafwijking daarvan is dus 0.

8. Een eenvoudige en voor het gestelde doel toepasselijke toets is de rangcorrelatietoets van KENDALL. De daarbij behorende toetsingsgrootheid S is gelijk aan het aantal paren waarnemingen, waarvan de tweede groter is dan de eerste, verminderd met het aantal paren, waarvoor het omgekeerde geldt.
Paren van gelijke waarnemingen geven een bijdrage 0.
De berekening kan bijv. als volgt uitgevoerd worden.
Achter de eerste waarneming staan er 15, die groter zijn en 4, die kleiner zijn. Achter de tweede 10, die groter en 8, die kleiner zijn, enz. Op deze wijze worden twee rijen getallen verkregen en S is gelijk aan het verschil van de totalen daarvan.
In dit geval wordt gevonden: S = 48.
Er zijn slechts twee paren van gelijke waarnemingen; dit betekent, dat hun invloed op de uitkomst van de

toets zeer gering is en verwaarloosd kan worden.
Uit een tabel van kritieke waarden van S valt af te lezen, dat (voor n=20) de tweezijdige overschrijdingskans van de waarde 48 tussen 0,10 en 0,20 ligt (de bij deze drempels behorende kritieke waarden zijn 52 en 42).
Is geen tabel beschikbaar dan make men gebruik van de volgende formules voor verwachting en variantie van $\underline{S}$, onder H_0.

$$\mu = 0; \quad \sigma^2 = \frac{1}{18} n(n-1)(2n+5) = 950.$$

Daar S (in dit geval) alleen even waarden aan kan nemen bedraagt de continuïteitscorrectie -1.
Men berekent dus

$$\frac{S-1}{\sigma} = \frac{47}{30,8} = 1,53$$

en, de normale verdeling als benadering gebruikend, vindt men voor de bijbehorende tweezijdige overschrijdingskans de waarde $2.0,0630 \approx 0,13$.
Er is dus geen reden om tot een verloop met de tijd te concluderen.

9. a. De correlatiecoëfficient is 0. Dit volgt uit het feit, dat de verdeling symmetrisch is ten opzichte van de vertikale lijn x=2, die door het zwaartepunt gaat. (Ook de symmetrie ten opzichte van de horizontale lijn y=3 leidt tot dit resultaat.) Voor dergelijke symmetrische tweedimensionale verdelingen is nl. de covariantie gelijk aan 0.

   b. De grootheden $\underline{x}$ en $\underline{y}$ zijn stochastisch afhankelijk. Dit blijkt uit het feit, dat de waarden, die $\underline{y}$ bij gegeven x aan kan nemen, voor verschillende waarden van x niet dezelfde zijn. De voorwaardelijke verdeling van $\underline{y}$, bij gegeven x, is dus niet onafhankelijk van x. Analoog voor de voorwaardelijke verdeling van $\underline{x}$ bij gegeven y.

Zesde examen, november 1958

1. Gaat men er vanuit, dat de personeelsbezetting van de afdeling niet zo klein is, dat deze tengevolge van een ongeval merkbaar kleiner wordt, dan mag men veronderstellen, dat het aantal ongevallen per jaar een POISSON-verdeling bezit.
   In de voorafgaande periode was de verwachting van deze verdeling gelijk aan 1. De hypothese moet nu getoetst worden, dat dit in de 3 jaar na het in gebruik nemen van de nieuwe machines nog zo is. De verwachting van de POISSON-verdeling van het aantal ernstige ongevallen in 3 jaar is dan 3.
   De rechter-overschrijdingskans van het aantal 10 is dan, volgens een tabel van de POISSON-verdeling, gelijk aan 0,0011. Deze kleine overschrijdingskans (die ook, voor tweezijdige toetsing, na vermenigvuldiging met 2 nog zeer klein is) leidt tot verwerping van de getoetste hypothese. De conclusie is dus, dat het werken met de nieuwe machines inderdaad gevaarlijker is dan met de oude.
   Dat hier geen sprake is van een kortstondig groter gevaar tengevolge van onbekendheid met de nieuwe machines, blijkt uit het feit, dat in het tweede en derde jaar na het in bedrijf stellen van deze machines vrijwel evenveel ernstige ongevallen zijn voorgekomen als in het eerste jaar.
   De exacte grootte van de personeelsbezetting behoeft niet bekend te zijn. Het gegeven, dat deze niet van omvang is veranderd, is essentieel.
   Zou de personeelsbezetting zo gering zijn, dat een ongeval aanzienlijk meer werk voor de rest van het personeel zou betekenen, zodat de kans op verdere ongevallen daardoor vergroot zou worden, dan volgt het aantal ongevallen per jaar geen POISSON-verdeling meer en is de

bovenstaande redenering niet langer juist.

2. De klassen van het inkomen en de uitgaven worden als volgt getransformeerd:

x (inkomen) oorsprong 6699,5 eenheid 600 gld.
y (uitgaven) " 2999,5 " 400 gld.

De tabel wordt dan:

Uitgaven (klassemidden)	Inkomen (klassemidden)								
	-3	-2	-1	0	1	2	3	4	totaal
-3	-	1	1	1	-	-	-	-	3
-2	6	1	2	1	1	-	-	-	11
-1	1	5	9	4	2	-	1	-	22
0	-	3	4	8	6	3	-	1	25
1	-	-	2	3	5	3	2	1	16
2	-	-	-	2	1	3	3	-	9
3	-	-	-	-	1	-	2	1	4
Totaal	7	10	18	19	16	9	8	3	90

$$\bar{x} = \frac{7(-3) + 10(-2) + \ldots}{90} = \frac{11}{90}$$

$$\bar{y} = \frac{3(-3) + 11(-2) + \ldots}{90} = -\frac{7}{90}.$$

$$\Sigma(x-\bar{x})^2 = 7(-3)^2 + 10(-2)^2 + \ldots - \frac{(11)^2}{90} = 293 - 1{,}34 =$$

$$= 291{,}66.$$

$$\Sigma(y-\bar{y})^2 = 3(-3)^2 + 11(-2)^2 + \ldots - \frac{(-7)^2}{90} = 181 - 0{,}54 =$$

$$= 180{,}46.$$

$$\Sigma(x-\bar{x})(y-\bar{y}) = 152 - \frac{(11)(-7)}{90} = 152{,}86.$$

$$r = \frac{152,86}{\sqrt{291,66 \times 180,46}} = 0,67.$$

Opmerkingen:

1) De transformatie van x en y heeft geen invloed op de grootte van de correlatiecoëfficiënt. Het is dus niet nodig terug te transformeren (dit zou wel het geval zijn indien ook de regressiecoëfficiënten zouden moeten worden berekend).
2) De bovenstaande rekenwijze is het eenvoudigst indien men niet over een rekenmachine beschikt. Indien dit wel het geval is, kan het de voorkeur verdienen de berekening enigszins anders uit te voeren (afhankelijk van het type rekenmachine).
3) Geen rekening is gehouden met het feit dat de laatste klassen open zijn. Wegens de lage frequenties in deze klassen kan de hierdoor ontstane fout niet groot zijn.
4) De correctie van SHEPPARD is niet toegepast. In dit geval is deze nl. toch veel kleiner dan de steekproeffluctuaties.

3. a. Indien de totale tijd door tandarts i aan de in de vraag bedoelde consulten besteed aangegeven wordt met t_i en het aantal van dergelijke consulten met n_i, dan is de gemiddelde duur van een consult bij deze tandarts

$$d_i = \frac{t_i}{n_i}.$$

Zijn k tandartsen in de steekproef opgenomen, dan wordt het algemeen gemiddelde van de duur van een consult $\overline{d}_I$ berekend volgens methode I, gegeven door

$$\overline{d}_I = \frac{\sum_{i=1}^{i=k} d_i}{k}.$$

Het algemeen gemiddelde van de duur van een consult $\bar{d}_{II}$, berekend volgens methode II wordt gegeven door

$$\bar{d}_{II} = \frac{\sum_{i=1}^{i=k} t_i}{\sum_{i=1}^{i=k} n_i} = \frac{\sum_{i=1}^{i=k} n_i d_i}{\sum_{i=1}^{i=k} n_i} .$$

Hieruit volgt, dat $\bar{d}_I$ een ongewogen gemiddelde van de grootheden d_i en $\bar{d}_{II}$ een gewogen gemiddelde - met gewichten n_i, het aantal consulten in een jaar - van de grootheden d_i is.
Het algemeen gemiddelde $\bar{d}_I$ zal dus in het algemeen ongelijk zijn aan het algemeen gemiddelde $\bar{d}_{II}$.

b. In de vraag wordt gesteld, dat het honorarium per consult zodanig moet zijn, dat bij een redelijk geachte werktijd een redelijk inkomen door de tandartsen verkregen kan worden. Daar verder gesteld werd, dat de geënquêteerde tandartsen een (aselecte) steekproef uit de populatie van tandartsen vormen, is er geen reden om aan de tandartsen een verschillend gewicht toe te kennen bij de berekening van het algemeen gemiddelde van de duur van een consult.
Dit heeft namelijk tot gevolg dat de eis van een redelijk honorarium bij de gemiddelde tandarts wordt bereikt.
Voor het in vraag b gestelde doel zal dus het algemeen gemiddelde volgens methode I berekend moeten worden.

Opmerkingen.
Bij de beantwoording van deze vraag is door vele candidaten als antwoord gegeven, dat voor de berekening van de gemiddelde duur van een consult in het onderhavige geval methode II gebruikt moet worden. Vaak ging men hierbij uit van de plausibele veronderstelling dat er een negatieve correlatie bestaat tussen de drukte van de practijk (= aantal consulten) en de gemiddelde duur van

een consult. Voorts werd dan meestal betoogd dat het onjuist zou zijn om de consultduur van een arts met een kleine practijk even zwaar te wegen als die van een arts met een grote practijk. Als een dergelijk antwoord redelijk werd geargumenteerd, is het niet onvoldoende gerekend, daar de onjuistheid ervan niet van statistische aard is.

4. a. Indien de grootte van de standaardafwijking niet afhangt van het werkelijk gehalte van het monster, dan is het redelijk als eis te stellen de duplobepaling over te doen, indien de absolute waarde van het verschil tussen twee duplo's een bepaalde waarde (in het vraagstuk 0,05) overschrijdt. Indien de grootte van de standaardafwijking evenredig is met het werkelijk gehalte (althans in het gebied wat in de practijk voorkomt), dan is het redelijk om als eis te stellen, dat de experimenten overgedaan worden, indien het verschil een bepaald percentage van het gemiddelde der twee bepalingen overschrijdt.
Het overdoen van een stel bepalingen, indien de spreidingsbreedte der uitkomsten een van te voren vastgesteld bedrag, resp. percentage overschrijdt, heeft alleen zin, indien de mogelijkheid van blunders niet uitgesloten is (deze opmerking geldt ook voor vraag b, c en d) en de waarnemingen alle een gelijke spreiding in absolute zin, resp. procentueel bezitten. Normaliteit der waarnemingen is wat dit voorschrift betreft niet nodig, wel als men met behulp van de factor 0,886 uit de spreidingsbreedten de standaardafwijking wil schatten. Indien blunders niet zouden kunnen optreden en de duplowaarnemingen onderling afhankelijk zijn, heeft het nauwelijks zin meer dan één waarneming te verrichten; voor de beantwoording van vraag a behoeft echter de veronderstelling van de onderlinge onafhankelijk-

heid niet gemaakt te worden.

b. De methode om door vermenigvuldigen van het gemiddelde der spreidingsbreedten van de niet weggeworpen duplobepalingen met een factor 0,886 een schatting van de standaardafwijking te verkrijgen, geeft geen juist beeld der werkelijke nauwkeurigheid. Immers deze berekening is alleen juist als elke waarneming een onafhankelijke (punt a) steekproef uit een normale verdeling (punt b) is, met dezelfde standaardafwijking (punt c). Bij een normale verdeling kan echter elke willekeurig grote spreidingsbreedte bij een duplobepaling theoretisch voorkomen. Het weglaten van duplobepalingen, waarvan de spreidingsbreedte een zekere waarde overschrijdt, geeft dus een onderschatting van de gemiddelde spreidingsbreedte en dus van de daaruit afgeleide standaardafwijking.

c. Het overdoen van duplobepalingen, indien het onderlinge verschil groter dan 0,05 is, heeft alleen zin, indien

1. een mogelijkheid van blunders bestaat (veronderstelling d);
2. de grens zodanig gekozen is, dat niet een te grote fractie der waarnemingen overgedaan moet worden wegens een door het toeval veroorzaakte spreidingsbreedte groter dan 0,05.

De kans op het optreden van een spreidingsbreedte groter dan 0,05 onder de veronderstelling van normale, onafhankelijke gelijkgespreide waarnemingen met gegeven standaardafwijking σ kan op de volgende wijze gevonden worden:
Het verschil van twee dergelijke waarnemingen is normaal verdeeld met gemiddelde nul en standaardafwijking $\sigma\sqrt{2}$.
De kans, dat de absolute waarde van dit verschil groter dan 0,05 is, is gelijk aan het dubbele van de kans, dat een normaal verdeelde stochastische grootheid u met gemiddelde 0 en standaardafwijking 1 de waarde $\frac{0,05}{\sigma\sqrt{2}}$ overschrijdt.

Is dus σ gelijk aan 0,012, dan is $\frac{0,05}{\sigma\sqrt{2}} = 2,95$ en de tweezijdige overschrijdingskans 0,003. Is σ gelijk aan 0,028, dan is $\frac{0,05}{\sigma\sqrt{2}} = 1,26$ en de tweezijdige overschrijdingskans is 0,21.
Als de standaardafwijking van de populatie 0,012 bedraagt, zullen dus circa 3 van de 1000 bepalingen overgedaan worden, als er geen blunders begaan werden; bij een standaardafwijking van 0,028 circa 21 op de 100. In het eerste geval is de methode dus redelijk, in het tweede geval niet.

d. Daar het wegwerpen van een stel waarnemingen niet afhankelijk is van de hoogte der uitkomsten, doch alleen van hun verschil, is het duidelijk, dat de methode een zuivere schatting van de gemiddelden geeft. Hierbij is er van uitgegaan, dat de waarnemingen onafhankelijk zijn, althans de overgedane waarnemingen niet afhankelijk zijn van de waarnemingen welke zij vervangen; in het bijzonder klemt dit als blunders mogelijk zijn. Normaliteit en gelijkgespreidheid behoeven niet aangenomen te worden.

5. a. Stel $\underline{v} = \underline{b} - \underline{k}$. Daar $\underline{b}$ en $\underline{k}$ beide normaal verdeeld zijn ondersteld, is $\underline{v}$ dit ook en wel aanvankelijk met

gemiddelde $\mu_v = \mu_b - \mu_k = 200 - \mu_k$, en

variantie $\sigma_v^2 = \sigma_b^2 + \sigma_k^2$ (b en k zijn onafhankelijk).

Daar breuk optreedt als $\underline{v} < 0$, is dus

$$P\left[\underline{v} < 0\right] = 0,03,$$

ofwel

$$P\left[\frac{\underline{v} - \mu_v}{\sigma_v} < \frac{-\mu_v}{\sigma_v}\right] = 0,03.$$

Hierbij heeft $\underline{u} = \frac{\underline{v} - \mu_v}{\sigma_v}$ dus een normale verdeling met gemiddelde 0 en spreiding 1.
Deze verdeling is symmetrisch om 0, zodat ook moet gelden

$$P\left[\underline{u} > \frac{\mu_v}{\sigma_v}\right] = 0{,}03.$$

In een tabel van de normale (0,1)-verdeling zoeken we nu de waarde op, die een kans 0,03 heeft, overschreden te worden en vinden daarvoor 1,88.

Dus $$\frac{\mu_v}{\sigma_v} = 1{,}88,$$

of $$\frac{200 - \mu_k}{\sigma_v} = 1{,}88. \qquad (1)$$

In de gewijzigde toestand heeft $\underline{v}$ wel dezelfde spreiding, maar een ander gemiddelde nl.

$$\mu_v' = \mu_b' - \mu_k = 231{,}5 - \mu_k.$$

Op dezelfde wijze redenerend vinden we dan, uitgaande van $P\left[\underline{v} < 0\right] = 0{,}006$, de betrekking:

$$\frac{231{,}5 - \mu_k}{\sigma_v} = 2{,}51. \qquad (2)$$

Lost men $\frac{\mu_k}{\sigma_v}$ op uit (2), en substitueert men de gevonden waarde in (1), dan volgt:

$$\frac{200}{\sigma_v} + \left(2{,}51 - \frac{231{,}5}{\sigma_v}\right) = 1{,}88$$

ofwel

$$31{,}5 = 0{,}63\ \sigma_v$$

zodat

$$\sigma_v = 50. \qquad (3)$$

Uit $\sigma_v^2 = \sigma_b^2 + \sigma_k^2$ volgt dan, daar gegeven is $\sigma_b = 40$,

$$\sigma_k^2 = 50^2 - 40^2 = 30^2,$$

dus $\underline{\sigma_k = 30.}$

Uit (1) en (3) volgt:

$$200 - \mu_k = 50 \times 1{,}88$$
$$\underline{\mu_k = 200 - 94 = 106.}$$

b. Men wijzigt dus de gemiddelde breuksterkte. Deze wordt μ_b''. Dan is

$$\mu_v'' = \mu_b'' - 106, \qquad (4)$$

terwijl nog steeds

$$\sigma_v = 50. \qquad (5)$$

De eis is nu dat $P[v < 0] = 0{,}011$. Volgens de redenering, die onder a. leidde tot (1), betekent dit dus dat $\frac{\mu_v''}{\sigma_v}$ gelijk moet zijn aan de waarde van een normaal (0,1)-verdeelde grootheid, die een kans van 0,011 heeft overschreden te worden. Uit de betreffende tabel volgt:

$$\frac{\mu_v''}{\sigma_v} = 2{,}29.$$

Invullen van (4) en (5) geeft

$$\underline{\mu_b'' = 220{,}5.}$$

6. Uit de cijfers blijkt, dat het percentage ernstige ongevallen bij de vrouwelijke bestuurders kleiner is dan

bij de mannelijke.
Hieruit kan echter niet geconcludeerd worden, dat vrouwen voorzichtiger rijden dan mannen. Immers, het gesignaleerde verschijnsel kan ontstaan zijn doordat de vrouwen naar verhouding van het aantal gereden kilometers minder ernstige ongevallen maken dan de mannen, maar ook doordat zij naar verhouding meer kleine ongevallen veroorzaken. Over de al of niet grotere voorzichtigheid der vrouwen zeggen de verstrekte gegevens dan ook in het geheel niets.

7. $$\sigma_y^2 = 4\,\sigma_x^2 .$$

Daar $\underline{x}$ en $\underline{y}$ onafhankelijk zijn, is verder:

$$\sigma_z^2 = 4\,\sigma_x^2 + \sigma_y^2 = 8\,\sigma_x^2.$$

Vervolgens berekenen wij:

$$\operatorname{cov}(\underline{x},\underline{z}) = E(\underline{x}-\mu_x)(2\underline{x}+\underline{y}-2\mu_x-\mu_y) =$$

$$= 2E(\underline{x}-\mu_x)^2 = 2\,\sigma_x^2 ,$$

daar $E(\underline{x}-\mu_x)(\underline{y}-\mu_y) = 0$ wegens de onafhankelijkheid van $\underline{x}$ en $\underline{y}$.

$$\rho(\underline{x},\underline{z}) = \frac{\operatorname{cov}(\underline{x},\underline{z})}{\sigma_x\,\sigma_z} = \frac{2\,\sigma_x^2}{\sigma_x\,\sigma_z} .$$

Wegens $\sigma_z = 2\,\sigma_x\sqrt{2}$ volgt hieruit:

$$\rho(\underline{x},\underline{z}) = \frac{1}{\sqrt{2}} = \tfrac{1}{2}\sqrt{2}.$$

8. Wij kennen de 17 werkstukken aselect getallen onder 100 toe, gebruik makende van de eerste twee kolommen van de tabel en vervolgens van de 3e en 4e kolom. Reeds eerder voorgekomen nummers slaan wij over.

Dus:

nummer werkstuk	aselect getal onder 100	
1	71	C
2	36	A
3	79	C
4	22	A
5	67	B
6	41	A
7	70	C
8	65	B
9	06	A
10	60	B
11	37	A
12	97	C
13	08	A
14	68	B
15	59	B
16	44	A
17	92	C

De 7 laagste aselecte getallen worden A (dat zijn dus de werkstukken nr 2, 4, 6, 9, 11, 13 en 16); de 5 daarop volgende worden B (nr 5, 8, 10, 14 en 15) en de rest C (de nummers 1, 3, 7, 12 en 17).

Opmerkingen.
Bij dit vraagstuk komt het er op aan een zodanige keuze uit de gegeven aselecte getallen te doen, dat ook de 17 exemplaren aselect over de drie bewerkingsmethoden werden verdeeld. Een methode waarbij men de som van twee cijfers als nummertrekking van de exemplaren behandelt is dus fout, omdat de verdeling van deze sommen niet rechthoekig, maar driehoekig is, zodat de getallen in de buurt van 10 meer kans hebben om te worden getrokken dan de getallen 1 of 17 bijv. Men maakt deze fout iets kleiner door een dubbele loting volgens dit systeem toe te passen, nl. één loting voor de exemplaarnummers en één loting voor de methoden. Een goede oplossing is de z.g. rechtstreekse methode, waarbij men ter bepaling van de werkstukken volgens

methode C, uit de tabel 5 getallen van twee cijfers neemt ≤ 85 en deze deelt door 17, waarna dan de rest direct het exemplaarnummer aangeeft. Dezelfde procedure herhaalt men voor methode B, waarna de overblijvende exemplaren aan methode A worden toegewezen.

Behalve op aselectheid komt het in dit vraagstuk aan op een zekere efficiency in de methode. Men zou ook de rechtstreekse methode in die zin kunnen toepassen, dat men in de tabel getallen beneden de achttien opzoekt. Dit is wel zuiver maar onpraktisch omdat men dan het gros van de getallen moet overslaan. Ook is het bij deze methode ondoelmatig om eerst zeven exemplaren voor methode A te loten, dan vijf exemplaren voor methode B en tenslotte vijf exemplaren voor methode C. Het is natuurlijk voldoende om slechts twee van deze drie methoden "vol te loten" omdat de overblijvende exemplaren dan automatisch in de derde methode terechtkomen. Het snelst werkt men dus door 10 exemplaren te loten voor B en C en de overblijvende 7 exemplaren bij A onder te brengen.

9. a. De worpen kunnen in drie klassen van combinaties worden onderscheiden, nl.

aaa, uitkomst = a,
aab, uitkomst = a + b
en abc, uitkomst = a + b + c,

waarin a, b en c verschillende aantallen ogen zijn.
In de klasse aaa kan a vier verschillende waarden aannemen, nl. 1, 2, 3 en 4. Er zijn dus vier combinaties met ieder 1 permutatie.
In de klasse aab kan a vier waarden aannemen en bij iedere a kan b drie waarden aannemen. Er zijn dus $4 \times 3 = 12$ combinaties met ieder drie permutaties, hetgeen 36 variaties oplevert. Deze 12 combinaties hebben paarsgewijs dezelfde uitkomst (nl. aab en bba). De laag-

ste uitkomst is 1 + 2 = 3 en de hoogste 3 + 4 = 7. Er zijn vier combinaties die de uitkomst 5 opleveren, nl. 1 + 4, 4 + 1, 2 + 3 en 3 + 2. Deze uitkomst omvat dus 4 x 3 = 12 variaties. De overige uitkomsten (3, 4, 6 en 7) omvatten 2 x 3 = 6 variaties.

De klasse abc omvat $\frac{4!}{3!1!}$ = 4 combinaties met ieder 3! = 6 permutaties, tezamen dus 24 variaties. De uitkomsten zijn (a+b+c+d) - d = 10 - d. De mogelijke uitkomsten zijn dus 6, 7, 8 en 9, die ieder zes variaties omvatten.

Aannemende dat de dobbelstenen "zuiver" zijn, is de kans op een bepaalde variatie $1/4^3$ = 1/64. Deze kans vermenigvuldigd met de aantallen variaties uit de laatste kolom van de tabel op blz. 104 geeft de kansen voor de verschillende uitkomsten (9 in getal).

b. Als men één steen mag overgooien na een worp 1/2/4, zal men de beste mogelijkheden hebben met overgooien van de 1. Gooit men opnieuw 1, dan blijft de uitkomst 7. Gooit men 2 of 4, dan wordt de nieuwe uitkomst 6. Gooit men 3, dan verbetert de uitkomst tot 9. Ieder dezer worpen heeft een kans 1/4. De verwachting van de uitkomst na overgooien bedraagt dus

$$2/4 \times 6 + 1/4 \times 7 + 1/4 \times 9 = 7,$$

hetgeen gelijk is aan de verkregen uitkomst. Als het er om gaat zoveel mogelijk punten te behalen, dan is het indifferent, of men overgooit of niet.

Hangt daarentegen het al of niet winnen af van het al of niet hoger gooien dan een reeds bekende worp van de tegenpartij, dan is er geen algemeen antwoord mogelijk. Heeft de tegenpartij een worp tussen 1 en 5 gedaan, dan betekenen niet overgooien en wel overgooien beide een zekere winst. Na worp 6 betekent niet

overgooien een zekere winst. Wel overgooien kan tot gelijk spel leiden. In dat geval dus niet overgooien. Na worp 7 van de tegenpartij speelt men gelijk door niets doen. Door overgooien heeft men kans 2/4 te verliezen tegen kans 1/4 op gelijk spel en kans 1/4 op winst. In dat geval dus evenmin overgooien. Is een worp 8 of 9 gedaan, dan neemt men door niet overgooien een zeker verlies. Door overgooien heeft men kans 1/4 op winst, subs. gelijk spel. In deze situatie moet men dus overgooien.
In tabelvorm zien de berekeningen sub a er als volgt uit:

Uitkomst	Aantal variaties			
	aaa	aab	abc	Totaal
1	1			1
2	1			1
3	1	6		7
4	1	6		7
5		12		12
6		6	6	12
7		6	6	12
8			6	6
9			6	6
Totaal	4	36	24	64
Combinaties:	4	12	4	20
Permutaties:	1	3	6	

c. Als de vlakken der stenen niet van 1, 2, 3 en 4 ogen, maar van 0, 1, 2 en 3 ogen waren voorzien, zou de frequentie-verdeling er anders uitzien. De laagste uitkomst zou dan 0 worden (bij combinatie 0/0/0) en de

hoogste 6 (bij combinatie 1/2/3). In totaal dus 7 klassen, tegen 9 klassen bij de oorspronkelijke inrichting.

10. a. $E(\underline{m}) = \mu$, want m is een zuivere schatting van μ. (Voor $E(\underline{m})$ schrijft men ook wel $\mu(\underline{m})$.)

    b. $E(\underline{s}^2) = \sigma^2 = \mu$, want $\underline{s}^2$ is een zuivere schatting van σ^2, terwijl bij de Poisson-verdeling voorts geldt $\sigma^2 = \mu$.

    c. $\text{var}(\underline{m}) = \frac{\sigma^2}{n} = \frac{\mu}{n}$. Voor $\text{var}(\underline{m})$ mag men ook schrijven van $\sigma^2(\underline{m})$.

11. a. Zie figuur 1. Aangezien de gewichten tot op hectogrammen nauwkeurig zijn opgegeven, moet aangenomen worden, dat de klassegrenzen bij 34,95 resp. 36,95 enz. liggen.

    b. Zie figuur 2.

    c. Zie de in figuur 2 getrokken rechte lijn. De relatieve nauwkeurigheid van het eerste punt bij klassegrens 34,95, dat bij een percentage van 2,5% behoort, is minder dan die van de vijf volgende punten; daarom is de lijn in hoofdzaak aan de vijf volgende punten aangepast. Vijf kinderen hadden een gewicht van meer dan 40,95 kg; dit geldt ook voor de aangepaste verdeling.

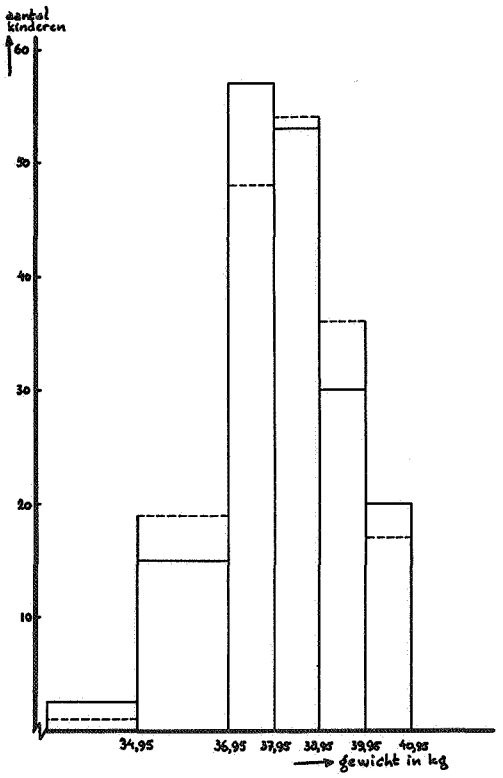


Fig. 1. Frequentieverdeling van de gewichten van 200 schoolkinderen.

——— histogram van de waargenomen verdeling.

····· histogram van de aangepaste normale verdeling.

d. Uit de volgens c aangepaste lijn volgen de in onderstaande tabel 1 aangegeven cumulatieve, en daaruit in de volgende tabel 2 afgeleide niet cumulatieve

relatieve frequenties. Door vermenigvuldiging met 200 kunnen hieruit de in de laatste kolom aangegeven absolute frequenties berekend worden. Deze zijn aangegeven in figuur 1.

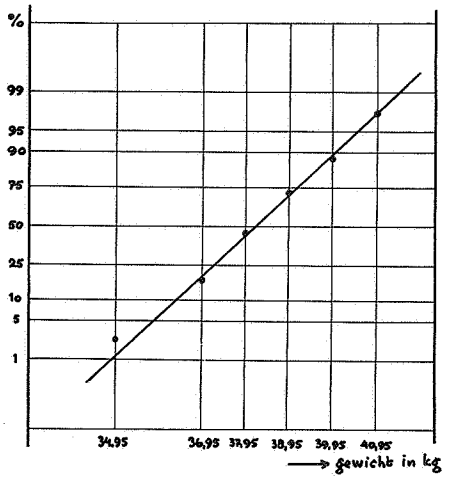


Fig. 2. Cumulatieve frequentieverdeling van de gewichten van 200 schoolkinderen met op het oog aangepaste normale verdeling.

TABEL 1

Klassegrens	Cumulatieve frequentie in %
34,95	1,2
36,95	20
37,95	44
38,95	71
39,95	89
40,95	97,5

Tabel 2

Klasse	Relatieve frequentie in %	Absolute frequentie (afgerond)
32,95 - 34,95	1,2	2
34,95 - 36,95	18,8	38
36,95 - 37,95	24	48
37,95 - 38,95	27	54
38,95 - 39,95	18	36
39,95 - 40,95	8,5	17
> 40,95	2,5	5

Opmerkingen.
De meest voorkomende fout bij dit vraagstuk was dat in het histogram de hoogten (en niet de oppervlakken) evenredig gesteld werden met de aantallen in de diverse gewichtsklassen.

Zevende examen, oktober 1959

1. a. χ_a^2 wordt op de bekende wijze berekend:

$$\chi_a^2 = \frac{(24-10)^2}{10} + \frac{(8-10)^2}{10} + \ldots + \frac{(6-10)^2}{10} = \frac{245}{10} = 24,5.$$

In een χ^2-tabel zien we, bij 9 vrijheidsgraden, dat de bijbehorende overschrijdingskans kleiner is dan 0,005. De onder a. gestelde hypothese moeten we dus ten duidelijkste verwerpen.

b. $\chi_b^2 = \frac{(9-10)^2}{10} + \frac{(7-10)^2}{10} + \ldots + \frac{(5-10)^2}{10} = \frac{82}{10} = 8,2.$

De bijbehorende overschrijdingskans is ongeveer 0,50. Deze hypothese wordt dus niet verworpen.

c. De verwachte frequenties zijn in dit geval voor beide reeksen:

cijfer	0	1	2	3	4	5	6	7	8	9
verwachte frequentie	$16\frac{1}{2}$	$7\frac{1}{2}$	8	$10\frac{1}{2}$	$10\frac{1}{2}$	$9\frac{1}{2}$	$11\frac{1}{2}$	$9\frac{1}{2}$	11	$5\frac{1}{2}$

Dus $\chi^2_c = 2\left\{\frac{(24-16\frac{1}{2})^2}{16\frac{1}{2}} + \frac{(8-7\frac{1}{2})^2}{7\frac{1}{2}} + \ldots + \frac{(6-5\frac{1}{2})^2}{5\frac{1}{2}}\right\} =$

$= 2 \times 5{,}78 = 11{,}56.$

De bijbehorende overschrijdingskans, opnieuw bij 9 vrijheidsgraden, is ongeveer 0,25, hetgeen dus niet tot verwerping van de hypothese leidt.

d. We kunnen dus concluderen dat de laatste cijfers van de telefoonnummers niet onafhankelijk homogeen verdeeld zijn. Van de op één na laatste cijfers is geen duidelijke afwijking van homogeniteit geconstateerd, hetgeen niet wil zeggen dat bewezen is dat de kansen op de diverse cijfers nu ook precies aan elkaar gelijk zijn. Onder c. is geen significant verschil tussen beide reeksen gevonden. Ook hier is niet geconcludeerd dat de verdelingen gelijk zijn. De drie conclusies zijn dus niet met elkaar in strijd.

2. Daar bekend is dat de spreiding niet voor alle waarnemingen dezelfde is, komt voor de beantwoording van de gestelde vraag toepassing van variantie-analyse niet in aanmerking.
Hier moet de methode der m rangschikkingen worden toegepast op de 3 rangschikkingen van de 7 laboratoria.
We krijgen dan het volgende schema van rangnummers en kolomtotalen:

	Laboratoria						
	1	2	3	4	5	6	7
Monster: A	4	5	3	6	7	1	2
B	2	7	4	5	6	1	3
C	4	6	3	5	7	2	1
Totalen	10	18	10	16	20	4	6
Afwijkingen van gemiddelde (12)	-2	+6	-2	+4	+8	-8	-6

De som van de kwadraten van de afwijkingen van het gemiddelde kolomtotaal is 224. De kritieke waarde bij een onbetrouwbaarheidsdrempel van 0,01 is 185,6. We kunnen dus concluderen dat de laboratoria duidelijk verschillen.

3. a. Zie onderstaande figuur. Opgemerkt dient te worden dat de vijfde waarneming niet juist kan zijn. De candidaat dient op te merken dat x en y hier vermoedelijk verwisseld zijn of deze waarneming weg te laten; dit laatste is in de grafiek gedaan.

   b. De correlatieberekening, op één der gebruikelijke wijzen uitgevoerd, leidt tot de regressievergelijking:

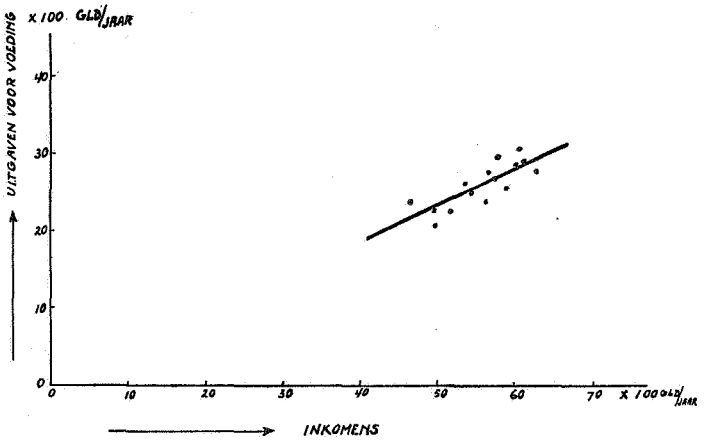
$$y = 0,48x - 0,91.$$

   Hieruit volgt dat een inkomensstijging van 100 gld zal leiden tot een stijging van de uitgaven voor voeding met gemiddeld 48 gld.

   c. Voor de correlatiecoëfficient wordt gevonden met weglating van de 5e waarneming:

$$r = 0,78.$$

   Wanneer de candidaat heeft opgemerkt dat x en y verwisseld zijn in de 5e waarneming en deze weder in de berekening heeft betrokken, wordt een iets andere uitkomst voor de correlatiecoëfficient gevonden.



4. Dit vraagstuk bevat toepassing van het theorema van BAYES. Stel:

   P(A) = kans dat een exemplaar van machine A afkomstig is
   P(B) = idem van B
   P(C) = idem van C
   P(D) = kans dat een exemplaar defect is
   P(D|A) = kans dat als een exemplaar van machine A afkomstig is, dit exemplaar defect is
   P(D|B) = idem van machine B
   P(D|C) = idem van machine C
   P(A|D) = kans dat een defect exemplaar van machine A afkomstig is (de gevraagde kans)

De regel van BAYES zegt:

$$P(A|D) = \frac{P(D|A)P(A)}{P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)} .$$

Gegeven is:

$P(A) = \frac{3017}{7286} = 0,414$

$P(B) = \frac{2655}{7286} = 0,364$

$P(C) = \frac{1614}{7286} = 0,222$

$P(D|A) = 0,036$

$P(D|B) = 0,011$

$P(D|C) = 0,016$

Dus:

$$P(A|D) = \frac{0,036 \times 0,414}{0,036 \times 0,414 + 0,011 \times 0,364 + 0,016 \times 0,222}$$

$$= \frac{0,01490}{0,02246}$$

$$= 0,663 \text{ of } 66,3\%.$$

5. Noem de verwachtingswaarde van het aantal deeltjes per cm^2 μ. Dan is de verwachtingswaarde van het aantal deeltjes per gezichtsveld van 3,03 mm^2 gelijk aan $\frac{3,03}{100}\mu$. Aangezien de aantallen deeltjes per gezichtsveld verdeeld zijn volgens een Poisson-verdeling, zal de verdeling van de som van de aantallen deeltjes, geteld in elf gezichtsvelden, ook verdeeld zijn volgens een Poisson-verdeling en wel met gemiddelde $\frac{11 \times 3,03}{100}\mu = \frac{33,33}{100}\mu$. De gevonden som 117+103+...+108 = = 1100 is dus de beste schatting voor $\frac{33,3}{100}\mu$. De beste schatting voor μ is dus $1100 \times \frac{100}{33,33} = 3300$.

De beste schatting voor de variantie van de som van de elf waarnemingen is, daar deze som zelf volgens een

Poisson-verdeling verdeeld is, gelijk aan deze som. Deze beste schatting is dus 1100. De beste schatting voor de variantie van het gemiddelde μ, dat gelijk is aan $\frac{100}{33,33} = 3,00$ maal bedoelde som, is dus $(3,00)^2 \times 1100 = 9900$.

Daar bij een Poisson-verdeling gemiddelde en variantie gelijk moeten zijn, kan het al dan niet volgens een Poisson-verdeling verdeeld zijn gecontroleerd worden door het gemiddelde per gezichtsveld te vergelijken met de variantie, berekend uit de kwadraatsom der afwijkingen van het gemiddelde.

In dit geval is het gemiddelde 100 en de uit de elf individuele waarden berekende variantie 97,8. Deze laatste berekening werd niet gevraagd.

6. a. De prijsindex wordt berekend met behulp van de formule van LASPEYRES:

$$P_{01} = \frac{10 \times 0,60 + 5 \times 0,50}{10 \times 0,50 + 5 \times 0,40} \times 100 = 121,4.$$

In verband met de uitbreiding van het goederenpakket in periode 1 is het noodzakelijk op een nieuw indexcijfer over te gaan met die periode als basis, en dit indexcijfer vervolgens te koppelen aan het voorafgaande:

$$P_{12} = \frac{10 \times 0,55 + 5 \times 0,55 + 5 \times 0,75}{10 \times 0,60 + 5 \times 0,50 + 5 \times 1,00} \times 100 = 88,9.$$

Bijgevolg is: $P_{02} = P_{01} \times P_{12} : 100 = 107,9$.

b. Voor de berekening van de volume-index kan de formule van PAASCHE met als basisperiode 0 niet gebruikt worden in verband met de uitbreiding van het goederenpakket in periode 1. Wanneer men veronderstelt dat de prijsontwikkeling van goed C van periode 0 tot 1 dezelfde is geweest als aangegeven

wordt door het prijsindexcijfer van LASPEYRES, kan de volume-index berekend worden door het waardecijfer uit te drukken als een indexcijfer, en dit vervolgens te delen door de bovengevonden prijsindex.

Periode	Uitgaven per hoofd		Prijsindex	Volume-index
	in gld	index-cijfer		
0	7,00	100	100	100
1	13,50	192,9	121,4	158,8
2	12,20	174,3	107,9	161,5

Opmerking. Men kan ook een enigszins andere methode toepassen, nl. een prijsindexcijfer volgens FISHER berekenen voor de perioden 0 en 1, en voor de perioden 1 en 2, en vervolgens deze beide indexcijfers koppelen. De volume-index wordt dan op dezelfde wijze als boven gevonden door de index van de uitgaven te delen door de prijsindex. Deze berekening is uiteraard iets bewerkelijker. Men vindt:

Periode	Prijsindices			Volume-index
	LASPEYRES	PAASCHE	FISHER	(waarde: FISHER-index)
0	100	100	100	100
1	121,4	121,4	121,4	158,8
2	107,9	105,8	106,9	163,0

7. Volgens de opgave zijn de fouten, die gemaakt worden door het afkappen van de bedragen kleiner dan een gulden, voor iedere gemeente stochastische grootheden $\underline{x}$, met een multinomiale verdeling.

$$P(\underline{x}=a) = \frac{1}{100} . \quad a = 0{,}00;\ 0{,}01;\ 0{,}02;\ \ldots;\ 0{,}99.$$

Hieruit volgt:

$$E(\underline{x}) = \frac{1}{100} \sum_{i=0}^{i=99} \frac{i}{100} = \frac{1}{100} \cdot \frac{99 \times 100}{2.100} = 0{,}495.$$

$$\text{var}(\underline{x}) = E\left[\underline{x}-E(x)\right]^2 = Ex^2 - \left[E(\underline{x})\right]^2$$

$$= \frac{1}{100} \sum_{i=0}^{i=99} \frac{i^2}{100^2} - 0{,}495^2 =$$

$$= \frac{1}{100} \frac{99(99+1)(2.99+1)}{6.100^2} - 0{,}495^2$$

$$= \frac{1}{100^2} \times \frac{99 \times 199}{6} - \frac{1}{100^2} \times \frac{99^2}{2^2} =$$

$$= \frac{99}{100^2.2^2} \times \frac{398-297}{3}$$

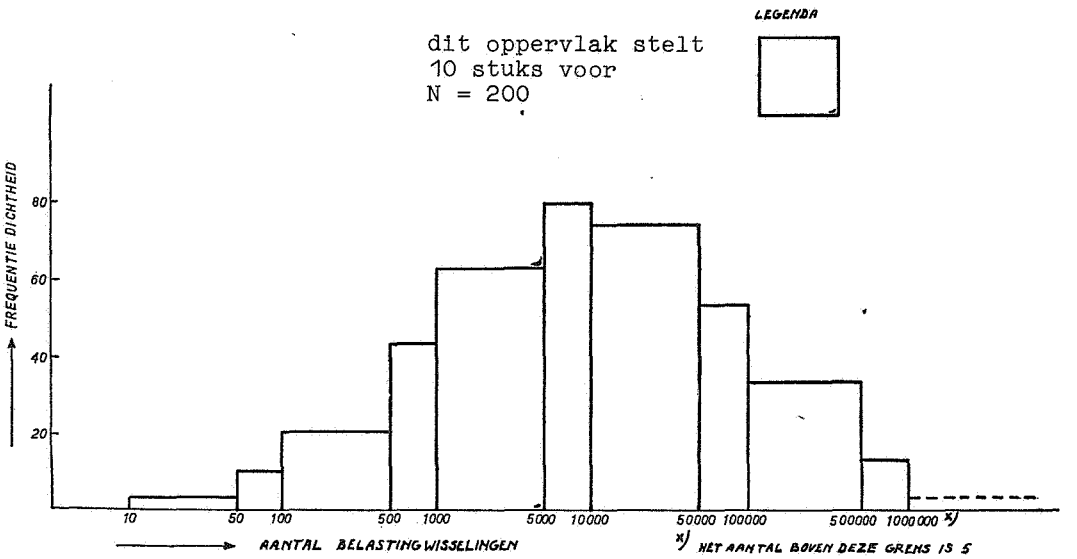
$$= \frac{99 \times 101}{200^2 \times 3} = \frac{3333}{200^2} = 0{,}083325.$$

Volgens de centrale limietstelling is de som van n stochastische variabelen met eenzelfde verdeling indien n groot is, praktisch normaal verdeeld, met als gemiddelde n maal het gemiddelde en als variantie n maal de variantie der gemeenschappelijke verdeling. De som der 793 afkappingsfouten is dus bij benadering normaal verdeeld met gemiddelde 793 x 0,495 = 392,535 en variantie $\frac{793 \times 3333}{200^2} = \frac{2643069}{200^2}$

De standaardafwijking is dus $\frac{\sqrt{2643069}}{200} = 8{,}129$. De grenzen voor de totale opbrengst van de vermakelijkheidsbelasting zijn dus (26.074.281 + 392,535) + $\pm$ 1,96 x 8,129, dus *f* 26.074.657,60 en *f* 26.074.689,47.

Een zeer goede benadering voor de variantie kon worden verkregen door de verdeling van de weggelaten bijdragen als een (continue) rechthoekige verdeling met een spreidingsbreedte 1(gulden) te beschouwen. De variantie van een continue rechthoekige verdeling bedraagt $\frac{1}{12}$ maal de spreidingsbreedte, in dit geval dus $\frac{1}{12}$ x 0,083333. De uitkomst volgens deze wijze van berekenen komt tot op de cent overeen met de bovengegeven exacte uitkomst.

8. Zie bijgaande figuur.



9. a. Voor iedere frequentieverdeling geldt per definitie dat het oppervlak tussen de mediaan (het tweede kwartiel) en het derde kwartiel 0,25 bedraagt.
   b. Bij de beantwoording van a is geen der vermelde gegevens nodig.
   Wel is nodig dat de verdeling continu is, anders zijn mediaan en kwartielen niet noodzakelijkerwijs gedefinieerd.

10. Een standaardoplossing in absolute zin is voor dit vraagstuk uiteraard niet te geven. Als mogelijke doelmatige oplossing volgt hieronder de tekst van een brief aan de personeelsleden (I) met vragenformulier (II) alsmede nog enkele opmerkingen over de organisatie (III).

I: brief

Personeelsvereniging

DE KAAYMANNEN

datum

L.S.

Onze personeelsvereniging wil een groots Sinterklaasfeest organiseren, compleet met surprises en al, op ... december a.s. voor u en uw gezinsleden van 4 jaar en ouder. Wij denken aan afzonderlijke bijeenkomsten voor kinderen van 4 t/m 11 jaar, voor jongeren van 12 t/m 17 jaar en voor de volwassenen. Daarvoor moeten wij weten op hoeveel deelnemers wij voor elk van deze drie bijeenkomsten kunnen rekenen. Uw medewerking is daarvoor nodig. Deze kunt u ons geven door op bijgaande vragen kort, duidelijk en leesbaar het antwoord in te vullen en dit formulier zo vlug mogelijk maar in

elk geval vóór ... november in te sturen aan de secretaris van "De Kaaymannen", de heer ....., afd.....
Bij voorbaat dank.

(voorzitter)

II: vragenlijst

in hokje dat van toepassing is antwoord invullen of kruisje zetten.

Vraag	Antwoord
1 Geboortedatum van uzelf	
2 Bent u gehuwd of niet?	gehuwd ☐ niet gehuwd ☐
3a Hoeveel kinderen heeft u van 4 t/m 11 jaar?	geen ☐ wel, nl. (aantal invullen) ☐
b Hoeveel daarvan zullen deelnemen aan het feest?	geen ☐ wel, nl. (aantal invullen) ☐
4a Hoeveel kinderen heeft u van 12 t/m 17 jaar?	geen ☐ wel, nl. (aantal invullen) ☐
b Hoeveel zullen daarvan deelnemen aan het feest?	geen ☐ wel, nl. (aantal invullen) ☐
5a Denkt u zelf deel te nemen aan het feest?	ja ☐ neen ☐
b Hoeveel gezinsleden van 18 jaar of ouder zult u meenemen, uzelf niet meegerekend?	geen ☐ wel, nl. (aantal invullen) ☐
c Hoeveel daarvan zijn mannen en hoeveel vrouwen?	aantal mannen ☐ aantal vrouwen ☐
6a Wat is uw naam?	
b Uw adres?	

III: organisatie

1. Met het oog op te verwachten non-response (bij schriftelijke enquêtes soms 85% of meer!) dient rekening gehouden te worden met de noodzaak tot het zenden van een of twee rappels na bijv. 14 dagen en 4 weken. Deze kunnen gelijkluidend zijn aan het oorspronkelijke verzoek. Alleen met opdruk "herhaald verzoek". Een speciale tekst is ook mogelijk.
2. Zo nodig zullen zij die na het tweede rappel niet reageerden nog persoonlijk moeten worden bewerkt (of een steekproef uit deze groep). Men kan dit echter ook nalaten (eventueel ook het tweede rappel) en dan aannemen dat de niet-reagerende groepen niet zullen deelnemen. Dit geeft een mogelijke vertekening waarvan de maximale omvang echter kan worden berekend.
3. Vraag 6c kan ook gesplitst worden naar
   echtgeno(o)t(e)
   kinderen en/of leeftijdsgroepen
   overigen.
4. Naar het geslacht van het personeelslid zelf is niet gevraagd. Aangenomen kan worden dat men dit bij een personeel van 150 leden wel weet.

Achtste examen, oktober 1960

1. De verdeling van de waarnemingen is kennelijk noch normaal, noch volgens Poisson. De twee-steekproeventoets van Wilcoxon is in dit geval de aangewezen methode.
   Ter berekening van de toetsingsgrootheid W stellen wij de volgende tabel op.

Rangnummer	Steekproef 1	Steekproef 2	Bijdrage tot W
0		0	0
1	1		
2		3	2
3	5		
4		6,9,12,16,18	20
5	24		
6		26	6
7	31		
8			
9	53		
10		58	10
11	61		
12			
13	62		
14		65,68	28
15	90		
16		96	16
17	149		
18		176	18
19	184		
20			
21	200		
22		239	22
23	267		
24			
25	270		
26			
27	417		
28			
29	430		
30			
Totaal	m = 15	n = 14	W = 122

De verwachting van W is $mn = 15 \times 14 = 210$.

De standaardafwijking is $\sqrt{1/3\ mn\ (m + n + 1)} =$
$= \sqrt{2100} = 45,8$.

De linkséénzijdige overschrijdingskans is bij benadering gelijk aan de linkséénzijdige overschrijdingskans van het punt

$$\frac{122 - 210 + 1}{45,8} = \frac{87}{45,8} = -1,90 .$$

Deze overschrijdingskans bij de normale verdeling met gemiddelde 0 en spreiding 1 is 0,03. De getoetste hypo-

these, dat de waarnemingen alle uit dezelfde verdeling komen, moet dus worden verworpen.

2. De tabel zou als volgt kunnen worden ingericht

eenheden: 1000 woningen

	1956	1957	1958	1959
1. Woningvoorraad op 1 jan.				
Vermeerdering door: a. nieuwbouw				
b. verbouwing en wijziging van bestemming				
2. Totaal (a+b)				
Vermindering door: a. afbraak, vernietiging				
b. onbewoonbaarverklaring				
c. verbouwing, enz.				
3. Totaal (a+b+c)				
4. Woningvoorraad op 31 dec. (1+2-3)				

3. Stel $\underline{b}$ = gewicht van een biscuit
$\underline{v}$ = gewicht van een verpakking
$\underline{r}$ = gewicht van een rol
$\Sigma\underline{b}$ = gewicht van 40 biscuits in één rol

De volgende betrekking geldt nu:

$$\underline{r} = \Sigma\underline{b} + \underline{v}$$

Gegeven is:

$$\mu_b = 3 \text{ g}$$
$$\mu_v = 15 \text{ g}$$

dus

$$\mu_r = 40 \times 3 + 15 = 135 \text{ g}$$

Verder is gegeven dat

$$\sigma_b = 0{,}2 \text{ g, dus } \sigma_b^2 = 0{,}04 \text{ g}^2$$
$$\sigma_v = 0{,}5 \text{ g, dus } \sigma_v^2 = 0{,}25 \text{ g}^2.$$

Hieruit volgt, onder aanname dat de gewichten van de biscuits en die van de verpakking alle onderling onafhankelijk zijn:

$$\sigma_r^2 = 40\,\sigma_b^2 + \sigma_v^2$$
$$= 40 \times 0{,}04 + 0{,}25$$
$$= 1{,}85 \text{ g}^2$$
$$\sigma_r = \sqrt{1{,}85} = 1{,}36 \text{ g}$$

a. Weliswaar heeft het gewicht van een biscuit geen normale verdeling, maar op grond van de centrale limietstelling heeft de som van 40 onafhankelijk gekozen gewichten wel bij benadering een normale verdeling. Ook het gewicht van een verpakking heeft een normale verdeling, zodat ook het gewicht van een rol voor praktische doeleinden als normaal verdeeld mag worden beschouwd.
Bij een gewicht van 131 gram vinden we voor $u = \frac{131 - \mu_r}{\sigma_r}$

$$u = \frac{131 - 135}{1{,}36} = -2{,}94 \ .$$

Hierbij behoort een eenzijdige overschrijdingskans van 0,00164. De kans dat een rol minder dan 131 gram weegt, bedraagt dus 0,16%.

b. Volgens de tabel (die men kan vinden als tabel 27 in Biometrika Tables for Statisticians, Vol.I) bedraagt de verwachting van de spreidingsbreedte bij een steekproef van 5 stuks:

$$E(\underline{W}) = 2{,}326\sigma.$$

Daar $\sigma = 1{,}36$ wordt dus voor de gevraagde verwachting

gevonden

$$E(\underline{W}) = 2,326 \times 1,36 = 3,16 \text{ g.}$$

$$\Sigma\ \underline{b} = \underline{r} - \underline{v}.$$

c. Volgens een bekende formule geldt nu:

$$40\,\sigma_b^2 = \sigma_r^2 + \sigma_v^2 - 2\,\sigma_r\,\sigma_v\,\rho_{rv}$$

of

$$\rho_{rv} = \frac{\sigma_r^2 + \sigma_v^2 - 40\,\sigma_b^2}{2\,\sigma_r\,\sigma_v}.$$

Substitueert men hierin de gegeven numerieke waarden, dan verkrijgt men:

$$\rho_{rv} = \frac{1,85 + 0,25 - 1,60}{2 \times 1,36 \times 0,5} = 0,368.$$

Dit is de gevraagde correlatie-coëfficient.
De correlatiecoëfficient kan ook rechtstreeks als volgt uit de definitie worden berekend:

$$\rho_{rv} = \frac{cor(r,v)}{\sigma_r\,\sigma_v} = \frac{cov(\Sigma b+v, v)}{\sigma_r\,\sigma_v}.$$

Daar $\Sigma\underline{b}$ en $\underline{v}$ onafhankelijk verdeeld zijn gaat dit over in:

$$\rho_{rv} = \frac{\sigma_v^2}{\sigma_r\,\sigma_v} = \frac{\sigma_v}{\sigma_r} = \frac{0,50}{1,36} = 0,368.$$

4. De uit de waarnemingen van de beide analisten berekende geschatte varianties bedragen respectievelijk 0,743 en 4,195. Passen we hierop de F-toets toe dan blijkt het quotient $\frac{4,195}{0,743} = 5,65$ te liggen tussen de 0,95 en 0,975 fractielen van de F-verdeling met 4 en 5 vrijheidsgraden die respectievelijk gelijk zijn aan 5,19 en 7.39. Daar hier tweezijdig getoetst moet worden, is de kritieke waarde 7,39 en verwerpen we dus de hypothese dat de beide analisten even nauwkeurig werken <u>niet</u>.

5. Stel dat in de in beschouwing genomen bevolkingsgroep N ergelijk belaste twee-kindergezinnen aanwezig zijn en dat de kans dat een kind in een erfelijk belast gezin de ziekte vertoont, p is. De verwachtingswaarde van het aantal gezinnen met nul zieke kinderen is dan $N(1-p)^2$, die van het aantal gezinnen met één kind ziek $2\,N\,p(1-p)$ en die van het aantal gezinnen met twee zieke kinderen $N\,p^2$. Worden nu de verwachtingswaarden gelijk gesteld aan de gevonden aantallen, dan worden twee vergelijkingen verkregen, waaruit zowel p als ook N opgelost kunnen worden, waarmee dus een schatting van p (en van N) verkregen wordt:

$$2\,N\,p\,(1-p) = 108$$
$$N\,p^2 = 27$$

Dus $$\frac{2\,N\,p(1-p)}{N\,p^2} = \frac{2\,(1-p)}{p} = 4.$$

$p = \frac{1}{3}$

$N = 243.$

Uit deze uitkomst kan dus ook afgeleid worden, dat er naar schatting $243 \times (1-\frac{1}{3})^2 = 108$ erfelijk belaste twee-kindergezinnen waren, die niet als zodanig herkend werden, daar geen van beide kinderen ziek waren.

6. Volgens een bekende formule is:

$$\Sigma(x-\bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n} .$$

De uitdrukking in het lin erlid zal altijd positief moeten zijn (of nul). Voor het rechterlid vindt men echter

$$1030 - \frac{(414)^2}{100} = -\,684.$$

Deze uitkomst is negatief. Derhalve kan de opgave niet juist zijn.

7. a) $\underline{x} + \underline{y} + \underline{z} = 4$, dus $\underline{w} = 2 \times 4 - 3\underline{y} = 8 - 3\,\underline{y}$.
De verdeling van $\underline{y}$ is binomiaal met kans
$p = 1/3$ en $n = 4$, dus $P(\underline{y} = y) = \binom{4}{y}(1/3)^y(2/3)^{4-y}$.
Deze verdeling is dus:

$\underline{y}$	0	1	2	3	4
kans	16/81	32/81	24/81	8/81	1/81

De verdeling van $\underline{w} = 8 - 3\,\underline{y}$ is dus:

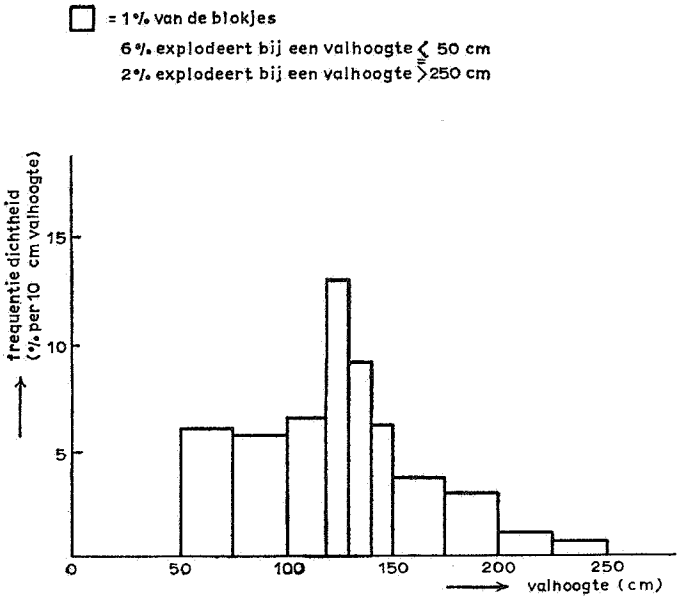
$\underline{w}$	8	5	2	-1	-4
kans	16/81	32/81	24/81	8/81	1/81

b) De variantie van $\underline{y}$ is: $n\,p\,(1 - p) = 4.\ 1/3.\ 2/3 = 8/9$
Dus var $(\underline{w}) = 3^2 \times 8/9 = 8$.

8. Een schatting van de fractie der blokjes, die juist exploderen in een bepaald interval, wordt gevonden door de fractie der blokjes, die bij de valhoogte gelijk aan de bovengrens van het interval exploderen te verminderen met de fractie der blokjes, die bij de ondergrens exploderen. Immers het feit, dat een blokje explodeert bij een experiment met valhoogte a bewijst slechts, dat er een valhoogte bestaat, die hoogstens gelijk is aan a, waarbij het blokje juist zou exploderen.
Uit de gegevens kan op deze wijze de volgende tabel afgeleid worden:

Valhoogte	fractie der blokjes, die juist exploderen in het interval in kolom 1 aangegeven
0 - 50 cm	$\frac{3}{50} = 0,06$
50 - 75 cm	$\frac{21}{100} - \frac{3}{50} = 0,15$
75 - 100 cm	$\frac{35}{100} - \frac{21}{100} = 0,14$
100 - 120 cm	$\frac{48}{100} - \frac{35}{100} = 0,13$
120 - 130 cm	$\frac{61}{100} - \frac{48}{100} = 0,13$
130 - 140 cm	$\frac{70}{100} - \frac{61}{100} = 0,09$
140 - 150 cm	$\frac{76}{100} - \frac{70}{100} = 0,06$
150 - 175 cm	$\frac{85}{100} - \frac{76}{100} = 0,09$
175 - 200 cm	$\frac{92}{100} - \frac{85}{100} = 0,07$
200 - 225 cm	$\frac{48}{50} - \frac{92}{100} = 0,04$
225 - 250 cm	$\frac{49}{50} - \frac{48}{50} = 0,02$
> 250 cm	$1 - \frac{49}{50} = 0,02$

Bij het uitzetten van deze uitkomsten in een histogram moet er rekening mee gehouden worden, dat de klassebreedten ongelijk zijn, zodat de hoogten der diverse rechthoeken niet gelijk aan de gevonden fracties genomen mogen worden. De hoogten moeten evenredig zijn aan de fracties gedeeld door de klasse-breedte.
Het gevraagde histogram is getekend in onderstaande figuur.
Voor het interval 0-50 cm zou men weliswaar een staaf kunnen tekenen (ter hoogte van 6%/5=1,2%), doch dit is niet realistisch, omdat men kan aannemen, dat beneden een bepaalde valhoogte geen explosie tot stand komt.



Frequentieverdeling van de valhoogte waarbij het onderzochte explosief juist explodeert

a-b) In de grafische voorstelling wordt de oorspronkelijke reeks zowel als de reeks der voortschrijdende vier-kwartaalsgemiddelden afgebeeld. Voor een dergelijke korte reeks, waarbij de voortschrijdende gemiddelden slechts weinig fluctueren, is het niet mogelijk nauwkeurig vast te stellen of de amplitude der seizoenbewegingen additief is of multiplicatief (d.w.z. een vast percentage is van de trendconjunctuurcomponent). Een multiplicatief patroon is meer aannemelijk.

c) De reeks der voortschrijdende 4-kwartaalsgemiddelden wordt verkregen door telkens twee opeenvolgende vier-kwartaalsommen te middelen en door vier te delen. Men vindt het volgende:

Vier-kwartaalsgemiddelden (voortschr.)	1956	1957	1958	1959
Februari		121.1	116.8	118.0
Mei		122.1	115.2	120.4
Augustus	116.9	121.7	114.4	
November	119.5	119.3	115.6	

Afwijkingen van de vier-kwartaals-gemiddelden	1956	1957	1958	1959
Februari		-15.1	-16.8	-19.0
Mei		+12.9	+ 6.8	+12.6
Augustus	+10.1	+10.3	+10.6	
November	- 5.5	- 2.3	- 4.6	

Idem, in % van het voortschrijdende gemiddelde	1956	1957	1958	1959	gem.
Februari	-	-12.5	-14.4	-16.1	-14.3
Mei	-	+10.6	+ 5.9	+10.5	+ 9.0
Augustus	+8.6	+ 8.5	+ 9.3	-	+ 8.8
November	-4.6	- 1.9	- 4.0	-	- 3.5

De seizoencoefficienten zijn bijgevolg	1956	1957	1958	1959
	85.7	109.0	108.8	96.5

d) Het voor seizoenbeweging gecorrigeerde indexcijfer voor het eerste kwartaal 1960 is bijgevolg:

$\frac{108}{85,7} = 126.$

Het examen Statistisch Analist 1961

Opgave

A 1

Aan 13 typistes werd gevraagd om in 5 minuten zoveel mogelijk te typen van een bepaald tijdschriftartikel. Aan het einde van deze 5 minuten werd het aantal aanslagen geteld en eveneens het aantal foutieve aanslagen.

De resultaten waren als volgt:

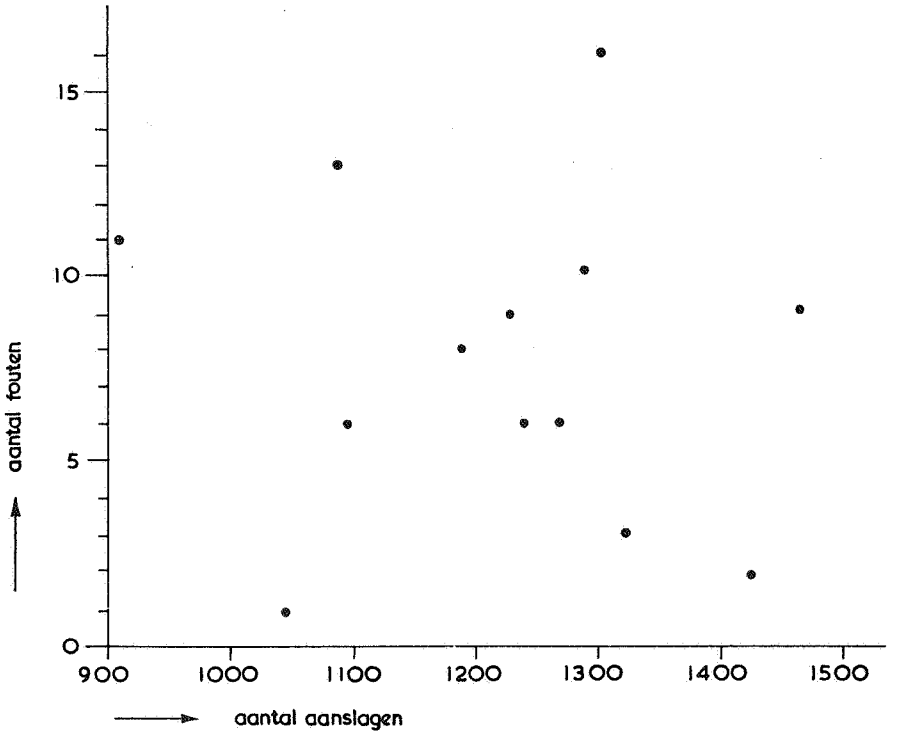
Aantal aanslagen:	*Aantal fouten:*
1425	2
1188	8
1090	13
1238	6
1307	16
911	11
1293	10
1271	6
1320	3
1043	1
1464	9
1095	6
1226	9

Vraag: Toets of er een verband bestaat tussen het aantal aanslagen en het aantal fouten. Motiveer Uw keuze van de gevolgde methode.

Standaardantwoord

In een geval als dit is het verstandig te beginnen met het tekenen van een spreidingsdiagram. Dit is in bijgaande figuur gedaan. Hieruit blijkt reeds dat er

waarschijnlijk geen verband bestaat tussen het aantal aanslagen en het aantal fouten.



Er is geen aanleiding om aan te nemen dat het aantal fouten normaal verdeeld is.

Voor de toetsing is daarom een rangcorrelatieberekening de beste methode. De rangcorrelatiecoëfficiënt van Spearman is - 0,094.

De variantie is $\frac{1}{12}$, dus de standaardafwijking bedraagt $\frac{1}{\sqrt{12}} = 0{,}29$.

Het gevonden resultaat is dus verre van significant. Bij deze berekening is de correctie voor gelijke waarnemingen niet toegepast. Dit maakt overigens zeer weinig verschil.

De rangcorrelatiecoëfficiënt volgens Kendall is - 0,077, de S-score is -6 met als standaardafwijking $\sqrt{\frac{13.12.31}{18}} = 16{,}4$. Ook hier is de correctie voorgelijke waarnemingen verwaarloosd.

De conclusie luidt dus in beide gevallen dat de hypothese, dat er géén verband tussen de beide waargenomen grootheden bestaat, niet kan worden verworpen.

Berekening van de normale correlatiecoëfficiënt leidt tot de zelfde conclusie. Deze correlatiecoëfficiënt is namelijk —0,107. Gezien de aard van de waarnemingen moet deze methode echter minder verantwoord worden geacht.

A 2

Gegeven een populatie bestaande uit zes kegels. De afmetingen van deze kegels zijn:

Nr.	Diameter van het grondvlak x in cm.	Hoogte y in cm.
1	3	6
2	5	4
3	7	2
4	9	2
5	11	4
6	13	6

Vraag: a. Bereken de (populatie)-variantie van x en die van y.
b. Bereken de (gewone) correlatiecoëfficiënt ϱ tussen x en y in de populatie.

Door met een (zuivere) dobbelsteen te werpen en telkens de kegel te kiezen wiens nummer overeenkomt met het resultaat van de worp, worden de grootheden x en y tot stochastische veranderlijken $\underline{x}$ en $\underline{y}$.
(De trekking der kegels geschiedt met teruglegging).

Vraag: c. Zijn $\underline{x}$ en $\underline{y}$ stochastisch onafhankelijk? Licht Uw antwoord toe en betrek in Uw beschouwing ook Uw uitkomst van vraag b.

Standaardantwoord

Vraag a.

$\Sigma x = 48 \qquad \bar{x} = 8$
$\Sigma y = 24 \qquad \bar{y} = 4$

Nr. kegel	$x - \bar{x}$	$(x - \bar{x})^2$	$(y - \bar{y})$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	—5	25	2	4	—10
2	—3	9	0	0	0
3	—1	1	—2	4	2
4	1	1	—2	4	—2
5	3	9	0	0	0
6	5	25	2	4	10
Som	0	70	0	16	0

Aangezien naar de populatie-variantie van x en y gevraagd wordt, moeten $\Sigma (x - \bar{x})^2$ en $\Sigma (y - \bar{y})^2$ gedeeld worden door $n = 6$, het aantal individuen in de populatie, en niet door $n - 1 = 5$. Derhalve is:

$$\text{var } (x) = \tfrac{70}{6} = 11{,}67.$$
$$\text{var } (y) = \tfrac{16}{6} = 2{,}67.$$

Vraag b.

Daar de covariantie van x en y gelijk is aan nul is correlatie-coëfficiënt ϱ eveneens gelijk aan nul ($\varrho = 0$).

Vraag c.

De (onvoorwaardelijke) waarschijnlijkheid voor het optreden van elk der drie mogelijke waarden van $\underline{y}$ is 1/3. De voorwaardelijke waarschijnlijkheid, gegeven de waarde van $\underline{x} = x$ is voor één waarde van y gelijk aan 1 en voor alle overige gelijk aan 0. Daar de onvoorwaardelijke en de voorwaardelijke waarschijnlijkheden ongelijk zijn, zijn $\underline{x}$ en $\underline{y}$ stochastisch afhankelijk.

Daar in het stochastisch model de kans op het trekken van ieder der zes kegels gelijk is aan de relatieve frequentie in de onder a) en b) beschouwde populatie is in dit model de correlatie-coëfficiënt eveneens nul, hetgeen betekent dat er geen *lineair* verband bestaat tussen $\underline{x}$ en $\underline{y}$. Dit is niet in tegenspraak met het bestaan van stochastische afhankelijkheid; een grafische voorstelling toont dat de stochastische afhankelijkheid hier berust op een niet-lineair verband.

A 3

Sigmanië is verdeeld in 100 provincies met zeer uiteenlopende dichtheid van bevolking (aantal inwoners per vierkante km).

Het Statistisch Zakboek van Sigmanië vermeldt voor 1960 daaromtrent het volgende:

Aantal inwoners per km².	Aantal provincies
2 — 5	5
5 — 20	37
20 — 50	35
50 — 200	22
200 — 500	1
Totaal	100

Vraag: a. Laat grafisch zien, dat de frequentieverdeling bij benadering logaritmisch-normaal is (log 2 = 0,3).

b. Bereken een aangepaste log-normale verdeling voor de volgende klasse-indeling:

	< 1	inwoner	per	km².
1 —	4	inwoners	„	„
4 —	16	„	„	„
16 —	64	„	„	„
64 —	256	„	„	„
256 —	1024	„	„	„
	> 1024	„	„	„

Standaardantwoord

Stel X = aantal inwoners per km²
$x = 2 \log X$
f = frequentie

Klassegrenzen $x_1 — x_2$	Klasse-midden x	f	$\sum_{-\infty}^{x_2} f$	fx	fx^2
0,6 — 1,4	1	5	5	5	5
1,4 — 2,6	2	37	42	74	148
2,6 — 3,4	3	35	77	105	315
3,4 — 4,6	4	22	99	88	352
4,6 — 5,4	5	1	100	5	25
		100		277	845
				$277^2 / 100$ =	767,29
					77,71

$\bar{x} = \frac{277}{100} = 2{,}77$ = schatting van μ

$s^2 = \frac{77{,}71}{99} = 0{,}785$

$s = 0{,}886$ = schatting van σ.

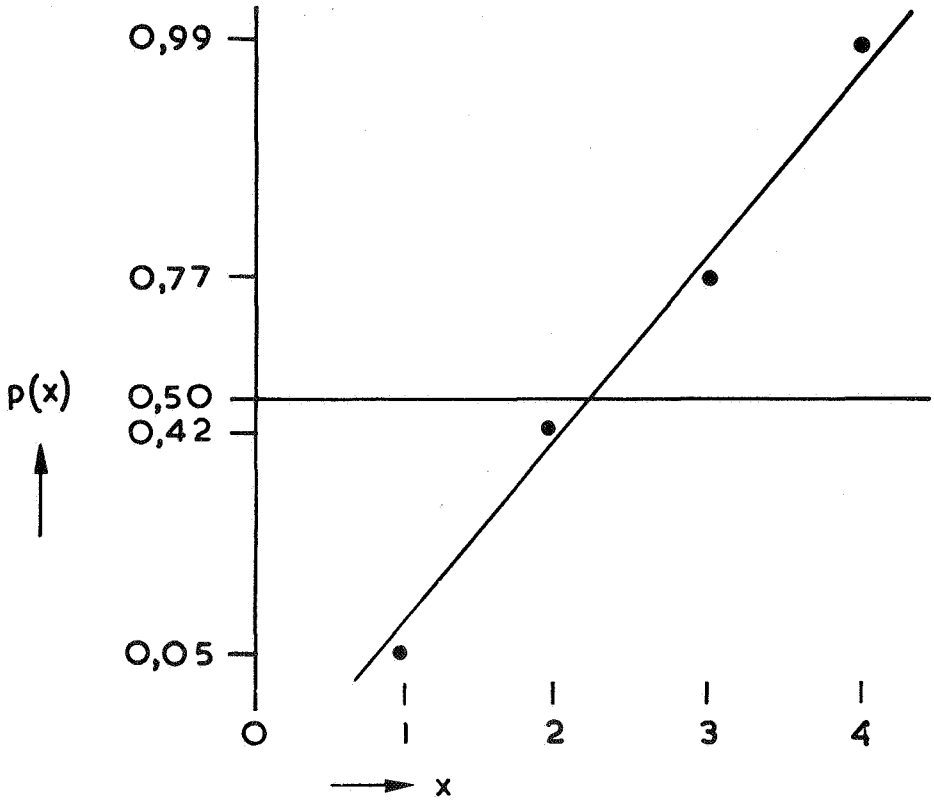
Vraag a.

Onderstaande grafiek op waarschijnlijkheidspapier laat zien dat de uitgezette waarden van de cumulatieve frequentieverdeling van x vrijwel op een rechte lijn liggen. De verdeling van x is dus praktisch normaal en daarmee de verdeling van X log-normaal.

Vraag b.

Aanpassing van een log-normale verdeling aan X is identiek met aanpassing van een normale verdeling aan x.

-2-



Stel voor de normale verdeling

$$\mu = \bar{x} = 2{,}77$$
$$\sigma = s = 0{,}886$$

Klassegrenzen $x_1 - x_2$	$u = \frac{x - \mu}{\sigma}$	$F(u)$	$P(x_1 - x_2)$	f
< 0			0,001	0,1
0 — 1,2	— 3,13	0,001	0,037	3,7
1,2 — 2,4	— 1,77	0,038	0,299	29,9
2,4 — 3,6	— 0,42	0,337	0,489	48,9
3,6 — 4,8	0,94	0,826	0,163	16,3
4,8 — 6,0	2,29	0,989	0,011	1,1
$> 6{,}0$	3,65	1,000	0,000	0,0

De laatste kolom geeft de gevraagde aangepaste verdeling bij de gegeven klasse-indeling.

A 4

Voor het rekenkundig gemiddelde ($\bar{x}$) en de standaardafwijking (s) van een groep van 50 waarnemingen heeft men gevonden:

$$\bar{x} = 6{,}81 \qquad\qquad s = 3{,}10$$

Later bleek, dat bij de berekening één der waarnemingen foutief was overgenomen. Er moest staan 2,7 in plaats van 7,2.

Vraag: Bereken de juiste waarden van het rekenkundig gemiddelde en van de standaardafwijking.

Standaardantwoord

De correctie voor het rekenkundig gemiddelde wordt als volgt aangebracht:

gevonden waarde van $\bar{x}$	=	6,81
correctie: $\frac{2{,}7 - 7{,}2}{50}$	=	− 0,09
gecorrigeerde waarde van $\bar{x}$	=	6,72

Voor de standaardafwijking is de gang der berekening:

netto quadraatsom:	$49\ s^2$	470,89
correctie bruto quadraatsom:		
bijtellen $2{,}7^2$		
aftrekken $7{,}2^2$		
verschil 44,55		− 44,55
		426,34
Herstellen van de correctieterm voor het gemiddelde:		
bijtellen $50\ (6{,}81)^2$		
aftrekken $50\ (6{,}72)^2$		
verschil + 60,885		+ 60,885
		487,225
gecorrigeerde standaardafwijking,	s =	$\sqrt{\frac{487{,}225}{49}}$
	s =	$\sqrt{9{,}9434}$
	s =	3,15

A 5

Een fabriek heeft twee automatische draaibanken, die ingesteld kunnen worden op de vervaardiging van assen van gelijke diameter. De asdiameters zijn blijkens de ervaring normaal verdeeld en de diameters van opeenvolgend, op één draaibank, vervaardigde assen vertonen geen verloop of periodiciteit. Gedurende de productie worden elk uur drie assen aselect gekozen, waarvan de diameters worden gemeten; het gemiddelde van de drie waarnemingen wordt gerapporteerd.

Na vijf uur productie door beide draaibanken, die zo goed mogelijk op dezelfde asdiameter zijn ingesteld, wenst men een betrouwbaarheidsinterval te bepalen voor het eventuele verschil in gemiddelde asdiameter tussen de beide draaibanken. Van draaibank A waren alle vijftien afzonderlijke waarnemingen nog beschikbaar, van draaibank B slechts de vijf uurgemiddelden.

Deze worden hieronder gegeven.

Draaibank A (afzonderlijke waarnemingen) in cm.				Draaibank B. (gemiddelden van de 3 per uur verrichte waarnemingen) in cm.
7,41	7,34	7,24	7,36	7,307
7,33	7,27	7,28	7,46	7,287
7,34	7,25	7,29	7,08	7,280
7,24	7,32	7,38		7,403
				7,283

Vraag: a. Toets of de varianties van de diameters der assen afkomstig van A significant verschillen van die afkomstig van B, bij een 5% onbetrouwbaarheidsdrempel.

Bij vraag b moet men voor de berekening van de standaardafwijking gebruik maken van beide steekproeven. Ongeacht het antwoord op vraag a moet worden gehandeld alsof de populatie-varianties van A en B *niet* verschillen. Voorts moet van de *t*-verdeling gebruik gemaakt worden.

Vraag: b. Bereken een betrouwbaarheidsinterval voor het verschil tussen de gemiddelde asdiameters bij A en bij B, met 5% onbetrouwbaarheid. Leid uit dit betrouwbaarheidsinterval af of de gevonden gemiddelden bij A en B significant verschillen bij een 5% onbetrouwbaarheidsdrempel.

Standaardantwoord

x_1 = (individuele) uitkomsten van draaibank A
x_2 = (gemiddelde) uitkomsten van draaibank B
} normaal verdeeld

$\bar{x}_1 = \Sigma x_1/15 = 7{,}306$
$\bar{x}_2 = \Sigma x_2/5 = 7{,}3120$
$s_1^2 = \Sigma(x_1 - \bar{x}_1)^2/14 = 0{,}00794$
$s_2^2 = \Sigma(x_2 - \bar{x}_2)^2/4 = 0{,}002699$

Vraag a.

Als

σ_1^2 = variantie van de verdeling van machine A
σ_2^2 = variantie van de verdeling van machine B (individuele cijfers)

dan is

$E(s_1^2) = \sigma_1^2$ en $E(s_2^2) = \sigma_2^2/3$ of $E(3s_2^2) = \sigma_2^2$
$H_0 : \sigma_1^2 = \sigma_2^2$

$$F = \frac{3s_2^2}{s_1^2} = \frac{0{,}00801}{0{,}00794} = 1{,}01 \text{ met 4 en 14 vrijheidsgraden.}$$

Dit is kennelijk niet significant, zodat geen aanwijzingen aanwezig zijn voor verschil in variantie.

Vraag b.

Om het verschil $\bar{x}_2 - \bar{x}_1$ te toetsen met behulp van de t-toets, moet eerst de beste schatting s^2 van $\sigma^2 = \sigma_1^2 = \sigma_2^2$ worden bepaald.

$$s^2 = \frac{14\, s_1^2 + 4\,(3s_2^2)}{18} = 0{,}00797$$

Beide gemiddelden berusten op 15 waarnemingen. De variantie van hun verschil, s_v^2, bedraagt dus $s_v^2 = \frac{2}{15} s^2 = 0{,}00106$
$s_v = 0{,}0326.$

Het betrouwbaarheidsinterval voor het verschil tussen de gemiddelde asdiameter bij A en bij B met 5% onbetrouwbaarheid heeft als grenzen:

$(\bar{x}_2 - \bar{x}_1) - t \times s_v = 0{,}006 - 2{,}101 \times 0{,}0326 = -0{,}0625$
en
$(\bar{x}_2 - \bar{x}_1) + t \times s_v = 0{,}006 + 2{,}101 \times 0{,}0326 = +0{,}0745;$

De waarde van t (2,101) wordt gevonden in de t-tabel bij 18 graden van vrijheid. Aangezien het betrouwbaarheidsinterval nul omsluit verschillen de gevonden gemiddelden bij A en B niet significant bij een 5% onbetrouwbaarheidsdrempel.

A 6

Een keuringsdienst voor waren maakt bij elke keuring van melk gebruik van twee methoden A en B om vast te stellen of deze besmet is met een bepaald soort bacteriën.

Beide methoden geven een negatieve uitkomst als melk niet besmet is. Bij beide methoden bestaat er een, overigens ongelijke, kans dat met besmette melk een negatieve uitkomst verkregen wordt, zodat dus (ten onrechte) geconcludeerd wordt, dat de melk niet besmet is. Elk melkmonster wordt met beide methoden onderzocht; is de uitkomst volgens één van beide methoden, danwel volgens beide methoden positief, dan is dus vastgesteld, dat de melk, waaruit dit monster getrokken was, besmet was. Is de uitkomst negatief, dan bestaat een mogelijkheid, dat de melk nochtans besmet was. De uitkomst van 10.000 achtereenvolgende analyses staat in de hieronder gegeven tabel.

		Methode A		Totaal.
		neg.	pos.	
Methode B	neg.	9877	18	9895
	pos.	15	90	105
Totaal		9892	108	1000

Aangenomen mag worden:

1e. dat voor elk besmet monster de kans p_a om met methode A als besmet herkend te worden dezelfde is;

2e. dat voor elk besmet monster de kans p_b die niet gelijk behoeft te zijn aan p_a om met methode B als besmet herkend te worden dezelfde is. De uitkomst bij toepassing van methode A is dus onafhankelijk van die bij toepassing van methode B.

Vraag: Geef een schatting van p_a, van p_b en van het aantal niet besmette monsters onder deze 10.000 onderzochte monsters.

Standaardantwoord

Om de kans p_a te schatten beschouwen wij de 105 monsters die met methode B als besmet zijn herkend. Daar de uitkomsten met methode A en met methode B onafhankelijk van elkaar zijn heeft elk van deze 105 monsters een kans p_a om met methode A als besmet herkend te worden.
Een schatting van p_a is dus: $\frac{90}{105} = \frac{6}{7}$.
Geheel analoog kan worden beredeneerd dat een schatting van p_b is: $\frac{90}{108} = \frac{5}{6}$.
Van de 10.000 onderzochte monsters zijn $15 + 90 + 18 = 123$ zeker besmet.

De kans dat een besmet monster noch met methode A, noch met methode B wordt ontdekt is $(1 - p_a)(1 - p_b)$, dus geschat: $\frac{1}{7} \times \frac{1}{6} = \frac{1}{42}$.
M.a.w. van alle besmette monsters wordt een fractie $\frac{1}{42}$ niet, $\frac{41}{42}$ wel ontdekt. Naar schatting zijn er dus nog $\frac{1}{41} \times 123 = 3$ besmette monsters onder de 9877 die beide methoden gepasseerd zijn.
Een schatting van het gevraagde aantal niet besmette monsters is dus $9877 - 3 = 9874$.

Het examen Statistisch Analist 1962

Opgave

A 1

In een stad is het aantal verkeersongevallen per dag verdeeld volgens een Poissonverdeling met gemiddelde 1.

Vraag: Beschouw alleen die dagen, waarop zich tenminste één ongeval voordeed. Wat is voor deze dagen het gemiddelde aantal ongevallen per dag?

Standaardantwoord

Noem $\underline{x}$ het aantal ongevallen per dag. Er geldt dus:

$$P\{\underline{x} = k\} = \frac{e^{-1}\,1^k}{k!} = \frac{e^{-1}}{k!} \qquad (k = 0, 1, \ldots)$$

Noem $\underline{y}$ het aantal ongevallen per dag onder de voorwaarde dat $\underline{x} > 0$ is. Dus:

$$P\{\underline{y} = k\} = P\{\underline{x} = k \mid \underline{x} > 0\} = P\{\underline{x} = k\}.\ [P\{\underline{x} > 0\}]^{-1} = \frac{e^{-1}}{k!}.(1 - e^{-1})^{-1} \quad (k = 1, 2, \ldots)$$

De verwachting van y is:

$$\mathscr{E}\underline{y} = \sum_{k=1}^{\infty} \frac{k \,.\, e^{-1}}{k!}(1 - e^{-1})^{-1} = (1 - e^{-1})^{-1} \sum_{k=1}^{\infty} \frac{k \,.\, e^{-1}}{k!} =$$

$$= (1 - e^{-1})^{-1} \sum_{k=0}^{\infty} \frac{k \,.\, e^{-1}}{k!} = (1 - e^{-1})^{-1}\, \mathscr{E}\underline{x} = (1 - e^{-1})^{-1}.$$

Opgave

A 2

De eenheden in een populatie volgen een alternatieve verdeling met de onbekende kans op succes van p. De uitkomsten van aselecte steekproeven (met teruglegging) uit deze populatie volgen dus de binomiale verdeling.

Men heeft een steekproef getrokken van 100 waarnemingen. De relatieve frequentie van succes hieronder bedraagt 0,30.

Men wil de hypothese H_0 toetsen dat $p = 0{,}25$ tegenover de alternatieve hypothese H_1 dat $p > 0{,}25$.

Vraag:
a. Bepaal het kritieke gebied van de toets als men wil toetsen met een onbetrouwbaarheid van 0,05.
b. Welke is het onderscheidingsvermogen van de toets in het geval dat $p = 0{,}40$?
c. Formuleer de conclusie waartoe de toepassing van de toets in dit geval leidt.

Standaardantwoord

a. Het kritieke gebied van een toets is dat gedeelte van de „steekproefruimte" waar de te toetsen hypothese wordt verworpen.

Daar $n = 100$ mag de binomiale verdeling door een normale verdeling worden benaderd. Voor de grens van het éénzijdige kritieke gebied vindt men met continuïteitscorrectie:

$$25 + 1{,}645\sqrt{npq} + 0{,}5 = 25 + 1{,}645\sqrt{\frac{100 \,.\, 1 \,.\, 3}{4 \,.\, 4}} + 0{,}5 = 32{,}6.$$

Dit betekent dat het kritieke gebied bestaat uit aantallen successen $x > 33$, daar 33 het kleinste gehele getal is dat groter is dan 32,6. De onbetrouwbaarheid is dus kleiner dan de drempelwaarde 0,05.

b. Het onderscheidingsvermogen van een toets is de kans als functie van de parameter p, dat het steekproefpunt in het kritieke gebied valt. In het gegeven geval vindt men (voor $p = 0{,}40$):

$$\sqrt{npq} = \sqrt{100 \,.\, 0{,}4 \,.\, 0{,}6} = 4{,}90$$

De kans dat $\underline{x} > 33$ is berekent men nu als volgt:

$$u = \frac{np - (33 - 0{,}5)}{\sqrt{npq}} = \frac{40 - 32{,}5}{4{,}9} = 1{,}53.$$

De bijbehorende kans is 0,937.

c. De waargenomen waarde $x = 30$ ligt niet in de kritieke zone en leidt dus niet tot verwerping van de hypothese $p = 0{,}25$.

Opgave

A 3

Een vulmachine voor een poedervormig produkt werkt met 6 vulpotten I, II . . . VI, die op een draaiende schijf zijn gemonteerd. De door deze machine gevulde pakjes zijn achtereenvolgens afkomstig van de potten I, II, III . . . VI, I, II . . . enzovoorts.

Van de lopende produktie werden 30 achtereenvolgende pakjes gewogen. De volgende gewichten werden gevonden (in grammen).

Pakje no.	gewicht
1	52,3
2	53,6
3	51,5
4	53,8
5	51,2
6	50,9
7	55,0
8	52,4
9	52,3
10	55,9

Pakje no.	gewicht
11	53,0
12	50,8
13	50,4
14	51,0
15	52,7
16	55,3
17	51,6
18	51,3
19	53,8
20	52,7

Pakje no.	gewicht
21	54,0
22	52,9
23	48,4
24	51,6
25	52,8
26	52,9
27	53,0
28	53,7
29	50,4
30	52,5

Vraag: Ga met de methode der variantie-analyse na of er systematische verschillen bestaan tussen de gewichten van de pakjes die door verschillende potten zijn gevuld.

Standaardantwoord

De gewichten worden eerst zodanig gerangschikt, dat per kolom juist die gewichten staan, die afkomstig zijn van dezelfde vulpot. Ter vereenvoudiging van de berekening wordt tegelijkertijd van alle getallen 50 afgetrokken; dit geeft dezelfde variantie-analyse als met de oorspronkelijke getallen. Het onderstaande tableau ontstaat:

2,3	3,6	1,5	3,8	1,2	0,9
5,0	2,4	2,3	5,9	3,0	0,8
0,4	1,0	2,7	5,3	1,6	1,3
3,8	2,7	4,0	2,9	—1,6	1,6
2,8	2,9	3,0	3,7	0,4	2,5

Dit vraagstuk kan het best worden opgevat als een variantie-analyse met twee criteria, nl. vulpotten (kolommen) en tijdsinvloed (rijen).

Het volgende schema ontstaat:

	kwadraatsom	vr. gr.	gem. kwadraatsom	F	P
tussen vulpotten	35,6697	5	7,1339	4,32	0,01
tussen „tijden" (rijen)	5,3087	4	1,3272	0,80	0,5
rest	33,0553	20	1,6528		
totaal	74,0537	29			

Er is dus een verschil aantoonbaar (op het 1% punt) tussen de vulpotten. De gemiddelden per vulpot hebben de volgende waarden (in deze gemiddelden is het getal 50 weer opgenomen):

52,86
52,52
52,70
54,32
50,92
51,42

Deze uitkomsten suggereren, dat de vierde pot te hoge gewichten oplevert en de vijfde en zesde pot te lage gewichten. De variantie-analyse geeft hierover geen uitsluitsel; hiervoor zouden contrast-toetsen (Scheffé, Tukey e.a.) nodig zijn (omdat niet gegeven is van welk potnummer het 1e pakje afkomstig is, kan niet worden nagegaan welk potnummer behoort bij de potten, de vierde pot behoeft dus niet pot IV te zijn).

Deze variantie-analyse geeft geen aanwijzing van een tijdsinvloed.

Opgave

A 4

In een populatie bestaat een verdeling, die de eigenschap heeft dat de verwachting gelijk is aan de variantie van die verdeling.

Bekend is dat deze verwachting een geheel getal is (n).

Voorts is gegeven dat de kans 0,00135 bedraagt dat het rekenkundige gemiddelde van een steekproef van n waarnemingen, die met teruglegging uit de populatie worden getrokken, groter is dan 39.

Vraag: Bereken n.

Standaardantwoord

Het rekenkundig gemiddelde is bij benadering normaal verdeeld (centrale limietstelling).

De overschrijdingskans $P(M > 39) = 0{,}00135$.

Uit de tabel van de normale verdeling blijkt dat deze kans behoort bij een waarde van

$$u = 3{,}00.$$

Bijgevolg is:

$$u = 3 = \frac{39 - \mu}{\sigma : \sqrt{n}}$$

Gegeven is dat $\mu = n$ en $\sigma = \sqrt{n}$.

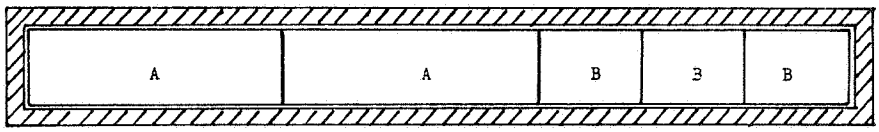
Bijgevolg is $n = 36$.

Opgave

A 5

In een speelgoedfabriek maakt men eenvoudige blokkendozen, die 2 blokken bevatten van type A en 3 blokken van type B (zie figuur).

Bovenaanzicht van de geopende doos.



lengterichting→

De blokken worden gezaagd uit latten, die een breedte hebben van b cm met een te verwaarlozen spreiding.

De lengte van de blokken A is normaal verdeeld met een gemiddelde 5 cm en een standaarddeviatie 0,15 cm.

Bij de blokken B, die eveneens een normaal verdeelde lengte hebben, is de gemiddelde lengte 2 cm en de standaarddeviatie 0,05 cm. De inwendige lengte-afmeting van de doos heeft een normale verdeling met een gemiddelde van 16,4 cm en een standaarddeviatie van 0,10 cm.

Vraag: Bereken het percentage van de dozen, waarbij de lengte binnenwerks te klein zal blijken te zijn, indien de blokken aselect over de dozen verdeeld.

Standaardantwoord

We maken gebruik van de eigenschap, dat de som (of het verschil) van een aantal onafhankelijke normale verdelingen weer een normale verdeling oplevert met als gemiddelde de som (of het verschil) van de gemiddelden en als variantie de som (altijd de som!) van de varianties.

De ruimte, die gemiddeld overblijft, indien twee blokken A en drie blokken B achter elkaar in de doos worden gelegd, is

$$16{,}4 - (2 \times 5) - (2 \times 3) = 0{,}4 \text{ cm}.$$

De variantie van deze overblijvende ruimte is

$$(0{,}1)^2 + 2 \times (0{,}15)^2 + 3 \times (0{,}05)^2 = 0{,}0625$$

en de standaardafwijking is dus 0,25 cm.

De gevraagde kans is gelijk aan de kans, dat de overblijvende ruimte kleiner is dan nul; dat is de kans, dat een normaal verdeelde stochastische variabele met gemiddelde 0 en standaardafwijking 1 groter is dan $\frac{0{,}4}{0{,}25} = 1{,}6$. Deze kans is gelijk aan 0,0548.

Opgave

A 6

Een machine, bestemd voor een kansspel, is als volgt geconstrueerd. De speler drukt op een knop. Er verschijnt dan in een venster op aselecte wijze een vierkant dat rood of wit is, het eerste met een kans 0,8, het tweede met een kans 0,2. Is de verschijnende kleur wit, dan heeft de speler gewonnen. Is de kleur rood, dan mag hij nogmaals op de knop drukken, waarbij de beide mogelijkheden zich met dezelfde kansen herhalen. Is ook de tweede maal de kleur rood, dan mag de speler nog een derde en laatste maal op de knop drukken.

Is de uitslag dan wederom rood, dan heeft hij verloren.

Vraag: Bereken de kans dat de speler wint.

Standaardantwoord

De kans dat de speler verliest bedraagt $0{,}8^3 = 0{,}512$.
De kans dat de speler wint bedraagt dus $1 - 0{,}512 = 0{,}488$.

Opgave

A 7

Uit een continue rechthoekige verdeling wordt een steekproef getrokken. Na een klassificatie worden de onderstaande uitkomsten verkregen:

klasse	aantal waarnemingen
$> 0{,}9$	24
$0{,}9 - 1{,}0$	40
$1{,}0 - 1{,}1$	51
$1{,}1 - 1{,}2$	42
$1{,}2 - 1{,}3$	43
$1{,}3 - 1{,}4$	38
$1{,}4 - 1{,}5$	45
$1{,}5 - 1{,}6$	25

Vraag: a. Teken van deze tabel een histogram.

b. Hoe kan dit histogram worden verbeterd, als bovendien gegeven is, dat de rechthoekige verdeling, waaruit de steekproef getrokken is, een (populatie-) gemiddelde μ 1,2 en een (populatie-) standaardafwijking $\sigma = \frac{7}{60}\sqrt{3}$ heeft?

c. Teken de cumulatieve frequentieverdeling van de steekproef en zet in dezelfde figuur ter vergelijking de cumulatieve kansverdeling, genoemd onder *b*.

d. Toets de hypothese, dat de steekproef getrokken is uit de onder *b* genoemde populatie.

Standaardantwoord

a. Zie bijgaande figuur 1 (getrokken lijnen).

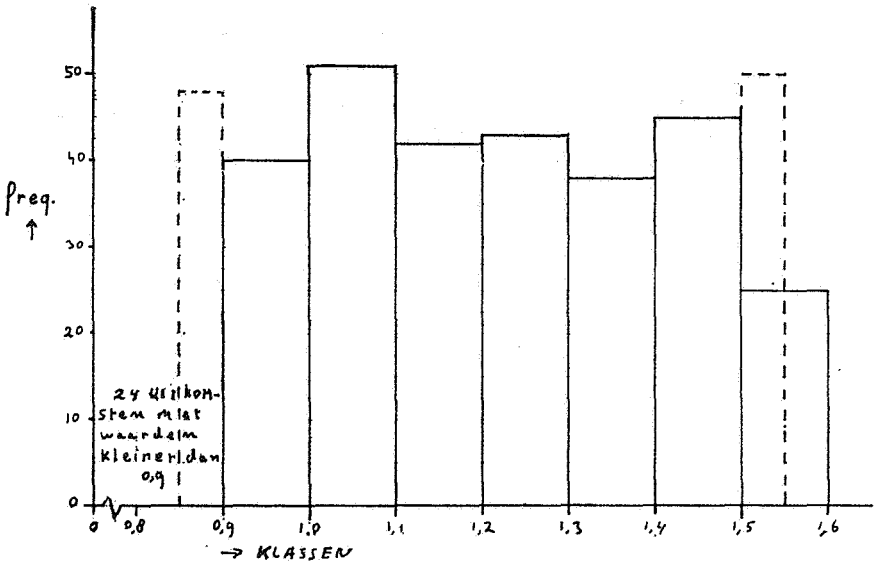
b. Stel dat de verdeling zich uitstrekt van a tot $a + h$. Het is bekend dat de variantie σ^2 gelijk is aan $\frac{1}{12}h^2$. Dus:

$$\tfrac{1}{12}h^2 = \tfrac{49}{1200},\ h^2 = \tfrac{49}{100},\ h^2 = 0{,}49,\ h = 0{,}70.$$

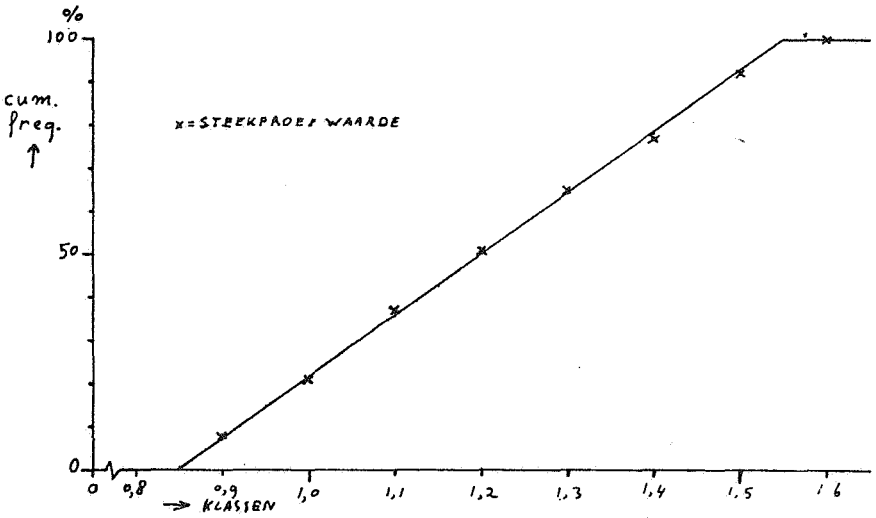
Er is verder gegeven dat $\mu = 1{,}2$, dus de verdeling loopt van $1{,}2 - 0{,}35$ tot $1{,}2 + 0{,}35$, dus van 0,85 tot 1,55. De eerste en de laatste klasse kunnen nu volgens de stippellijnen in figuur 1 worden uitgetekend.

A 7

FIGUUR 1



FIGUUR 2



c. De cumulatieve frequentieverdeling is als volgt:

	frequentie	in procenten
$< 0{,}9$	24	8
$< 1{,}0$	64	21
$< 1{,}1$	115	37
$< 1{,}2$	157	51
$< 1{,}3$	200	65
$< 1{,}4$	238	77
$< 1{,}5$	283	92
$< 1{,}6$	308	100

Deze punten zijn met kruisjes in figuur 2 aangegeven. De getrokken lijn geeft de onder *b* genoemde verdeling aan.

d. Er zijn zes klassen met een breedte van 0,1 en twee met een breedte van 0,05. Het totale aantal waarnemingen is 308. Per klasse met een breedte 0,1 is het verwachte aantal $\frac{308}{7} = 44$. De beide klassen met halve breedte hebben een verwachting van 22.

De hypothese dat de steekproef uit deze verdeling komt, wordt getoetst met de χ^2 toets voor aanpassing. De toetsingsgrootheid wordt als volgt berekend:

Klasse	gevonden aantal	verwachting	Bijdrage tot χ^2
0,85 — 0,9	24	22	$\frac{(24-22)^2}{22} = 0{,}18$
0,9 — 1,0	40	44	0,36
1,0 — 1,1	51	44	1,11
1,1 — 1,2	42	44	0,09
1,2 — 1,3	43	44	0,02
1,3 — 1,4	38	44	0,82
1,4 — 1,5	45	44	0,02
1,5 — 1,55	25	22	0,41
	308	308	3,01

onder de nulhypothese is de waarde 3,01 een trekking uit een χ^2 verdeling met zeven vrijheidsgraden. De overschrijdingskans van deze waarde is meer dan 50%. Er is dus geen enkele reden de hypothese te verwerpen.

Het examen Statistisch Analist 1963

Opgave

A 1

Vijf hardlopers houden een wedstrijd. Hun volgorde van aankomst wordt vastgesteld. Op de drie daaropvolgende dagen herhalen zij dezelfde wedstrijd. De volgordes van aankomst staan in onderstaande tabel:

Loper	Wedstrijd			
	1	2	3	4
A	2	3	5	4
B	5	5	3	5
C	3	1	4	3
D	1	4	1	2
E	4	2	2	1

Vraag: Is de conclusie gewettigd dat de vijf hardlopers niet allen even snel lopen?

Standaardantwoord

De vraag kan worden beantwoord met de toets van Friedman voor $m = 4$ rangschikkingen en $k = 5$ individuen.

De som van de rangnummers per hardloper zijn:

Hardloper	A	B	C	D	E	Totaal
Som	14	18	11	8	9	60 (controle!)

De toetsingsgrootheid $S = 14^2 + \ldots + 9^2 - 60^2/5 = 66$.

Onder de alternatieve hypothese, dat de ene hardloper beter is dan de andere, zal $E(S)$ groter zijn dan onder de nulhypothese. Er moet dus eenzijdig worden getoetst. De tafel vermeldt $S_{0,95} = 89$. De werkelijke S is kleiner, zodat de conclusie, dat de ene hardloper harder kan lopen dan de andere niet uit dit cijfermateriaal getrokken kan worden.

Opgave

A 2

Van een bepaald type electrische batterijen is op grond van lange ervaring bekend, dat de levensduren aan de volgende frequentie-verdeling voldoen:

minder dan 3	maanden	25%
3–6	maanden	50%
méér dan 6	maanden	25%

1. Van een partij van 600 batterijen kon worden vastgesteld dat de frequentie-verdeling der levensduren als volgt is geweest:

minder dan 3	maanden	155 stuks
3–6	maanden	322 stuks
méér dan 6	maanden	123 stuks

2. Voor een volgende partij bestaande uit 400 batterijen bleken de levensduren als volgt te zijn verdeeld:

minder dan 3	maanden	98 stuks
3–6	maanden	179 stuks
méér dan 6	maanden	123 stuks

Vraag: a. Toets met een onbetrouwbaarheid van 1% of de onder 1 gevonden uitkomsten in overeenstemming zijn met de in de eerste alinea genoemde theoretische frequentieverdeling.
Doe hetzelfde voor de onder 2 genoemde uitkomsten.

b. Hoe stelt ge vast of de onder a verkregen toetsingsgrootheden tezamen beschouwd erop wijzen dat in het productieproces fluctuaties optreden welke leiden tot significante afwijkingen van de in de eerste alinea genoemde theoretische verdeling?
Ook hier een onbetrouwbaarheid van 1% aanhouden.
Licht Uw antwoord toe.

Standaardantwoord

a. De toetsingsgrootheid van de χ^2 toets is in geval 1

$$\chi^2 = \frac{(155-150)^2}{150} + \frac{(322-300)^2}{300} + \frac{(123-150)^2}{150} = \frac{996}{150} = 6{,}64$$

De bijbehorende overschrijdingskans bij 2 vrijheidsgraden ligt tussen 5% en 2½%. De theoretische verdeling kan dus met een onbetrouwbaarheid van 1% niet worden verworpen.
In geval 2 wordt gevonden

$$\chi^2 = \frac{(98-100)^2}{100} + \frac{(179-200)^2}{200} + \frac{(123-100)^2}{100} = \frac{753{,}5}{100} = 7{,}54$$

Hierbij hoort een overschrijdingskans van ongeveer $2\frac{1}{2}\%$. Ook hier wordt de nulhypothese niet verworpen.

b. De onder a gevonden uitkomsten kunnen worden gecombineerd door ze op te tellen. De som van twee onderling onafhankelijke χ^2 verdeelde grootheden heeft nl. opnieuw een χ^2 verdeling met als aantal vrijheidsgraden de som van de vrijheidsgraden van de termen.
De uitkomst 14,18 heeft bij 4 vrijheidsgraden een overschrijdingskans tussen 1% en 0,5%. Op grond van de gecombineerde resultaten moet de nulhypothese dus wel worden verworpen. Dit wijst dus inderdaad sterk op fluctuaties in het productieproces. De afwijking in de groep van 3–6 maanden is namelijk eerst positief en daarna negatief.

Opgave

A 3

Voor drie provincies van Nederland zijn de volgende gegevens bekend over het gemiddelde verbruik per gezin van een consumptiegoed M, in kg per jaar;

Provincie	Aantal gezinnen	Gemiddeld verbruik in kg per jaar	standaardafwijking van het verbruik in kg
A	500.000	10	1
B	750.000	16	2
C	750.000	12	1

Vraag: a. Men neemt een enkelvoudige aselecte steekproef, bestaande uit 400 gezinnen, uit de bevolking van de drie provincies tezamen.
Bereken de standaardafwijking van het gemiddelde verbruik per gezin afgeleid uit de steekproef.

b. Waarom zou een gelede steekproef in dit geval de voorkeur kunnen verdienen boven een enkelvoudige steekproef?
Er worden geen berekeningen gevraagd.

Standaardantwoord

a. Tussen de verwachting μ, de variantie σ^2 en de verwachting van het kwadraat van een stochastische variabele $\underset{\sim}{x}$ bestaat het volgende verband:

$$\sigma^2 = \mathscr{E}(\underset{\sim}{x}^2) - \mu^2,$$
$$\text{of } \mathscr{E}(\underset{\sim}{x}^2) = \sigma^2 + \mu^2.$$

Van de drie provincies A, B en C geldt dus:

$$\text{A } \mathscr{E}(\underset{\sim}{x}^2) = 10^2 + 1^2 = 101$$
$$\text{B } \mathscr{E}(\underset{\sim}{x}^2) = 16^2 + 2^2 = 260$$
$$\text{C } \mathscr{E}(\underset{\sim}{x}^2) = 12^2 + 1^2 = 145$$

als $\underset{\sim}{x}$ telkens het verbruik van een familie in een provincie voorstelt. Voor een aselect gekozen familie uit de drie provincies tezamen geldt dus:

$$\mathcal{E}(\underline{x}^2) = \frac{0{,}50}{2{,}00} \cdot 101 + \frac{0{,}75}{2{,}00} \cdot 260 + \frac{0{,}75}{2{,}00} \cdot 145 = 177{,}125$$

$$\mu = \frac{0{,}50}{2{,}00} \cdot 10 + \frac{0{,}75}{2{,}00} \cdot 16 + \frac{0{,}75}{2{,}00} \cdot 12 = 13$$

Dus $\sigma^2 = \mathcal{E}(\underline{x}^2) - \mu^2 = 177{,}125 - 169 = 8{,}125$

$\sigma = \sqrt{8{,}125} = 2{,}85.$

Het gemiddelde van een steekproef van 400 gezinnen heeft dus als standaardafwijking:

$$\sigma_{\bar{\underline{x}}} = \frac{2{,}85}{\sqrt{400}} = 0{,}1425$$

b. Een gelede steekproef zou in dit geval de voorkeur verdienen, omdat de gemiddelde verbruiken per provincie vrij sterk uiteenlopen. Daarom geeft een gelede steekproef een kleinere standaardafwijking voor het geschatte gemiddelde. Deze standaardafwijking wordt nog kleiner als provincie B in de steekproef meer dan evenredig wordt vertegenwoordigd, omdat in deze provincie de spreiding het grootst is.

Opgave

A 4

Men heeft van 10 varkens de hoeveelheden opgenomen voedsel (x) en de toename in gewicht (y) over een bepaalde tijd vastgesteld. De resultaten (in kg) zijn als volgt:

varken	x	y
1	173	32
2	152	30
3	176	38
4	143	21
5	145	28
6	181	39
7	162	34
8	161	33
9	156	27
10	154	32

Vraag: Schat met behulp van een lineaire regressievergelijking de gemiddelde hoeveelheid opgenomen varkensvoer bij varkens met een gewichtstoename van 32 kg. Om de berekening gemakkelijker te maken wordt gegeven dat

$$\sum x^2 = 258461$$
$$\sum y^2 = 10112$$
$$\sum xy = 50867.$$

Standaardantwoord

Om de vraag te kunnen beantwoorden moet men de lineaire regressie van x op y berekenen. Voor de regressievergelijking wordt gevonden:

$$(x - 160{,}3) = 2{,}1109\,(y - 31{,}4)$$

Door substitutie volgt, dat bij een gewichtstoename $y = 32$ een hoeveelheid opgenomen voedsel $x = 161{,}57$ behoort.

Opgave

A 5

Met een zuivere dobbelsteen worden vier worpen gedaan met als uitkomsten respectievelijk x_1, x_2, x_3, x_4.

Vraag: a. Wat is de kans dat de hoogste worp gelijk is aan m? Bereken deze kans voor $m = 1, 2, 3, 4, 5, 6$.

b. Als nog eens vier worpen worden gedaan is dan de kans om weer x_1, x_2, x_3, x_4 te gooien, nu echter ongeacht de volgorde, onafhankelijk van de numerieke waarden van x_1, x_2, x_3, x_4?
Motiveer Uw antwoord.

Standaardantwoord

a. De kans dat de hoogste worp m is ($m = 1, 2, \ldots 6$), is de kans dat in vier worpen geen hoger aantal ogen dan m wordt geworpen, verminderd met de kans dat geen hoger aantal ogen dan $m - 1$ wordt geworpen.
Deze kans bedraagt dus

$$\left(\frac{m}{6}\right)^4 - \left(\frac{m-1}{6}\right)^4$$

De kansen voor de verschillende waarden van m zijn dus:

m	P (hoogste $= m$)
1	1/1296 = 0,0008
2	15/1296 = 0,0116
3	65/1296 = 0,0502
4	175/1296 = 0,1350
5	369/1296 = 0,2847
6	671/1296 = 0,5177
Totaal	1,0000

b. De kans om x_1, x_2, x_3, x_4 te gooien is wel degelijk afhankelijk van de numerieke waarden van x_1, x_2, x_3 en x_4. Als b.v. $x_1 = x_2 = x_3 = x_4 = 1$, is de kans $(\frac{1}{6})^4$.
Als $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$, dan is de kans $24 \,.\, (\frac{1}{6})^4$.

Opgave

A 6

Drie onafhankelijke trekkingen uit eenzelfde Poissonverdeling leveren de waarden 3, 2 en 4 op.

Vraag: Stel een betrouwbaarheidsinterval op voor het gemiddelde van deze verdeling, waarbij de bovengrens en de ondergrens ieder een onbetrouwbaarheid 0,05 hebben.

Standaardantwoord

De som van drie onafhankelijke trekkingen uit een Poissonverdeling met gemiddelde μ heeft zelf ook een Poissonverdeling, doch nu met gemiddelde 3μ. In de steekproef levert de som van de drie trekkingen de waarde 9 op, welk getal kan worden beschouwd als één trekking uit de afgeleide Poissonverdeling.

De ondergrens van het betrouwbaarheidsinterval (voor 3μ) is die waarde voor het gemiddelde, waarbij de kans op 9 of meer (of 1 minus de kans op 8 of minder) gelijk is aan 0,05. Dit is 4,7. De bovengrens is die waarde, waarbij de kans op 9 of minder (of 1 minus de kans op 10 of meer) gelijk is aan 0,05. Dit levert op 15,7.

Om een betrouwbaarheidsinterval voor μ te vinden moeten beide waarden door 3 worden gedeeld. Het gevraagde interval is dus $1{,}6 < \mu < 5{,}2$.

Opgave

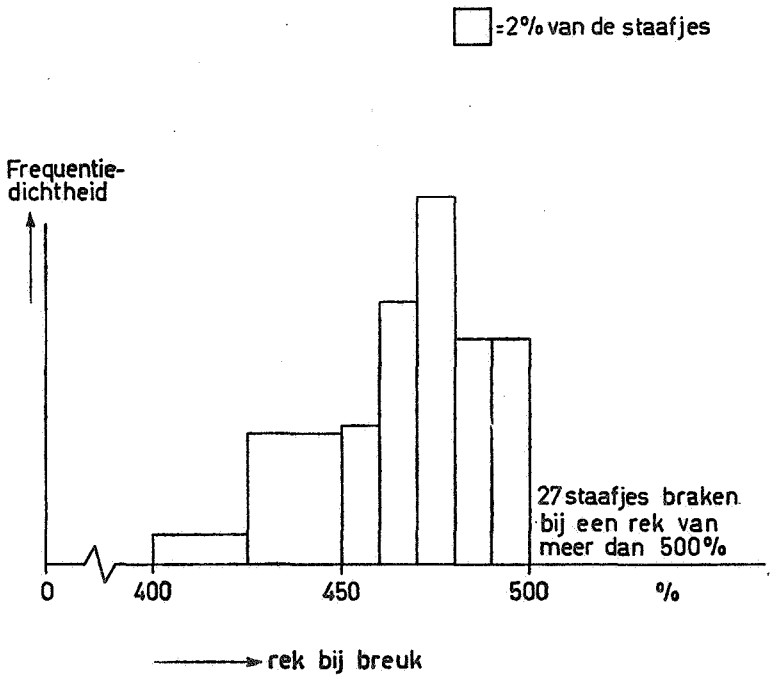
A 7

Van een partij van 200 rubberstaafjes werd de rek bij breuk gemeten, dat wil zeggen de uitrekking in procenten van de oorspronkelijke lengte, die een staafje heeft op het moment dat het kapot getrokken wordt.

De volgende tabel geeft de frequentieverdeling van de uitkomsten.

rek bij breuk	aantal staafjes
400 – 425	8
425 – 450	35
450 – 460	15
460 – 470	28
470 – 480	39
480 – 490	24
490 – 500	24
500 of meer	27

Vraag: Teken een histogram, dat de frequentieverdeling weergeeft.



Histogram van 200 destructieve rekmetingen aan rubberstaafjes

Opgave

A 8

Een fabrikant heeft de beschikking over een machine, waarmee toffees kunnen worden gemaakt. Op grond van een onderzoek heeft hij geconstateerd, dat de frequentieverdeling van het gewicht van de toffees logarithmisch normaal is: zij x het gewicht (in grammen) van een toffee, dan heeft $\log x$ een normale verdeling met gemiddelde 0,69897 en standaardafwijking 0,1.

De fabrikant wil voor 1 gram van de toffee-substantie 0,4 cent ontvangen. Daar de logarithme van het gewicht als gemiddelde 0,69897 = log 5 heeft, stelt hij de prijs van elke toffee op $5 \times 0{,}4$ cent = 2 cent.

Na enige miljoenen toffees te hebben verkocht, blijkt de fabrikant een gemiddelde prijs van 2,012 cent per toffee te hebben ontvangen. Hij realiseert zich, dat ten gevolge van statistische fluctuaties dit gemiddelde niet precies gelijk behoeft te zijn aan de door hem verwachte 2 cent. Bij toetsing blijkt evenwel, dat het gemiddelde van 2,012 cent zeer sterk significant afwijkt van 2.

De productieomstandigheden zijn niet veranderd.

Vraag: Geef een verklaring van de slechte overeenstemming tussen het gevonden gemiddelde van 2,012 cent en de 2 cent, welke de fabrikant verwacht. U behoeft deze verklaring niet met een berekening toe te lichten.

Standaardantwoord

De door de fabrikant verwachte waarde van 2 cent per toffee is het geldbedrag behorende bij het *meetkundige gemiddelde* van de gewichten van de toffees.

Dit meetkundige gemiddelde is altijd lager dan het gewone rekenkundige gemiddelde, waarbij de prijs van 2,012 cent behoort.

Het Examen Statistisch Analist 1964

Opgave

A 1

Aan 37 huisvrouwen werd gevraagd, welke van 2 soorten soep zij het lekkerst vonden.

De uitkomsten waren:

46 % prefereerde soep A
40,5 % had geen oordeel of vond ze gelijk
13,5 % prefereerde soep B

Vraag: Onderzoek of de voorkeur voor soep A significant is.

Standaardantwoord

De gegeven percentages worden eerst teruggerekend op aantallen:

17 huisvrouwen prefereerden *A*
15 huisvrouwen hadden geen voorkeur
5 huisvrouwen prefereerden *B*

37

Als de huisvrouwen zonder voorkeur buiten beschouwing worden gelaten, kan de tekentoets worden gebruikt om uit te maken of de voorkeur voor soep *A* significant is.
Onder de nulhypothese dat geen voorkeur aanwezig is, is bij $n = 22$ de tweezijdige overschrijdingskans van de waarde 17 iets kleiner dan 2%. De voorkeur voor soep *A* is dus significant.

De vraag „Onderzoek of de voorkeur voor soep *A* significant is" kan ook worden opgevat als een aanduiding dat eenzijdig moet worden getoetst, hoewel dit uit de redactie van het begin van het vraagstuk niet blijkt. Dit antwoord is hier daarom ook goed gerekend. De overschrijdingskans is dan kleiner dan 1% en de conclusie is dezelfde.

Commentaar

Verschillende candidaten pasten niet de tekentoets (binomiale verdeling), maar de chi-kwadraatbenadering toe, of, wat bij 1 vrijheidsgraad op hetzelfde neerkomt, de normale benadering. De benaderingsmethode is hier tijdrovender dan de exacte methode en deze oplossingen zijn dus minder goed.

Sommige candidaten stelden „voorkeur voor soep *A*" tegenover „geen voorkeur voor soep *A*". Dit is alleen te verdedigen als men rechtseenzijdig toetst. Gezien de numerieke uitkomsten 17 *A* en 20 niet-*A* behoeft men dan niets uit te rekenen omdat de overschrijdingskans van 17 kennelijk $> 50\%$ is.

Een nog moeilijker te verdedigen procedure is om de 15 huisvrouwen zonder voorkeur gelijkelijk te verdelen over *A* en *B*. De verdeling onder de nulhypothese is dan niet meer binomiaal (wat wel het geval zou zijn als men verdeelde door loting met kans $\frac{1}{2}$), terwijl het onderscheidingsvermogen van de toets afneemt.

Opgave

A 2

In een kogellagerfabriek, die beschikt over vijf machines om kogels te vervaardigen, werd van een aantal volgens toeval uitgezochte kogels de diameter bepaald. De uitkomsten van de metingen staan in de onderstaande tabel.

nummer van de machine waarmee de kogel werd gemaakt	diameter in mm van de kogel				
1	15,281	15,325	51,305	15,292	15,317
2	15,360	15,337			
3	15,325	15,348	15,316	15,303	
4	15,305	15,327			
5	15,333	15,340	15,321		

Vraag: Toets met behulp van variantie-analyse de nulhypothese, dat alle machines gemiddeld dezelfde diameter opleveren tegen de alternatieve hypothese, dat er een verschil is tussen de werkelijke gemiddelden per machine. Neem 5% als onbetrouwbaarheidsdrempel.

Standaardantwoord

Alle uitkomsten worden ter vereenvoudiging van het rekenwerk eerst met een constante verminderd (b.v. 15,3) en dan met 1000 vermenigvuldigd, waardoor de volgende getallen ontstaan

Machine nr.					
1	—19	25	5	—8	17
2	60	37			
3	25	48	16	3	
4	5	27			
5	33	40	21		

De gevraagde variantie-analyse op de bovenstaande getallen ziet er als volgt uit:

Bron van variatie	*Kwadraatsom*	*Vr. gr.*	*Gem. kwadr. som*	*F*
tussen machines	3343,77	4	835,94	3,01
binnen machines	3053,17	11	277,56	
totaal	6396,94	15		

De rechtseenzijdige kritieke waarde van de F-verdeling met 4 en 11 vrijheidsgraden bij een onbetrouwbaarheid van 5% is 3,36. De nulhypothese wordt dus niet verworpen, zodat een verschil „tussen machines" niet is aangetoond.

Opgave

A 3

Uit beschikbare productiestatistieken over een aantal jaren heeft men afgeleid dat de trend in de afzet van een bedrijfstak kan worden voorgesteld door de volgende vergelijking:

$$y = 1200 + 32\,t.$$

Hierin is:

y de afzet per jaar, uitgedrukt in miljoenen guldens;
t de tijd, uitgedrukt in jaren, waarbij 1960 overeenkomt met $t = 0$.

De kwartaalcijfers blijken een seizoenfluctuatie te tonen, welke men met behulp van de multiplicatieve methode heeft geanalyseerd. Voor de eerste drie kwartalen werden voor de seizoenindices de volgende getallenwaarden gevonden:

98 85 111

Vraag: a. Bereken de trendwaarden van de afzet in elk der kwartalen van het jaar 1962.

b. Bereken de werkelijke afzet in elk der vier kwartalen van 1962, als nog is gegeven dat de toevallige fluctuaties verwaarloosbaar klein zijn.

c. Is bij toepassing van een multiplicatief seizoenpatroon in een geval als bovenbedoeld de algebraïsche som van de seizoenafwijkingen over de vier kwartalen van een jaar gelijk aan nul? Licht uw antwoord toe.

Standaardantwoord

a. Uit de opgave volgt dat $t = 0$ overeenkomt met het midden van 1960, m.a.w. $t = 0 =$ 1 juli 1960. Het midden van het jaar 1962 komt overeen met $t = 2$. De middens van de vier kwartalen van 1962 vallen op de tijdstippen:

$$1\tfrac{5}{8},\ 1\tfrac{7}{8},\ 2\tfrac{1}{8},\ 2\tfrac{3}{8}.$$

Deze waarden van t ingevuld in de vergelijking voor de trend geeft de volgende omzetcijfers voor de vier kwartalen van 1962 op jaarbasis:

$y_1 = 1252$ $\quad$ $y_3 = 1268$
$y_2 = 1260$ $\quad$ $y_4 = 1276$

De trendwaarden van de kwartaalomzetten worden verkregen door deze cijfers door 4 te delen. Men vindt:

1962, 1e kwartaal 313
1962, 2e kwartaal 315
1962, 3e kwartaal 317
1962, 4e kwartaal 319.

b. De som van de vier seizoenindices moet gelijk zijn aan 400. De seizoenindex voor het vierde kwartaal is dus gelijk aan

$$400 - (98 + 85 + 111) = 106.$$

De werkelijke afzet in elk der kwartalen van 1962 is dus gelijk aan de berekende trendwaarden van de kwartaalomzetten vermenigvuldigd met de seizoenindices, gedeeld door 100. Men vindt:

$313 \times 0{,}98 = 306{,}74$
$315 \times 0{,}85 = 267{,}75$
$317 \times 1{,}11 = 351{,}87$
$319 \times 1{,}06 = 338{,}14$

c. De som van de seizoenindices is gelijk aan 400. De algebraische som van de afwijkingen van de seizoenindices vanaf 100 is gelijk aan nul.
De som van de seizoenafwijkingen is bij een stijgende trend evenwel niet gelijk aan nul, omdat de seizoenafwijking voor latere kwartalen een grotere amplitude heeft. In het bovengenoemde geval is de algebraische som van de seizoenafwijkingen gelijk aan:

$$-6{,}26 - 47{,}25 + 34{,}87 + 19{,}14 = +0{,}50.$$

Commentaar

Enkele candidaten hebben $t = 0$ gesteld voor het eerste kwartaal van 1960, of voor januari 1960, of per 1 januari 1960.

De in het standaardantwoord gegeven interpretatie is gebruikelijk bij tijdreeksen.

Ad a. De trendwaarden van de kwartaalomzetten mogen ook op jaarbasis worden weergegeven.

Ad b. Er worden hier kwartaalcijfers gevraagd. Wanneer onder *a* cijfers op jaarbasis zijn gegeven, moet nu nog door 4 worden gedeeld.

Opgave

A 4

Het hoofd van een lagere school vroeg aan al zijn leerlingen, met hoeveel kinderen zij thuis waren (inwonende grote broers en zusters en zichzelf meegeteld). Hij verkreeg uit deze informatie de volgende frequentieverdeling:

aantal kinderen	%	(1) × (2)
(1)	(2)	(3)
1	9,8	9,8
2	25,0	50,0
3	21,7	65,1
4	16,8	67,2
5	9,6	48,0
> 5[1])	17,1	136,8
	100.0	376.9

[1]) gemiddeld 8

Dus gemiddeld 3,77 kinderen per gezin. Voor de gehele gemeente bedroeg het gemiddelde aantal kinderen per gezin 1,89.

Vraag: Hoe moet dit verschil worden verklaard? (Geen berekeningen).

Standaardantwoord

Het schoolhoofd hanteerde als eenheid van telling niet een gezin, maar een informant. Onder een informant kan in dit verband worden verstaan een kind op de lagere-schoolleeftijd. Het aantal informanten zal gemiddeld groter zijn naarmate het gezin groter is. Bij gezinnen met 0 kinderen is de kans zelfs 0 dat het in de steekproef wordt opgenomen.

Daarom komen grote gezinnen te veel en kleine gezinnen te weinig in de steekproef voor, waarmee het verschil tussen de twee gemiddelden is verklaard.

Commentaar

Van de candidaten werd geen uitvoerige motivering gevraagd. De essentie diende echter wel te worden weergegeven, met name dat er een verschuiving in teleenheid had plaats gevonden en dat daardoor het aantal meldingen per gezinsgrootte (min of meer evenredig) toenam met het kindertal. Slechts zeer weinig candidaten hebben dit scherp ingezien.

Een grote groep wees op het ontbreken van de gezinnen met 0 kinderen en op aanwezige dubbeltellingen van de gezinnen met 2 en meer kinderen doordat meerdere kinderen uit één gezin op de betreffende school konden zitten. Dit antwoord, mits behoorlijk geformuleerd, werd als goed beschouwd.

Opgemerkt moet echter worden dat ook bij het weglaten van dubbeltellingen nog een zware accentverschuiving naar de grote gezinnen optreedt. Ieder gezin kan nu nog maar éénmaal voorkomen, maar de kans is voor een groot gezin nog steeds groter dan voor een klein gezin.

Als van de eenvoudige veronderstelling wordt uitgegaan dat ieder kind uit een gezin dezelfde kans p heeft informant te zijn, ongeacht de gezinsgrootte, zal de verwachting van het aantal informanten $\mathcal{E}\,\underline{x}$ per gezin evenredig zijn met de gezinsgrootte n. In formule:

$$\mathcal{E}\,(\underline{x} \mid n) = n\,p$$

Neemt men slechts kinderen van één klas of één leeftijd zodat dubbeltellingen niet kunnen voorkomen dan zal nog een uitkomst in de buurt van 3,77 worden verkregen. Immers nog

steeds zal de verwachting van het aantal keren dat het gezin in de steekproef voorkomt (dat nu slechts 0 of 1 kan zijn) evenredig zijn met het aantal kinderen n.

Is de gemaakte veronderstelling juist dat de kans informant te zijn, onafhankelijk is van de gezinsgrootte? Waarschijnlijk niet. Onder de gezinnen met weinig of geen inwonende kinderen bevinden zich relatief veel pas beginnende gezinnen met zeer jonge kinderen en gezinnen met volwassen kinderen die ten dele reeds uit huis zijn. In die laatste gezinnen zullen de overblijvende kinderen meestal ook boven de lagere-schoolleeftijd zijn. Deze verschijnselen drukken het aantal informanten uit kleine gezinnen en versterken dus het gevonden effect. Hiertegenover staat dat ook de zeer grote gezinnen wellicht iets te weinig informanten zullen opleveren omdat daar het maximale aantal informanten op fysieke gronden gelimiteerd is.

Een antwoord dat alleen wees op weglaten van de nul-kindergevallen of alleen op de dubbeltellingen, werd als ontoereikend beschouwd.

Kwalitatief nog slechter waren de antwoorden die het slechts zochten in de selectie uit bepaalde godsdienstige of sociale groepen. Deze verklaring is, behoudens zeer abnormale gevallen, ontoereikend en raakt niet de essentie van de vertekening.

De uitdrukking „niet-aselecte steekproef" werd vele malen misbruikt. De steekproef van het schoolhoofd kan een volkomen aselecte steekproef zijn, weliswaar niet uit de gezinnen, maar wel uit de informanten. Als het de enige dorpsschool is, zal de steekproef zelfs identiek met de populatie zijn.

Opgave

A 5

Uit een urn gevuld met rode en witte balletjes worden zonder teruglegging balletjes getrokken. De kans, dat het eerste balletje rood is, is 1/5. De kans, dat het tweede rood is, onder voorwaarde dat het eerste rood is, is 1/6.

Vraag: Wat is de kans dat het tweede balletje rood is, als het eerste balletje wit is?

Standaardantwoord

Stel dat de urn n balletjes bevat, waarvan k rode.
De kans dat het eerste balletje rood is, is

$$P(r_1) = \frac{k}{n} = \frac{1}{5}$$

Als het eerste balletje getrokken is en het blijkt rood te zijn, zijn er nog $n - 1$ balletjes over, waarvan $k - 1$ rode. Uit de opgave volgt dus:

$$P(r_2 \mid r_1) = \frac{k-1}{n-1} = \frac{1}{6}$$

Oplossing van deze twee vergelijkingen levert

$$k = 5 \text{ en } n = 25$$

De gevraagde kans is dus

$$P(r_2 \mid w_1) = \frac{k}{n-1} = \frac{5}{24}$$

De gevraagde kans kan ook worden gevonden zonder de aantallen witte en rode balletjes te berekenen.

Uit de definitie van kans als de verhouding van het aantal gunstige en het aantal mogelijke gevallen volgt:

$$P(r_2) = P(r_1) \tag{1}$$

Eventueel is dit als volgt te bewijzen:

$$P(r_2) = P(r_2 \mid r_1)\,P(r_1) + P(r_2 \mid w_1)\,P(w_1) =$$
$$= \frac{k-1}{n-1} \cdot \frac{k}{n} + \frac{k}{n-1} \cdot \frac{n-k}{n} = \frac{k}{n}\left(\frac{(k-1)+(n-k)}{n-1}\right) = \frac{k}{n} \tag{2}$$

Uit (1) en (2) volgt:

$$P(r_2 \mid w_1) = \frac{P(r_1) - P(r_2 \mid r_1)\,P(r_1)}{P(w_1)} = \frac{\frac{1}{5} - \frac{1}{6} \cdot \frac{1}{5}}{\frac{4}{5}} = \frac{5}{24}$$

Commentaar

De candidaten die de tweede weg trachtten te volgen kwamen gedeeltelijk wel tot het eerste gedeelte van vergelijking (2), maar zij zagen niet in dat $P(r_2) = P(r_1)$ en zij kwamen zodoende niet tot de eindoplossing.

Opgave

A 6

De volgende getallen zijn onderling onafhankelijke trekkingen uit een normale verdeling:

1,65	1,96
2,36	1,68
1,87	1,38
1,55	2,08
1,29	1,70

Vraag: Bereken, met gebruikmaking van de χ^2-verdeling, een tweezijdig betrouwbaarheidsinterval voor de standaardafwijking van deze verdeling met onbetrouwbaarheid 10%.

Standaardantwoord

De grootheid $\dfrac{(n-1)s^2}{\sigma^2}$ heeft een χ^2-verdeling met $(n-1)$ vrijheidsgraden.

In ons geval is $n - 1 = 9$

$$(n-1)s^2 = \Sigma(x_i - \bar{x})^2 = 0{,}96876$$

Een betrouwbaarheidsinterval voor σ^2 wordt dus gevonden uit de ongelijkheden

$$\chi^2_{0,05} < \frac{0,96876}{\sigma^2} < \chi^2_{0,95}$$

waarin $\chi^2_{0,05}$ het linker en $\chi^2_{0,95}$ het rechter 5% punt is van de χ^2-verdeling met 9 vrijheidsgraden. Deze waarden zijn resp. 3,325 en 16,919. Het interval voor σ^2 is dus

$$\frac{0,96876}{16,919} < \sigma^2 < \frac{0,96876}{3,325}$$

of

$$0,05726 < \sigma^2 < 0,2914$$

En dus voor de standaardafwijking:

$$0,239 < \sigma < 0,540$$

Commentaar

In enkele gevallen werd de normale benadering voor χ^2 toegepast, hetgeen echter onnodig is en minder juist omdat n klein is.

De volgende fouten werden gesignaleerd.

Sommige candidaten bepaalden het betrouwbaarheidsinterval voor σ^2 in plaats van dat voor σ. Andere candidaten berekenden een eenzijdig betrouwbaarheidsinterval.

In enkele gevallen was het principe van het betrouwbaarheidsinterval niet begrepen en werd in de formule

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

voor σ^2 de uit de steekproef geschatte s^2 ingevuld waarna voor s^2 een interval werd gevonden. Dergelijke misverstanden zijn waarschijnlijk het gevolg van een niet duidelijk onderscheiden van populatiewaarden van parameters en de uit steekproeven berekende schatters daarvoor.

Opgave

A 7

Door een instituut werd een enquête gehouden onder bezitters van personenauto's op een bepaalde datum. Doel van het onderzoek was gegevens te verkrijgen over het aantal per jaar gereden kilometers gemiddeld per auto, onderscheiden naar gebruik in het beroep, gebruik voor rijden van en naar kantoor of fabriek, gebruik in de vrije tijd, en gebruik voor vacantie. Hierbij werd een vacantie gedefinieerd als een afwezigheid van meer dan drie dagen.

De gevraagde gegevens hadden betrekking op de afgelopen 12 maanden. Voor hen die korter dan 12 maanden auto-bezitter waren geweest, dienden de gegevens alleen op de periode gedurende welke zij in het bezit van de auto waren geweest, te worden opgegeven. Daarbij werd gevraagd deze tijd op te geven in maanden.

Voor het onderzoek werd van de steekproefmethode gebruik gemaakt. De

enquête had betrekking op 6000 auto-bezitters, wier adressen op aselecte wijze werden getrokken uit het kaartsysteem waarover de overheid beschikt. De enquête-formulieren werden per post toegezonden. Het aantal terugontvangen formulieren bedroeg 5000.

Vraag: Ontwerp één tabel waarin de uitkomsten van het onderzoek uitgedrukt als gemiddelden per auto, kunnen worden opgenomen. De invloed van de duur van het auto-bezit moet daarbij in de tabel tot uitdrukking komen.

Standaardantwoord

De tabel zou als volgt kunnen worden ingericht:

Aantal maanden v. autobezit	Aantal auto's	Gemiddeld aantal kilometers per auto gereden [1])				
		Totaal	In het beroep	Van en naar kantoor of fabriek	In vrije tijd	Tijdens vacantie
12 mnd. en meer						
11 mnd.						
10 ,,						
9 ,,						
8 ,,						
7 ,,						
6 ,,						
5 ,,						
4 ,,						
3 ,,						
2 ,,						
1 ,,						
Totaal	5.000					

[1]) Gedurende de laatste 12 maanden of gedurende de tijd dat de auto in het bezit was van de ondervraagde op het tijdstip van het onderzoek.

Opmerkingen

Aan bovenstaande tabel zou een tweede kunnen worden toegevoegd met de opgaven van de gemiddelden per auto en per maand. Deze gemiddelden worden verkregen door de cijfers in de kolommen 3-7 te delen door het aantal maanden vermeld in kolom 1. De cijfers in de eerste rij moeten daarbij worden gedeeld door 12. Deling van de cijfers in kolom 7 door het aantal maanden heeft minder betekenis omdat bij korte duur van het bezit de auto misschien nog niet voor vacantie werd gebruikt. Het zou voorts aanbeveling verdienen de standaardafwijkingen van de gemiddelden in kolom 3 te berekenen (eventueel ook van de cijfers in de andere kolommen) en deze in een afzonderlijke kolom op te nemen.

In verband met de vrij hoge „non-response" zijn de uitkomsten onzeker. *Deze* onzekerheid komt uiteraard niet tot uitdrukking in de berekende standaardafwijkingen.

Commentaar

De gemiddelde kilometrages moeten over alle auto's worden berekend, niet alleen over die auto's welke voor een of meer der opgegeven doeleinden zijn gebruikt. Hieruit volgt dat de gemiddelden in kolom 3 gelijk zijn aan de sommen der gemiddelden in de kolommen 4-7.

Eventueel zouden in een afzonderlijke tabel de aantallen auto's kunnen worden opgegeven, die werkelijk zijn gebruikt voor de vier doeleinden.

Samentrekking van de onderscheiding naar het aantal maanden van het autobezit in b.v. twee categorieen (meer dan zes maanden, resp. zes maanden en minder) is niet in overeenstemming met de vraagstelling.

Op een juiste omschrijving van de hoofden der kolommen en vermelding van de eenheden moet worden gelet.

Vermelding van het aantal auto's (kolom 2) is noodzakelijk.

Opgemerkt kan nog worden dat de vergelijking van de gemiddelden per maand voor hen, die de auto minder dan 12 maanden in hun bezit hadden, niet goed mogelijk is in verband met de seizoeninvloeden waaromtrent in de opgave echter geen inlichtingen worden gegeven.

Voor hen die de auto minder dan 12 maanden in hun bezit hadden, zijn de cijfers over het gemiddelde aantal kilometers gereden in de vacantie zonder nadere informatie moeilijk te interpreteren.