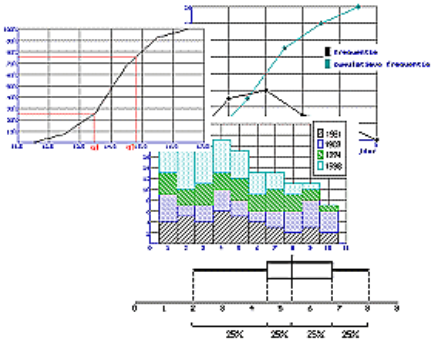
Reader statistiek voor de minor



Lerarenopleiding wiskunde
Willem van Ravenstein

oktober 2009
bijgewerkt maart 2011

https://www.wiswijzer.nl

Inhoudsopgave

Hoofdstuk 1 – Voorkennis en herhaling

Inleiding

Opgave 1

Hiernaast zie je een frequentieverdeling van de gemiddelde cijfers van een groep leerlingen.

gemiddeld cijfer	frequentie
1	1
2	1
3	3
4	4
5	4
6	8
7	5
8	4
9	3
10	1

a. Bereken het gemiddelde en de standaarddeviatie op 1 decimaal nauwkeurig.
b. Teken een somfrequentiepolygoon.
c. Bepaal m.b.v. het somfrequentiepolygoon de mediaan, q_1 en q_3.
d. Teken het boxplot.
e. Geef de modus.

Opgave 2

In een vaas zitten 10 knikkers. Er zijn 4 rode knikkers, 3 blauwe knikkers, 2 groene knikkers en 1 witte knikker. Je haalt 3 knikkers uit de vaas **zonder** terugleggen.

a. Bereken de kans op minstens 1 rode knikker
b. Bereken de kans op 3 rode knikkers.

We bekijken de volgende gebeurtenissen:

A: er is minstens 1 knikker rood (zie a.)
B: de knikkers zijn rood (zie b.)

c. Bereken P(A|B) en P(B|A)
d. Zijn de gebeurtenissen A en B onafhankelijk? (Leg uit!)

Opgave 3

- Beantwoord dezelfde vragen als bij opgave 2 maar nu **met** terugleggen.

Opgave 4

Een gemeenteraad bestaat uit negen leden. Er zijn vier leden van het CDA, twee leden van de PvdA, twee leden van de VVD en één lid van Groen Links. Er moet een commissie van drie leden worden samengesteld.

a. Hoeveel commissies kan je samenstellen?
b. In hoeveel van die commissies zitten geen leden van dezelfde partij?
c. Bereken de kans dat er in een willekeurig gekozen commissie minstens twee leden van dezelfde partij zitten.

Opgave 5

Sjors schaakt 3 keer tegen zijn ouders, afwisselend pa en ma. Als hij 2 keer achter elkaar wint, dan krijgt hij meer zakgeld. Hij mag dus kiezen ma-pa-ma of pa-ma-pa.

- Welke keus moet hij maken als gegeven is dat pa beter schaakt dan ma? (leg uit)

Opgave 6

Een vliegtuigmaatschappij heeft bij een vlucht waarbij er 116 passagiers mee kunnen 120 tickets verkocht. Er wordt verondersteld dat er algemeen 2,5% van de passagiers niet komt opdagen.

a. Bereken op 3 decimalen nauwkeurig de kans dat er meer dan 116 passagiers op komen dagen.
b. Bereken op 3 decimalen nauwkeurig de kans op minimaal 1 lege plaats tijdens de vlucht.

Opgave 7

Je trekt zonder terugleggen 4 kaarten uit een volledig kaartspel van 52 kaarten. Voor **elk** plaatje (boer, vrouw, heer of aas) wordt er 10€ uitgekeerd.

- Bereken de verwachtingswaarde van het uit te keren bedrag.

Opgave 8

De buurman van Joanneke geeft bijles. Aan 't einde van zo'n bijles gooit de leraar met een munt. Hij telt het aantal keren dat hij moet gooien tot hij kop gooit. Het aantal keren dat hij dan moet gooien vermenigvuldigd met 10 is de vergoeding in euro die Joanneke moet betalen voor de bijles.

- Bereken op 3 decimalen nauwkeurig de kans dat Joanneke meer dan 40 euro moet betalen.

Opgave 9

Op pakjes margarine staat meestal 250 gr e. Dit betekent dat volgens Europese norm niet meer dan 5% van die pakjes minder dan 250 gram mag bevatten.

a. De gewichten van pakjes Bona zijn normaal verdeeld en hebben een standaarddeviatie van 7 gram.
Bereken het gemiddelde gewicht zodat precies voldaan wordt aan de Europese norm op 1 decimaal nauwkeurig.
b. De pakjes margarine van de firma Fide hebben een gemiddeld gewicht van 256 gram.
Ook de gewichten van deze pakjes zijn normaal verdeeld en voldoen precies aan de Europese norm.
Bereken de standaarddeviatie op 2 decimalen nauwkeurig.

Opgave 10

Bij een productieproces in een fabriek moeten telkens twee buizen A en B aan elkaar worden gelast. Van de buizen A is de lengte normaal verdeeld met een gemiddelde van 75 cm en een standaardafwijking van 3 cm. Van buizen B is de lengte ook normaal verdeeld met een gemiddelde van 50 cm en een standaardafwijking van 2 cm.

Er worden telkens twee willekeurige buizen A en B gepakt.

- Bereken op 4 decimalen nauwkeurig de kans dat de totale lengte van de aan elkaar gelaste buizen meer is dan 130 cm.

Opgave 11

In het algemeen wordt er van uit gegaan dat het IQ van de 'gemiddelde Nederlander' normaal verdeeld is met een gemiddelde van 100 en een standaarddeviatie van 15. Ik heb altijd gedacht dat het IQ van een wiskundestudent op de lerarenopleiding hoger is dan 100. Om mijn vermoeden nader te onderzoeken neem ik een willekeurige steekproef van 25 studenten en onderwerp ze aan een IQ-test. Het gemiddelde IQ uit deze steekproef noemen we μ.

a. Leg uit dat de nulhypothese $\mu = 100$ moet zijn en dat je H_1 gelijk aan $\mu > 100$ neemt.
b. Bij welke grenswaarde (bij een significantieniveau van 0,05) kan ik H_0 verwerpen?

Uit mijn onderzoek blijkt het gemiddelde IQ in de steekproef gelijk te zijn aan 107.

c. Bereken de kans dat ik ten onrechte H_0 niet verwerp.

Opgave 12

In een textielfabriek worden rollen stof vervaardigd met een lengte van 50 meter per rol. Het aantal weeffouten per rol is Poisson-verdeeld met een bijbehorende verwachtingswaarde van 1 weeffout per rol.

Bij de kwalitatieve keuring van de rollen stof worden deze gescheiden in rollen van “A-kwaliteit” (met 0 of 1 weeffout per rol) en rollen van “B-kwaliteit” (met twee of meer weeffouten per rol).

a. Bereken de kans dat een willekeurige rol de aanduiding “B-kwaliteit” krijgt.
b. De productieomvang per dag is gelijk aan 2000 meter stof. Hoe groot is de kans dat er op een willekeurige dag tenminste 30 rollen met “A-kwaliteit” worden gemaakt?

Inleveropdracht A

Deze inleveropdracht bestaat uit twee opgaven.

Opgave 1

Een doosje sterrenmix thee bevat 15 zakjes. De inhoud van de zakjes is normaal verdeeld met gemiddelde 2.2 gram en een standaard afwijking 0.5 gram. Bereken de kans dat:

a. Een doosje minder dan 28 gram thee bevat
b. Het gemiddelde gewicht van 20 zakjes minder is dan 1.9 gram
c. Een doosje meer dan tien zakjes bevat die elk meer dan 2.1 gram thee bevatten.

Martin controleert net zo lang zakjes sterrenmix totdat hij eentje heeft gevonden die meer dan 3 gram thee bevat

d. Bereken de kans dat hij minder dan 20 zakjes hoeft te controleren

De fabrikant wil een nieuwe doosje op de markt brengen. Op de doosje staat dat de zakjes gemiddeld 2 gram thee bevatten.

e. Hoeveel zakjes moet de fabrikant minimaal in een doosje stoppen opdat deze bewering 99% van de doosjes klopt.

Opgave 2

Hier onder zie je de resultaten van 12 leerlingen op twee toetsen:

leerling	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1e toets	7	4	4	4	8	6	6	8	7	8	3	6
2e toets	8	8	4	7	7	9	9	5	9	6	7	9

- Onderzoek of je met een significantieniveau van 5% aannemelijk kan maken dat de tweede toets beter gemaakt is dan de eerste toets.

einde hoofdstuk 1

Hoofdstuk 2 – Telproblemen, kansrekening en kansverdeling

Telproblemen

Veel kansproblemen hebben te maken met **tellen**. Stel jezelf, voor je begint, de volgende **twee vragen**:

1. Is het met of zonder teruglegging?
2. Is de volgorde belangrijk?

Dit levert vervolgens 4 verschillende soorten telproblemen op:

		Met terugleggen?	
		Nee	**Ja**
Volgorde belangrijk?	**Ja**	**Permutaties** $(n)_k = \frac{n!}{(n-k)!}$ **faculteitsboom**	**Rangschikkingen met herhaling** $aantal = n^k$ **machtsboom**
	Nee	**Combinaties** $\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$ **ja-nee rooster**	**Herhalingscombinaties** $aantal = \binom{n-1+k}{k}$

Opgave 1

Kies eerst het juiste telmodel en bereken vervolgens het aantal mogelijkheden

a. Hoeveel verschillende dominostenen zijn er?[1]
b. Hoeveel verschillende getallen van 6 cijfers kan je vormen met de cijfers 1 t/m 6 waarbij je elk cijfer maximaal 1 keer gebruikt?
c. Op hoeveel manieren kan je 7 (verschillende) studenten kiezen uit een groep van 15?
d. Hoeveel pincodes kan je maken van 4 cijfers als er verder geen beperkingen zijn?

Opgave 2

a. Je hebt een (open) ketting met 7 rode, 4 witte en 3 blauwe kralen. Hoeveel verschillende kettingen kan je daarmee maken?
b. In een vaas zitten 10 rode, 15 witte en 20 blauwe knikkers. Je trekt 9 knikkers **met** terugleggen. Bereken de kans op 2 rode, 3 witte en 4 blauwe knikkers.

[1] Op een dominosteen stelt het aantal ogen op iedere helft van een steen een getal voor. De getallen kunnen 0,1,2,3,4,5 of 6 zijn. Alle mogelijke verschillende stenen komen in het spel voor.

Kansrekening

In veel schoolboeken kun je de volgende formule tegenkomen (Laplace):

$$\text{kans} = \frac{\text{aantal gunstige uitkomsten}}{\text{aantal mogelijke uitkomsten}}$$

Hierbij ga je er vanuit dat alle uitkomsten een gelijke kans hebben. Voor de berekening van dat aantal gunstige of mogelijke uitkomsten is het noodzakelijk dat je weet hoe je handig kunt tellen.

Opgave 3

Je gooit met 10 munten. Wat is de kans dat je 3 keer een kop gooit ?

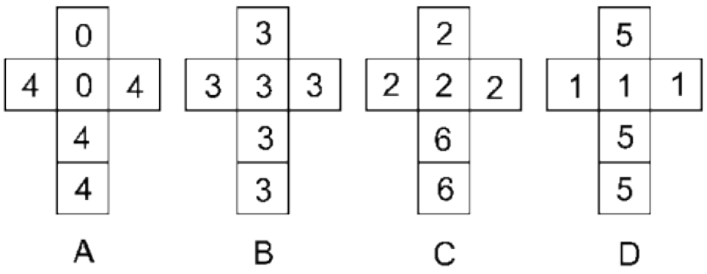
Opgave 4

Jan moet een toets maken van 20 meerkeuzevragen. Bij elke vraag kun je kiezen uit 4 antwoorden. Bij 12 vragen weet Jan het antwoord. Bij de andere vragen twijfelt hij tussen 2 antwoorden en moet dus gokken. Je mag ervan uitgaan dat hij bij de vragen die hij zeker weet alle antwoorden goed zijn en dat bij de vragen waarbij hij twijfelt tussen 2 antwoorden het goede antwoord er inderdaad bij zit. Om voor de toets voldoende te halen moet Jan minimaal 15 vragen goed hebben.

- Wat is de kans dat Jan een voldoende haalt?

Opgave 5

Je ziet hier de uitslagen van 4 dobbelstenen:



Laat iemand een dobbelsteen uitkiezen en wed vervolgens om een euro dat je hoger zult gooien.

Als je tegenstander A kiest, kies jij D. Kiest hij of zij B, dan kies je A. Mocht je tegenstander C kiezen, dan kies je B. Als hij of zij D kiest, dan kies je C. Geloof het of niet, je zult uiteindelijk vaker winnen, dan je tegenstander!

- Laat zien dat 'A' wint van 'B', 'B' van 'C', 'C' van 'D' en 'D' van 'A'.

Opgave 6

In een vaas zitten 8 witte, 4 blauwe en 2 rode ballen. We trekken steeds drie ballen uit de vaas zonder terugleggen.

- Bereken de kans op 2 witte ballen op 3 verschillende manieren.

Kansen optellen en/of vermenigvuldigen

Opgave 7

Op de vraag **"Je gooit met een munt en een dobbelsteen. Wat is de kans dat je kop of een vijf gooit?"** geeft iemand als antwoord:

> "De kans op kop is een $^1/_2$, de kans op vijf is $^1/_6$. Het is een **of-kans**, dus is de kans op een kop of vijf gelijk aan $^1/_2+^1/_6=^2/_3$."

a. Welke 'denkfout' maakt deze persoon?
b. Hoe zou je zelf de gevraagde kans berekenen?

Noem de gebeurtenis 'kop gooien' A en de gebeurtenis 'vijf gooien' B.

c. Laat zien dat geldt: $\mathrm{P(A\ of\ B)} = \mathrm{P(A)} + \mathrm{P(B)} - \mathrm{P(A\ en\ B)}$

Opgave 8

Je gooit met twee dobbelstenen. Je kijkt naar het totaal aantal ogen. We onderscheiden twee gebeurtenissen:

A: het totaal aantal ogen is even.
B: je gooit met beide dobbelstenen hetzelfde aantal ogen.

a. Bereken P(A), P(B), P(A en B) en P(A of B).
b. Laat zien dat geldt: $\mathrm{P(A\ of\ B)} = \mathrm{P(A)} + \mathrm{P(B)} - \mathrm{P(A\ en\ B)}$.
c. Is P(A)×P(B)=P(A en B)? Welke conclusie kan je daaruit trekken?
d. Bereken P(A|B) en P(B|A).

Opgave 9

We gooien met een rode en blauwe dobbelstenen. We onderscheiden de volgende gebeurtenissen:
A: blauw is meer dan rood B: rood en blauw zijn samen 7.

a. Bereken de kans P(A), P(B) en P(A en B)
b. Bereken de kans P(A of B)
c. Sluiten de gebeurtenissen A en B elkaar uit?
d. Zijn A en B onafhankelijk?

Opgave 10

Voor twee gebeurtenissen A en B geldt: P(A|B) = 0,30 en P(A en B) = 0,20.

- Bereken P(B)

Opgave 11

Voor twee gebeurtenissen A en B geldt P(A) = 0,30 en P(B) = 0,70. Tevens geldt dat de gebeurtenissen A en B onderling onafhankelijk zijn.

- Bereken P(A en B)

Opgave 12

- Bereken de kans om met 4 dobbelstenen een totaal van 10 ogen te gooien.

Inleveropdracht B

Opgave 1

Op tafel liggen 26 letters: A t/m Z. Hiermee leg ik, **zonder terugleggen**, 'woorden' van drie letters. Onder 'woorden' verstaan we alle mogelijke series van drie letters, dus ook onzinwoorden zijn toegestaan.

a. Hoeveel verschillende 'woorden' kan ik maken?
b. Hoeveel verschillende combinaties van drie letters kan je maken?
c. Wat is het 'verband' tussen de uitkomsten van **a.** en **b.**?

Opgave 2

Op tafel liggen 26 letters: A t/m Z. Hiermee leg ik, **met terugleggen**, 'woorden' van drie letters. Onder 'woorden' verstaan we alle mogelijke series van drie letters, dus ook onzinwoorden zijn toegestaan.

a. Hoeveel verschillende 'woorden' kan ik maken?
b. Hoeveel verschillende herhalingscombinaties van drie letters kan je maken?
c. Wat is het 'verband' tussen de uitkomsten van **a.** en **b.**?

Opgave 3

"Nederland is een mooi land, de helft van de tijd regent het, en als je met de auto gaat sta je 8 van de 10 keer in de file...."

- Kun je met deze gegevens berekenen wat de kans is dat als je morgen met de auto op weg gaat je met regen in de file staat? Zo ja, bereken die kans. Zo nee, leg uit.

Opgave 4

Er zijn drie dozen. In de eerste doos zitten twee witte ballen. In de tweede doos een witte en een zwarte bal. In de derde doos twee zwarte ballen. Eerst kiest men een doos en dan neemt men uit die doos een bal. Wat is de kans dat in de gekozen doos een zwarte bal overblijft, op voorwaarde dat een zwarte bal genomen is uit die doos?

einde hoofdstuk 2

Hoofdstuk 3 – Kansfuncties en kansverdelingen

De geometrische verdeling

We beschouwen een experiment waarbij de kans op 'succes' gelijk is aan p en de kans op mislukking gelijk is aan 1-p. We definiëren de stochast X als het aantal keren dat het experiment moeten doen totdat het eerste 'succes' optreedt.

Als de 'succes'-kansen van de opeenvolgende experimenten **onafhankelijk** van elkaar zijn, dan heeft X een **geometrische verdeling**.

Er geldt: $P(X = k) = (1 - p)^{k-1} \cdot p$

Voor de geometrische verdeling geldt: $E(X) = \frac{1}{p}$ en $Var(X) = \frac{1-p}{p^2}$

Opgave 1

De kans om met een dobbelsteen 'zes ogen' te gooien is $\frac{1}{6}$. We gooien met de dobbelsteen net zolang totdat we 'zes ogen' gooien.

a. Bereken de kans dat je na 5 keer gooien 'zes ogen' gooit.
b. Bereken de kans dat je meer dan 3 keer moet gooien.

De binomiale verdeling

We beschouwen n onafhankelijke experimenten met elk experiment een kans van p op 'succes'. De stochast X, die het totaal aantal 'successen' voorstelt, heeft een **binomiale verdeling** met parameters p en n.

Er geldt: $P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$

Voor de binomiale verdeling geldt: E(X)=n·p en Var(X)=n·p(1-p)

Opgave 2

Voor een tentamen algebra hebben zich 55 studenten ingeschreven. De docent heeft zich echter een beetje vertelt en heeft maar 50 toetsen gekopieerd. In het algemeen is bekend dat ongeveer 10% van de studenten die zich inschrijven voor een tentamen niet komt opdagen.

a. Bereken op 3 decimalen nauwkeurig de kans dat er meer dan 50 studenten op komen dagen.
b. Bereken op 3 decimalen nauwkeurig de kans dat er minimaal één toets overblijft.

Opgave 3

De laatste jaren worden er veel telefonische enquêtes gehouden. Uit onderzoek blijkt dat ongeveer 85% van de mensen bereid is mee te werken als ze gebeld worden. Voor een bepaald onderzoek wil men minstens 10 mensen ondervragen. Hoe vaak moet de enquêteur bellen om met een kans van 90% deze 10 mensen te bereiken?

Hypergeometrische verdeling

In een vaas bevinden zich **a** witte en **b** rode knikkers. Je pakt er **n** knikkers uit. De kans op **k** witte knikkers is dan gelijk aan:

$$P(X = k) = \frac{\binom{a}{k} \cdot \binom{b}{n-k}}{\binom{a+b}{n}}$$

Voorbeeld

In een klas zitten 12 jongens en 15 meisjes. Uit deze klas gaan 5 leerlingen een feest organiseren. Je kiest willekeurig 5 leerlingen. Wat is de kans dat er 3 jongens (en dus 2 meisjes) in dit comité zitten?

Oplossing

$$P(X = 3) = \frac{\binom{12}{3} \cdot \binom{15}{2}}{\binom{27}{5}} \approx 0,286$$

Opgave 4

In een werkgroep van 34 studenten zitten 20 mannen en 14 vrouwen. Uit deze groep worden willekeurig 5 kandidaten gekozen voor het organiseren van een werkweek.

a. Bereken op 3 decimalen nauwkeurig de kans dat precies 3 van de 5 kandidaten mannen zijn.
b. Bereken op 3 decimalen nauwkeurig de kans dat minimaal 3 van de 5 kandidaten mannen zijn.

Opgave 5

Bij het spel klaverjassen krijgt ieder van de 4 deelnemers 8 kaarten uit een spel van 32 kaarten. De stochast X is het aantal boeren dat Joost krijgt. Geef de kansverdeling van X .

Poissonverdeling

De Poisson-verdeling is een limiet geval van de binomiale verdeling (n groot en n·p vast). De kans op een bepaalde gebeurtenis bereken je met de volgende formule:

$$P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

De verdeling wordt alleen bepaald door de verwachtingswaarde λ. De standaarddeviatie is gelijk aan de wortel uit de verwachtingswaarde:

$$\sigma = \sqrt{\lambda}$$

Opgave 6

Gemiddeld worden 1 op 2000 huizen per jaar door brand vernield. Bereken de kans dat in een gemeente met 6000 huizen er 4 door brand vernield worden in 2003.

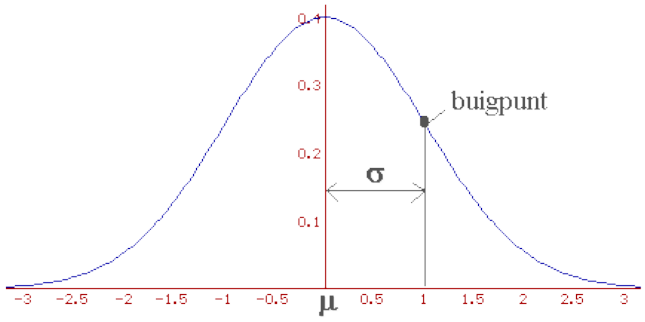
Normale verdeling

De normale verdeling is een continue kansverdeling. Kansverdelingen waarbij een continue variabele een rol speelt komen veel voor.

Een paar eigenschappen van een normale verdeling:

- klokvormig
- de oppervlakte onder de kromme komt overeen met 100% van de gegevens
- symmetrisch t.o.v. het gemiddelde
- gemiddelde, mediaan en modus vallen samen
- de verdeling wordt bepaald door de verwachtingswaarde en de standaarddeviatie
- vuistregels:
    - 68% van de gegevens wijkt op z'n hoogst één keer de standaarddeviatie af van de verwachtingswaarde
    - 95% van wijkt op z'n hoogst twee keer de standaarddeviatie af van de verwachtingswaarde

Voor een normale dichtheidskromme is het mogelijk de standaarddeviatie op het oog te schatten. De afstand van het buigpunt tot het centrum (gemiddelde en mediaan) is namelijk de standaarddeviatie.



Opgave 7

Een boomkweker koopt een grote partij jonge sparrenboompjes. Uit onderzoek is bekend dat de lengte van jonge sparrenboompjes bij benadering normaal verdeeld is met een gemiddelde van 25 cm en dat 5% van de boompjes korter is dan 20 cm. De partij jonge sparrenboompjes is te beschouwen als een aselecte steekproef.

a. Hoeveel procent van de boompjes is naar verwachting langer dan 30 cm? Licht je antwoord toe.
b. Bereken de standaardafwijking van de lengteverdeling van jonge sparrenboompjes. Geef je antwoord in twee decimalen nauwkeurig.

Na een aantal jaren wordt een groot aantal van deze sparrenboompjes voor de kerstverkoop gerooid. Je kunt er nu van uitgaan dat de lengte van deze partij bomen bij benadering normaal verdeeld is met een gemiddelde van 145 cm en een standaardafwijking van 15 cm.

c. Bereken de kans dat een aselect gekozen boom uit deze partij een lengte heeft die ligt tussen de 140 en de 170 cm. Rond je antwoord af op twee decimalen.

De bomen worden ingedeeld in twee prijsklassen, namelijk: kleine bomen van € 10,- per stuk en grote bomen van € 15,- per stuk. De kweker wil dat de te verwachten opbrengst per 100 bomen € 1300,- is.

d. Bereken bij welke lengte de grens tussen de beide prijsklassen dan moet liggen. Rond je antwoord af op hele centimeters.

Uit: het HAVO-examen wiskunde B1 van 2003

Negatief exponentiële verdeling

De negatief exponentiële verdeling is de **continue** variant van de geometrische verdeling. De negatief exponentiële verdeling kan je opvatten als 'hoe lang duurt het totdat een 'succes' optreedt. Een typisch voorbeeld is de tijd die zal verstrijken tot de eerstvolgende telefoonoproep wanneer er gemiddeld λ oproepen per tijdseenheid zijn.

Als T een (negatief) exponentiële verdeling heeft , dan is T een continue stochastische variabele met de volgende kansdichtheid:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} \; voor \; t \geq 0 \\ 0 \; voor \; t < 0 \end{cases}$$

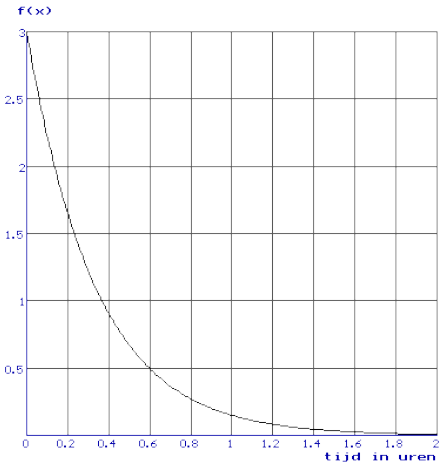
Met als verwachtingswaarde en variantie:

$$\mu = \frac{1}{\lambda}$$

$$\sigma^2 = \frac{1}{\lambda^2}$$

Opgave 8

Bij een telefonische hulpdiensten komen 3 gesprekken per uur binnen.



- Bereken de kans dat je langer dan 24 minuten moet wachten tussen twee oproepen.

Hoofdstuk 4 – Steekproeven, schatters en betrouwbaarheid

Bij beschrijvende statistiek verzamel je eerst gegevens die je vervolgens presenteert op een overzichtelijke en gestructureerde manier. De 'andere tak' van statistiek wordt 'verklarende' of 'inferentiële' statistiek genoemd. Daarbij worden juist vooraf bepaalde uitspraken gedaan over mogelijke uitkomsten van het onderzoek. Daarbij speelt kansrekening een belangrijke rol.

Wat is een steekproef?

Op grond van een steekproef proberen we conclusies te trekken over de populatie. De populatie kan je opvatten als de 'verzameling elementen' waarop het onderzoek betrekking heeft. Zo'n populatie kan uit een eindig aantal, maar ook uit een oneindig aantal elementen bestaan. Een aantal 'trekkingen' uit een populatie noemen we **steekproef**.

Voordat je gegevens gaat verzamelen zijn er 2 vragen die belangrijk zijn:

a. Welk verschijnsel, welke variabele gaan we onderzoeken?
b. Op welke manier moeten de gegevens verzameld worden?

Voor het 'opzetten' van een steekproef spelen steeds vier begrippen een rol:

1. **De populatie**
   De verzameling van alle elementen waarover het onderzoek een uitspraak gaat doen/
2. **Het steekproefkader**
   Dat is een administratieve weergave van de populatie. Je hoopt dat het steekproefkader overeenkomt met de werkelijke populatie. In de praktijk blijkt dat niet altijd het geval.
3. **De steekproef**
   De elementen die in de steekproef terecht komen worden getrokken uit het steekproefkader.
4. **De waarnemingen**
   De informatie die uit de steekproef naar voren komt. Die informatie kan onvolledig zijn, gewoon onjuist of niet aanwezig. Denk daarbij aan personen die het enquêteformulier niet invullen, niet terug sturen, e.d.

Zoals je ziet staat in de praktijk lang niet altijd vast dat de waarnemingen in een steekproef nog wel een goed beeld geven van de populatie.

Bij verklarende statistiek maken we onderscheid tussen **schattingsmethodes** en **toetsingsmethodes**. Als je een onbekende parameter van een kansverdeling onderzoekt dan spreken we van een **schattingsprobleem**. Als we op grond van voorkennis proberen een hypothese te controleren (=toetsen) dan spreken we van een **toetsingsprobleem**. In de hoofdstukken hiervoor heb je daar al voorbeelden van gezien.

Simpel gezegd moet een steekproef aan de volgende voorwaarden voldoen:

- De steekproef moet aselect zijn.
- De steekproef moet voldoende groot zijn.

Daarmee is al veel gezegd, maar weinig opgelost. Wat is aselect precies? Hoe doe je dat dan? Wat is dan voldoende groot? Waar hangt dat van af?

Met 'aselect' wordt bedoeld, dat bij de steekproeftrekking ieder individu of iedere eenheid in de populatie dezelfde kans heeft om in de steekproef terecht te komen. We spreken dan van een aselecte steekproef.

Grotere steekproeven geven betrouwbaarder schattingen dan kleinere steekproeven. Logisch ook wel, want bij kleinere steekproeven is de kans groot dat je toevallig...

Opdracht 1

Ik wil graag weten hoeveel deeltijdstudenten van jaar 1 al voor de klas staan. Voor de cursus 'Vakproject wiskunde en ICT' hebben zich 30 studenten aangemeld op Osiris. Op basis van die lijst kies ik **willekeurig** 5 studenten die ik ga vragen of ze al les geven. Het blijkt dan 3 van de 5 al les geven.

a. Wat is de populatie? Wat is het steekproefkader? Wat is mijn steekproef?
b. Dit is natuurlijk geen goede werkwijze. Leg uit waarom.

Neem aan dat in werkelijkheid 20% van de deeltijdstudenten van jaar 1 al voor de klas staat.

c. Is er op grond van onderzoek reden om te veronderstellen dat bovenstaande aanname **niet** klopt? (Neem α=0,05)

Opdracht 2

Een onderzoeksbureau kreeg van een gemeenteraad de vraag 'Hoe vol zijn de bussen in onze stad?' Men vroeg zich af 'Moeten we misschien meer bussen laten rijden?' Het bureau besloot een enquête te houden onder buspassagiers. Aan een groot aantal mensen die rondliepen op het busstation vroeg men simpelweg: 'Met hoeveel mensen zat U in de bus?' Het gemiddelde antwoord van al deze mensen zou dan immers wel de gemiddelde busbezetting zijn.....? Of toch niet? Wat vind je er van?

Schatters

Na het verzamelen van de gegevens door middel van een steekproef kan je de zogenaamde steekproefgrootheden worden bepaald. Je kunt daarbij denken aan het steekproefgemiddelde, de steekproefvariantie of de steekproeffractie.

Gegeven: $x_1, x_2, ..., x_n$

- Het steekproefgemiddelde: $\bar{x} = \frac{\sum x_i}{n}$
- De steekproefvariantie: $s^2 = \frac{\sum (x_i - \bar{x})}{n-1}$
- De steekproeffractie: $p = \frac{k}{n}$

Deze grootheden spelen een belangrijke rol bij schattingsproblemen, omdat verwacht mag worden dat bij een continue verdeling $\bar{x}$ en s iets zeggen over μ en σ van de populatie. Bij de binomiale verdeling zegt p iets over de onbekende populatiefractie π.

De bovengenoemde grootheden kan je opvatten als **schatters** van de onbekende parameters van de populatie.

Betrouwbaarheid en schattingsinterval

Meestal wordt zo'n schatter niet gegeven als één waarde (puntschatter) maar als interval. We spreken dan van een schattingsinterval of betrouwbaarheidsinterval. De breedte van zo'n interval hangt af van de standaarddeviatie en de gewenste kans op een goede schatting. De kans op een goede schatting heet **betrouwbaarheid** van het interval. Meestal kiezen we daarvoor 95% of 99%.

De breedte van een schattingsinterval geeft, bij een gegeven betrouwbaarheid, de nauwkeurigheid van de schatting aan.

Voorbeeld 1

De levensduur van een bepaald type batterij heeft een onbekende verwachtingswaarde μ en een standaarddeviatie van σ=5 uur. Voor n=100 batterijen wordt vervolgens de gebruiksduur bepaald. Dit levert $\bar{x}$ = 46.00 uur op.

We willen het schattingsinterval bepalen met een betrouwbaarheid van 95%. In het interval ligt met een kans van 95% die onbekende μ. Bij een kans van 95% hoort een z-score van -1,96 en 1,96. Als je daar ook nog bij bedenkt dat $s = \frac{5}{\sqrt{100}}$ kan je grenzen als volgt berekenen:

$$\left.\begin{array}{l} \dfrac{46-\mu}{\frac{5}{\sqrt{100}}} = -1.96 \Rightarrow \mu = 46.98 \\ \dfrac{46-\mu}{\frac{5}{\sqrt{100}}} = 1.96 \Rightarrow \mu = 45.02 \end{array}\right\} \Rightarrow 45.02 < \mu < 46.98$$

Je kan ook zeggen dat het betrouwbaarheidsinterval bestaat uit alle waarde van μ die een afstand tot $\bar{x}$ die hoogstens $z \cdot \frac{\sigma}{\sqrt{n}}$ bedraagt.

FORMULE

$$\bar{x} - z \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z \cdot \frac{\sigma}{\sqrt{n}}$$

Opdracht 3

Bereken het betrouwbaarheidsinterval van het voorbeeld bij een betrouwbaarheid van 99%.
(tip: welke z=waarde hoort er bij een kans van 99%?)

Opdracht 4

Een machine vult zakken met meel. Het gewicht van de meel in een willekeurige zak is te beschouwen als een trekking uit een normale verdeling met onbekende μ en een standaarddeviatie van 100 gram.

Om het schattingsinterval te berekenen voor μ weegt men de inhoud van 25 zakken meel. Het gemiddelde van de meel per zak bedroeg 20,142 kg.

a. Bereken een schattingsinterval voor μ dat een betrouwbaarheid heeft van 95%.
b. Bereken een schattingsinterval voor μ dat een betrouwbaarheid heeft van 99%.

Normaal gesproken zou je dan nu moeten kijken naar betrouwbaarheidsintervallen voor fracties. Dat slaan we in even over. Ook gaan we niet kijken naar betrouwbaarheidsintervallen van bijvoorbeeld de poissonverdeling.

Berekening van de steekproefomvang

Je kunt aan de **formule** goed zien dat hoe groter je n kiest hoe 'smaller' het schattingsinterval wordt:

$$\bar{x} - z \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z \cdot \frac{\sigma}{\sqrt{n}}$$

Anders gezegd: $\left|\bar{x} - \mu\right| \leq z \cdot \frac{\sigma}{\sqrt{n}}$

Neem aan dat we μ bepalen met een toegestane afwijking **a** (bij een gegeven betrouwbaarheid). Dan geldt:

$$z \cdot \frac{\sigma}{\sqrt{n}} \leq a$$

Daarmee kan je de steekproefomvang **n** berekenen:

$$n \geq \frac{z^2 \sigma^2}{a^2}$$

Opdracht 5

Ter bepaling van het koolmonoxidegehalte van een gasmengsel is een vrij onnauwkeurige bepalingsmethode beschikbaar. Per proefmonster kan de uitkomst beschouwd worden als een kansvariabele X met verwachtingswaarde μ% (het koolmonoxidegehalte van het gasmengsel waaruit de monsters genomen worden) en een standaarddeviatie van 4%.

- Gevraagd wordt om een 95%-betrouwbaarheidsinterval voor het gehalte koolmonoxide van het gasmengsel, waarbij de toegestane marge 1% naar boven en naar beneden mag zijn. Bereken daartoe het aantal benodigde monsters om op grond van het gemiddelde koolmonoxidegehalte van deze monsters een schattingsinterval van de geëiste nauwkeurigheid te kunnen berekenen.

Schatten van de variantie

Het bepalen van de steekproefomvang van een normaal verdeelde stochast X met bekende σ is niet zo'n probleem. In de praktijk komt dat echter maar weinig voor. Meestal zijn μ en σ allebei onbekend. Het bepalen van een schattingsinterval wordt dan moeilijker omdat je eerst (op grond van dezelfde steekproefgegevens) de variantie moet schatten.

We bestuderen een willekeurige continu variabele X met onbekende verwachtingswaarde E(X)= μ en onbekende variantie Var(X)= σ. We doen een steekproef waarvan de uitkomsten $x_1, x_2, ..., x_n$ zijn.

We berekenen eerst $\bar{x} = \frac{\sum x_i}{n}$. De afwijking die de afzonderlijke waarnemingen x_i vertonen ten opzichte van $\bar{x}$ geven ons dat een indruk van de spreiding van de variabele X.

$$s^2 = \frac{\sum (x_i - \bar{x})}{n-1}$$

De grootheid s^2 is een puntschatter, maar soms kan je ook een betrouwbaarheidsinterval aangeven voor σ^2. Maar dat gaan we niet doen…☺

We hebben nu 'slechts' één geschatte standaarddeviatie **s**. Bij een herhaling van de steekproeftrekking zal je niet dezelfde waarde vinden. Voor het bepalen van het betrouwbaarheidsinterval van μ heb je te maken met een extra element van onzekerheid.

De t-verdeling

De t-verdeling wordt als volgt gedefinieerd:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \text{ waarin } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Dit lijkt erg veel op de normale verdeling die we gebruikte bij het bepalen van betrouwbaarheidsintervallen bij een gegeven σ. Het verschil is dat we bij de t-verdeling een 'geschatte standaarddeviatie' gebruiken terwijl we bij de normale verdeling een exact bekende standaarddeviatie gebruikte.

Op dezelfde manier kunnen we nu het betrouwbaarheidsinterval voor μ bij een onbekende standaarddeviatie schrijven als:

$$\bar{x} - t \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t \cdot \frac{s}{\sqrt{n}}$$

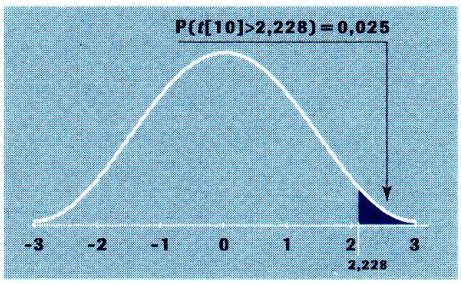
Zoek de verschillen! ☺

Van de kansverdeling van de variabele **t** zijn tabellen opgesteld. Als $n \to \infty$ gaat de t-verdeling naar de normale verdeling. Voor elke steekproefomvang heb je een andere t-verdeling. Men spreekt hier van v=n-1 als het aantal **vrijheidsgraden** waarmee σ^2 geschat is.

In de tabel op de volgende pagina kan je bij een gegeven overschrijdingskans (bijvoorbeeld 5% of $2\frac{1}{2}$%) een t-waarde aflezen als het aantal vrijheidsgraden **v** gegeven is.

Tabel F: de *t*-verdeling

Figuur A.6



Aantal vrijheidsgraden v	rechteroverschrijdingskans				
	.1	.05	.025	.01	.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Dat kan ook met je TI83/TI83+/TI84:

- Command Summary: Calculates the Student's t probability between lower and upper for specified degrees of freedom.
- Command Syntax: **tcdf(lower, upper, df)**

Opdracht 6

Zoek naar de relatie tussen de functie op je GR en de tabel.

Voorbeeld

Hieronder zie je het alcoholpercentages van 10 monsters van een grote partij wijn:

12,4 11,8 12,0 11,7 12,1 12,3 11,9 11,6 11,9 en 12,3.

Gevraagd: een 95%-schattingsinterval voor μ.

Uitwerking

Bepaal eerst het gemiddelde en standaarddeviatie van de steekproef.

$\bar{x} = 12$ en $s^2 = 0,073$, dus $s = 0,27$

$s_{\bar{x}} = \frac{0,27}{\sqrt{10}} = 0,086$

Het gaat om een t-verdeling met v=n-1=9 vrijheidsgraden. De t-verdeling geeft bij 5% overschrijdingskans een tabelwaarde van 2,26.

$$\bar{x} - t_{0,025} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0,025} \cdot \frac{s}{\sqrt{n}}$$

$$12 - 2.26 \cdot 0.086 < \mu < 12 - 2.26 \cdot 0.086$$

$$11.81 < \mu < 12.19$$

De t-verdeling is belangrijk in tal van toepassingsgebieden. In feite berust dit op de gangbare praktijk bij bijna alle onderzoeken omdat men meestal de standaarddeviatie van de variabelen die men onderzoek niet kent.

Opdracht 7

Bij een postorderbedrijf wil men een schatting maken van de gemiddelde tijdsduur van telefoongesprekken met klanten. We gaan ervan uit dat de gespreksduur **t** een kansvariabele is met een normale verdeling. Voor 9 gesprekken werd als tijdsduur gevonden (in seconden):

218, 225, 230, 240, 214, 202, 204, 195 en 180.

- Bereken een 95%-betrouwbaarheidsinterval voor μ: gemiddelde tijdsduur.

Opdracht 8

De handvatten van koffers moeten van zodanige kwaliteit zijn dat ze niet breken bij een forse belasting van de koffers. Ter controle van de kwaliteit van de handvatten werden er proeven genomen waarbij het gewicht dat aan een handvat hing net zolang werd opgevoerd tot het handvat brak. Er werden 9 handvaten getest waarbij de maximale belasting werd vastgesteld.

handvat	1	2	3	4	5	6	7	8	9
breuk bij een gewicht (in kg) van	84	87	81	85	90	93	86	88	80

a. Geef een schatting van de variantie.
b. Geef een betrouwbaarheidsinterval voor μ_x : de verwachtingswaarde van de maximale belasting van een handvat. Kies een 99%-betrouwbaarheidsinterval.

Wanneer de t-verdeling?

De t-verdeling wordt gerekend tot de kansverdelingen die een bepaalde band met de normale verdeling hebben. De t-verdeling typeert men wel als de verdeling voor kleine steekproeven bij onbekende σ. In de praktijk werkt men met de t-verdeling bij 30 of minder waarnemingen. Bij n>30 gebruikt men meestal de normale verdeling omdat er dan nauwelijks verschil bestaat tussen de twee verdelingen.

Inleveropdracht C

Hier onder zie je de resultaten van 12 leerlingen op twee toetsen:

leerling	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1e toets	7	4	4	4	8	6	6	8	7	8	3	6
2e toets	8	8	4	7	7	9	9	5	9	6	7	9

- Onderzoek of je met een significantieniveau van 5% aannemelijk kan maken dat de tweede toets **gemiddeld** beter gemaakt is dan de eerste toets.

Hoofdstuk 5 – Toetsen, chi-kwadraatverdeling en verschiltoetsen

Bij hypothesetoetsen gaat het steeds om de beslissing of je gegevens voldoende grond geven om de nulhypothese te kunnen verwerpen. Op basis van de nulhypothese kun je dan een **toetsingsgrootheid** formuleren en kun je allerlei kansen uitrekenen. Je kunt (net als bij betrouwbaarheidsintervallen) bepalen welke uitkomsten je nog vindt passen onder de nulhypothese en welke uitkomsten als uitzonderlijk moeten worden bestempeld (gegeven H_0).

Toetsen, voorspellingsinterval en kritieke gebied

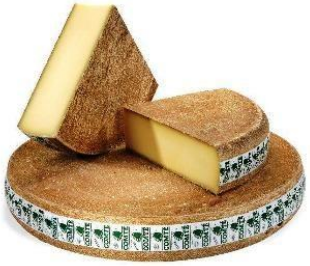
Met de toetsingsgrootheid op basis van de nulhypothese kan je een **voorspellingsinterval** bepalen. Meestal is dat 95% en alle uitkomsten die in het voorspellingsinterval vallen zijn geen reden om de nulhypothese te verwerpen. Wat overblijft is een kans van 0,05. Deze kans geven we meestal aan met α. Dat is de kans dat je onder H_0 een uitkomst vindt die buiten het voorspellingsinterval valt.

We spreken in plaats van voorspellingsinterval ook wel over **acceptatiegebied**. De verzameling uitkomsten die leidt tot verwerpen van de nulhypothese noemen we **kritieke gebied Z**.

Je hebt al eerder hypothesetoetsen gedaan met de binomiale verdeling en de normale verdeling. Maar je kunt natuurlijk ook hypothesetoetsen met de Poissonverdeling of de t-verdeling.

Opdracht 1

Volgens een richtlijn van het ministerie mag het gemiddelde aantal van een bepaald type bacterie per monster Franse kaas hoogstens 900 zijn. Van een grote partij werden 10 monsters genomen. De leverden als onderzoeksresultaat:



920, 960, 915, 910, 935, 965, 930, 970, 945 en 950.

- Onderzoek of deze resultaten uitwijzen dat aan de kwaliteitseis is voldaan. Kies α=0,01

Opdracht 2

Een fabrikant van een bepaald geneesmiddel beweert dat dit geneesmiddel in minstens 99% van de gevallen doeltreffend is. Om deze garantiebepaling te controleren worden 200 proefpersonen ondervraagd over de werkzaamheid van dit middel. In 8 gevallen had het middel geen succes.

- Onderzoek of de fabrikant desondanks gelijk kan hebben. Kies α=0,01.

Opdracht 3

We gooien 60 keer met een dobbelsteen. Van die 60 keer gooi je 13 keer een ‘zes’ (zes ogen).

- Is er reden om aan te nemen dat deze dobbelsteen niet zuiver is? α=0,05.

De chi-kwadraatverdeling

De chi-kwadraat wordt veel gebruikt. Met name bij het analyseren van kruistabellen waarin verbanden tussen (meestal nominale) variabelen zichtbaar worden gemaakt.

Voorbeeld 1

Bij een bankkantoor onderzoekt men de gang van zaken bij de afbetaling van persoonlijke leningen. Voor 100 verstrekte leningen werd bekeken of deze volledig volgens de voorwaarden afgelost zijn. Bij 30 leningen bleken er problemen te zijn geweest. We noemen dit wanbetalers, bij 70 leningen is de afbetaling correct verlopen. De bank wil onderzoeken of het betaalgedrag afhangt van de leeftijd van de leners. Daarom is tevens vastgesteld wat de leeftijd van iedere lener was.

Leeftijd van de lener	betaalgedrag		totaal
	wanbetalers	correct	
Jonger dan 40	24	36	60
40 of ouder	6	34	40
Totaal	30	70	100

De tabel hierboven noemen we de **observed** tabel. Om te onderzoeken in hoeverre deze tabel afwijkt van wat je zou verwachten als betaalgedrag en leeftijd onafhankelijk zouden zijn bepalen we de **expected** tabel.

Leeftijd van de lener	betaalgedrag		totaal
	wanbetalers	correct	
Jonger dan 40	18	42	60
40 of ouder	12	28	40
Totaal	30	70	100

Het vergelijken van **theoretische** en de **waargenomen** frequenties doen we volgens een methode die men de χ^2-toets noemt. Bij deze methode wordt een toetsingsgrootheid χ^2 berekend volgen de volgende formule:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$ met:

O_i=de waargenomen frequentie (observed)
E_i=de frequentie volgens de nulhypothese (expected)

Deze grootheid heeft (gegeven de nulhypothese) bij benadering een χ^2-verdeling met v=(n-1)(m-1) vrijheidsgraden. Hierbij is n het aantal keuze mogelijkheden bij de eerste variabele en m het aantal keuzemogelijkheden bij de tweede indeling.

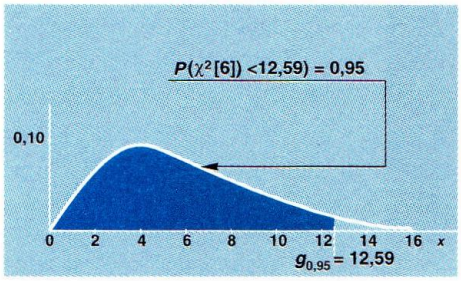
H_0: de variabelen zijn onafhankelijk
H_1: ze zijn niet onafhankelijk

$$\chi^2 = \frac{(24-18)^2}{18} + \frac{(36-42)^2}{42} + \frac{(6-12)^2}{12} + \frac{(34-28)^2}{28} = 2 + 0{,}86 + 3 + 1{,}29 = 7{,}14$$

Het aantal vrijheidsgraden: v=(2-1)(2-1)=1

Tabel G: enkele kritieke grenzen voor de χ^2-verdeling

Aangegeven zijn: grenswaarden g_α met $P(\chi^2 \leqq g_\alpha) = \alpha$
v is het aantal vrijheidsgraden



v	1	2	3	4	5	6	7	8	9	10
$g_{0,01}$	–	0,02	0,11	0,30	0,55	0,87	1,24	1,65	2,09	2,56
$g_{0,025}$	–	0,05	0,22	0,48	0,83	1,24	1,69	2,18	2,70	3 25
$g_{0,05}$	–	0,10	0,35	0,71	1,15	1,64	2,17	2,73	3,33	3,94
$g_{0,95}$	3,84	5,99	7,81	9,49	11,07	12,59	14,07	15,51	16,92	18,31
$g_{0,975}$	5,02	7,38	9,35	11,41	12,83	14,45	16,01	17,53	19,02	20,48
$g_{0,99}$	6,63	9,21	11,34	13,28	15,09	16,81	18,48	20,09	21,67	23,21

v	12	14	16	18	20	25	30	50	100
$g_{0,01}$	3,57	4,66	5,81	7,01	8,26	11,52	14,95	29,71	70,06
$g_{0,025}$	4,40	5,63	6,91	8,23	9,59	13,12	16,79	32,36	74,22
$g_{0,05}$	5,23	6,57	7,96	9,39	10,85	14,61	18,49	34,76	77,93
$g_{0,95}$	21,03	23,68	26,30	28,87	31,41	37,65	43,77	67,50	124,34
$g_{0,975}$	23,34	26,12	28,85	31,53	34,17	40,65	46,98	71,42	129,56
$g_{0,99}$	26,22	29,14	32,00	34,81	37,57	44,31	50,89	76,15	135,81

We gaan nu het kritieke gebied bepalen bij α=0,05. Voor 1 vrijheidsgraden levert de tabel voor de χ^2-verdeling een grenswaarde $g_{0,95} = 3,84$.

Het kritieke gebied Z={χ^2| χ^2>3,84}. We hadden 7,14 berekend. De nulhypothese wordt verworpen. We hebben aangetoond dat betaalgedrag en leeftijd afhankelijk zijn.

Opgave 4

Hieronder zie je een tabel met voor- en tegenstanders van de Euro verdeeld naar politieke voorkeur.

	CDA	D'66	PvdA	VVD	Totaal
Voor de Euro	96	84	106	154	440
Tegen de Euro	154	38	100	128	420
Totaal	250	122	206	282	**860**

- Onderzoek of de twee genoemde indelingen onderling onafhankelijk zijn. Kies α=0,05.

Voorbeeld 3

Een personeelsdirecteur van een groot bedrijf wil graag nagaan hoe groot het ziekteverzuim op de verschillende dagen van de week is. Een bepaalde week leverde voor het aantal ziekteverzuimdagen de volgende resultaten:

Dag	Aantal zieken
Maandag	20
Dinsdag	14
Woensdag	14
Donderdag	12
Vrijdag	20
Totaal	80

We willen toetsen of het aantal ziektedagen gelijkmatig over de dagen van de week verdeeld is. Met 80 zieken en 5 dagen zou je 'verwachten' dat als het gelijkmatig verdeeld is dat er elke dag 16 zieken moeten zijn.

Dag	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
Maandag	20	16	4	16	1
Dinsdag	14	16	-2	4	0,25
Woensdag	14	16	-2	4	0,25
Donderdag	12	16	-4	16	1
Vrijdag	20	16	+4	16	1
Totaal	80	16	0		**3,5**

We vinden $\chi^2=3,5$.

Het aantal vrijheidsgraden is 4.

Bij $\alpha=0,05$ vinden we een grenswaarde van $g_{0,95}=9,49$, dus $Z=\{\chi^2 \mid \chi^2>9,49\}$

De gevonden waarde ligt niet in het kritieke gebied, dus H_0 wordt niet verworpen.

Opgave 5

We gooien 60 keer met een dobbelsteen. Hieronder zie de resultaten.

Aantal ogen	1	2	3	4	5	6
Waargenomen frequentie	7	10	12	9	9	13

- Is er reden om aan te nemen dat deze dobbelsteen niet zuiver is? Neem $\alpha=0,05$.

Hier zou nog iets kunnen over 'aanpassing bij een normale verdeling'. Dat is wel geinig.

Verschiltoetsen voor μ

We komen nog terug op verschiltoetsen voor μ bij twee onafhankelijke steekproeven. Het gaat steeds om de vraag: kunnen de twee populaties waaruit de steekproeven getrokken zijn dezelfde waarde voor μ hebben?

1. Verschiltoets voor μ bij gegeven varianties
2. Verschiltoets voor μ bij onbekende varianties (variant A)
3. Verschiltoets voor μ bij onbekende varianties (variant B)

We kijken steeds naar twee kansvariabelen X en Y. Beide stochasten zijn normaal verdeeld. We kijken naar de verschilvariabel V waarvoor geldt: V = X – Y.

We nemen dan steeds als nulhypothese dat μ_X=0 (er is geen verschil) en de alternatieve hypothese $\mu_X \neq 0$ (er is **wel** verschil).

De 'kunst' is nu steeds om de standaarddeviatie van V te schatten. Dat is dan steeds bij bovengenoemde verschiltoetsen anders. Hieronder zie je daar een overzicht van met toelichting.

1. Verschiltoets voor μ bij gegeven varianties

   $$V \sim N\left(\mu_V = 0, \sigma_V = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right)$$

2. Verschiltoets voor μ bij onbekende varianties (variant A)
   Als je mag veronderstellen dat de varianties van X en Y aan elkaar gelijk zijn dan kan je de schatters s_X^2 en s_Y^2 combineren tot één schatter:

   $$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n-1+m-1}$$ (pooled variance)

   Je gebruikt dan de t-verdeling met v = n + m – 2.

   $$s_V = \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}$$

3. Verschiltoets voor μ bij onbekende varianties (variant B)
   Als de twee variantieschattingen onderling sterk verschillen (minstens een factor 4) dan mag je variant A niet gebruiken. Voor de standaarddeviatie van V gebruik je:

   $$s_V = \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

   En dan de t-verdeling met v = n + m – 2.

Opgave 6

De variabele X is normaal verdeeld met onbekende μ en σ=10. De variabele Y is normaal verdeeld met onbekende μ en $\sigma=\sqrt{30}$. Van de variabele X worden 10 trekkingen gedaan. Deze hebben een gemiddelde van 50. Er worden 5 trekkingen gedaan van Y met een gemiddelde van 55.

- Toets of beide variabelen dezelfde verwachtingswaarde kunnen hebben. Neem α=0,05.

Opgave 7

Van werknemers van een bepaalde bedrijfstak wordt het verband tussen jaarsalaris en het bezit van een vakdiploma onderzocht. Groep 1 is een steekproef uit de werknemers met diploma, groep 2 is een steekproef van werknemers zonder diploma. De resultaten (in duizenden guldens):

Groep 1	40	45	48	33	42	35	32	47	38	37	43
Groep 2	25	28	30	35	38	32	22				

Stel dat beide inkomensverdelingen beschouwd mogen worden als normale verdelingen met dezelfde variantie.

- Toets of $\mu_1 = \mu_2$. Neem α=0,05.

Opgave 8

Dezelfde vraag als bij 7, maar nu als je er van uit gaat dat de varianties van beide inkomensverdelingen sterk verschillen (variant B).

Hoofdstuk 6 – Meetniveau, regressie, correlatie en samenhang

Bij veel onderzoeken gaat het om het zoeken naar verbanden tussen variabelen. Is er een verband tussen geslacht en politieke voorkeur? Hangt inkomen samen met het niveau van onderwijs? Hebben bepaalde eetgewoontes een verband met bepaalde lichamelijke klachten? Enz.

Regressie en correlatie zijn belangrijke hulpmiddelen om dergelijke verbanden weer te geven. Bij regressie gaat het vooral om het aangeven van de **richting** van zo'n verband. Bij correlatie geven door middel van één getal de **sterkte** van zo'n verband aan. Er zijn verschillende maten voor correlatie die o.a. afhangen van het **meetniveau** van de onderzochte variabelen.

Meetniveaus

In de statistiek onderscheiden we doorgaans 4 verschillende soorten variabelen:

Nominale schaal

Hier heeft de waarde die een variabele kan aannemen alleen de betekenis van een naam. Er is geen sprake van een volgorde. Denk bijvoorbeeld aan rugnummers van een voetbalelftal. Je kunt niet op grond van het rugnummer beweren dat de speler met rugnummer 14 beter is dan de speler met rugnummer 7.

Ordinale schaal

Een waarde op ordinaal niveau geeft alleen een volgorde aan. Denk bijvoorbeeld aan opleiding.

Voorbeeld:
1=VO
2=MBO
3=HBO
4=UNIVERSITEIT

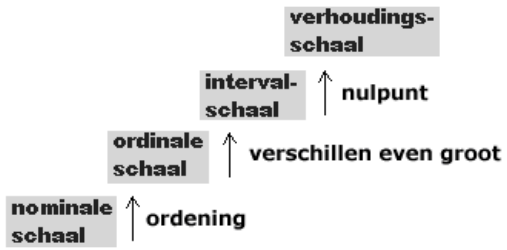
Er is wel sprake van **volgorde**, een hoger nummer duidt op een hogere opleiding, maar het verschil tussen bijvoorbeeld 3 en 4 is niet hetzelfde als het verschil tussen 1 en 2.

Intervalschaal

Bij variabelen op intervalniveau hebben verschillen wel een betekenis. Neem bijvoorbeeld temperatuur. Het verschil tussen 30° en 40° is hetzelfde als het verschil tussen 70° en 80°. Je kunt echter niet zeggen dat 80° twee keer zo warm is als 40°.

Verhoudingsschaal

Bij variabelen op rationiveau heb je altijd een nulpunt. Denk aan gewicht, lengte of het aantal verkochte exemplaren van een product. Je kunt uitspraken doen 'A scoort twee keer zo hoog als B'.



Opdracht 1

Neem over en vul in:

Variabele	Nominaal	Ordinaal	Interval	Verhouding
Geslacht				
Leeftijd in jaren				
Lengte in meter				
Gewicht in kg				
Hoogst afgeronde opleiding				
Soort rijbewijs				
Aantal kinderen				
Gemiddeld aantal glazen alcohol per week				
Aantal jaren in Nederland				
IQ				

Regressie

Als je te maken hebt met één variabele X die van invloed is op een andere variabele Y dan kan je proberen de invloed van X op Y te onderzoeken met behulp van **enkelvoudige regressie**. De meest toegepaste methode is te veronderstellen dat het om een lineair verband gaat. De rol die de variabelen spelen is niet symmetrisch. De ene variabele wordt beschouwt als het gevolg van de andere. Er is een **onafhankelijke** variabele X en een **afhankelijke** variabele Y.

Voorbeelden

- Bij een chemische reactie is de hoeveelheid omgezette stof (Y) afhankelijk van de temperatuur (X) in de reageerbuis.
- Het inkomen (Y) van iemand is afhankelijk van het aantal jaren onderwijs (X) dat deze persoon genoten heeft.
- De lengte van een persoon (Y) hangt af van het geslacht (X).

Het model

Bij enkelvoudige lineaire regressie gaan we op zoek naar een vergelijking van de **best passende lijn** met als vergelijking:

$$Y = \alpha + \beta \cdot X$$

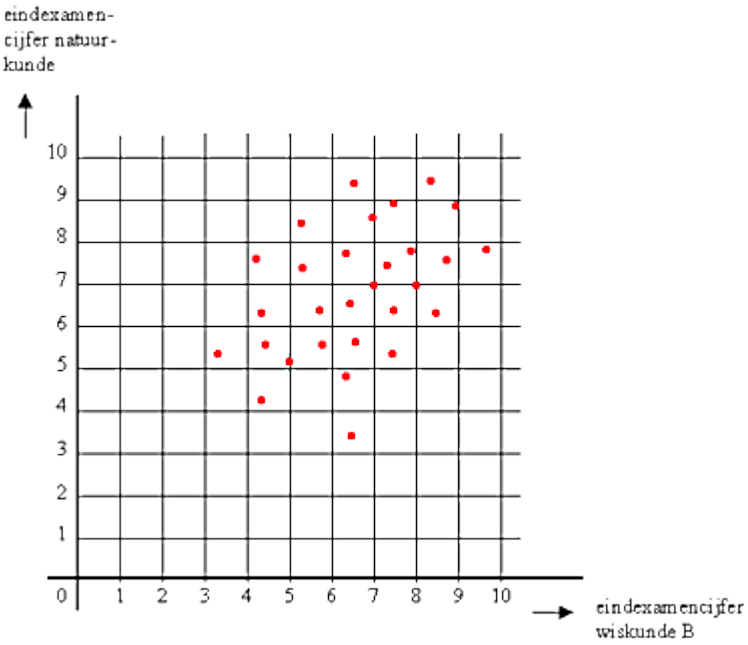
De 'kunst' is om de waarde van α en β zo goed mogelijk te schatten. Die α en β zijn kenmerken van de populatie. Meestal heb je echter te maken met een steekproef van getallenparen. Op basis van die getallenparen ga je proberen uitspraken te doen over de populatie. Dat is feitelijk een schattingsprobleem. Met de beschikbare gegevens gaan we op zoek naar de lijn Y=a+bX die het beste past bij de waargenomen getallenparen en schatten daarmee de α en β van de populatie.

Spreidingsdiagram

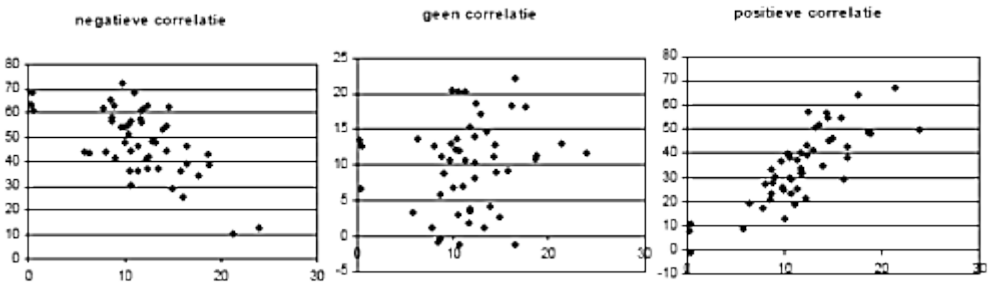
Om een indruk te krijgen van de samenhang tussen de variabelen X en Y kunnen we de waargenomen getallenparen in een grafiek weergeven. Zo'n grafiek heet een **spreidingsdiagram**. Zo'n diagram geeft je een idee over de aard van het verband. Meestal zijn we op zoek naar een lineair verband maar anders soorten verbanden kunnen natuurlijk ook.

Voorbeeld

Hieronder zie je het spreidingdiagram. Je ziet de getallenparen van de eindexamencijfers van een aantal leerlingen voor ‘wiskunde B’ (X) en ‘natuurkunde’ (Y). Elk ‘stipje’ stelt dus een gepaarde waarneming (de cijfers van een leerling) voor.



Hieronder zie je nog een aantal voorbeelden:



Je ziet dat een correlatie negatief, nul en positief kan zijn.

Opdracht 2

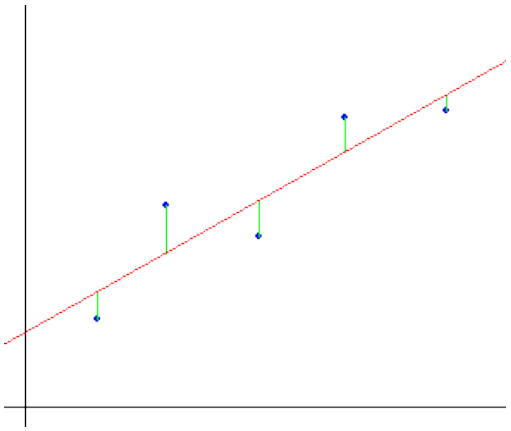
De resultaten van 10 studenten voor hun test (T) en hun examen (E) zijn gegeven in de onderstaande tabel:

T	10	12	8	13	9	10	7	14	11	6
E	11	14	9	13	9	9	8	14	10	6

- Teken een spreidingsdiagram en geef een vergelijking van de best passende lijn (op basis van je eigen gevoel).

De kleinste-kwadratenmethode

De vraag is nu: hoe bepaal je de **best passende lijn** door zomaar wat verspreide punten?
De methode daarvoor noemt men **kleinste kwadraten methode**.



De gevraagde lijn (rood) wordt zo getekend dat de kwadraten van de afwijkingen (de groene lijnstukjes) het kleinst is! Dit heeft alles te maken met variantie. Door de kwadraten van de verticale afwijkingen te minimaliseren maak je de 'onverklaarde variantie' het kleinst.

Er zijn verschillende manieren om de waarden voor 'a' en 'b' te berekenen. Met je GR kan dat heel goed (dat doen we straks) maar er zijn ook manieren waarop je dat met 'de hand' kunt. Dat zullen we dan doen aan de hand van gegevens van opdracht 2.

Formules

$$a = \frac{n \cdot \sum XY - \sum X \sum Y}{n \cdot \sum X^2 - \left(\sum X\right)^2}$$

$$b = \frac{\sum Y - a \cdot \sum X}{n}$$

Voorbeeld

De resultaten van 10 studenten voor hun test (T) en hun examen (E):

T	10	12	8	13	9	10	7	14	11	6
E	11	14	9	13	9	9	8	14	10	6

Je kunt dan de verschillende 'sommaties' uitrekenen. Dat kan bijvoorbeeld heel handig met Excel. In onderstaande tabel kan je daar een voorbeeld van vinden.

	T	E	T^2	E^2	T·E
	10	11	100	121	110
	12	14	144	196	168
	8	9	64	81	72
	13	13	169	169	169
	9	9	81	81	81
	10	9	100	81	90
	7	8	49	64	56
	14	14	196	196	196
	11	10	121	100	110
	6	6	36	36	36
som	**100,0**	**103,0**	**1060,0**	**1125,0**	**1088,0**
n=	**10**		**a=**	**0,966667**	
			b=	**0,633333**	

Opdracht 3

- Controleer met bovenstaande formules of de berekening van ‘a’ en ‘b’ in de tabel juist zijn.

Opdracht 4

In een regio met veel industrie wordt dagelijks de luchtverontreiniging bepaald. De hoeveelheid vervuiling in lucht neemt doorgaans af (vergeleken met 24 uur eerder) als het ondertussen geregend heeft in het gebied. Voor zeven willekeurige dagen heeft men vastgesteld hoeveel regen er gevallen is in de voorgaande 24 uur en hoe groot de daling is van de hoeveelheid luchtverontreiniging per monster. Deze daling is gemeten in vervuilingeenheden (VE’s).

Dag	Regen (in mm)	Daling (in VE)
1	12	125
2	2	30
3	3	43
4	5	62
5	10	108
6	9	102
7	8	90

- Waarom is ‘regen’ de onafhankelijk variabele en ‘daling’ de afhankelijke variabele?
- Geef een vergelijking van de enkelvoudige lineaire regressielijn (zie opdracht 3).

Correlatie

De correlatiecoëfficiënt is een maat voor de lineaire samenhang van twee variabelen (op minimaal intervalniveau). De officiële naam is ‘**product-moment-correlatiecoëfficiënt van Pearson’**. Maar in normaal spraakgebruik spreken we over de ‘**correlatiecoëfficiënt’** en soms zelf over ‘**correlatie’**.

De correlatiecoëfficiënt kan waarden aannemen van -1 tot 1. De waarde -1 of 1 geeft een perfecte lineaire samenhang aan. Dit komt in de praktijk maar weinig voor. Hoe dichter de correlatiecoëfficiënt bij nul ligt hoe minder sterk het verband is.

Formule

$$r = \frac{n \cdot \sum XY - \sum X \cdot \sum Y}{\sqrt{\left(n \cdot \sum X^2 - \left(\sum X\right)^2\right)\left(n \cdot \sum Y^2 - \left(\sum Y\right)^2\right)}}$$

Voor deze formule gebruiken we dezelfde ‘sommaties’ als bij de berekening van de vergelijking van de regressielijn.

Opdracht 5

Bereken de correlatiecoëfficiënt bij de tabel van opdracht 3.

Opdracht 6

Hieronder zie de uitslagen op twee verschillende tests van 7 willekeurig gekozen kandidaten:

Test 1	0	2	4	1	3	5	6
Test 2	2	3	6	1	4	5	5

- Bereken de correlatiecoëfficiënt.

De grafische rekenmachine

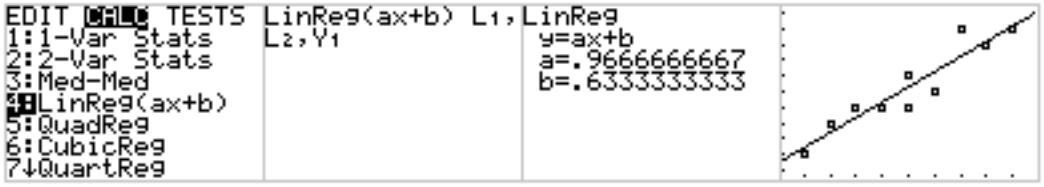
Dit is voor de Ti83/83+/84

De resultaten van 10 studenten voor hun test (T) en hun examen (E) zijn gegeven in de onderstaande tabel:

T	10	12	8	13	9	10	7	14	11	6
E	11	14	9	13	9	9	8	14	10	6

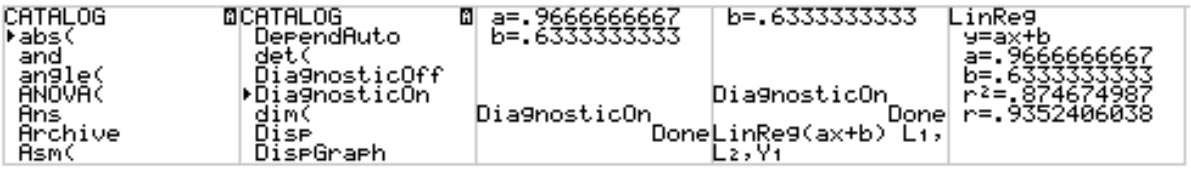
We willen de samenhang onderzoeken en gaan een puntenwolk plotten en de correlatie berekenen met de GR. Via onderstaande aanpak kan je het spreidingsdiagram plotten. Eerst de data in **L1** en **L2** zetten en dan via **[STATPLOT]**.

Via **[STAT]** en **Calc** kies je dan voor **LinReg(ax+b)**. Je kan dit zonder parameters doen, je GR kiest dan zelf **L1** en **L2**, maar je kan naast de lijsten ook meteen een 'functie' opgeven waar de regressievergelijking moet worden opgeslagen. Dat kan met **LinReg(ax+b)L1,L2,Y1** maar dat kan ook met **LinReg(ax+b)Y1**. De **Y1** kan je vinden via **[VARS]** en dan **Y-vars**.



Je kunt nu een voorspelling doen over een student die op de test 12 punten haalt. Uit de tabel (via **[TABLE]**) kan je opmaken dat Y1=12,2.

Nu heb ik wel de regressievergelijking en de waarde van a en b maar de correlatiecoëfficiënt hebben we nog niet. Om die te krijgen moeten we de GR instellen op **DiagnosticsOn**. De diepere bedoeling daarvan ontgaat mij een beetje, maar dat kan je instellen via **[CATALOG]**. Je moet het maar weten:



Het blijkt dat r=0,94.

Zie de bijlage voor de berekeningen met de CASIO.

Opdracht 7

- Controleer je antwoord van opdracht 6 met je GR.

Opdracht 8

In onderstaande tabel vind je data over 10 aselect gekozen studenten die vorig jaar in eerste zittijd slaagden in TI2.

OEF = aantal ingediende oefeningen in de loop van het jaar

RES = resultaat van het examen statistiek in de 1e zittijd

OEF	25	30	45	40	50	60	55	60	75	75
RES	7	9	10	11	12	12	13	14	15	16

a. Bereken de correlatiecoëfficiënt en bepaal een vergelijking voor de regressielijn.
Rond af op 2 decimalen.

b. Een student heeft in de loop van het jaar 65 oefeningen ingediend. Doe een voorspelling (1 decimaal) van het resultaat van het examen.

De rangcorrelatiecoëfficiënt van Spearman

Tot nu toe ging het bepalen van de correlatie om getallenparen (X,Y) waarbij X en Y minimaal van intervalniveau waren. Een bijzonder vorm van correlatie is rangcorrelatie. Dit wordt toegepast als de variabelen van ordinaal niveau zijn. Eén van die rangcorrelaties is de **'rangcorrelatiecoëfficiënt van Spearman'**.

Hierbij gaat het niet om de waargenomen paren van uitkomsten zelf maar om de **rangnummers** van uitkomsten.

Voorbeeld

Door de consumentenbond zijn 8 verschillende merken CD-spelers getest en voorzien van een beoordeling. Na het onderzoeken van allerlei kenmerken zoals bedieningsgemak, veiligheid en vormgeving is een ranglijst opgesteld waarbij nummer 1 de beste CD-speler is, zo oplopend tot 8 als slechtste score.

Merk CD-speler	A	B	C	D	E	F	G	H
Rangnummer beoordeling	5	1	6	2	3	8	7	4
Prijs	398	530	495	595	449	369	475	565

Het vermoeden is dat er een relatie is tussen het oordeel over de CD-speler en de prijs. We berekenen de rangcorrelatiecoëfficiënt van Spearman. Voor de beoordeling hoeven we geen bewerkingen te doen omdat de gegevens al in de vorm van een rangnummers gegeven zijn. De prijzen moeten nog wel vertaald worden tot rangnummers.

Merk CD-speler	A	B	C	D	E	F	G	H
Rangnummer beoordeling	5	1	6	2	3	8	7	4
Prijs	7	3	4	1	6	8	5	2

We kijken naar het verschil in rangnummer van X en Y.

Merk CD-speler	A	B	C	D	E	F	G	H
Rangnummer beoordeling	5	1	6	2	3	8	7	4
Prijs	7	3	4	1	6	8	5	2
d_i	-2	-2	2	1	-3	0	2	2
d_i^2	4	4	4	1	9	0	4	4

Formule

$$r_S = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

n = het aantal waargenomen paren (X,Y)

d_i = het verschil in rangnummer van X en Y

Berekening

$$\sum d_i^2 = 4+4+4+1+9+0+4+4 = 30$$

$$r_S = 1 - \frac{6 \cdot 30}{8^3 - 8} = 0{,}64$$

Opdracht 9

In onderstaande tabel is van 6 willekeurig gekozen medewerkers het hoogst genoten type onderwijs uitgezet tegen het inkomen.

	1	2	3	4	5	6
Hoogst genoten onderwijs	VO	VO	MBO	HBO	HBO	UNIVERSITEIT
Inkomen	32.000	44.000	36.000	37.500	42.000	41.000

De 'volgorde' van de opleiding is VO, MBO, HBO en UNIVERSITEIT.
Zoals je ziet zijn er wat opleiding betreft medewerkers met hetzelfde rangnummer. In dat geval deel je wel de rangnummers 1 t/m 6 uit en geeft bij 'gelijke waarden' het gemiddelde van die rangnummers.

	1	2	3	4	5	6
Hoogst genoten onderwijs	VO	VO	MBO	HBO	HBO	UNIVERSITEIT
Rangnummers	1	2	3	4	5	6
Rangnummers na correctie	1,5	1,5	3	4,5	4,5	6
Inkomen	32.000	44.000	36.000	37.500	42.000	41.000

- Bereken r_S

In **tabel G** (hiernaast) kan je de kritieke waarden vinden voor de correlatiecoëfficiënt van Spearman. Hier kan je het kritieke gebied vinden waarin r_S zou moeten liggen wil je de nulhypothese 'er is geen correlatie' kunnen verwerpen.

Table G

Critical values of ρ, the Spearman rank correlation coefficient

	Significance level (one-tailed test)	
N	**.05**	**.01**
4	1.000	
5	.900	1.000
6	.829	.943
7	.714	.893
8	.643	.833
9	.600	.783
10	.564	.746
12	.506	.712
14	.456	.645
16	.425	.601
18	.399	.564
20	.377	.534
22	.359	.508
24	.343	.485
26	.329	.465
28	.317	.448
30	.306	.432

Opdracht 10

Is je resultaat bij opdracht 9 grond genoeg om de nulhypothese 'er is geen correlatie' te kunnen verwerpen?

Opdracht 11

Is r_S=0,64 bij 't voorbeeld over de CD-spelers grond genoeg om de nulhypothese 'er is geen correlatie' te kunnen verwerpen?

Associatiemaatstaven

Je zou ook bij nominale variabelen wel 's behoefte kunnen krijgen voor een maat voor samenhang. De vraagstelling is dan of bepaalde combinaties van kenmerken vaker voorkomen dan anderen. We noemen dat **associatie**.

Voorbeeld

Aan een groep van 60 economiestudenten (mannen en vrouwen) is gevraagd of ze zelfstandig wonen of bij hun ouders.

	Zelfstandig	Bij ouders	Totaal
Man	16 (a)	24 (b)	40
Vrouw	12 (c)	8 (d)	20
	28	32	60

Een maat voor associatie is de associatiemaat van Yule:

$$\text{Yule} = \frac{ad - bc}{ad + bc} = \frac{16 \cdot 8 - 12 \cdot 24}{16 \cdot 8 + 12 \cdot 24} = -0,385$$

Een andere maat van associatie is **phi**. Deze maat is gebaseerd op de **chi-kwadraatverdeling** die we al eerder zagen bij het toetsen van onafhankelijkheid bij kruistabellen. Deze wordt gedefinieerd als:

$\phi = \sqrt{\frac{\chi^2}{n}}$ met n=aantal waarnemingen. ϕ neemt waarden aan tussen 0 en 1 bij toepassen van 2x2 tabellen. Voor grotere tabellen (met p rijen en q kolommen) wordt meestal **Cramer's V** gebruikt. Deze is gedefinieerd als:

$V = \sqrt{\frac{\chi^2}{n(k-1)}}$ met k=min(p,q)

Opdracht 12

- Bereken ϕ en V voor de tabel uit het voorbeeld hierboven.

Beperkingen en valkuilen

Correlatie tussen twee variabelen wil nog niet zeggen, dat de verschijnselen, die door die variabelen worden gemeten, een **causaal** verband hebben. Soms berust het geheel op toeval en men spreekt dan wel van een schijncorrelatie.

Een klassiek voorbeeld is de correlatie tussen het aantal ooievaars en het aantal geboren kinderen. Er is wel een correlatie maar geen verband! De verklaring was dat er op het platteland meer kinderen geboren worden (traditionelere levensopvatting?) en ooievaars meestal niet in de stad rond vliegen (behalve in Den Haag dan...:-).

Opdracht 13

In de tabel hiernaast vindt u omzetcijfers uit 1994, 1995 en 1996.

	X	Y
1994		
	I	22
	II	32
	III	34
	IV	26
1995		
	I	24
	II	35
	III	38
	IV	30
1996		
	I	28
	II	38
	III	41
	IV	..

a. Voorspel via lineair regressie-analyse de omzet voor het IV-de kwartaal van 1996.
b. Vind je dit een 'betrouwbare' voorspelling? Leg uit.

TIP: kies voor de verschillende kwartalen bijvoorbeeld de getallen 1 t/m 11. De vraag is dan kan je voorspellen wat de omzet zal zijn bij x=12.

Opdracht 14

In Den Haag heeft men voor verschillende dagen de volgende gegevens verzameld. De **F** staat voor de hoeveelheid frisdrank die verkocht wordt in Scheveningen en de **O** staat voor verkeersdrukte in de regio Den Haag.

F	0,50	1,60	2,05	1,00	1,10	0,85	1,50	1,65	0,85	0,70	1,00	1,90	2,20	1,30	2,00
O	1,25	3,10	3,00	1,80	2,30	1,75	2,60	2,70	0,85	1,25	1,25	3,10	2,75	1,55	2,60

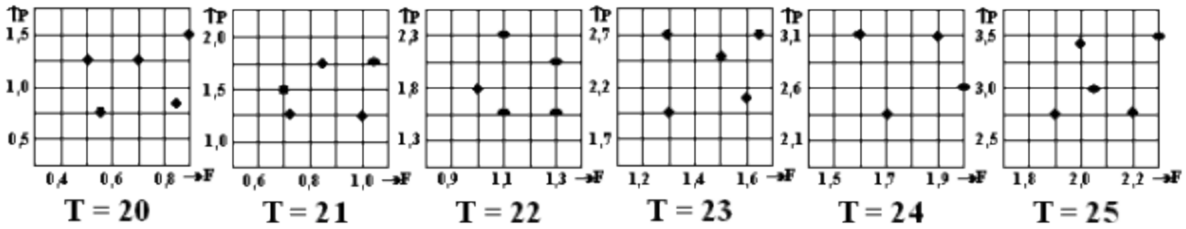
a. Plot een spreidingsdiagram en bereken de correlatiecoëfficiënt.
b. Welke conclusie kan je trekken uit het resultaat van a.?

Om dit nader te onderzoeken kijken we naar een derde factor, de temperatuur **T**.

F	0,50	1,60	2,05	1,00	1,10	0,85	1,50	1,65	0,85	0,70	1,00	1,90	2,20	1,30	2,00
O	1,25	3,10	3,00	1,80	2,30	1,75	2,60	2,70	0,85	1,25	1,25	3,10	2,75	1,55	2,60
T	20	24	25	22	22	21	23	23	20	20	21	24	25	22	24

c. Laat zien dat er een positieve correlatie bestaat tussen zowel **T** en **F** als tussen **T** en **O**.

In het onderzoek heeft men met extra gegevens gekeken naar de samenhang voor verschillende waarden van **T**. De gegevens zijn als volgt in beeld gebracht:



d. Welke conclusie kan je trekken t.a.v. de resultaten van vraag a.?

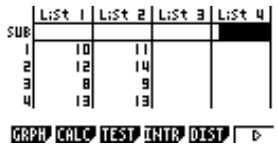
EINDE

Bijlage A – regressie en correlatie op de CASIO

De resultaten van 10 studenten voor hun test (T) en hun examen (E) zijn gegeven in de onderstaande tabel:

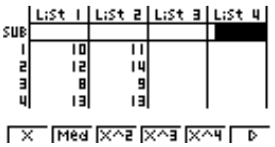
T	10	12	8	13	9	10	7	14	11	6
E	11	14	9	13	9	9	8	14	10	6

We willen de samenhang onderzoeken en gaan een puntenwolk plotten en de correlatie berekenen met de GR. Via onderstaande aanpak kan je de vergelijking van de regressielijn en de correlatiecoëfficiënt bepalen.

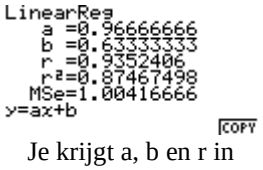


Zet eerst de gegeven in List 1 en List 2

Kies voor CALC en vervolgens voor REG



Kies voor X (lin.regr.)

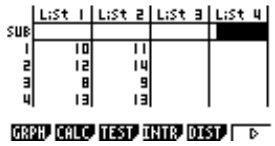


Je krijgt a, b en r in beeld.

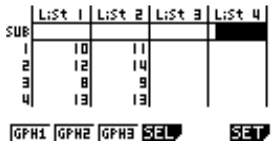
We zien r=0,94.

Via **COPY** kan je de vergelijking van de regressielijn in **Y1** zetten. Je kunt dan via **TABLE** een tabel maken van waarden waarmee je bijvoorbeeld een voorspelling doen over een student die op de test 12 punten haalt. Y is dan 12,2.

Je kunt ook de regressielijn plotten:

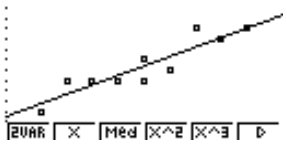


Kies voor GRPH



Kies voor GPH1

Via CALC en X en DRAW



Het spreidingsdiagram met de regressielijn

Bronvermelding

- **Statistiek om mee te werken** – Dr. A. Buijs – 6e druk 1998
- **Statistiek om mee te werken** – Dr. A Buijs - opgaven
- **Statistiek om mee te werken** – Ir K de Bont - uitwerkingen
- Wiswijzer.nl | WisFaq.nl | Wiskundeleraar.nl