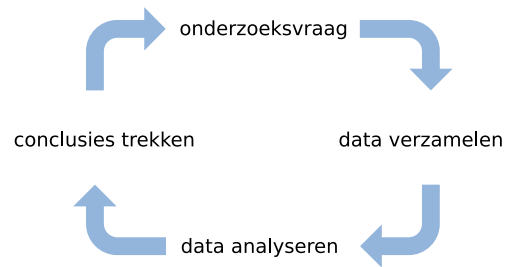


6.4 Samenhang tussen variabelen

Inleiding

Tussen statistische variabelen bestaat soms een statistisch verband. Dat wil nog niet zeggen dat de éne variabele dan automatisch de andere veroorzaakt.



Figuur 1

Je leert in dit onderwerp

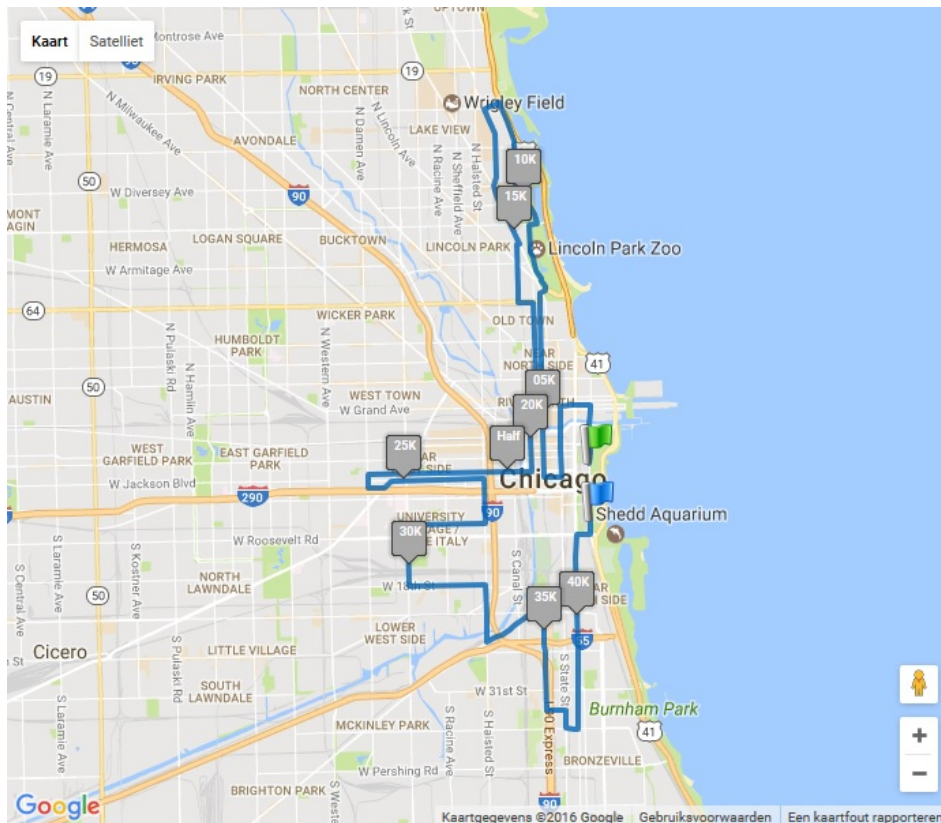
- de (statistische) samenhang tussen twee variabelen te beschrijven en te interpreteren met behulp van correlatiecoëfficiënt en trendlijn.

Voorkennis

- de beschrijvende statistiek: het maken van (som)frequentietabellen, diagrammen en het berekenen van centrummaten en spreidingsmaten;
- werken met de normale verdeling, met normaal waarschijnlijkheidspapier en met de wortel-n-wet.
- data leren ordenen in verband met de gestelde onderzoeksvraag en de aard van de gekozen variabelen.

Verkennen

Opgave V1



Figuur 2

Bij de Chicago Marathon 2016 kun je je afvragen of er een verband is tussen de tijd die iemand halverwege het parcours (H) klokt en de eindtijd (E).

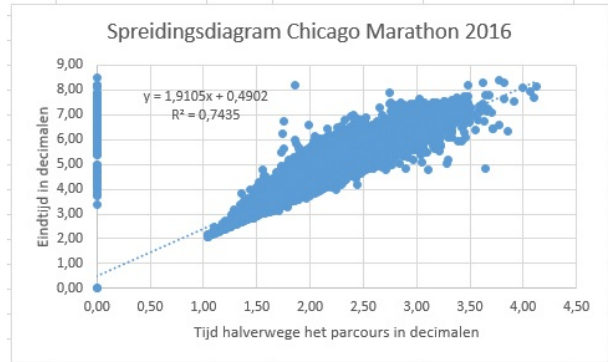
- a Wat voor soort diagram stel je je dan voor?

Gebruik deze [gegevens van de Chicago Marathon 2016](#).

- b Zoek uit hoe je in Excel het gewenste diagram kunt maken.
- c Is er een verband tussen H en E ? En is dit alleen een statistisch verband of is het een oorzakelijk verband?
- d Als je een verband tussen twee statistische variabelen zoekt, moeten dit dan kwantitatieve variabelen zijn?

Uitleg

Bij de **gegevens van de Chicago Marathon 2016** heb je voor wat betreft de tijden halverwege en de eindtijden te maken met continue kwantitatieve variabelen. Je zou verwachten dat er een zeker verband tussen beide bestaat, maar zeker is dit niet. Of een dergelijk verband bestaat kun je onderzoeken met behulp van een puntenwolk of spreidingsdiagram. Daarin zet je de éne variabele - in dit geval de tijd halverwege het parcours - op de horizontale as en de andere - hier de eindtijd - op de verticale as. Merk op dat beide variabelen in decimalen zijn omgerekend.



Figuur 3

Je kunt Excel in het spreidingsdiagram een zogenaamde trendlijn laten tekenen. Die trendlijn beschrijft de samenhang tussen beide variabelen.

Hier is y de eindtijd en x de halverwege tijd.

De formule van de trendlijn wordt dan $y = 1,9105x + 0,4902$.

Dit zou betekenen dat iemand die na 2,5 uur halverwege is, een eindtijd zou moeten halen van $1,9105 \cdot 2 + 0,4902 = 4,3112$ uur.

Maar hoe betrouwbaar is dit verband? De punten in het diagram liggen echt niet allemaal op de rechte lijn.

Daarom staat in de figuur ook $R^2 = 0,7453$.

Dit getal is het kwadraat van de correlatiecoëfficiënt R . En R is een maat voor de kwaliteit van de samenhang.

Als $R = 0$ is er geen samenhang, een vrijwel ronde puntenwolk en er is geen trendlijn te maken.

Als $R = 1$ is er een perfecte correlatie, alle punten liggen op één lijn met een positieve helling.

Als $R = -1$ is er ook een perfecte correlatie, alle punten liggen op één lijn met een negatieve helling.

Hier is $R \approx 0,8633$. En dat is veel dichterbij 1 dan bij 0. De correlatie er daarom redelijk en de trendlijn geeft redelijk goed het verloop van de puntenwolk weer.

Opgave 1

Bekijk in de **Uitleg** hoe wordt onderzocht of er een verband bestaat tussen de tijd die halverwege het parcours wordt gemeten en de eindtijd van de Chicago Marathon in 2016.

- Maak zelf zo'n puntenwolk. Het omrekenen van de tijd in uu:mm:ss naar decimalen gaat met de formule $=([celnummer]-INTEGER([celnummer]))*24$.
- Zet vervolgens de trendlijn in je figuur, evenals de correlatiecoëfficiënt R .
- Welke betekenis hebben de stippen op de verticale as?

Opgave 2

Bekijk in de **Uitleg** het spreidingsdiagram bij de Chicago Marathon in 2016.

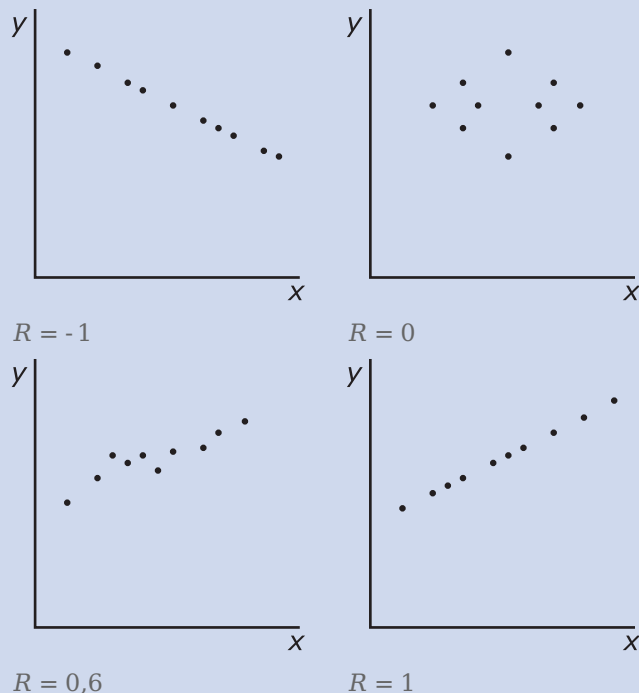
- Ga na, dat het punt $(\bar{x}, \bar{y})$ op de trendlijn ligt.
- Uit de waarde voor R blijkt een redelijk goed statistisch verband tussen beide variabelen. Licht dit toe en leg ook uit waarom het alleen een statistisch verband is en geen oorzakelijk verband.
- Maak nu zelf een puntenwolk met een trendlijn en met R^2 voor een deelgroep van deelnemers. Neem bijvoorbeeld de vrouwen, of de mannen, of de vrouwen van 25 - 29 jaar, of nog iets anders.

Theorie en voorbeelden

Om te onthouden

Bij twee statistische (meestal verschillende) variabelen kan naar een **statistisch verband** worden gezocht. Om dit verband zichtbaar te maken wordt een **puntenwolk** of (andere naam) **spreidingsdiagram** gebruikt.

Puntenwolken kunnen verschillende vormen hebben. Bekijk de afbeeldingen.



Figuur 4

Als de punten van de puntenwolk ongeveer op een rechte (of kromme) lijn liggen, is er een statistisch verband. De **correlatiecoëfficiënt** R bepaalt in hoeverre dit het geval is. In een spreidingsdiagram met statistische samenhang kan met een **trendlijn** de (lineaire) samenhang worden weergegeven. Er geldt:

- Als $R = 0$ is er geen samenhang, een vrijwel ronde puntenwolk en er is geen trendlijn te maken.
- Als $R = 1$ is er een perfecte samenhang, alle punten liggen op een lijn met een positieve helling.
- Als $R = -1$ is er een perfecte samenhang, alle punten liggen op een lijn met een negatieve helling.
- Als $0,5 \leq R < 0,7$ of $-0,7 < R \leq -0,5$ is er een middelmatige correlatie.
- Als $0,7 \leq R < 0,9$ of $-0,9 < R \leq -0,7$ is er een hoge correlatie.
- Als $0,9 \leq R < 1,0$ of $-1,0 < R \leq -0,9$ is er een zeer hoge correlatie.
- In alle andere gevallen is de correlatie mager en is er geen trendlijn te maken.

Alleen als het veranderen van een van de variabelen de oorzaak ervan is dat de andere variabele verandert, is er een **oorzakelijk verband**.

Voorbeeld 1

Je doet onderzoek naar bijvoorbeeld de examenresultaten van je eigen school. De gegevens betreffende de examenresultaten bevatten per individu het geslacht, het profiel, de gevolgde vakken en de eindcijfers SE en CE1.

Je onderzoeksvraag is nu "Is er een verband tussen het SE-cijfer en het CE1-cijfer?".

Hoe ga je die vraag nu beantwoorden met de examenresultaten die je van je eigen school hebt verzameld? Of, als je nog niets hebt, gebruik bijvoorbeeld deze [resultaten van het VWO op CE1 in 2013](#).

Antwoord

Natuurlijk kies je één of meer vakken om afzonderlijk te bekijken, misschien doe je dit wel voor alle vakken.

Neem bijvoorbeeld het vak natuurkunde en maak een puntenwolk (spreidingsdiagram) van de SE-cijfers tegen de CE1-cijfers.

Zorg er voor dat in je diagram zowel de vergelijking van de trendlijn als (het kwadraat van) de correlatiecoëfficiënt R komen te staan.

Trek vervolgens een conclusie gebaseerd op de vuistregels in de **Theorie**.

Opgave 3

Bekijk **Voorbeeld 1**.

- Gebruik de dataset die je daar vindt of je eigen dataset en maak de puntenwolk.
- Voeg de trendlijn met bijbehorende formule en R^2 toe.
- Welke conclusie trek je? Welk cijfer verwacht je als je op het SE een 6,5 staat?
- Is hier sprake van een statistisch verband of een oorzakelijk verband? En is dat voor alle vakken zo?

Opgave 4

Je kunt ook kijken of er verband is tussen de CE1-cijfers van twee verschillende vakken, bijvoorbeeld Frans en wiskunde A, of wiskunde B en natuurkunde.

- Kies zo'n combinatie van twee vakken uit en voer het onderzoek naar een verband tussen beide uit.
- Waarom zal hier waarschijnlijk sprake zijn van een statistisch verband? Of zijn er ook combinaties van vakken te bedenken waarbij sprake kan zijn van een oorzakelijk verband?

Voorbeeld 2

Bekijk het bestand **TK-Verkiezingen-NL-2012**.

Je ziet hier een kruistabel waarin de variabelen 'Links_rechts_zelfplaatsing' en 'Conservatief_progressief_zelfplaatsing' zijn weergegeven.

Cons_prog_zelfplaatsing: 1 = Conservatief, 10 = Progressief		Links_rechts_zelfplaatsing: 1 = Links, 10 = Rechts										Eindtotaal
Aantal van Conservatief_progressief_zelfplaatsing		Aantal										
Rijlabels	Coloml	1	2	3	4	5	6	7	8	9	10	
1	4	1	1		1				4	2	6	19
2	1	2	1	1	1	1			9	9	2	27
3	3			11	4	4	2	14	12	3		53
4	2	1	9	12	12	19	27	20	6			108
5	5	4	17	23	50	29	34	28	7	4		201
6	8	6	20	36	29	33	40	28	4	4		208
7	8	16	62	81	38	35	70	43	6	5		364
8	22	24	94	75	48	35	33	30	7	1		369
9	5	13	28	26	12	6	7	2	3	1		103
10	32	12	16	12	10	6	4	2		4		98
Eindtotaal		90	79	259	270	205	166	229	178	47	27	1550

Figuur 5

Dit zijn twee kwalitatieve variabelen die echter wel in getallen zijn uitgedrukt. Daarom kun je beide variabelen ook in een puntenwolk tegen elkaar uitzetten.

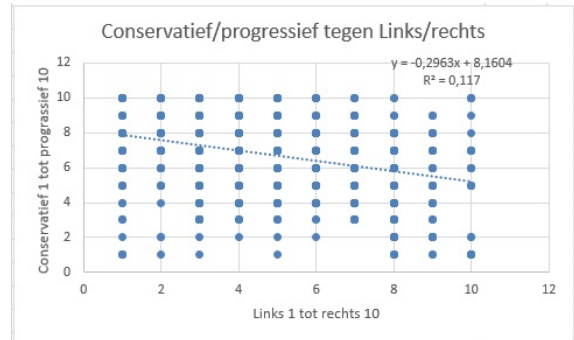
Wat gaat er dan fout?

Antwoord

Je krijgt wel een plaatje met punten, maar geen inzicht hoeveel mensen er bij elk punt horen. Dit komt vooral omdat het aantal mogelijke waarden klein is (1 t/m 10) en de steekproef behoorlijk groot (1550 mensen).

In de kruistabel zie je bijvoorbeeld dat de combinatie (7,4) wel 81 keer voorkomt en de combinatie (2,4) maar 1 keer.

De gegeven trendlijn en correlatiecoëfficiënt zijn echter wel gebaseerd op alle combinaties, ook al zie je die niet terug in je puntenwolk.



Figuur 6

Opgave 5

Bekijk in **Voorbeeld 2** de vergelijking van 'Links_rechts_zelfplaatsing' en 'Conservatief_progressief_zelfplaatsing'.

- Waarom geeft de kruistabel dit beter weer dan het spreidingsdiagram?
- Laat zien dat je een correlatie krijgt die nogal klein is. Is dat in overeenstemming met je kruistabel?

Verwerken

Opgave 6

Je bent bezig met een eigen onderzoek opzetten en je bent op zoek naar een verband tussen twee variabelen.

- Maak een spreidingsdiagram met daarin de trendlijn en de correlatiecoëfficiënt.
- Trek een conclusie.
- Keer met je conclusies nog een keer terug naar je oorspronkelijke statistische probleem. Zijn je onderzoeksvragen beantwoord? En zijn die antwoorden enigszins realistisch en/of te verklaren?
- Schrijf een onderzoeksrapport waarin je de statistische cyclus duidelijk herkent.



© 2024

Deze paragraaf is een onderdeel van het Math4All wiskundemateriaal.

Math4All stelt het op prijs als onvolkomenheden in het materiaal worden gemeld en ideeën voor verbeteringen in de content of dienstverlening kenbaar worden gemaakt. Klik op  in de marge bij de betreffende opgave. Uw mailprogramma wordt dan geopend waarbij het emailadres en onderwerp al zijn ingevuld. U hoeft alleen uw opmerkingen nog maar in te voeren.

Email: f.spijkers@math4all.nl

Met de Math4All Foliostraat kunnen complete readers worden samengesteld en toetsen worden gegenereerd. Docenten kunnen bij a.f.otten@math4all.nl een gratis inlog voor de maatwerkdienst aanvragen.
