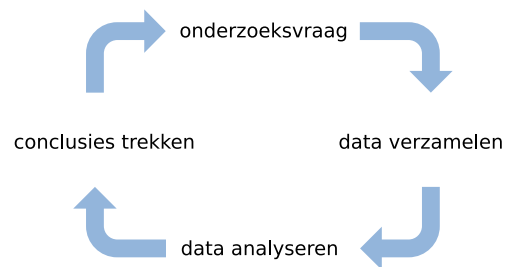


6.2 Data analyseren

Inleiding

Afhankelijk van de gestelde vraag, het te onderzoeken probleem, voer je een data-analyse uit. Daarbij gebruik je vaak centrummaten, spreidingsmaten en diagrammen. Welke maten en welke diagrammen dat zijn, hangt sterk af van de aard van de variabelen die je bekijkt. Bij kwalitatieve variabelen werk je anders dan bij kwantitatieve variabelen.



Figuur 1

Je leert in dit onderwerp

- data leren ordenen in verband met de gestelde onderzoeksvraag en de aard van de gekozen variabelen.

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.

Verkennen

Opgave V1

Je wilt de data die je hebt natuurlijk analyseren. Tabellen en diagrammen helpen daarbij.

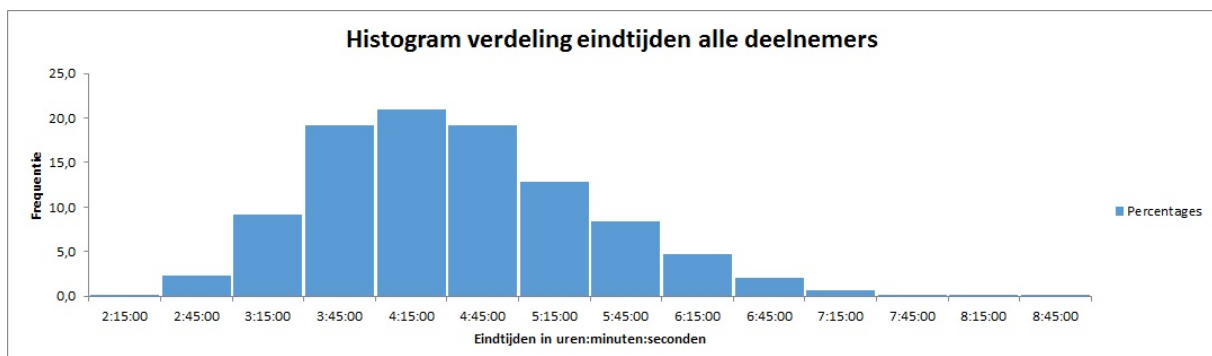
- Welke soorten tabellen gebruik je bij data-analyse? En waar dienen ze voor?
- Welke soorten diagrammen ken je en wanneer gebruik je ze?

Uitleg 1

Bij de **gegevens van de Chicago Marathon 2016** heb je voor wat betreft de eindtijden te maken met een continue kwantitatieve variabele. Dergelijke variabelen vat je samen met behulp van een frequentietabel met aaneengesloten klassen en een bijbehorend histogram of frequentiepolygoon.

Wil je daarmee twee deelgroepen (zoals mannen en vrouwen, of leeftijdscategorieën) vergelijken, dan gebruik je altijd relatieve frequenties (percentages).

Vaak worden gemiddelde als centrummaat en standaardafwijking als spreidingsmaat gebruikt.



Figuur 2

Wil je twee deelgroepen vergelijken, dan kun je hun histogrammen vergelijken. Maar om dan goede uitspraken te doen is dan nogal lastig: je hebt niet alleen met verschillende gemiddeldes te maken,

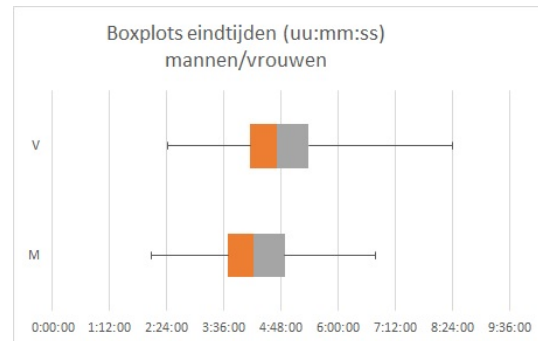
maar ook met verschillende standaardafwijkingen. En hoe interpreteer je dat? Dat zou bijvoorbeeld kunnen als ze zijn te benaderen door normale verdelingen. Hoe je die kunt vergelijken zie je in het volgende onderdeel.

Een manier om kwantitatieve variabelen te vergelijken is met behulp van boxplots.

Een boxplot is een diagram waarbij je de gegevens in vier kwarten verdeeld, je hebt er minimum, eerste kwartiel, mediaan, derde kwartiel en maximum voor nodig.

Wil je daarmee twee deelgroepen (zoals mannen en vrouwen, of leeftijdscategorieën) vergelijken, dan gebruik je voor beide boxplots dezelfde schaalverdeling.

Wil je de samenhang tussen twee kwantitatieve variabelen vergelijken, dan maak je een puntenwolk. Daarover meer in het laatste onderdeel.



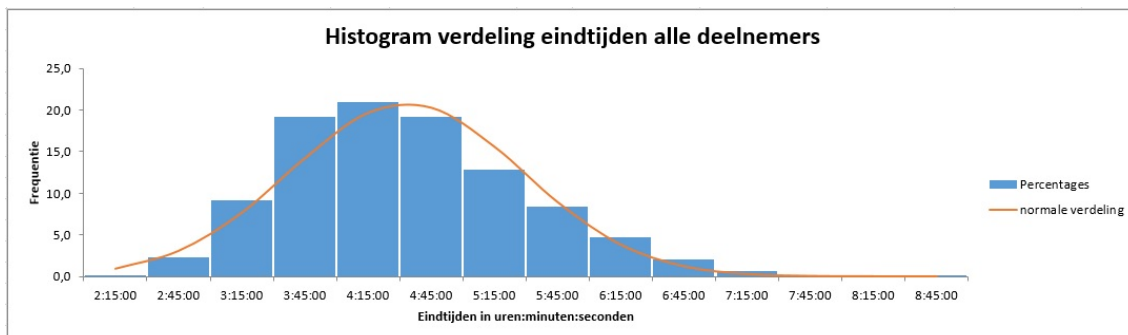
Figuur 3

Opgave 1

Bekijk in **Uitleg 1** het verhaal van de Chicago Marathon in 2016. Je ziet een histogram van de relatieve frequenties van alle eindtijden met een klassenindeling $2:00 - \leq 2:30$, $2:30 - \leq 3:00$, etc. De tijden hebben de indeling u:mm:ss (uren/minuten/seconden).

- Wat is het verschil tussen een histogram en een staafdiagram?
- Open het databestand en maak zelf zo'n histogram voor de mannelijke en de vrouwelijke hardlopers afzonderlijk.
- Waarom is het nu echt noodzakelijk om met relatieve frequenties (procenten) te werken?

Hier zie je het histogram van de eindtijden nog een keer. Er is een normale verdeling in getekend die hetzelfde gemiddelde en dezelfde standaardafwijking heeft.



Figuur 4

- Zijn de eindtijden normaal verdeeld? Hoe zou je dit nauwkeuriger kunnen onderzoeken?

Opgave 2

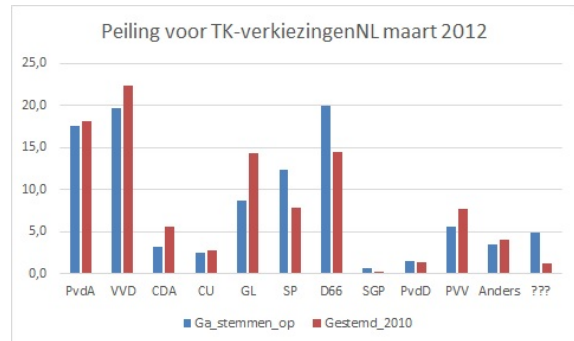
Bekijk in **Uitleg 1** de boxplots van de eindtijden van de Chicago Marathon 2016.

- Maak zelf dergelijke boxplots in Excel.
- Kun je zonder meer zeggen dat de mannen sneller zijn dan de vrouwen?

Uitleg 2

In het bestand **TK-Verkiezingen-NL-2012** vind je de gegevens van een aselechte steekproef van 1550 personen uit het enorme databestand van KiesKompas. Het gaat hierbij vooral om kwalitatieve variabelen. Daarom werk je voor het vergelijken van twee variabelen met staafdiagrammen, of kruistabellen.

Het staafdiagram hiernaast laat per partij zien wat de verschillen zijn tussen wat mensen in 2010 hebben gestemd en wat ze in 2012 van plan waren om te gaan stemmen op 12 september. Maar je zou hierbij ook een kruistabel kunnen maken.



Figuur 5

Zo'n kruistabel zie je hieronder voor de twee variabelen 'Ga_stemmen_op' en 'Zekerheid_keuze'.

Aantal van Ga_stemmen_op	Kol																
Rijlabels	PvdA	VVD	CDA	CU	GL	SP	D66	SGP	PvdD	PVV	50P	PP	Anders	???	Eindtotaal		
Zeker op welke partij	94	159	16	17	50	83	111	6	11	33	2	6	5	2	595		
Twijfel	152	125	28	18	75	90	167	3	11	39	11	8	1	39	767		
Weet niet	27	21	6	3	8	17	30	1	2	14	8	5	4	33	179		
Ik ga niet stemmen					1	2	1						1	3	1	9	
Eindtotaal	273	305	50	38	134	192	309	10	24	86	21	20	13	75	1550		

Figuur 6

Je kunt hierin bijvoorbeeld aflezen dat het aantal twijfelaars bij de PvdA groter is dan bij de VVD, ook naar verhouding. Kun je dan een duidelijke conclusie trekken? Daarover gaat het volgende onderdeel.

Opgave 3

Bekijk in **Uitleg 2** het verhaal van de NL verkiezingen in 2012. Om statistische variabelen goed in beeld te krijgen is het verstandig om een afzonderlijk werkblad te maken voor de variabele(n) die je wilt bekijken.

- Maak een werkblad waarop de variabelen 'Ga_stemmen_op' en 'Gestemd_2010' voorkomen. Maak daarop twee frequentietabellen en het staafdiagram dat je in de uitleg ziet.
- Maak een werkblad voor de variabelen 'Ga_stemmen_op' en 'Zekerheid_keuze'. Maak de kruistabel die je in de uitleg ziet.

Opgave 4

Bij de verkiezingen 2012 dataset van KiesKompas kun je verschillende onderzoeksvragen stellen. Bijvoorbeeld:

- “Vinden mannen veiligheid belangrijker dan vrouwen?”
- “Is er verschil in stemgedrag tussen mannen en vrouwen?”
- “Is er verband tussen jezelf plaatsen op de 1 - 10 schaal Links/Rechts en jezelf plaatsen op de 1 - 10 schaal Conservatief/Progressief?”
- “Stemmen mensen die ‘Zorg’ een belangrijk onderwerp vinden anders dan mensen die dit niet doen?”

- Schrijf bij elk van deze vragen op hoe je het onderzoek op basis van de beschikbare gegevens aan zou pakken.
- Kies twee van deze vragen uit en begin een bijbehorende data-analyse.

Theorie en voorbeelden

Om te onthouden

Hoe je gegevens samenvat is sterk afhankelijk van de soort gegevens:

Bij **kwantitatieve variabelen** werk je meestal met:

- Frequentietabellen, staafdiagrammen en lijndiagrammen bij discrete kwantitatieve variabelen, zoals cijfers, leeftijden, e.d.
Hierbij horen centrummaten als modus, gemiddelde en spreidingsmaten als spreidingsbreedte.
- Frequentietabellen met aaneengesloten even brede klassen, histogrammen en frequentiepolygo-
nen bij continue kwantitatieve variabelen, zoals lengte, gewicht, tijdsduur, e.d.
Hierbij horen centrummaten als gemiddelde en spreidingsmaten als standaardafwijking.
- Boxplots waarvoor je minimum, eerste kwartiel, mediaan, derde kwartiel en maximum nodig hebt.
- Puntenwolken (spreidingsdiagrammen) als je een verband tussen twee kwalitatieve variabelen wilt zoeken.

Bij **kwalitatieve variabelen** werk je meestal met:

- Frequentietabellen, staafdiagrammen, lijndiagrammen, beelddiagrammen, cirkeldiagrammen bij kwalitatieve variabelen, zoals vakkenprofiel, aantal per maand, e.d.
Hierbij horen centrummaten zoals modus en mediaan.
- Kruistabellen als je twee kwalitatieve variabelen wilt vergelijken.

Voorbeeld 1

Stel je wilt onderzoek doen naar de examenresultaten van je eigen school. De gegevens betreffende de examenresultaten moeten in ieder geval per individu het geslacht, het profiel, de gevolgde vakken en de eindcijfers SE en CE1 bevatten.

Je onderzoeksvraag is bijvoorbeeld “Zijn jongens beter in de exacte vakken dan meisjes?”.

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

Hopelijk heb je wel ingezien dat het weinig zinvol is vakken als wiskunde, natuurkunde, scheikunde, biologie, e.d., op één hoop te gooien. Er zijn immers maar weinig leerlingen die die vakken allemaal hebben, en dan heb je nog wiskunde A, B en D, hoe weeg je dat mee?

Je kunt beter naar afzonderlijke vakken kijken en dan bijvoorbeeld de CE1-cijfers vergelijken, want die worden door iedereen in dezelfde omstandigheden gemaakt (geen herkansingen nog meegeteld, geen werkstukken).

Opgave 5

Bekijk in [Voorbeeld 1](#) de examenresultaten ofwel van je eigen school ofwel die in het daar aangeboden databestand.

- Maak een werkblad met daarop de CE-resultaten van (bijvoorbeeld) natuurkunde. Maak afzonderlijke frequentietabellen voor de jongens en de meisjes. Welke klassenindeling gebruik je?
- Stel je gebruikt de klassenindeling $1,0- < 1,5$, $1,5- < 2,5$, $2,5- < 3,5$, etc. Je wilt een histogram maken. Welke moeilijkheid doet zich dan voor? En hoe los je dat op?
- Maak nu twee afzonderlijke histogrammen voor de jongens en voor de meisjes en bereken de bijpassende centrummaten.
- Je zou ook naar de eindcijfers voor natuurkunde kunnen kijken en daar boxplots voor de jongens en de meisjes kunnen maken. Doe dat en beschrijf de voor- en nadelen daarvan.

Opgave 6

In de vorige opgave heb je het vak natuurkunde bekeken om iets te kunnen zeggen over de vraag “Zijn jongens beter in de exacte vakken dan meisjes?”

- a** Hoe zou je nu het onderzoek naar die vraag voortzetten?

Je zou natuurlijk ook heel andere vragen kunnen stellen, zoals: “Wordt bij engels het SE net zo goed gemaakt als het CE?”

- b** Hoe zou je zo'n onderzoeksvraag aanpakken?

- c** Probeer zelf een andere onderzoeksvraag te bedenken bij de gegeven dataset en beschrijf welke vorm van data-analyse je erbij zou gaan uitvoeren.

Voorbeeld 2

Bekijk het bestand [TK-Verkiezingen-NL-2012](#).

Je wilt onderzoeken of ‘Veiligheid’ bij mannen een belangrijker onderwerp is dan bij vrouwen. Je maakt daarbij deze kruistabel. Hoe zou je deze vergelijking kunnen maken?

Aantal van Veiligheid_belangrijk	Kolomlabel ▼		
Rijlabels ▼	0	1	Eindtotaal
M	504	706	1210
V	148	182	330
(leeg)	3	7	10
Eindtotaal	655	895	1550

Figuur 7

Antwoord

Hier valt niets te vergelijken omdat het over absolute aantallen gaat.

Je moet in ieder geval de gegevens omrekenen naar percentages. Maar ja, percentages waarvan?

Omdat je het percentage van de mannen dat veiligheid een belangrijk thema vindt wilt vergelijken met het percentage van de vrouwen dat dit vindt, is het totaal aantal mannen 100% en is ook het totaal aantal vrouwen 100%. Reken nu de getallen om naar de juiste percentages en probeer een uitspraak te doen.

Opgave 7

Bekijk in [Voorbeeld 2](#) het verhaal van de NL verkiezingen in 2012 en de daar gestelde onderzoeksvraag.

- a** Zijn er nog een andere manieren denkbaar om een data-analyse uit te voeren?
- b** Laat zien dat het percentage mannen dat veiligheid belangrijk vindt ongeveer 58,3 is en bereken dit percentage ook voor de vrouwen. Is er veel verschil?

Opgave 8

Bekijk de dataset van [Voorbeeld 2](#) nog eens.

- a** Maak zelf zo'n kruistabel bij de vraag of ‘Zorg’ bij mannen een belangrijker onderwerp is dan bij vrouwen.
- b** Reken je getallen om naar percentages. Is er nu een groot verschil?

Verwerken

Opgave 9

Je bent bezig met een eigen onderzoek opzetten. Je plan van aanpak is klaar, de gegevens zijn verzameld.

- a** Ga na wat voor soort statistische variabelen je hebt en wat dit betekent voor de data-analyse.
- b** Voer de data-analyse uit, maak dus geschikte tabellen en diagrammen en bereken relevante centrummaten.



© 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.
