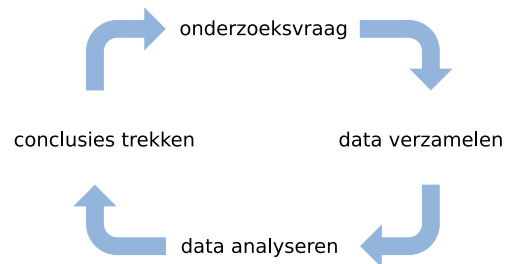


6.3 Variabelen vergelijken

Inleiding

Nu ga je statistische variabelen vergelijken. Er bestaan veel manieren om dat te doen, veel hangt af van de soort variabelen waar je mee te maken hebt.



Figuur 1

Je leert in dit onderwerp

- het verschil tussen twee kwantitatieve variabelen beschrijven en interpreteren door boxplots vergelijken of normale verdelingen te gebruiken;
- het verschil tussen twee kwalitatieve variabelen beschrijven en interpreteren met kruistabellen en phi, of met de max.Vcp.

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

Je hebt de data geanalyseerd, tabellen en diagrammen gemaakt afhankelijk van de soort variabele.

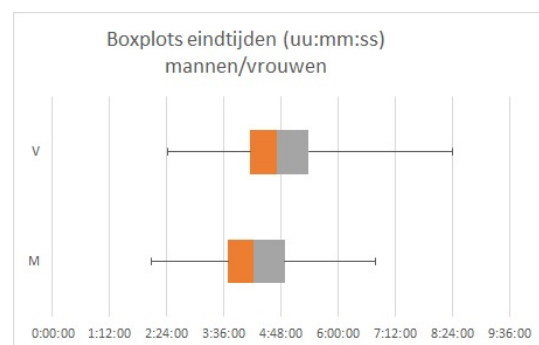
- Hoe kun je frequentietabellen vergelijken?
- Kun je iets bedenken om boxplots te vergelijken?
- Kun je iets bedenken om histogrammen te vergelijken?
- En hoe vergelijk je kwalitatieve variabelen?

Uitleg 1

Bij de **gegevens van de Chicago Marathon 2016** heb je voor wat betreft de eindtijden te maken met een continue kwantitatieve variabele.

Een manier om dergelijke variabelen te vergelijken is met behulp van boxplots. Hier zie je de boxplots van de eindtijden van de mannen en de vrouwen. Er is enig verschil, maar ook veel overlap van deze boxplots. Om een uitspraak te kunnen doen hanteer je de volgende afspraken:

- Als de boxen elkaar niet overlappen, dan zeg je "Het verschil is groot".
- Als de boxen elkaar wel overlappen en minstens één mediaan buiten de box van de andere boxplot ligt, dan zeg je "Het verschil is middelmatig".
- In alle andere gevallen zeg je "Het verschil is gering".



Figuur 2

Bij het vergelijken van de boxplots van de eindtijden van de mannen en de vrouwen moet je op grond daarvan concluderen dat het verschil gering is.

Een andere manier om kwantitatieve variabelen te vergelijken is met behulp van de normale verdeling.

Je gaat dan eerst na of beide variabelen bij benadering normaal zijn verdeeld. In dit geval klopt dat wel ongeveer (vooral als je de erg langzame lopers buiten beschouwing laat). Je zou dus kunnen zeggen:

- De eindtijden van de mannen zijn normaal verdeeld met $\mu(M) \approx 4,32$ en $\sigma(M) \approx 0,87$
- De eindtijden van de vrouwen zijn normaal verdeeld met $\mu(V) \approx 4,80$ en $\sigma(V) \approx 0,89$

Je ziet dat de gemiddelden verschillen, de standaardafwijking nauwelijks. Nu kun je twee normale verdelingen vergelijken door naar hun verschilverdeling te kijken.

Als twee normaal verdeelde variabelen X en Y hetzelfde gemiddelde hebben, heeft hun verschilverdeling een gemiddelde van $\mu = \mu(X) - \mu(Y) = 0$ en een standaardafwijking van $\sigma = \sqrt{(\sigma(X))^2 + (\sigma(Y))^2}$. Als echter $\mu(X) \neq \mu(Y)$ dan zal het gemiddelde van de verschilverdeling ook ongelijk aan 0 zijn. Hoe meer μ van 0 afwijkt, hoe groter het verschil.

Voor de verschilverdeling van de eindtijden van de mannen en de vrouwen geldt $\mu(E) = \mu(V) - \mu(M) \approx 0,48$ en $\sigma(E) \approx \sqrt{0,87^2 + 0,89^2} \approx 1,3$.

$\mu(E)$ wijkt af van $\mu = 0$, maar niet veel.

Je kunt dat zien aan de kans dat $\mu(E)$ nog verder van 0 zit: $P(\mu(E) \geq 0,48 | \mu = 0 \text{ en } \sigma = 1,3) \approx 0,356$. In de normale verdeling met $\mu = 0$ en $\sigma = 1,3$ is deze overschrijdingskans zo groot dat $\mu(E)$ naar verhouding dicht bij 0 zit. Meestal spreek je vooraf af hoe groot die overschrijdingskans maximaal mag zijn, dat is vaak 5%, dus 0,05. Dit heet het significantieniveau.

Omdat in dit geval de overschrijdingskans veel groter is dan het significantieniveau, zeg je dat het verschil niet significant is.

Opgave 1

Bekijk in **Uitleg 1** hoe de eindtijden van de mannen en de vrouwen van de Chicago Marathon 2016 worden vergeleken met behulp van boxplots.

- Leg uit waarom hier de conclusie moet worden getrokken dat het verschil gering is.
- Kies zelf twee andere deelgroepen (bijvoorbeeld twee leeftijdscategorieën) waarvan je de eindtijden wilt vergelijken. Gebruik weer boxplots.

Opgave 2

Bekijk in **Uitleg 1** hoe de eindtijden van de mannen en de vrouwen van de Chicago Marathon 2016 worden vergeleken met behulp van normale verdelingen.

- Bereken zowel voor eindtijden van de mannen als die van de vrouwen zowel het gemiddelde als de standaarddeviatie.
- Hoe kun je nagaan of van beide deelgroepen de eindtijden bij benadering normaal zijn verdeeld? Voer dit ook zelf uit.
- Je maakt een verschilverdeling van de eindtijden van de mannen en de vrouwen. Daarna bereken je een overschrijdingskans. Geef in een schets van die normale verschilverdeling deze overschrijdingskans door arceren aan en bereken hem ook zelf.
- Licht met behulp van je figuur toe dat als die overschrijdingskans veel kleiner zou zijn geweest, het verschil tussen $\mu(E)$ en $\mu = 0$ behoorlijk groter zou zijn geweest.
- Het significantieniveau is de maximale overschrijdingskans. Als het significantieniveau 0,05 is, hoe groot zou $\mu(E)$ dan minimaal moeten zijn als het verschil significant is?
- Kies zelf twee andere deelgroepen (bijvoorbeeld twee leeftijdscategorieën) waarvan je de eindtijden wilt vergelijken. Gebruik nu bijbehorende normale verdelingen.

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.

Hier zie je een kruistabel 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 3

	PvdA		VVD		
	p	cp	p	cp	Vcp
Zeker op welke partij	34,4	34,4	52,1	52,1	17,7
Twijfel	55,7	90,1	41,0	93,1	3,0
Weet niet	9,9	100,0	6,9	100,0	0,0
Ik ga niet stemmen	0,0	100,0	0,0	100,0	0,0

Figuur 4

Je kunt hierin bijvoorbeeld aflezen dat het aantal twijfelaars bij de PvdA groter is dan bij de VVD, ook naar verhouding. Je kunt de variabele 'Zekerheid_keuze' voor deze twee partijen vergelijken met behulp van cumulatieve percentages. Je ziet dat in de tweede tabel. In de laatste kolom vind je het verschil van de cumulatieve percentages V_{cp} . Het maximum hiervan is $\max V_{cp} = 17,7\%$ en daarvoor kun je de volgende vuistregels gebruiken:

- Als $\max V_{cp} > 40\%$, dan zeg je "Het verschil is groot".
- Als $20\% < \max V_{cp} \leq 40\%$, dan zeg je "Het verschil is middelmatig".
- Als $\max V_{cp} \leq 20\%$, zeg je "Het verschil is gering".

Het verschil tussen de keuzezekerheid bij deze twee partijen is dus gering.

	Zorg_belangrijk		
Geslacht	nee	ja	
M	104	1106	1210
V	20	310	330
	124	1416	1540

Figuur 5

Een speciaal geval is nog de twee bij twee kruistabel die je kunt maken als je bijvoorbeeld kijkt naar het verschil tussen mannen en vrouwen bij de variabele 'Zorg_belangrijk'.

Hierbij kun je mannen en vrouwen vergelijken met behulp van het getal ϕ :

$$\phi = \frac{104 \cdot 310 - 1106 \cdot 20}{\sqrt{1210 \cdot 330 \cdot 124 \cdot 1416}} \approx 0,0382.$$

Ook hierbij horen vuistregels:

- Als $\phi < -0,4$ of $\phi > 0,4$, dan zeg je "Het verschil is groot".
- Als $-0,4 \leq \phi < -0,2$ of $0,2 < \phi \leq 0,4$, dan zeg je "Het verschil is middelmatig".
- Als $-0,2 \leq \phi \leq 0,2$, zeg je "Het verschil is gering".

Ook hier is het verschil gering.

Opgave 3

Bekijk in **Uitleg 2** het vergelijken van twee partijen bij de variabele 'Zekerheid_keuze' bij het bestand NL verkiezingen in 2012.

- a 'Zekerheid_keuze' heet wel een ordinale kwalitatieve variabele. Hierbij geeft ordinaal aan dat er sprake is van een ordening, een duidelijke volgorde. Licht dit toe.

- b Maak zelf de vergelijkingstabel voor de PVV en GL van 'Zekerheid_keuze'. Probeer een conclusie te trekken met behulp van de $\max V_{cp}$.
- c Bedenk nog twee andere variabelen uit de dataset TK-Verkiezingen-NL-2012 die je met behulp van $\max V_{cp}$ kunt vergelijken.

Opgave 4

Bekijk in **Uitleg 2** hoe je bij een twee keer twee kruistabel het getal ϕ kunt gebruiken om te vergelijken of er verschil is tussen de antwoorden van mannen en vrouwen betreffende de variabele 'Zorg_belangrijk'.

- a Maakt het verschil welke variabele horizontaal en welke verticaal staat?
- b Reken zelf de waarde van ϕ na.

Theorie en voorbeelden

Om te onthouden

Overzicht van manieren om variabelen te vergelijken. Leer dit niet uit het hoofd, maar zoek het steeds op als je het nodig hebt.

Kwantitatieve variabelen vergelijken kan

- met behulp van boxplots. Daarbij hoort de afspraak:
 - Als de boxen elkaar niet overlappen, dan zeg je "Het verschil is groot".
 - Als de boxen elkaar wel overlappen en minstens één mediaan buiten de box van de andere boxplot ligt, dan zeg je "Het verschil is middelmatig".
 - In alle andere gevallen zeg je "Het verschil is gering".
- met behulp van de (normale) verschilverdeling van hun normale verdelingen. Daarbij ga je uit van het feit dat als de gemiddelden van de verdelingen van X en Y niet verschillen, voor de verschilverdeling geldt $\mu = \mu(X) - \mu(Y) = 0$ met $\sigma = \sqrt{(\sigma(X))^2 + (\sigma(Y))^2}$. Als $\mu(X - Y) \neq 0$ bereken je de overschrijdingskans hierbij. Is die overschrijdingskans groter dan het significantieniveau (de maximaal afgesproken overschrijdingskans) dan is het verschil niet significant (niet betekenisvol). Meestal neem je een significantieniveau van 5% ofwel 0,05.

Kwalitatieve variabelen vergelijken kan

- bij een twee bij twee kruistabel met behulp van ϕ :

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Daarbij hoort de afspraak:

- Als $\phi < -0,4$ of $\phi > 0,4$, dan zeg je "Het verschil is groot".
- Als $-0,4 \leq \phi < -0,2$ of $0,2 < \phi \leq 0,4$, dan zeg je "Het verschil is middelmatig".
- Als $-0,2 \leq \phi \leq 0,2$, zeg je "Het verschil is gering".
- bij een kruistabel waarin per categorie uitkomsten zijn die een duidelijke volgorde hebben (ordinale variabele) met behulp van $\max V_{cp}$, het maximale cumulatieve percentage. Daarbij hoort de afspraak:
 - Als $\max V_{cp} > 40\%$, dan zeg je "Het verschil is groot".
 - Als $20\% < \max V_{cp} \leq 40\%$, dan zeg je "Het verschil is middelmatig".
 - Als $\max V_{cp} \leq 20\%$, zeg je "Het verschil is gering".

	wel	niet	
groep A	a	b	a+b
groep B	c	d	c+d
	a+c	b+d	totaal

Figuur 6

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

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

Neem bijvoorbeeld het vak natuurkunde en bekijk de cijfers van de meisjes en de jongens afzonderlijk.

Je hebt dan te maken met een continue kwantitatieve variabele, dus je kunt bijvoorbeeld boxplots gaan vergelijken. En dan kun je met behulp van de vuistregels een conclusie trekken.

Opgave 5

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

- Maak boxplots van de resultaten van de meisjes en de jongens bij het CE1-cijfer voor het vak natuurkunde.
- Trek conclusies met behulp van de vuistregels voor het vergelijken van boxplots.
- Op deze wijze kun je ook andere exacte vakken vergelijken. Probeer iets zinnigs te zeggen over de onderzoeksvraag.

Opgave 6

Neem weer de cijfers voor natuurkunde uit het databestand uit [Voorbeeld 1](#). Of werk met andere geschikte gegevens van je eigen school.

- Bereken zowel van de jongens als van de meisjes het gemiddelde en de standaardafwijking.
Ga er van uit, dat de CE1-cijfers van de jongens en de meisjes normaal zijn verdeeld. (In werkelijkheid zou je dit natuurlijk zelf eerst onderzoeken!) Je gaat beide normale verdelingen vergelijken met behulp van hun verschilverdeling.
- Welk gemiddelde en welke standaardafwijking heeft die verschilverdeling?
- Als beide deelgroepen het examen natuurkunde even goed hebben gemaakt, verwacht je voor de verschilverdeling $\mu = 0$. Onderzoek met een significantieniveau van 10% of het gevonden gemiddelde daar significant van afwijkt.

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 kun je een conclusie trekken?

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

Omdat het hier kwalitatieve variabelen betreft moet je ofwel naar de $\max V_{cp}$ kijken, ofwel werken met ϕ .

Op het eerste gezicht lijkt dat laatste niet mogelijk, maar je kunt natuurlijk beslissen om de 10 mensen die deze vragen niet hebben beantwoord, weg te laten.

Je hebt dan een twee bij twee kruistabel en daarvan kun je ϕ berekenen.

Opgave 7

Bekijk in **Voorbeeld 2** het verhaal van de NL verkiezingen in 2012 en de gestelde onderzoeksvraag.

- a Laat uit de tabel de rij '(Geen)' weg en bereken ϕ .
- b Welke conclusie hoort hier bij?

Opgave 8

Bekijk in **Voorbeeld 2** het verhaal van de NL verkiezingen in 2012. Let op de variabelen 'Ge-hoor_bij_politiek' en 'Ga_stemmen_op'.

- a Maak hierbij een kruistabel.
- b Vergelijk de twee partijen PvdA en PVV op dit gebied met behulp van $\max V_{cp}$.

Verwerken

Opgave 9

Je bent bezig met een eigen onderzoek opzetten. Je hebt je data geanalyseerd, tijd voor conclusies.

- a Ga na wat voor soort statistische variabelen je hebt en wat dit betekent voor de manier waarop je conclusies kunt trekken.
- b Trek conclusies.
- c 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?
- d 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.
