Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

De statistische achtergrond van de vuistregels op het havo eindexamen wiskunde A

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

FRÉDÈRIQUE BAAIJ

Delft, Nederland
Augustus 2019

Inhoudsopgave

1 Inleiding

Sinds het schooljaar 2015-2016 krijgen leerlingen op de havo die wiskunde A volgen een formuleblad bij hun eindexamen. Op dit blad staan een aantal statistische vuistregels omschreven om te bepalen of het verschil tussen twee groepen *groot*, *middelmatig*, of *gering* is. Waar deze vuistregels statistisch gezien vandaan komen, wordt de leerlingen niet verteld. Mijn interesse is gewekt door de achtergrond van deze regels en daar heb ik mij dan ook in dit onderzoek in verdiept.

Bij het uitzoeken van de wiskundige achtergrond van de vuistregels gaan we er overal van uit dat de data normaal verdeeld is, tenzij dit expliciet anders staat vermeld. Normaal verdeeld houdt in "verdeeld volgens de *normale verdeling*", waarbij links en rechts van het gemiddelde evenveel waarnemingen liggen die in een klokvorm zijn verdeeld als we deze zouden visualiseren.

2 Onderzoeksvraag

"Wat is de statistische achtergrond van de vuistregels op het formuleblad bij het eindexamen havo wiskunde A?"

3 Het formuleblad

FORMULEBLAD

Vuistregels voor de grootte van het verschil van twee groepen

2×2 kruistabel $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, met $phi = \dfrac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$

- als $phi < -0,4$ of $phi > 0,4$, dan zeggen we “het verschil is groot”,
- als $-0,4 \leq phi < -0,2$ of $0,2 < phi \leq 0,4$, dan zeggen we “het verschil is middelmatig”,
- als $-0,2 \leq phi \leq 0,2$, dan zeggen we “het verschil is gering”.

Maximaal verschil in cumulatief percentage ($\max V_{\text{cp}}$) (met steekproefomvang $n > 100$)

- als $\max V_{\text{cp}} > 40$, dan zeggen we “het verschil is groot”,
- als $20 < \max V_{\text{cp}} \leq 40$, dan zeggen we “het verschil is middelmatig”,
- als $\max V_{\text{cp}} \leq 20$, dan zeggen we “het verschil is gering”.

Effectgrootte $E = \dfrac{\overline{X}_1 - \overline{X}_2}{\frac{1}{2}(S_1 + S_2)}$, met $\overline{X}_1$ en $\overline{X}_2$ de steekproefgemiddelden ($\overline{X}_1 \geq \overline{X}_2$), S_1 en S_2 de steekproefstandaardafwijkingen

- als $E > 0,8$, dan zeggen we “het verschil is groot”,
- als $0,4 < E \leq 0,8$, dan zeggen we “het verschil is middelmatig”,
- als $E \leq 0,4$, dan zeggen we “het verschil is gering”.

Twee boxplots vergelijken

- als de boxen[1)] elkaar niet overlappen, dan zeggen we “het verschil is groot”,
- als de boxen elkaar wel overlappen en een mediaan van een boxplot buiten de box van de andere boxplot ligt, dan zeggen we “het verschil is middelmatig”,
- in alle andere gevallen zeggen we “het verschil is gering”.

noot 1 De ‘box’ is het interval vanaf het eerste kwartiel tot en met het derde kwartiel.

4 De 2 × 2 kruistabel

Als eerste vuistregel bekijken we de regel rond de 2 × 2 kruistabel. Deze vuistregel is als volgt geformuleerd:

2×2 kruistabel $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, met $phi = \dfrac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$

- als $phi < -0{,}4$ of $phi > 0{,}4$, dan zeggen we "het verschil is groot",
- als $-0{,}4 \leq phi < -0{,}2$ of $0{,}2 < phi \leq 0{,}4$, dan zeggen we "het verschil is middelmatig",
- als $-0{,}2 \leq phi \leq 0{,}2$, dan zeggen we "het verschil is gering".

4.1 Wiskundige achtergrond

Rond 1904 heeft Karl Pearson een aantal grootheden geïntroduceerd. Een van deze grootheden is de Pearson correlation coefficient ϕ, welke hij heeft gedefinieerd aan de hand van de volgende kruistabel waarbij N staat voor de volledige populatie:

$$\begin{array}{cc|c} a & b & a+b \\ c & d & c+d \\ \hline a+c & b+d & N \end{array}$$

Uit deze kruistabel heeft hij de Pearson correlation coefficient ϕ^2 als volgt afgeleid:

$$\phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Als we uit deze formule de ϕ zelf willen bepalen, vinden we het volgende:

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

Om de grenswaarden voor een *gering*, *middelmatig* of *groot* verschil te bepalen, kijken we naar twee andere door Pearson geïntroduceerde grootheden, namelijk de mean square contingency, r en de mean contingency, ψ.

De mean square contingency, welke een soort schaling van ϕ naar het interval $[0, 1]$ is, is aan de hand van het bovengenoemde door Pearson als volgt gedefinieerd:

$$r = \sqrt{\frac{\phi^2}{1 + \phi^2}} \tag{1}$$

Verder kijken we naar de mean contingency ψ. Deze is als volgt door Pearson gedefinieerd:

$$\psi = \frac{2(ad - bc)}{N^2}$$

Voor de mean square contingency r vinden we de volgende beslisregel van Pearson:

- De overeenkomst tussen de variabelen is klein wanneer de waarde van r dichtbij 1 ligt *(het verschil is groot)*

- Er is veel overeenkomst tussen de variabelen als de waarde van r dichtbij 0 ligt *(het verschil is gering)*

We vinden in "On the theory of contingency and its relation to association and normal correlation" de onderstaande tabel voor r en ψ.
De grootheden I en Q zoals deze hier in de tabel zijn genoemd, zijn voor ons op dit moment niet van relevantie dus deze laten we buiten beschouwing.

TABLE I.—Table of Integrals I_r, Q_r, and the Contingency ψ for Values of r.

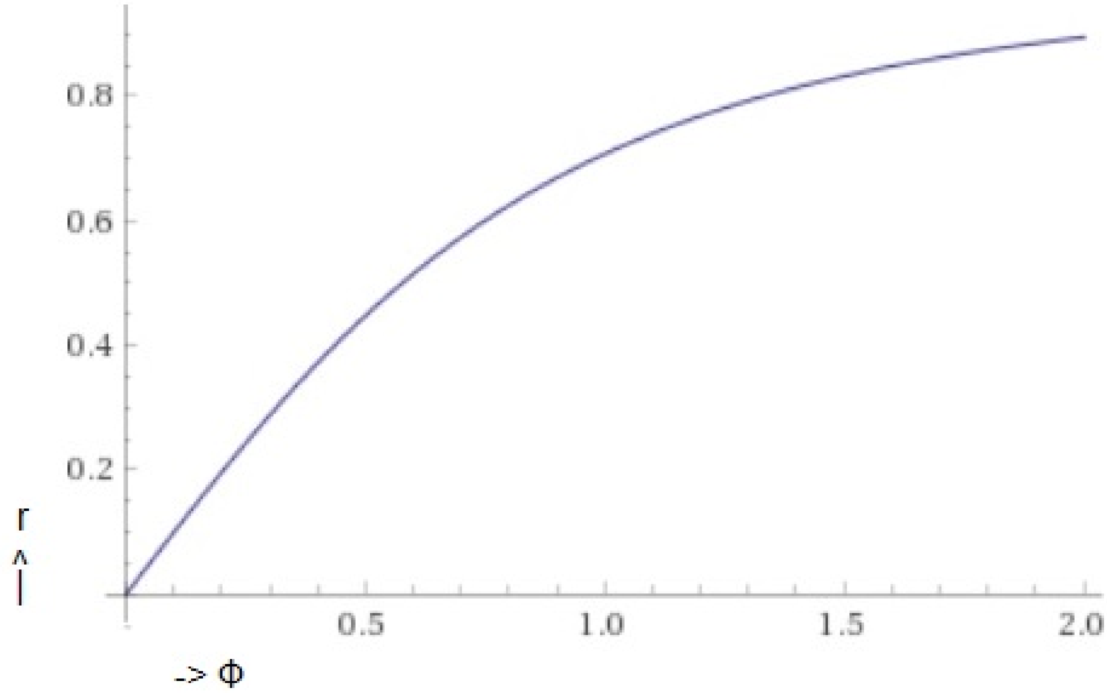
r.	I_r.	Q_r.	ψ.
0·00	·5000	·5000	·0000
·05	·4620	·4762	·0142
·10	·4342	·4652	·0310
·20	·3895	·4536	·0641
·30	·3501	·4498	·0996
·40	·3162	·4547	·1385
·50	·2830	·4643	·1813
·60	·2489	·4814	·2325
·70	·2128	·5106	·2978
·80	·1700	·5524	·3824
·90	·1186	·6279	·5093
·95	·0796	·7009	·6213
1·00	·0000	1·0000	1·0000

Voor een ϕ van 0.2 vinden we een r van $\sqrt{\frac{0.2^2}{1+0.2^2}} \approx 0.2$

Voor een ϕ van 0.4 vinden we een r van $\sqrt{\frac{0.4^2}{1+0.4^2}} \approx 0.4$

Wanneer we kijken naar een ϕ van bijvoorbeeld 0.5, zegt Pearson niets. De hierbij behorende r is $\sqrt{\frac{0.5^2}{1+0.5^2}} \approx 0.45$, dit is niet dichtbij 0 of dichtbij 1.
Tot slot beschouwen we een r dicht bij 1, zeg $r = 0.9$. Hierbij vinden we een ϕ van $\sqrt{\frac{0.9^2}{1-0.9^2}} \approx 2.1$

Wanneer we ϕ op de horizontale as uitzetten tegen r op de verticale as, zien we het volgende verband:



Hieruit kunnen we afleiden dat hoe groter r wordt, hoe groter ϕ wordt. Dus voor een grote waarde van ϕ is het verschil *groot* en voor een kleine waarde van ϕ is het verschil *gering.*

Deze Pearson correlation coefficient ϕ is ook te berekenen voor grotere matrices dan een 2×2-matrix, waarover meer te lezen is in de volgende terzijde. Dit is een geavanceerder gedeelte dan het voorafgaande, wat over kan worden geslagen zonder consequenties voor het begrip van de rest van het onderwerp.

4.1.1 Terzijde: de Pearson correlation coefficient voor grotere matrices

Wanneer wij ons verder verdiepen in Pearson's "On the theory of contingency and its relation to association and normal correlation", vinden we dat de Pearson correlation coefficient ook is gedefinieerd voor grotere matrices dan een 2×2-matrix. We gebruiken dan in de betreffende $k \times m$-matrix niet de letters van het alfabet, maar de abstractere notatie $a_{i,j}$ voor de waardes in de matrix, waarbij de i staat voor de rij waarin de waarde staat die we bekijken en de j voor de kolom van deze waarde.

De Pearson correlation coefficient ϕ^2 kunnen we nu uitrekenen aan de hand van χ^2 en de populatiegrootte N. We definiëren allereerst χ^2 als volgt, zoals deze ook in de χ^2-toets, welke wij hier niet gebruiken, is gedefinieerd:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(a_{i,j} - \frac{a_{i,+}a_{+,j}}{n})^2}{\frac{a_{i,+}a_{+,j}}{n}}$$

Hierbij is $a_{i,+}$ de som van alle waardes in rij i en $a_{+,j}$ de som van alle waardes in kolom j

en n het totaal aantal waarnemingen in de matrix.

Nu kunnen we ϕ^2 uitrekenen:

$$\phi^2 = \frac{\chi^2}{N}$$

We gaan hier verder niet in op een toepassing aangezien dit te abstract is voor het niveau dat een examenleerling zou moeten beheersen.

4.2 Toepassing op het eindexamen

Voordat we verder naar de wiskundige achtergrond van deze vuistregel kijken, bekijken we eerst hoe deze op het eindexamen toegepast kan worden. Hiertoe analyseren we de opgave "Zorginfecties" uit het eindexamen havo wiskunde A tijdvak 2 2016.

In de periode 2007 tot en met 2012 is een steekproef gehouden onder een deel van de Nederlandse ziekenhuizen. Enkele resultaten hiervan zijn in de tabel te zien.

tabel

	aantal
patiënten	95 299
patiënten die een zorginfectie hebben opgelopen	4694
geopereerde patiënten	32 664
geopereerde patiënten die een zorginfectie hebben opgelopen	1286

Het is mogelijk om op basis van de tabel een kruistabel te maken. Op de uitwerkbijlage is een begin gemaakt met deze kruistabel. Met behulp van de ingevulde kruistabel kun je bepalen of het verschil in het krijgen van een zorginfectie tussen geopereerde en niet-geopereerde patiënten groot, middelmatig of gering is.

6 Vul de kruistabel op de uitwerkbijlage in en bepaal daarmee, en met behulp van een vuistregel op het formuleblad, of het genoemde verschil groot, middelmatig of gering is.

4.2.1 De aanpak zoals een leerling het zou doen

Om meer inzicht in de vraag te krijgen, gaan we deze eerst oplossen zoals een leerling dit zou doen. We maken hierbij enkel gebruik van de gegeven vuistregel en de gegevens uit de vraag.

We vullen allereerst de kruistabel in:

		geopereerd		
		wel	niet	**totaal**
zorginfectie opgelopen	wel	1286	3408	4694
	niet	31378	59227	90605
	totaal	32664	62635	95299

Aan de hand van de gegeven vuistregel, gaan we nu de waarde van ϕ bij deze kruistabel berekenen. We krijgen

$$\begin{aligned}
\phi &= \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \\
\phi &= \frac{1286 * 59227 - 3408 * 31378}{\sqrt{(1286+3408)(1286+31378)(3408+59227)(31378+59227)}} \\
\phi &\approx -0.033
\end{aligned}$$

De gevonden ϕ voldoet aan $-0.2 \leq \phi \leq 0.2$ dus op het formuleblad lezen we af dat het verschil *gering* is.

4.2.2 De aanpak zoals een wiskundige het zou doen

We berekenen nu ϕ via de Pearson correlation coefficient. Onze observaties zijn te vinden in de kruistabel, we krijgen de volgende waarden voor a, b, c en d:

$$\begin{aligned}
a &= 1286 \\
b &= 3408 \\
c &= 31378 \\
d &= 59227 \\
a+b &= 4694 \\
c+d &= 90605 \\
a+c &= 32664 \\
b+d &= 62635 \\
N &= 95299
\end{aligned}$$

Dit geeft ons

$$\begin{aligned}
\phi &= \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \\
\phi &\approx -0.033
\end{aligned}$$

We vullen nu ϕ in (1) in om r te bepalen:

$$r \approx 0.033$$

Als we nu kijken naar de vuistregel, zien we dat ook volgens Pearson het verschil *gering* is.

4.3 Conclusie over deze vuistregel

Inmiddels is het duidelijk geworden dat de vuistregel van de Pearson contingency coefficient komt. De beslisregel komt van de r waardes, maar waarom er precies voor de grenzen van $\phi = 0.2$ en $\phi = 0.4$ is gekozen, is mij niet geheel duidelijk geworden.

5 Maximaal verschil in cumulatief percentage

Maximaal verschil in cumulatief percentage ($\max V_{cp}$) (met steekproefomvang $n > 100$)

- als $\max V_{cp} > 40$, dan zeggen we "het verschil is groot",
- als $20 < \max V_{cp} \leq 40$, dan zeggen we "het verschil is middelmatig",
- als $\max V_{cp} \leq 20$, dan zeggen we "het verschil is gering".

Het maximaal verschil in cumulatief percentage is tot op heden nog niet voorgekomen op het eindexamen, dus bouwen we hier enkel de wiskundige achtergrond op. Toepassing op een examenvraag blijft hier dus achterwege.

5.1 Wiskundige achtergrond

Wanneer wij een tabel met frequenties hebben, kunnen we hier bijvoorbeeld percentages mee berekenen. Hieronder hebben we een tabel met frequenties die we voor dit voorbeeld zullen gebruiken, met aan de bovenkant de waarnemingen en eronder de frequentie bij iedere waarneming (de frequentie geeft weer hoe vaak een waarneming voorkomt). We kijken naar het aantal SMS'jes dat per week binnenkomt, de frequentie is hierbij het aantal leerlingen waarbij een bepaald aantal SMS'jes binnenkomt. De tabel berust puur op fictieve waarden:

Aantal SMS'jes per week bij klas 5H

Aantal	0	1	2	3	4	5	>5
Frequentie	32	20	28	20	12	4	4

We zien in deze tabel dat er in totaal 120 leerlingen zijn ondervraagd. Deze 120 leerlingen zullen bij het berekenen van percentages ons *totaal* gaan vormen.
We kunnen berekenen bij hoeveel procent van de leerlingen een bepaald aantal SMS'jes binnenkomt, dit berekenen we door:

$$\text{percentage bij frequentie} = \frac{\text{frequentie}}{\text{totaal}} \times 100\%$$

Als we bij de waarnemingen van bovenstaande frequentietabel de percentages per frequentie uit gaan rekenen, krijgen we de volgende tabel:

De frequenties en percentages bij het aantal SMS'jes van klas 5H

Aantal	Frequentie	Percentage
0	32	$\frac{32}{120} \times 100\% \approx 26.7\%$
1	20	$\frac{20}{120} \times 100\% \approx 16.7\%$
2	28	$\frac{28}{120} \times 100\% \approx 23.3\%$
3	20	$\frac{20}{120} \times 100\% \approx 16.7\%$
4	12	$\frac{12}{120} \times 100\% = 10.0\%$
5	4	$\frac{4}{120} \times 100\% \approx 3.3\%$
>5	4	$\frac{4}{120} \times 100\% \approx 3.3\%$

Aan de hand van de frequenties kunnen we vervolgens de cumulatieve frequentie uitrekenen, dit is de frequentie van alle waarnemingen tot en met het getal van de waarneming (de cumulatieve frequentie van het getal 2 is de som van de frequenties van 0, 1 en 2). Vervolgens kunnen we met deze cumulatieve frequenties het cumulatief percentage uitrekenen, wat het percentage is behorend bij de cumulatieve frequentie, welke als volgt gedefinieerd is:

$$\text{cumulatief percentage} = \frac{\text{cumulatieve frequentie}}{\text{totaal}} \times 100\%$$

Wanneer we verder kijken naar het voorbeeld met het aantal SMS'jes van klas 5H, kunnen we de bovenstaande tabel nu uitbreiden met de cumulatieve frequentie en het cumulatief percentage:

De (cumulatieve) frequenties en (cumulatieve) percentages bij het aantal SMS'jes van klas 5H

Aantal	Frequentie	Cum. frequentie	Percentage	Cum. percentage
0	32	32	$\frac{32}{120} \times 100\% \approx 26.7\%$	$\frac{32}{120} \times 100\% \approx 26.7\%$
1	20	52	$\frac{20}{120} \times 100\% \approx 16.7\%$	$\frac{52}{120} \times 100\% \approx 43.3\%$
2	28	80	$\frac{28}{120} \times 100\% \approx 23.3\%$	$\frac{80}{120} \times 100\% \approx 66.7\%$
3	20	100	$\frac{20}{120} \times 100\% \approx 16.7\%$	$\frac{100}{120} \times 100\% \approx 83.3\%$
4	12	112	$\frac{12}{120} \times 100\% = 10.0\%$	$\frac{112}{120} \times 100\% \approx 93.3\%$
5	4	116	$\frac{4}{120} \times 100\% \approx 3.3\%$	$\frac{116}{120} \times 100\% \approx 96.7\%$
>5	4	120	$\frac{4}{120} \times 100\% \approx 3.3\%$	$\frac{120}{120} \times 100\% = 100\%$

Aangezien we geïnteresseerd zijn in het verschil tussen twee groepen, hebben we gegevens van een tweede groep nodig. We kijken hiervoor ook naar het aantal binnengekomen SMS'jes per week bij klas 4H:

Aantal SMS'jes per week bij klas 4H

Aantal	0	1	2	3	4	5	>5
Frequentie	40	28	4	12	20	4	12

Nu maken we ook voor 4H een tabel met de percentages en cumulatieve percentages zoals we dit eerder voor 5H deden:

De (cumulatieve) frequenties en (cumulatieve) percentages bij het aantal SMS'jes van klas 4H

Aantal	Frequentie	Cum. frequentie	Percentage	Cum. percentage
0	40	40	$\frac{40}{120} \times 100\% \approx 33.3\%$	$\frac{40}{120} \times 100\% \approx 33.3\%$
1	28	68	$\frac{28}{120} \times 100\% \approx 23.3\%$	$\frac{52}{120} \times 100\% \approx 56.7\%$
2	4	72	$\frac{4}{120} \times 100\% \approx 3.3\%$	$\frac{72}{120} \times 100\% = 60\%$
3	12	84	$\frac{12}{120} \times 100\% = 10\%$	$\frac{84}{120} \times 100\% = 70\%$
4	20	104	$\frac{20}{120} \times 100\% \approx 16.7\%$	$\frac{104}{120} \times 100\% \approx 86.7\%$
5	4	108	$\frac{4}{120} \times 100\% \approx 3.3\%$	$\frac{108}{120} \times 100\% = 90\%$
>5	12	120	$\frac{12}{120} \times 100\% = 10\%$	$\frac{120}{120} \times 100\% = 100\%$

We kunnen de twee groepen met elkaar vergelijken aan de hand van het maximale verschil in cumulatief percentage. Dit verschil in cumulatief percentage is voor rij n (waarbij n het rijnummer in de tabel is) als volgt gedefinieerd, waarbij de verticale strepen aangeven dat we de absolute waarde nemen zodat ons antwoord altijd positief is:

$$V_{cp}(n) = |\text{cum. percentage van rij } n \text{ bij dataset } 1 - \text{cum. percentage van rij } n \text{ bij dataset } 2|$$

Aan de hand van de gegevens uit beide tabellen kunnen we nu het V_{cp} berekenen. We stellen hiervoor een tabel op met de cumulatieve frequenties en cumulatieve percentages van de beide klassen en berekenen hierbij het V_{cp} op de manier die hierboven gedefinieerd is:

Vergelijking van klas 5H met klas 4H

	Klas 5H		**Klas 4H**		
Aantal	Cum. freq.	Cum. perc.	Cum. freq.	Cum. perc.	V_{cp}
0	32	26.7%	40	33.3%	33.3 - 26.7 = 6.6%
1	52	43.3%	68	56.7%	56.7 - 43.3 = 13.4%
2	80	66.7%	72	60%	66.7 - 60 = 6.7%
3	100	83.3%	84	70%	83.3 - 70 = 13.3%
4	112	93.3%	104	86.7%	93.3 - 86.7 = 6.6%
5	116	96.7%	108	90%	96.7 - 90 = 6.7%
> 5	120	100%	120	100%	100 - 100 = 0%

Het maximaal verschil in cumulatief percentage (*max* V_{cp}) vinden we hier bij het aantal 1 (dus in rij 2), het V_{cp} is hier 13.4%. Als we nu kijken naar de vuistregel, zien we dat het verschil *gering* is.

5.1.1 Terzijde: het *max* V_{cp}

Strikt wiskundig gezien is het *max* V_{cp} de maximumnorm (de $||\cdot||_\infty$ - norm) op $\mathbb{R}^2$. Deze norm is voor een rij waarden $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, $i \in \{1, \cdots, n\}$ gedefinieerd als

$$||\mathbf{x}||_\infty = \max_{1\leq i\leq n} |x_i|$$

In dit geval kijken we dus naar het maximum van alle waarden voor het V_{cp}, welke we het $max\ V_{cp}$ noemen.
Deze maximumnorm is wel gevoelig voor uitschieters, het is hierom interessant om te kijken naar wat er zou gebeuren als we de 1-norm of de Euclidische norm zouden nemen.

Wanneer we de 1-norm nemen, welke is gedefinieerd als

$$||x||_1 = \sum_{i=1}^{n} |x_i|,$$

dan tellen we alle verschillen in cumulatief percentage bij elkaar op. In dit geval zouden we dan een "totaal V_{cp}" krijgen van

$$||V_{cp}||_1 = 6.6 + 13.4 + 6.7 + 13.3 + 6.6 + 6.7 = 53.3$$

Wanneer we naar de Euclidische norm kijken, welke is gedefinieerd als

$$||x||_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}},$$

tellen we alle V_{cp}'s in het kwadraat bij elkaar op en nemen hier de wortel van. In ons geval krijgen we dan een "totaal V_{cp}" van

$$||V_{cp}||_2 = \left(6.6^2 + 13.4^2 + 6.7^2 + 13.3^2 + 6.6^2 + 6.7^2\right)^{\frac{1}{2}} \approx 23.1$$

Om een echt $max\ V_{cp}$ te krijgen, zullen we een maximum moeten nemen. Om deze reden kan ik begrijpen dat voor het $max\ V_{cp}$ is gekozen voor de maximumnorm in plaats van een van de andere twee normen. Wel ben ik van mening dat de Euclidische norm een robuustere maat is om naar de verschillende cumulatieve percentageverschillen te kijken, alleen zou dit ook gepaard moeten gaan met andere grenswaarden.

5.2 De grenzen bij de vuistregel

Ondanks uitvoerig literatuuronderzoek is het onvindbaar waar de grenzen bij deze vuistregel vandaan komen. In afwachting van antwoord van het College Voor Toetsing en Examens kan ik hier nog niets over zeggen, de grenzen lijken willekeurig te zijn gekozen.

5.3 Conclusie over deze vuistregel

Deze vuistregel is nog nooit voorgekomen op het eindexamen, waardoor het moeilijk is om toepassingen te vergelijken. Verder lijkt er geen stevige wiskundige grondslag te liggen aan de grenzen van de vuistregel, wat deze vuistregel zeer dubieus maakt. De regel lijkt eerder uit de lucht gegrepen dan stevig wiskundig gefundeerd.

6 Effectgrootte

Effectgrootte $E = \dfrac{\overline{X}_1 - \overline{X}_2}{\frac{1}{2}(S_1 + S_2)}$, met $\overline{X}_1$ en $\overline{X}_2$ de steekproefgemiddelden ($\overline{X}_1 \geq \overline{X}_2$), S_1 en S_2 de steekproefstandaardafwijkingen

- als $E > 0{,}8$, dan zeggen we "het verschil is groot",
- als $0{,}4 < E \leq 0{,}8$, dan zeggen we "het verschil is middelmatig",
- als $E \leq 0{,}4$, dan zeggen we "het verschil is gering".

Wanneer we kijken naar de effectgrootte, bepalen we de mate waarin twee groepen van elkaar verschillen. Een voorbeeld hierbij is de effectiviteit van een nieuw medicijn testen, we kunnen dan de testgroep en een controlegroep met elkaar vergelijken om te zien of het medicijn effect heeft.

6.1 Wiskundige achtergrond

In het boek "Statistical Power Analysis for the Behavioral Sciences" uit 1967 beschrijft Jacob Cohen de effectgrootte.
De effectgroote bij Cohen, Cohen's d, is als volgt gedefinieerd:

$$d = \frac{\overline{X_1} - \overline{X_2}}{s_p}$$

Hierbij zijn $\overline{X_1}$ en $\overline{X_2}$ de gemiddeldes van respectievelijk de groepen 1 en 2 en s_p is de gezamenlijke standaarddeviatie die als volgt is gedefinieerd:

$$s_p = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

In bovenstaande definitie voor s_p zijn de n_1 en n_2 het aantal waarnemingen in groep 1 en groep 2 en de σ_1 en σ_2 zijn de standaardafwijkingen in de groepen 1 en 2.

Bij Cohen vinden we de volgende grenswaarden voor een *klein*, *middelmatig* of *groot* verschil terug:

- Het verschil is *klein* als $d = 0.2$
- Het verschil is *middelmatig* als $d = 0.5$
- Het verschil is *groot* als $d = 0.8$

Opvallend hierbij is dat Cohen niets zegt over de waarden die tussen deze drie getallen in liggen.

We zien hier hele andere grenzen dan op het examen worden gegeven, maar dit is niet vreemd aangezien we een zeer vereenvoudigde versie van Cohen's d hier krijgen. Op het formuleblad bij het examen is de waarde voor s_p onder de breukstreep sterk vereenvoudigd, hier wordt puur het gemiddelde genomen van de twee standaardafwijkingen zonder rekening te houden met de groepsgroottes.

De grenzen die worden gebruikt op het formuleblad lijken wel enigszins op de vuistregels die Cohen heeft gegeven voor de interpretatie van Cohen's d, maar naar alle waarschijnlijkheid hebben de examenmakers gekeken naar de grenswaarden van Cohen en hier vervolgens hun eigen interpretatie aan gegeven gezien de andere noemer van d.

Wanneer populatie 1 veel groter zou zijn dan populatie 2, dus als $n_1 >> n_2$, dan zien we dat het versimpelen van beide noemers leidt tot $\frac{1}{2}(\sigma_1 + \sigma_2) < \sqrt{\frac{1}{2}\sigma_1^2 + \frac{1}{2}\sigma_2^2}$, wat een mogelijke verklaring is voor het verschil tussen de grenzen van Cohen's d en die van de examenmakers. Opvallend hierbij blijft wel dat bij Cohen alleen een aantal waarden wordt genoemd en niets wordt gezegd over de waarden hiertussen, terwijl de examenmakers ook waarden tussen hun grenzen meenemen.

6.2 Toepassing op het eindexamen

In het eindexamen van 2017 vinden we een vraag over de scores op het gebied van rekenvaardigheid van leerlingen uit verschillende landen. In de vraag treffen we onderstaande tabel aan, met de vraag erbij om op twee verschillende manieren te bepalen of het verschil in de scores tussen twee landen *gering*, *middelmatig* of *groot* is.

In de tabel staan de **percentielen** van de scores van enkele deelnemende landen. Je kunt bijvoorbeeld aflezen dat het 75e percentiel van Australië 305,4 is. Dit betekent dat 75% van de Australische deelnemers een score van 305,4 of lager had.

tabel

land	gemiddelde score	standaard-afwijking	percentiel						
			5	10	25	50	75	90	95
Australië	267,6	56,6	169,3	197,7	234,7	271,9	305,4	334,3	351,6
Canada	265,5	55,5	169,2	194,2	230,8	269,8	303,9	332,4	349,3
Finland	282,2	52,2	193,6	217,4	250,8	285,8	317,3	345,0	360,8
Frankrijk	254,2	56,2	152,1	179,7	219,9	259,2	293,9	321,5	336,5
Duitsland	271,7	53,1	179,0	201,9	238,4	275,9	309,3	335,0	350,5
Italië	247,1	50,0	161,1	182,9	215,4	249,3	281,9	309,1	324,1
Japan	288,2	44,0	212,6	231,7	260,7	290,8	318,1	341,7	355,4
Nederland	280,3	51,1	188,6	214,6	251,0	285,8	315,3	339,7	354,2
Spanje	245,8	51,3	149,1	177,8	216,3	250,3	280,9	307,4	322,3
Zweden	279,1	54,9	181,7	209,9	249,2	284,0	316,0	342,8	358,4
USA	252,8	57,0	151,7	177,9	217,1	256,1	293,1	322,7	340,0
alle deel-nemers van de 23 landen	268,7	51,3	178,4	202,8	237,9	272,5	303,9	330,3	345,6

7 Bepaal met behulp van het formuleblad op twee verschillende manieren of het verschil tussen de scores die behaald zijn door de Canadese deelnemers en de scores die behaald zijn door de Spaanse deelnemers groot, middelmatig of gering is.

We kijken nu puur naar de aanpak aan de hand van de effectgrootte, hoewel we met de verschillende gegeven percentielen ook een boxplot zouden kunnen maken zoals we in de volgende paragraaf doen.
We vergelijken de testscores van leerlingen uit Canada met die van leerlingen uit Spanje, dit levert ons de volgende gegevens op:

Testscores van leerlingen uit Canada en Spanje

land	gemiddelde score	standaardafwijking
Canada	265.5	55.5
Spanje	245.8	51.3

6.2.1 De aanpak zoals een leerling het zou doen

Op het formuleblad vinden we de volgende formule:

$$E = \frac{\overline{X_1} - \overline{X_2}}{\frac{1}{2}(S_1 + S_2)}$$

Als groep 1 nemen we hier de gegevens van Canada, als groep 2 nemen we de gegevens van Spanje.
We hebben nu

- $\overline{X_1} = 265.5$
- $\overline{X_2} = 245.8$
- $S_1 = 55.5$
- $S_2 = 51.3$

Wanneer we deze gegevens invullen in bovenstaande formule, vinden we de volgende waarde voor de effectgrootte:

$$\begin{aligned} E &= \frac{265.5 - 245.8}{\frac{1}{2}(55.5 + 51.3)} \\ E &= \frac{19.7}{53.4} \approx 0.37 \end{aligned}$$

Als we nu kijken naar de grenswaarden van de vuistregel, zien we dat $E = 0.37 \leq 0.4$, dus het verschil tussen de testscores van de leerlingen uit Canada en Spanje is *gering*.

6.2.2 De aanpak zoals een wiskundige het zou doen

We passen Cohen's d toe op de gevonden data. We rangschikken allereerst de data die we nodig hebben, wederom beschouwen we de data van Canada als groep 1 en de data van Spanje als groep 2. De steekproefgemiddelden zijn gelijk aan de eerdergenoemde steekproefgemiddelden, we beschouwen bij deze steekproefgemiddelden de volgende standaarddeviaties:

- $\sigma_1 = 55.5$
- $\sigma_2 = 51.3$

We hebben bij bovenstaande data alleen één probleem: we weten niet hoeveel deelnemers er per groep waren. Het enige gegeven dat we hebben gekregen, is dat er totaal 5000 mensen uit 23 landen zijn ondervraagd. We gaan er hier van uit dat er in ieder land ongeveer evenveel mensen zijn ondervraagd, we nemen hierom $n_1 = 218$ en $n_2 = 217$ (immers, $\frac{5000}{23} \approx 217.4$). Wel heeft Spanje 10 miljoen inwoners meer dan Canada, dus of deze aanname correct is, is moeilijk te zeggen.

Aan de hand van bovenstaande gegevens kunnen we de gezamenlijke standaarddeviatie s_p uitrekenen:

$$
\begin{aligned}
s_p &= \sqrt{\frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{(n_1-1)+(n_2-1)}} \\
&= \sqrt{\frac{(218-1)*55.5^2+(217-1)*51.3^2}{(218-1)+(217-1)}} \\
&= \sqrt{\frac{1236859.29}{433}} \approx 53.446
\end{aligned}
$$

Vervolgens kunnen we nu Cohen's d berekenen, aan de hand waarvan we een conclusie kunnen gaan trekken:

$$
\begin{aligned}
d &= \frac{\overline{X_1}-\overline{X_2}}{s_p} \\
&= \frac{265.5-245.8}{53.664} \approx 0.37
\end{aligned}
$$

We hebben een d gevonden van ongeveer 0.37, deze waarde valt tussen de 0.2 en de 0.5 dus is het verschil tussen beide groepen *middelmatig*.

6.3 Conclusie over deze vuistregel

De vuistregel op het eindexamen is een versimpelde versie van Cohen's d. Wanneer we de gevonden data invullen in beide formules voor de effectgrootte, vinden we twee keer dezelfde waarde voor de effectgrootte. Alleen de grenzen zijn anders bij het eindexamen dan ze bij Cohen's d zijn, om deze reden vinden we via het formuleblad een *gering* verschil en via Cohen een *middelmatig* verschil. Wiskundig gezien is de manier van Cohen om de effectgrootte te bepalen nauwkeuriger, dit omdat hij ook rekening houdt met de groottes van beide groepen.

7 Twee boxplots vergelijken

Als laatste vuistregel beschouwen we de vuistregel over boxplots:

Twee boxplots vergelijken

- als de boxen[1)] elkaar niet overlappen, dan zeggen we "het verschil is groot",
- als de boxen elkaar wel overlappen en een mediaan van een boxplot buiten de box van de andere boxplot ligt, dan zeggen we "het verschil is middelmatig",
- in alle andere gevallen zeggen we "het verschil is gering".

7.1 Punt van kritiek

De boxplot is puur wiskundig gezien een grafische weergave van een dataset. Een grafisch hulpmiddel is bedoeld om iets visueel te ondersteunen, niet om conclusies aan te verbinden. Om deze reden ben ik van mening dat het gebruik van boxplots op deze wijze in het huidige eindexamen de leerlingen foutieve wiskunde aanleert. Zij leren namelijk conclusies te trekken aan de hand van een hulpmiddel, wat hetzelfde zou zijn als bijvoorbeeld een top van een grafiek aflezen in een zelfgemaakte schets. Er zijn nauwkeurigere middelen om het verschil tussen twee populaties te bekijken, bijvoorbeeld aan de hand van de kansdichtheden van beiden groepen.

7.2 Wiskundige achtergrond

Wanneer we kijken naar boxplots, beschouwen we vooral de grootte van de box met behulp van de interkwartielafstand en mediaan. De interkwartielafstand is de afstand tussen de waarneming die 25% van de dataset is en de waarneming die 75% van de dataset is, waarbij de dataset is gerangschikt van klein naar groot. De mediaan is gedefinieerd als de middelste waarneming in een van klein naar groot gerangschikte dataset. In het geval van een oneven aantal waarnemingen is dit het middelste getal, bij een even aantal waarnemingen is dit de som van de middelste twee waarnemingen gedeeld door 2.

Bij het vergelijken van twee boxplots beschouwen we de afstand tussen de medianen (de Distance Between Medians ofwel de DBM) en de Overall Visible Spread (OVS).

Om de OVS te kunnen bepalen, moeten we eerst weten wat de Q_1 en Q_3 van een boxplot zijn. De Q_1 is het punt waar 25% van de waarnemingen onder zit, de Q_3 is het punt waar 75% van de waarnemingen onder zit. Voor beide boxplots hebben we een verschillende Q_1 en Q_3, die van de eerste boxplot geven we aan als $Q_{1,1}$ en $Q_{3,1}$ en bij de tweede boxplot zijn dit $Q_{1,2}$ en $Q_{3,2}$.

De OVS is als volgt gedefinieërd:

$$\text{OVS} = \text{grootste } Q_3 \text{ van de twee boxplots - kleinste } Q_1 \text{ van de boxplots}$$

De DBM is als volgt gedefiniëerd:

$$\text{DBM} = \text{grootste mediaan - kleinste mediaan}$$

Aan de hand van deze grootheden kunnen we het procentuele verschil berekenen:

$$\text{procentueel verschil} = \frac{DBM}{OVS} \times 100\% \tag{2}$$

Als vuistregel vinden wij hierbij het volgende:

Er is waarschijnlijk een verschil tussen de groepen als:

- bij een sample size van 30 het procentuele verschil minimaal 33% is
- bij een sample size van 100 het procentuele verschil minimaal 20% is
- bij een sample size van 1000 het procentuele verschil minimaal 10% is

Het meest opvallende verschil hier is dus dat bij de vuistregel hier rekening wordt gehouden met de sample size, terwijl dat bij de vuistregel op het eindexamen niet zo is.

7.3 Toepassing op het eindexamen

Om te beginnen, pakken we de vraag over de testscores uit de vorige paragraaf nu aan via boxplots. We hebben de volgende gegevens:

De onderzoeksgegevens van Canada en Spanje

land	mediaan	Q_1	Q_3
Canada	269.8	230.8	303.9
Spanje	250.3	216.3	280.9

Aan de waarden van de eerste en derde kwartielen zien we dat de boxen elkaar overlappen. Verder ligt de mediaan van de box van Canada binnen de box van Spanje en de mediaan van de box van Spanje binnen de box van Canada. We vinden zo dat het verschil *gering* is, wat we ook al vonden toen we deze gegevens met de effectgrootte benaderden.

Verder kijken we naar de vraag "Uitvaltijd" uit het examen van 2017 tijdvak 2.

De directie berekent de mediaan, het eerste kwartiel en het derde kwartiel van de uitvaltijd voor elk van de drie diensten. Zie tabel 2.

tabel 2 uitvaltijd per dag- of nachtdienst in minuten

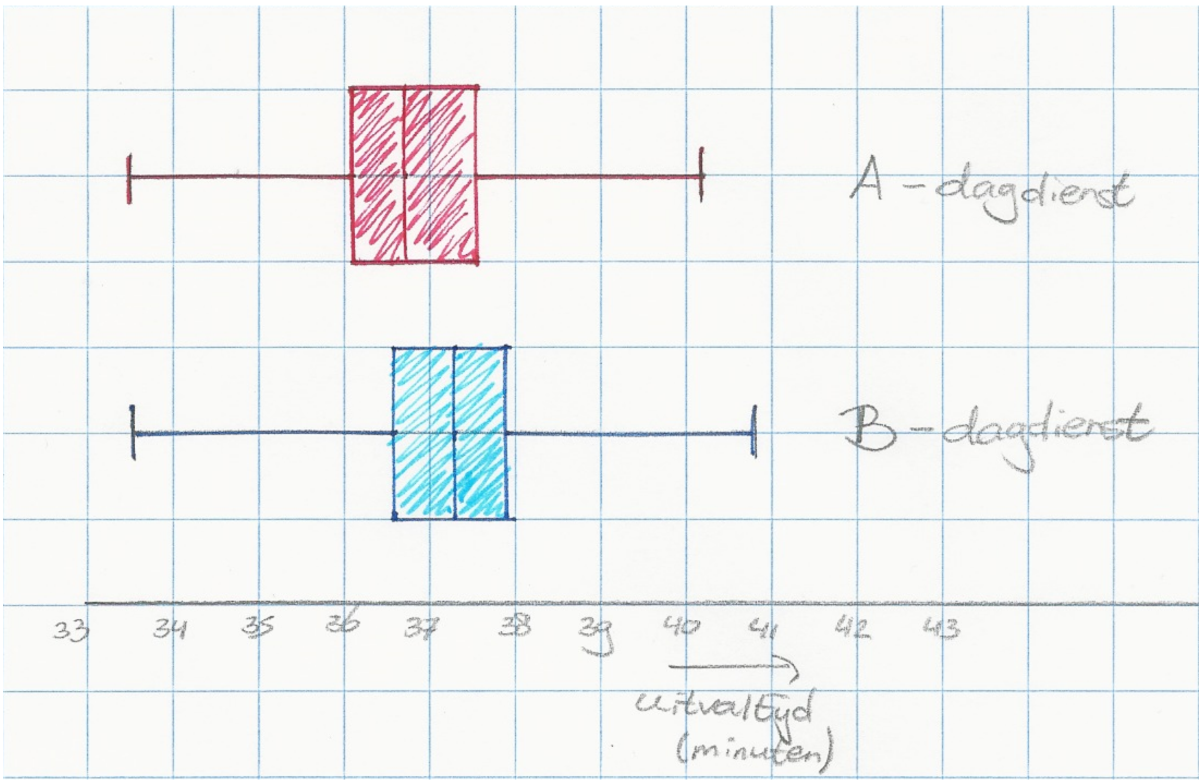
	mediaan	eerste kwartiel	derde kwartiel
dagdienst A (di-vr)	36,7	36,1	37,5
dagdienst B	37,3	36,6	37,9
nachtdienst	29,3	28,5	29,7

Eerst kijkt de directie alleen naar het verschil tussen dagdienst A en dagdienst B. Met behulp van boxplots kun je een uitspraak doen over het verschil tussen de uitvaltijden van de twee dagdiensten. Daarvoor hoeven de boxplots niet getekend te worden.

13 Bepaal met behulp van het formuleblad en tabel 2 of het verschil in uitvaltijd tussen dagdienst A en dagdienst B groot, middelmatig of gering is.

7.3.1 De aanpak zoals een leerling het zou doen

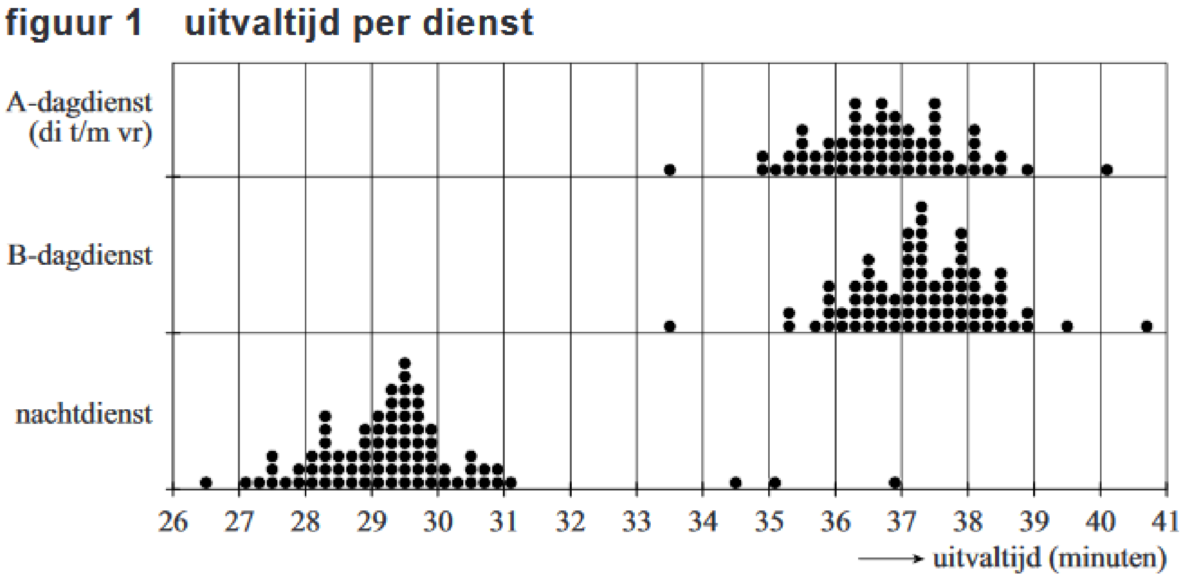
Aan de hand van de gegeven tabel kunnen we de boxplots tekenen:



We zien hier dat de medianen binnen elkaars boxen vallen. Wanneer we kijken naar de vuistregel, zien we nu dat het verschil *gering* is.

7.3.2 De aanpak zoals een wiskundige het zou doen

We beschouwen nu de vraag volgens de aanpak van het *procentuele verschil.* In de onderstaande dotplot die in een eerder onderdeel van de opgave is gegeven, lezen we het aantal waarnemingen af:



In deze dotplot zien we dat er 64 waarnemingen van A-dagdiensten zijn, 80 waarnemingen van B-dagdiensten en 80 waarnemingen van nachtdiensten.
We vergelijken de uitvalstijden van de A- en B-dagdiensten, waarbij we dus twee groepen met een ongelijk aantal waarnemingen vergelijken.

Aangezien er voor ons aantal waarnemingen geen directe beslisregel is over het procentuele verschil, leiden we deze beslisgrens af door middel van lineaire interpolatie van de eerder gegeven grenzen. Dit doen we door een formule op te stellen voor de lijn door de grenzen bij 30 en 100 waarnemingen en hier onze aantallen waarnemingen in in te vullen. De lijn van de vorm $y = ax + b$ bepalen we als volgt:

- Allereerst bepalen we de richtingscoëfficiënt a van de lijn die door de punten $(30, 33)$ en $(100, 20)$ gaat.

$$\begin{aligned} a &= \frac{\Delta y}{\Delta x} \\ a &= \frac{33-20}{30-100} \\ a &= -\frac{13}{70} \end{aligned}$$

- We hebben nu de formule $y = -\frac{13}{70}x + b$ welke door het punt $(30, 33)$ gaat.

- Nu gaan we b bepalen door $(30, 33)$ in te vullen in de bovenstaande formule:

$$
\begin{aligned}
33 &= -\frac{13}{70} * 30 + b \\
33 &= -\frac{39}{7} + b \\
b &= 33 + \frac{39}{7} \\
b &= 38\frac{4}{7}
\end{aligned}
$$

- We vinden dus $y = -\frac{13}{70}x + 38\frac{4}{7}$

Als we nu op de plaats van de x het aantal van 64 en 80 waarnemingen invullen, vinden we de procentuele verschillen zoals in onderstaande tabel:

Het aantal waarnemingen uitgezet tegen het procentuele verschil

Aantal waarnemingen	Procentueel verschil
30	33%
100	20%
64	26.7%
80	23.7%

Nu we voor beide aantallen waarnemingen de grens qua procentueel verschil hebben gevonden, kunnen we de beslisgrens afleiden voor als we deze twee groepen vergelijken:

$$
\text{procentueel verschil} = \frac{26.7 + 23.7}{2} \approx 25.2\%
$$

Er is dus een waarschijnlijk verschil (dat wil zeggen een middelmatig of groot verschil) tussen de groepen bij een procentueel verschil van minimaal 25.2%.

Om het procentuele verschil tussen de twee groepen te berekenen, hebben we eerst de DBM en OVS nodig. Deze berekenen we aan de hand van de eerder opgebouwde theorie:

$$
\begin{aligned}
\text{DBM} &= \text{grootste mediaan - kleinste mediaan} \\
&= 37.3 - 36.7 \\
&= 0.6
\end{aligned}
$$

Nu berekenen we de Overall Visible Spread:

$$
\begin{aligned}
\text{OVS} &= \text{grootste } Q_3 \text{ - kleinste } Q_1 \\
&= 37.9 - 36.1 \\
&= 1.8
\end{aligned}
$$

Aan de hand van deze gegevens berekenen we het procentuele verschil:

$$
\begin{aligned}
\text{procentueel verschil} &= \frac{DBM}{OVS} \times 100\% \\
&= \frac{0.6}{1.8} \times 100\% \\
&\approx 33.3\%
\end{aligned}
$$

Met behulp van de zojuist afgeleide beslisregel, kunnen we nu afleiden dat *er een waarschijnlijk verschil tussen de beiden groepen is.* Onze gevonden 33.3% is namelijk meer dan de 25.2% die we in de regel hadden gevonden. Er is dus een waarschijnlijk verschil tussen de groepen, alleen kan aan de hand van deze regel niet worden gezegd of dit verschil *middelmatig* of *groot* is.

7.4 Conclusie over deze vuistregel

Allereerst is zoals bij het "punt van kritiek" is aangegeven het gebruik van boxplots geen juiste manier van statistische conclusies trekken. Verder hebben we vanuit de wiskunde alleen een vuistregel gevonden die aangeeft of het verschil tussen twee groepen *waarschijnlijk* of *niet waarschijnlijk* is. Dit zouden we kunnen vertalen naar wanneer het verschil tussen de twee groepen *groot of middelmatig* of *gering* is. Wanneer het verschil precies *groot* is en wanneer het *middelmatig* is, hebben we helaas niet af kunnen leiden. Ook is het erg opvallend dat bij de vuistregel geen rekening wordt gehouden met de sample size, terwijl dit statistisch gezien wel wordt gedaan.

8 Conclusies

Kort samengevat is de vuistregel rondom de boxplots wiskundig gezien fout te noemen, omdat er conclusies worden verbonden aan een visualiseringsmiddel. Het is hierom ook niet wonderbaarlijk dat er niet precies terug te vinden is waar de criteria voor een *groot*, *middelmatig* of *gering* verschil vandaan komen.

De vuistregel rond de 2×2-kruistabel is wiskundig gezien de sterkste vuistregel van het formuleblad. Deze vuistregel is direct afgeleid van het werk van Pearson, alleen zijn de grenzen van Pearson niet overgenomen. Daar waar Pearson het heeft over waardes van r, gaat het in het examen over waardes van ϕ. Het is niet duidelijk waar precies de waarden 0.2 en 0.4 vandaan zijn gekomen in de vuistregel.

Rond de effectgrootte hebben we een versimpelde versie van Cohen's d gevonden. Ook hier wijken de grenzen weer af van de originele grenzen, dit wegens de vereenvoudiging. De vuistregel is hierom ook minder nauwkeurig dan de originele grootheid van Cohen.

De vaagste vuistregel is degene rond het maximale verschil in cumulatief percentage. Een stevige wiskundige fundering van deze test is niet te vinden, waardoor het ook een raadsel blijft waar de grenzen van de vuistregel vandaan komen.

Mijns inziens voegen de vuistregels niets toe aan het eindexamen, slechts twee van de vier regels hebben een redelijk solide wiskundige achtergrond. Bij alle vuistregels is niet precies duidelijk waarom er voor deze grenzen is gekozen, wat wiskundig gezien ronduit vaag is te noemen. De leerlingen wordt nu met de vuistregels alleen een "trucje" met opzoeken en invullen geleerd, terwijl dit niet is wat we met wiskunde willen bereiken.

9 Bronvermelding

Koolstra, G. (2018). ϕ als verschilmaat. *Euclides*, (7), 10-13.

Pearson, K. (1904). XIII. *On the theory of contingency and its relation to association and normal correlation* (pp. 6-8, 15, 16, 21). London: Dulau and co.

Salkind, N. (2011). *Statistics for people who (think they) hate statistics* (4th ed., pp. 197-199). SAGE

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Liebetrau, A. (1983). *Measures of association* (p. 12). Newbury Park, Calif.[u.a.]: Sage.

Carothers, N. (2000). *Real analysis* (p. 40). Cambridge: Cambridge University Press.

van den Berg, R. (2019). Cramér's V - Beginners Tutorial and Examples. Retrieved from https://www.spss-tutorials.com/cramers-v-what-and-why/

Ngo, L. (2018). More on how to compare box plots - BioTuring's Blog. Retrieved from https://blog.bioturing.com/2018/05/22/more-on-how-to-compare-box-plots/

Statistics: Power from Data! Analytical graphing: Cumulative percentage. (2017). Retrieved from https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch10/5214864-eng.htm

Hedge's g Statistic. Retrieved 23 July 2019, from https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/hedgeg.htm

Examen HAVO 2016 tijdvak 2 wiskunde A (pilot). (2016). [Ebook] (p. 6). Retrieved from https://static.examenblad.nl/9336116/d/ex2016/ha-1024-f-16-2-o.pdf

Examen HAVO 2017 tijdvak 1 wiskunde A. (2017). [Ebook] (pp. 7-8). Retrieved from https://static.examenblad.nl/9336117/d/ex2017/HA-1024-a-17-1-o.pdf

Examen HAVO 2017 tijdvak 2 wiskunde A. (2017). [Ebook] (pp. 8-9). Retrieved from https://static.examenblad.nl/9336117/d/ex2017/HA-1024-a-17-2-o.pdf