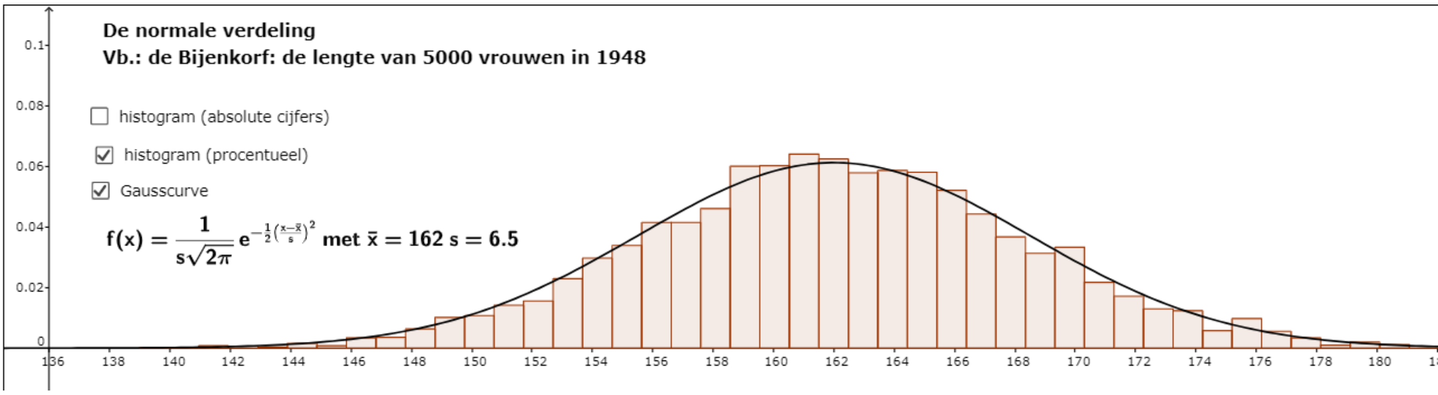
Normale verdeling binnen de statistiek

Karel Appeltans

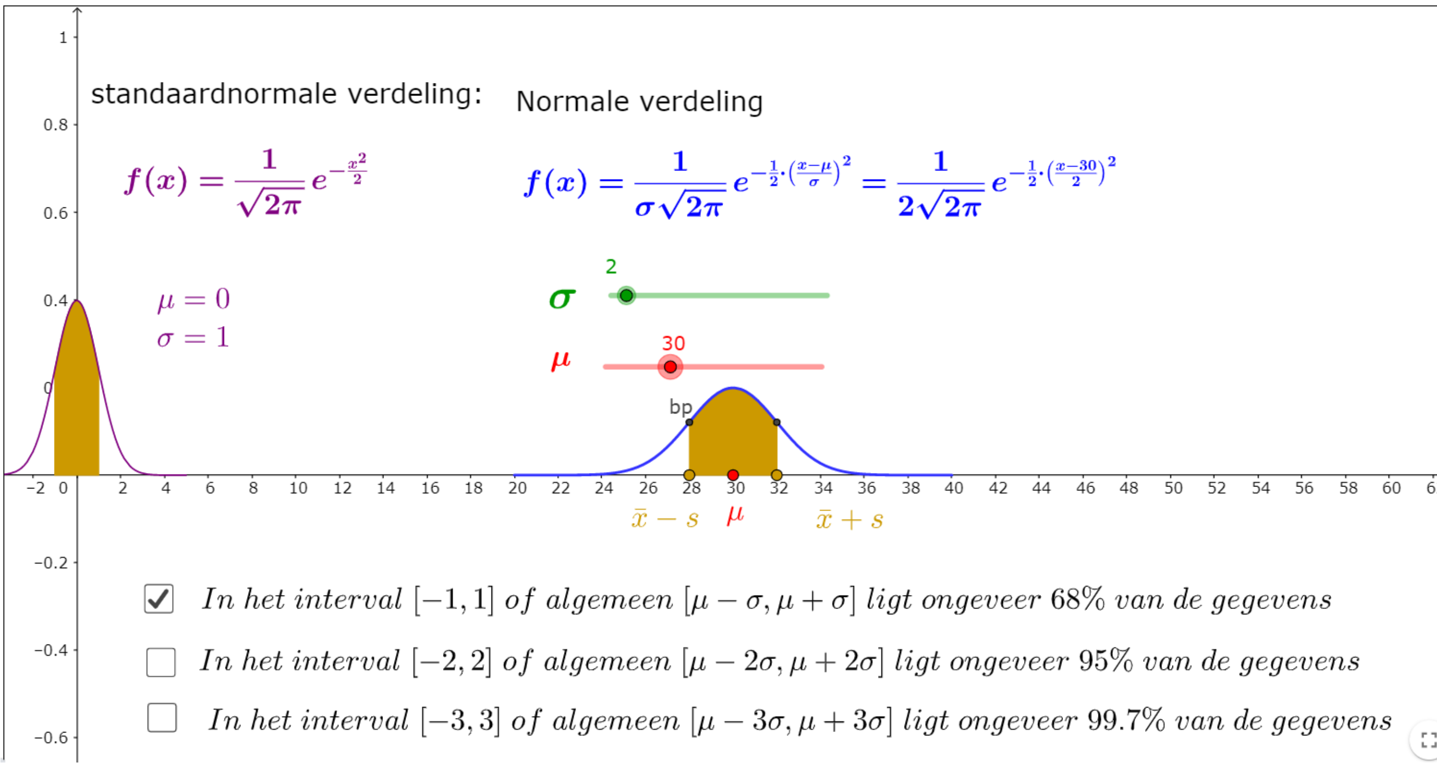
12 april 2021

1 Inleidend voorbeeld, de bijenkorf



Figuur 1: https://www.geogebra.org/m/muzjMBZq

2 (standaard) normale verdeling



Figuur 2: https://www.geogebra.org/m/muzjMBZq

3 z-score

3.1 begripsvorming

1.1 Een eerste voorbeeld: de punten van Pol

In een klas met 10 leerlingen zijn de resultaten op Nederlands, Frans en Duits als volgt:

Nederlands	7	5	8	7	6	3	5	2	4	3
Frans	9	7	5	7	6	9	10	4	5	8
Duits	7	5	7	8	7	7	8	7	7	7

Pol heeft op die drie vakken telkens 5 op 10 gehaald.

Er zijn nu verschillende mogelijkheden om de resultaten van Pol te interpreteren.

1.1.1 De ruwe score

Hierbij kijk je gewoon naar de behaalde punten, zonder enige verdere context. Je houdt hierbij geen rekening met wat de andere leerlingen gedaan hebben en ook niet met de manier waarop verschillende leerkrachten punten geven.
Pol haalde drie keer 5 op 10 en dus besluit je dat hij drie keer even goed presteerde.

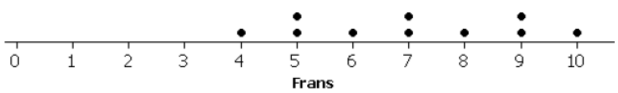
1.1.2 Vergelijken met het klasgemiddelde

Wat Pol presteerde is één ding, maar wat zijn medeleerlingen op diezelfde toetsen presteerden zegt toch ook iets. Je kan bijvoorbeeld kijken naar de globale prestatie van de hele klas en het klasgemiddelde als een referentiepunt nemen.
Op Nederlands haalde de klas een gemiddelde van 5, op Frans was dat 7 en op Duits ook 7.

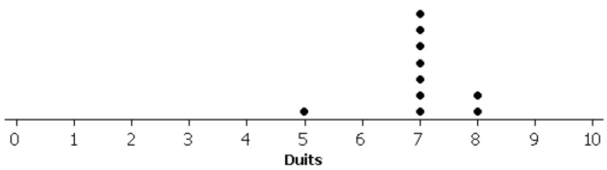
Zowel op Frans als op Duits scoorde Pol 2 punten lager dan het klasgemiddelde en dus heeft hij (in vergelijking met het gemiddelde) voor deze twee vakken op dezelfde manier gepresteerd. Zijn prestatie op Nederlands was beter want daar scoorde hij even hoog als het klasgemiddelde.

1.1.3 Variabiliteit rond het gemiddelde

Als je alleen het klasgemiddelde als referentie neemt, dan zie je bij Pol geen verschil tussen Frans en Duits (de score van Pol is telkens 5 en het klasgemiddelde is telkens 7). Zijn resultaten zal je nochtans heel anders interpreteren als je niet alleen naar het klasgemiddelde kijkt, maar ook naar de spreiding van de scores rond dat gemiddelde. Dat zie je op een eenvoudig puntendiagram.



Bij de toets Frans liggen de scores nogal gespreid. Twee leerlingen haalden een 5, er was ook een leerling met een 4 maar er waren er ook met 9 en 10. Voor de punten van die 10 leerlingen is het gemiddelde 7 en de standaardafwijking is 2.



De toets Duits ziet er helemaal anders uit. Iedereen haalde daar een 7 of een 8, behalve…. Pol, die had een 5. Bij deze toets is het gemiddelde 7 en de standaardafwijking is (afgerond) 0.8.

Een getal uit een dataset zomaar vergelijken met het gemiddelde vertelt niet het hele verhaal. Soms geeft dit zelfs een verkeerd beeld. De variabiliteit rond dat gemiddelde speelt ook een rol. Bij Frans behaalde Pol een score die 1 standaardafwijking onder het gemiddelde ligt, want $5 = 7 - (1)\times(2)$.
Bij Duits scoorde Pol 2.5 standaardafwijkingen onder het gemiddelde want $5 = 7 - (2.5)\times(0.8)$.

De standaardafwijking van een dataset is dikwijls een goede meetlat om punten uit die dataset te vergelijken met hun gemiddelde. Zo houd je ook rekening met de variabiliteit van de gegevens.

1.1.4 z-scores

Naar analogie met de benaming en de notatie bij populaties die normaal verdeeld zijn, spreekt men ook hier over z-scores wanneer je bepaalt hoeveel standaardafwijkingen een oorspronkelijke ruwe score verwijderd is (in positieve of negatieve zin) van het gemiddelde.

Als je de scores op Frans noteert als $x_1, x_2, x_3, \dots, x_{10}$ met gemiddelde $\overline{x} = 7$ en standaardafwijking $s = 2$ dan heb je voor de score van Pol (genoteerd als x) dat:

score van Pol = gemiddelde – **1** standaardafwijking –> 5 = 7 + (– **1**)(2) of $x = \overline{x} + z \cdot s$

zodat de z-score gelijk is aan: $z = \dfrac{x - \overline{x}}{s}$.

1.2 *Een tweede voorbeeld: de punten van Emma*

In een klas met 20 leerlingen zijn de resultaten op fysica en geschiedenis als volgt:

Fysica	7	7	8	5	7	9	7	8	8	6	7	7	7	7	6	7	5	6	9	7
Geschiedenis	8	8	7	8	6	5	7	5	7	8	8	8	6	7	6	8	6	8	8	6

Emma heeft zowel op fysica als op geschiedenis een 8 gehaald. Hoe vergelijk je die 2 resultaten?

Nota. Gebruik een notatie met x voor fysica en y voor geschiedenis.

1.2.1 De ruwe score

Volgens dit criterium zijn de 2 prestaties identiek want:

punt op fysica: $x = 8$
punt op geschiedenis: $y = 8$

1.2.2 Vergelijken met het klasgemiddelde

Volgens dit criterium zijn de 2 prestaties identiek want:

klasgemiddelde op fysica: $\overline{x} = 7$ zodat $x - \overline{x} = 8 - 7 = 1$.
Emma scoort 1 punt boven het klasgemiddelde bij fysica.
klasgemiddelde op geschiedenis: $\overline{y} = 7$ zodat $y - \overline{y} = 8 - 7 = 1$.
Emma scoort 1 punt boven het klasgemiddelde bij geschiedenis.

1.2.3 Variabiliteit rond het gemiddelde en z-scores

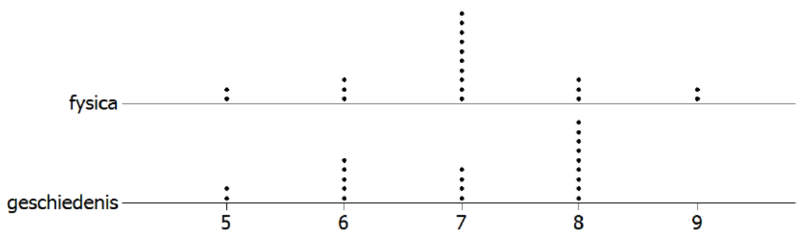
Volgens dit criterium zijn de 2 prestaties identiek want:

standaardafwijking bij fysica: $s_x = 1.08$ zodat $\dfrac{x - \overline{x}}{s_x} = \dfrac{8-7}{1.08} = 0.93$.
Op fysica haalt Emma een z-score van $z_x = 0.93$.

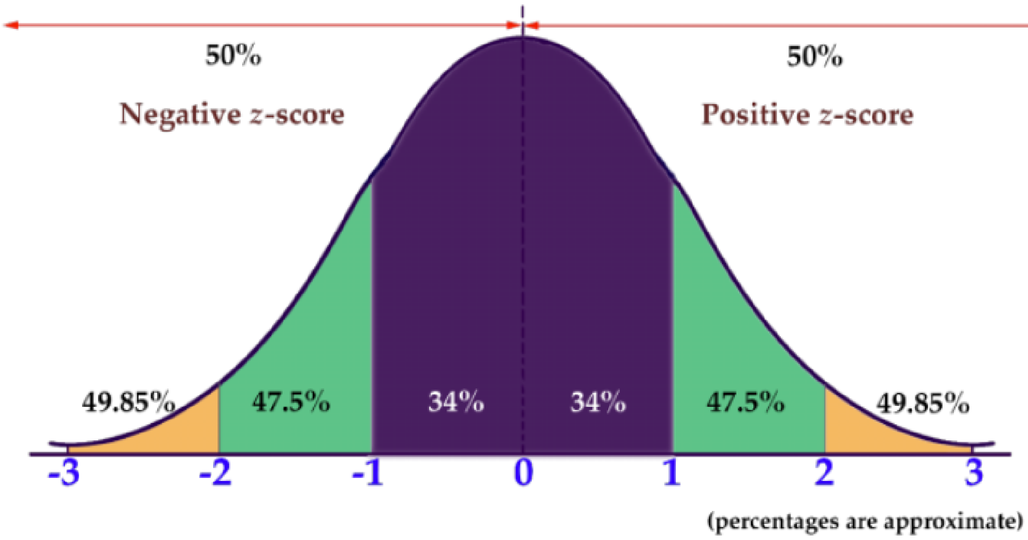
standaardafwijking bij geschiedenis: $s_y = 1.08$ zodat $\dfrac{y - \overline{y}}{s_y} = \dfrac{8-7}{1.08} = 0.93$.
Op geschiedenis haalt Emma een z-score van $z_y = 0.93$.

1.2.4 Hoort Emma bij dezelfde topgroep?

Zelfs als je rekening houdt met zowel het klasgemiddelde als met de variabiliteit rond dat gemiddelde, dan nog kan het gebeuren dat je het hele verhaal niet te pakken hebt. Ook de "vorm" van de puntenverdeling speelt een rol. Dat zie je goed op een figuur waar je de puntendiagrammen van beide vakken met elkaar vergelijkt.



3.2 z-score en normale verdeling

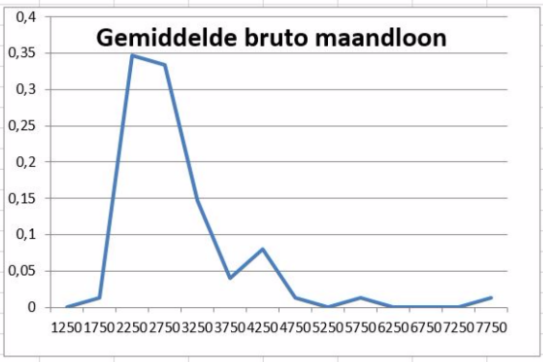


4 niet alles is normaal verdeeld

Niet alles is normaal verdeeld:

Gemiddelde bruto maandlonen van voltijdwerkers naar beroep in België (€) jaar 2008

grenzen		klassemidden	AF	RF
1000	1500	1250	0	0
1500	2000	1750	1	0,013333
2000	2500	2250	26	0,346667
2500	3000	2750	25	0,333333
3000	3500	3250	11	0,146667
3500	4000	3750	3	0,04
4000	4500	4250	6	0,08
4500	5000	4750	1	0,013333
5000	5500	5250	0	0
5500	6000	5750	1	0,013333
6000	6500	6250	0	0
6500	7000	6750	0	0
7000	7500	7250	0	0
7500	8000	7750	1	0,013333
			75	



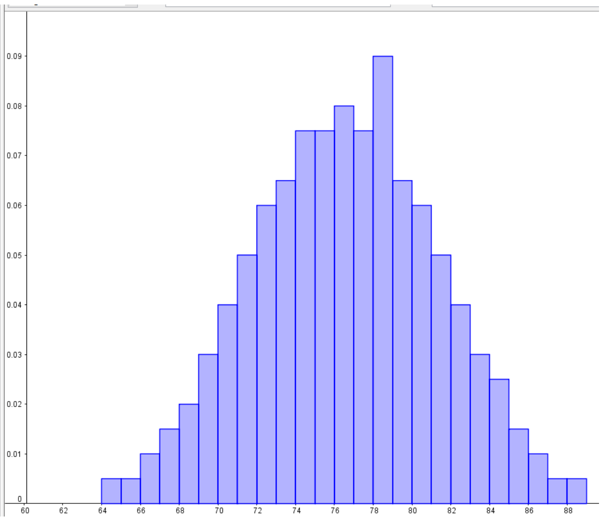
Figuur 3: https://www.geogebra.org/m/muzjMBZq

5 oefeningen

1. Bij visclub 'De Karper' werd de lengte van 200 opgeviste snoeks opgemeten. Onderstaande gegevens werden verzameld. Is deze lengte (bij benadering) normaal verdeeld?

64 - 65	0.005
65 - 66	0.005
66 - 67	0.01
67 - 68	0.015
68 - 69	0.02
69 - 70	0.03
70 - 71	0.04
71 - 72	0.05
72 - 73	0.06
73 - 74	0.065
74 - 75	0.075
75 - 76	0.075
76 - 77	0.08
77 - 78	0.075
78 - 79	0.09
79 - 80	0.065
80 - 81	0.06
81 - 82	0.05
82 - 83	0.04
83 - 84	0.03
84 - 85	0.025
85 - 86	0.015
86 - 87	0.01
87 - 88	0.005
88 - 89	0.005

n	200
Gemiddelde	76.54
σ	4.7339
s	4.7458
Σx	15308
Σx²	1176156.26
Min	64.5
Kwartiel1	73.15
Mediaan	76.75
Kwartiel3	79.85
Max	88.8

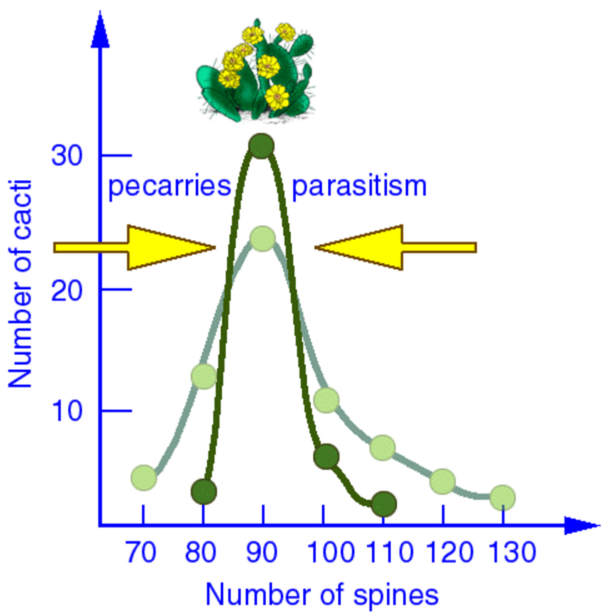


2. De Franse wiskundige Henri Poincaré werkte aan het eind van de 19de eeuw aan de Sorbonne in Parijs. Het verhaal gaat dat hij vermoedde dat de bakker in zijn straat de boel belazerde. De bakker verkocht broden van één kilo, maar Poincaré had het idee dat de broden meestal lichter waren. Een jaar lang kocht Poincaré elke dag een brood, legde dat thuis op een weegschaal en noteerde het gewicht. Aan het eind van het jaar tekende hij in een grafiek hoe vaak elk gewicht voorkwam. Het resultaat was een symmetrische grafiek met het hoogste punt bij 950 gram.. In wiskundige termen: wat Poincaré zag leek sterk op de grafiek van een ... met een ... van 950 gram en een ... van 50 gram.

   (a) Vul de juiste woorden in op de plaats van de puntjes
   (b) Schets zo nauwkeurig mogelijk een dergelijke grafiek. Vergeet de assen niet te benoemen.
   (c) Hoeveel procent van de broden weegt minder dan één kilo?

   Poincaré ging met zijn bewijsmateriaal naar de politie en de bakker kreeg een waarschuwing. Leg uit hoe Poincaré de politie kon overtuigen.

3. Cactussoorten zitten tussen twee vuren: bij te veel stekels nestelen er zich parasieten tussen de stekels waardoor de cactus ziek zal worden en bij te weinig stekels zullen dieren de cactus opeten als dorstlesser. Dit resulteert in onderstaande grafiek. Bespreek de verschillen tussen de twee cactussoorten.



4. Een vraag van het ingangsexamen geneeskunde

De score van een examen in eerste zittijd is normaal verdeeld met gemiddelde μ_1 en standaardafwijking σ_1. De score van het examen in tweede zittijd is ook normaal verdeeld met gemiddelde μ_2 en standaardafwijking σ_2. Stel dat $\mu_2 = \mu_1 + \sigma_1$ en $\sigma_2 = 2\sigma_1$ en beschouw de score $x = \mu_2 + \sigma_2$
Welke van de volgende vier uitspraken is waar?

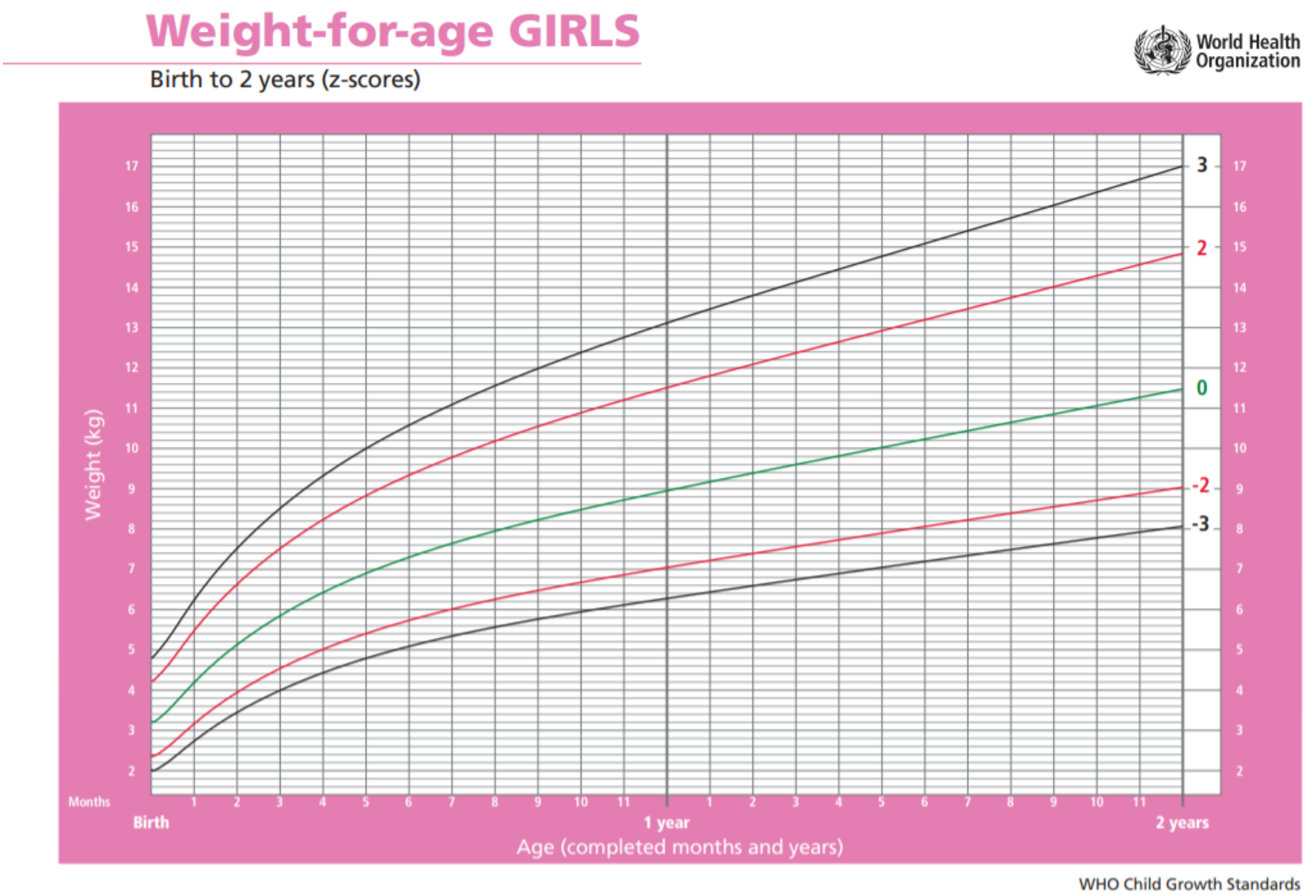
Vink alles aan wat van toepassing is

- ☐ De kans om in de tweede zittijd minstens de score x te behalen is ongeveer 10 keer kleiner dan de kans om in de eerste zittijd minstens de score x te behalen.
- ☐ De kans om in de tweede zittijd minstens de score x te behalen is ongeveer gelijk aan de kans om in de eerste zittijd minstens de score x te behalen.
- ☐ De kans om in de tweede zittijd minstens de score x te behalen is ongeveer 10 keer groter dan de kans om in de eerste zittijd minstens de score x te behalen.
- ☐ De kans om in de tweede zittijd minstens de score x te behalen is ongeveer 100 keer groter dan de kans om in de eerste zittijd minstens de score x te behalen.

5. Drie eerste bachelorstudenten (elke student zit in een andere richting) hebben zojuist hun punten op het examen statistiek teruggekregen (op 10). Probeer de situatie van de drie studenten te vergelijken en uit te rekenen wie het best gepresteerd heeft in zijn richting. Leg ook uit in woorden waarom de student het best presteerde.

Student	Resultaat	Gemiddelde	Standaardafwijking
Jana	8	9	0.7
Xavier	6	4.2	1.6
Mathijs	7	6	1

6. Bekijk volgende grafiek van de WHO



(a) Wat is het gemiddelde gewicht van een 1-jarig meisje?

(b) Wat is de standaardafwijking bij een 1-jarig meisje?

(c) Hoeveel moet een meisje wegen om bij $2,5\%$ zwaarste meisjes te zijn en dit op één-jarige leeftijd?

7. Waarom wordt er gebruik gemaakt van de z-score? (gelezen www.demorgen.be)

ZWARTE WEKEN

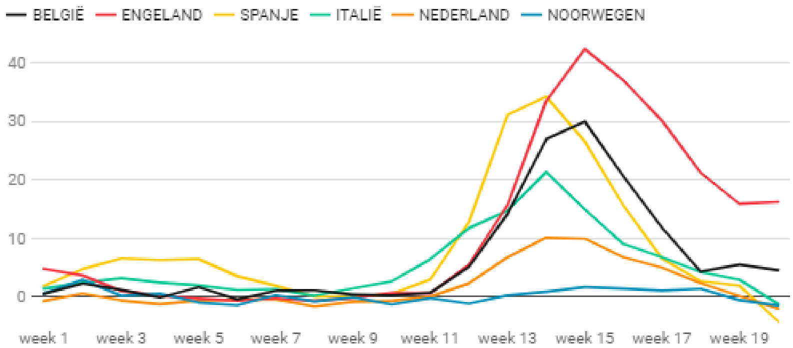
De zwaarste periode in Europa waren de dagen tussen 23 maart en 12 april; de weken 13, 14 en 15. Voor België ligt de piek precies in die drie weken. Dat leidt volgens Euromomo tot een zogeheten z-score tussen 25 en 30, wat behoorlijk hoog is. De z-score geeft per week aan hoe groot de afwijking is van de normale sterfte. Hoe hoger de z-score, des te hoger de sterfte aan covid.

Twee pieken steken erbovenuit: die van Engeland in week 15 (42,75) en Spanje in week 14 (34,41). Dan komt België al (29,91 in week 15), voor Italië (21,7 in week 14), Frankrijk (21,7 in week 14) en Nederland (21,13 in week 14).

De middengroep wordt gevormd door Zweden, Zwitserland, Noord-Ierland en Schotland. De overige landen kenden nauwelijks oversterfte.

Z-score sterfgevallen per land

De z-score geeft de afwijking aan van het aantal sterfgevallen t.o.v. het te verwachten aantal. Een score tussen -2 en 2 is normaal. Boven 15 is er sprake van een extreem hoog aantal sterfgevallen.



Grafiek: grafiek DM/TROUW • Bron: euromomo • Gegevens ophalen • Gecreëerd met Datawrapper