STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Populatiemodellen en normaal verdeelde populaties

3. Normaal verdeelde kansmodellen

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Normaal verdeelde kansmodellen

1. Een voorbeeld

Hier volgt een tekst uit Het Nieuwsblad van 17– 09 – 2004

Vlaming is 5 centimeter groter dan twintig jaar geleden - 17/09/2004

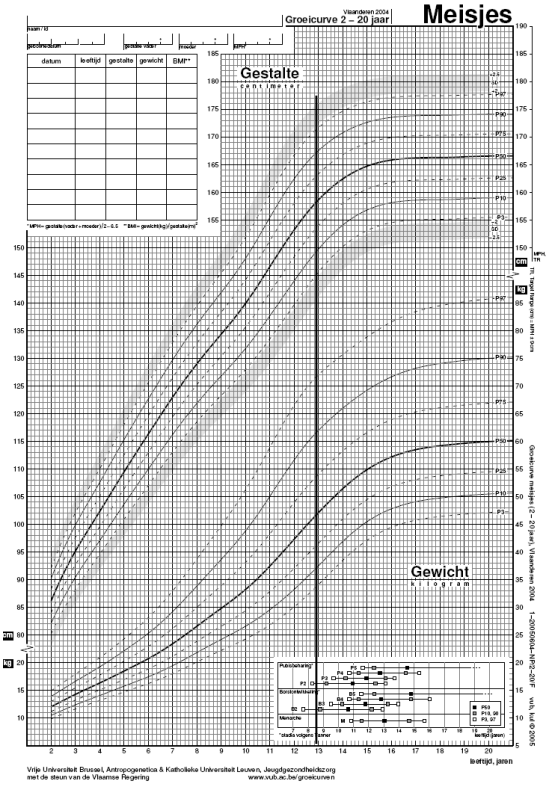
Vlaamse wetenschappers hebben nieuwe groeicurven ontwikkeld. Daarmee kunnen dokters nagaan of een kind een normale lengte en een correct gewicht heeft voor zijn leeftijd. De oude curven, die tot gisteren gebruikt werden, onderschatten de lengte van een volgroeide jongen en meisje met 4 tot 5centimeter.

Kinderen en volwassenen worden van de ene generatie op de andere alsmaar groter. Ze groeien ook sneller, waardoor ze vroeger in de puberteit komen en op steeds jongere leeftijd hun finale gestalte krijgen. Dat is een fenomeen dat zich in alle Europese landen voordoet en het is het gevolg van onze verbeterde levensomstandigheden.

Sedert 1960 worden de jongens in Vlaanderen, om de tien jaar, gemiddeld 1,5 centimeter groter dan de vorige generatie. De meisjes groeien om de tien jaar met gemiddeld 1,1 centimeter. Vooral de benen worden langer en dat gebeurt in de eerste vier levensjaren.

Ook het gewicht neemt generatie na generatie toe: 1,2 kg per decennium bij de jongens en 0,9 kg per decennium bij de meisjes. Met die evoluties hielden de groeicurven van 1960, die tot dusver gebruikt werden, geen rekening. Ze waren dringend aan vervanging toe. Wetenschappers van de universiteiten van Brussel en Leuven hebben een representatieve steekproef van 8.000 Vlaamse jongens en evenveel Vlaamse meisjes tussen 2 en 20 jaar nauwkeurig gemeten en gewogen. Op basis van die gegevens, ontwikkelden ze nieuwe groeicurven.

In 2004 waren in Vlaanderen meisjes van 13 gemiddeld 159 cm groot. Dat vind je uit de groeicurven op http://www.vub.ac.be/groeicurven (Antropogenetica VUB & Jeugdgezondheidszorg KUL).



De **populatie** die je nu wil bestuderen is de lengte van alle Vlaamse meisjes die 13 jaar zijn.

Er zijn nogal wat situaties te bedenken waar het interessant is om iets meer over die lengte te weten. Misschien is dat nuttig voor een fabrikant van jeans. Of voor een huisarts die denkt dat hij met een uitzonderlijke lengte wordt geconfronteerd maar daar toch niet zo zeker van is.

Typische vragen zijn hier:

- zijn “de meeste” meisjes ongeveer 159 cm of zitten daar enorme verschillen tussen?
- hoeveel percent van die meisjes is niet groter dan 146 cm?
- wat zijn extreem grote meisjes? Wanneer ben je bij de grootste 5 % van deze meisjes?

Op al deze vragen kan je niet antwoorden als je alleen maar het gemiddelde van die populatie kent. Je hebt daarvoor een kansmodel nodig.

Hoe maak je hier een kansmodel?

Begin met te bepalen welk soort kansmodel je nodig hebt. Je moet daarbij kiezen tussen een discreet model met een kansverdeling of een continu model met een dichtheidsfunctie.

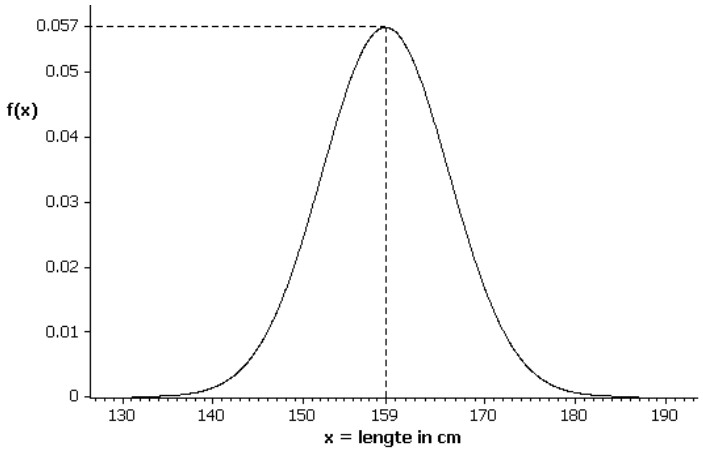
Het gaat hier over “lengte” en dat behandel je continu, zelfs al heb je de opmetingen afgerond tot op de centimeter (je moet altijd ergens afronden). Je moet dus op zoek gaan naar een dichtheidsfunctie.

Een verstandige en soms zeer snelle manier van werken is: “maak gebruik van wat al gekend is”. Jarenlange studies hebben aangetoond dat die fameuze “klokvormige” curve een goed kansmodel is om de populatie van “lichaamslengten” te beschrijven (per geslacht en per leeftijdsgroep). Je mag die kennis nu gebruiken.

De klokvormige curve die hier bedoeld wordt heet “Gausscurve” of “normale dichtheidsfunctie”. Als je die kan gebruiken om een populatie te beschrijven dan spreek je over **een normaal verdeelde populatie**. Het kansmodel voor de lengte van de meisjes van 13 zie je in figuur 1.

Opdracht 1

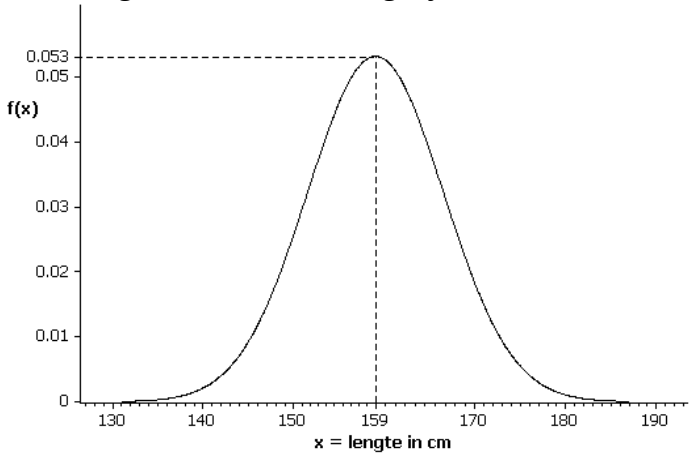
Een normale dichtheid is symmetrisch rond één top, en hier zie je een top bij een x-waarde van 159 cm. Had je dat verwacht? Wat zou je hier kunnen uit afleiden?



Figuur 1

Opdracht 2

Geloof het of niet, maar ook 13-jarige jongens zijn in Vlaanderen gemiddeld 159 cm groot. Het kansmodel voor hun lengte staat in figuur 2. Zoals je kon verwachten vind je ook daar de top bij een x-waarde van 159 cm. Maar die top is lager dan in figuur 1 (0.053 in plaats van 0.057). Hoe kan dat? Beide curven stellen een dichtheidsfunctie voor. Wat weet je van een dichtheidsfunctie? Wat moet daar dan uit volgen als je die 2 figuren met elkaar vergelijkt?

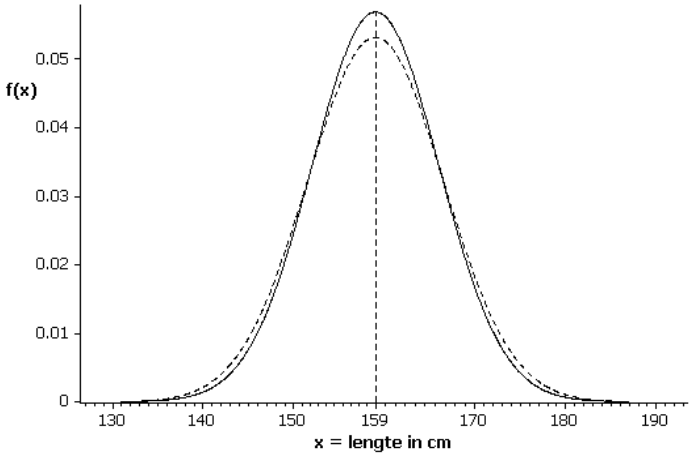


Figuur 2

Tot nu toe heb je, naast de globale vorm, één kenmerk van die populaties bekeken, namelijk het gemiddelde. Dat was in beide gevallen 159 cm. Maar er is natuurlijk ook spreiding rond dat gemiddelde. De standaardafwijking is een maat voor die spreiding. Bij de meisjes is die gelijk aan 7 cm en bij de jongens 7.5 cm. Dat is de reden waarom de normale curve van de jongens meer uitgespreid is. Je ziet dat goed als je beide curven over elkaar tekent zoals in figuur 3.

Opdracht 3

Voor welke populatie is de curve in stippellijn een kansmodel? Waarom?



Figuur 3

Er is blijkbaar meer dan één normale dichtheidsfunctie. Op de figuren zie je dat de curven hoger en spitser kunnen zijn of lager en breder. En natuurlijk hoeft de top niet altijd bij een x-waarde van 159 cm te liggen. Als de gemiddelde lengte van de populatie groter is dan 159 cm dan schuift de curve naar rechts, en anders naar links.

2. De normale familie

Op voorbeelden heb je ontdekt dat de curve van de dichtheidsfunctie waarmee je een normaal verdeelde populatie beschrijft de volgende eigenschappen heeft:

- ze is klokvormig
- haar top heeft ze bij een x-waarde die gelijk is aan het gemiddelde μ van de populatie
- ze is breder of smaller naarmate de standaardafwijking σ van de populatie groter of kleiner is

> Een populatie X die zich gedraagt als een normaal kansmodel met een gemiddelde μ en een standaardafwijking σ noteer je als
> $$X \sim N(\mu;\sigma).$$

Bij elk ander gemiddelde en bij elke andere standaardafwijking heb je een andere Gausscurve. Je hebt dus een hele familie normale dichtheidsfuncties (het zijn er oneindig veel).

Om een hele familie functies toch in 1 functievoorschrift te kunnen opschrijven werk je met parameters.
Dat ziet er voor de normale familie als volgt uit.

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{voor} \quad -\infty<x<+\infty \quad \text{en met} \begin{cases} -\infty<\mu<+\infty \\ 0<\sigma \end{cases} \qquad (*)$$

Je hoeft deze formule niet van buiten te kennen, maar je kan wel eens kijken wat ze wordt voor de lengte van de 13-jarige meisjes.

Als je de populatie van die lengten voorstelt door X dan schrijf je μ voor $E(X)$ en σ voor $sd(X)$ om de notatie voor populaties te gebruiken. Van die meisjes weet je dat $\mu=159$ en $\sigma=7$ zodat je voor (*) vindt:

$$f(x)=\frac{1}{7\sqrt{2\pi}}\,e^{-\frac{(x-159)^2}{2(7)^2}} \Rightarrow f(x)=0.057\,e^{-\frac{(x-159)^2}{98}} \qquad \text{voor} \quad -\infty<x<\infty$$

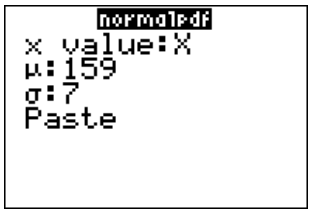
Hier heb je nu één welbepaalde functie. Het is een dichtheidsfunctie. Haar grafiek zie je in figuur 1.

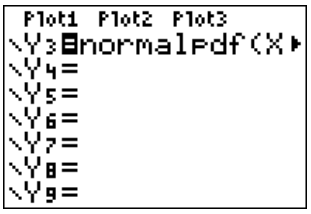
De populatie X van de lengte van die meisjes gedraagt zich als een normaal kansmodel met gemiddelde $\mu=159$ cm en standaardafwijking $\sigma=7$ cm. Je schrijft dat als $X \sim N(159;7)$.

Je GRM gebruikt dezelfde conventie. Een normale dichtheidsfunctie teken je als volgt.

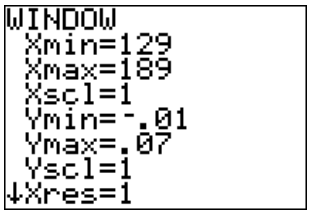
Zorg ervoor dat alle plots af staan. Druk [2nd][STAT PLOT] en dan 4:PlotsOff en dan [ENTER].

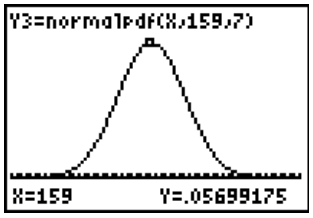
Druk dan [Y=] en zorg ervoor dat alle functies af staan. Loop naar een functie die nog vrij is, bijvoorbeeld \Y$_3$= , druk [2nd] [DISTR] en druk 1:normalpdf(. Dit staat voor *Normal Probability Density Function.* Vul het venster in zoals aangegeven (voor X druk je [X,T,Θ,n]), loop naar Paste en druk [ENTER].





Druk nu [WINDOW] en pas de vensterinstellingen aan. Druk dan onmiddellijk [TRACE]. Om de functiewaarde voor x=159 te zien tik je 159 en [ENTER].





Opdracht 4

Je hebt zopas de normale dichtheid voor de lengte van 13-jarige meisjes op je GRM getekend. Doe nu hetzelfde voor de normale dichtheid van 13-jarige jongens. Druk [2nd][QUIT] en dan [Y=] en laat de functie die je hebt ingevuld (bijvoorbeeld \Y$_3$=) aan staan. Loop naar een volgende functie die nog vrij is (bijvoorbeeld \Y$_4$=) en vul die in op de gepaste manier. Werk met dezelfde vensterinstelling. Druk [TRACE] en wacht tot beide functies getekend zijn.

Je hebt nu dezelfde figuur op je GRM als figuur 3 hierboven. Gebruik de pijltjes [▲] en [▼] om van de ene grafiek op de andere over te schakelen. De normale dichtheid met de grootste standaardafwijking (de jongens) heeft een lagere top en is breder dan de die van de meisjes.

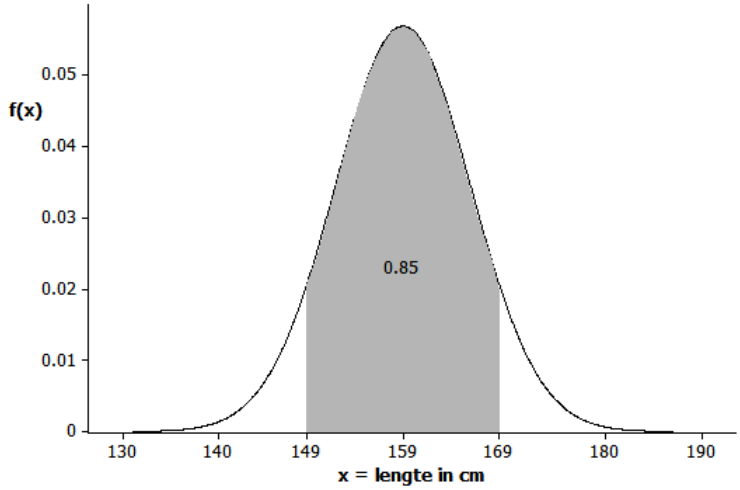
3. Rekenen met normale kansmodellen

3.1. Kansen

Een normaal kansmodel is gewoon een continu kansmodel en hoe je kansen moet berekenen bij continue modellen weet je al. Teken de dichtheidsfunctie en bereken de oppervlakte onder de curve voor een bepaald deelinterval. Die oppervlakte is de kans om in dat deelinterval terecht te komen. Die kans verandert niet als je de grenzen van dat deelinterval er al dan niet bij neemt.

Meisjes van 13 zijn gemiddeld 159 cm groot. Hoeveel percent van die meisjes wijkt daar niet meer dan 10 cm van af?

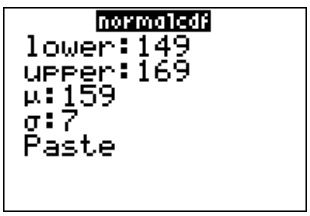
Vertaal die vraag eerst even in gekende notatie. Stel X = de lengte van meisjes van 13. Dan weet je uit vorig voorbeeld dat $X \sim N(159;7)$. Voor dit kansmodel wordt nu gevraagd naar $P(149 \le X \le 169)$. Dit is de kans dat een willekeurig gekozen meisje uit die populatie hoogstens 169 cm groot is maar ook niet kleiner is dan 149 cm. Of je kan ook zeggen dat dit het percent 13-jarige meisjes is dat niet groter dan 169 cm maar ook niet kleiner dan 149 cm is. Hoeveel percent is dat? Het antwoord zie je op de onderstaande figuur.

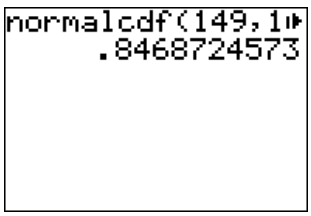


Figuur 4

Je GRM kan die kans ook berekenen. Hij gebruikt daarvoor het commando normalcdf(= *normal cumulative distribution function*.

Druk [2nd] [DISTR] en 2: normalcdf(. Vul het venster in zoals aangegeven, loop naar Paste en druk 2 keer [ENTER]. Het antwoord is 0.85 zodat $P(149 \le X \le 169) = 0.85$.



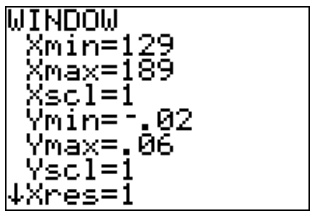


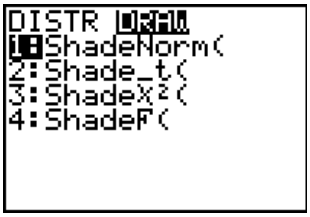
Per 100 meisjes zijn er dus toch nog 15 die meer dan 10 cm afwijken van de gemiddelde lengte van hun leeftijdsgenoten. Dat zijn er niet weinig.

Intuïtief kan je het resultaat van normalcdf(als volgt begrijpen. Je GRM tekent de gevraagde normale dichtheidsfunctie (bijvoorbeeld met gemiddelde $\mu = 159$ en standaardafwijking $\sigma = 7$) en berekent dan de oppervlakte onder die curve voor een interval waarvan jij de ondergrens (bijvoorbeeld 149) en de bovengrens (bijvoorbeeld 169) opgeeft.

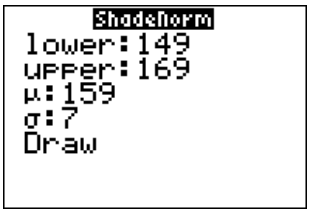
Je kan dit ook grafisch voorstellen. Controleer eerst met [Y=] of alle functies af staan. Indien nodig druk je ook vooraf [2nd] [DRAW] en 1:ClrDraw en [ENTER] .

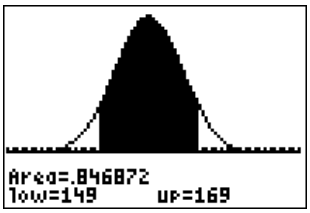
Pas de vensterinstellingen ([WINDOW]) aan zoals aangegeven. Druk dan [2nd] [DISTR] en selecteer de optie DRAW.





Druk 1:ShadeNorm(, vul het venster in, loop naar Draw en druk [ENTER].



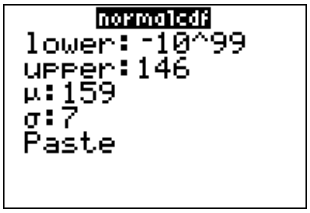


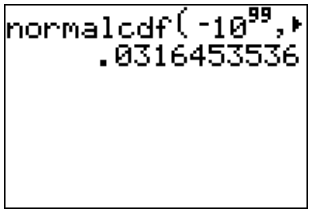
De oppervlakte boven het interval [149 ; 169] is gelijk aan 0.85. Dat betekent dat X met kans 0.85 in [149 ; 169] terechtkomt.

Hoeveel percent van die meisjes is niet groter dan 146 cm?
In verkorte notatie is de vraag als volgt: wat is $P(X \leq 146)$ wanneer $X \sim N(159;7)$?

Je moet nu de oppervlakte onder de normale curve berekenen voor het interval $]-\infty;146]$ want in theorie is elk normaal model gedefinieerd door een dichtheidsfunctie waarbij de x-waarden lopen van $-\infty$ tot $+\infty$. Je GRM kan niet met oneindig werken en de oplossing bestaat erin dat je $-\infty$ vervangt door [(-)] 10 [^] 99 en $+\infty$ door 10 [^] 99 .

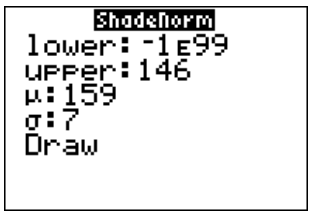
Druk [2nd] [DISTR] en 2: normalcdf(. Vul het venster in zoals aangegeven, loop naar Paste en druk 2 keer [ENTER]. Het antwoord is 0.03 zodat $P(X \leq 146) = 0.03$.

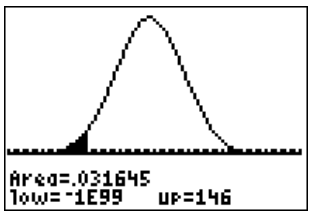




In die populatie zijn er slechts 3 % meisjes die hoogstens 146 cm groot zijn. Als je daar bij hoort dan ben je voor je leeftijd toch wel klein.

Voor een figuur druk je eerst [2nd] [DRAW] en 1:ClrDraw en [ENTER]. Dan [2nd] [DISTR] , selecteer DRAW, druk 1:ShadeNorm(, vervolledig het venster, loop naar Draw en druk [ENTER].





Opdracht 5

Denk je dat er in die populatie meisjes zijn van 180 cm of nog groter? Bereken de kans met je GRM en schrijf de uitkomst als een kansuitspraak in de juiste notatie.

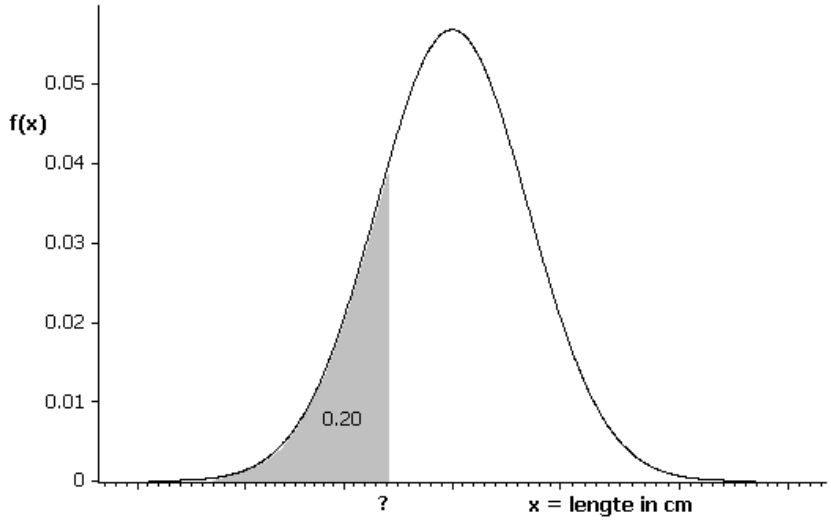
3.2. Kritische punten

Je weet nu hoe je met de GRM de kans kan berekenen om in een of ander interval terecht te komen wanneer je trekt uit een normaal verdeelde populatie. Je kan dus al antwoorden op vragen die eruit zien als: wat is $P(149 \le X \le 169)$ of hoeveel is $P(X \le 146)$ wanneer $X \sim N(159;7)$. Dit kan je allemaal vinden met het commando normalcdf(of met ShadeNorm(.

Soms moet je het antwoord zoeken op vragen die "omgekeerd" gesteld zijn.

In welk interval komt je lengte terecht als je hoort tot de kleinste 20 % van de meisjes van 13 ?
Deze vraag kan je schrijven als: wat is x als je weet dat $P(X \le x) = 0.20$ voor $X \sim N(159;7)$?
Je hebt nu een kans die gegeven is en jij moet op zoek naar een interval waarin je met die kans terechtkomt. Hier is dat interval van de vorm $]-\infty; x]$. Wat is x ?

Kijk eerst eens naar onderstaande figuur, zodat je zeker goed begrijpt waarover het hier gaat. Wat is er gegeven? Wat wordt er gevraagd?



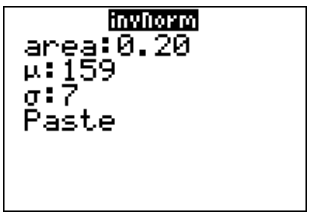
Figuur 5

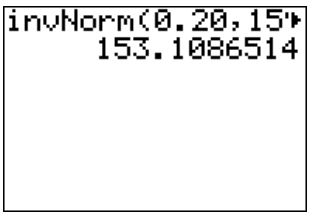
Met je GRM kan je de vraag oplossen met het commando invNorm(. Jij zegt wat de oppervlakte boven een linkerstaart is en invNorm(zegt je dan wat die staart is. Je krijgt dan de waarde x zodat je weet wat $]-\infty; x]$ is.

> *Let op.* Met het commando invNorm(kan je alleen **LINKERSTAARTEN** te weten komen als jij zegt wat de kans is om in zo'n linkerstaart te vallen.

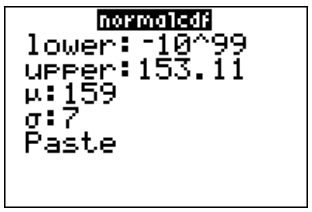
Als je weet dat de oppervlakte boven de linkerstaart 20 % moet zijn of, wat hetzelfde is, dat $P(X \le x) = 0.20$ voor $X \sim N(159;7)$, dan doe je het volgende om x te zoeken.

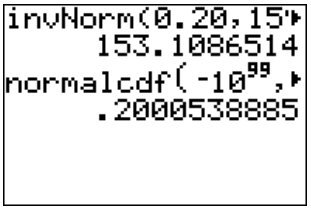
Druk [2nd] [DISTR] en 3:invNorm(. Op het scherm staat area:. Daar vul je de oppervlakte in boven de linkerstaart, wat ook de kans is om in die linkerstaart terecht te komen. Verder werk je nog altijd met een normale met gemiddelde $\mu = 159$ en standaardafwijking $\sigma = 7$. Loop dan naar Paste en druk 2 keer [ENTER]. Het antwoord is 153.11 zodat $P(X \leq 153.11) = 0.20$. Je valt met 20 % kans in de linkerstaart $]-\infty ; 153.11]$.





Als je even wil controleren of 153.11 wel het goede antwoord is dan kan je berekenen wat de oppervlakte boven de linkerstaart $]-\infty ; 153.11]$ juist is. Dat doe je met normalcdf(,dat ken je al. $P(-\infty < X \leq 153.11) = P(X \leq 153.11) = 0.20$.



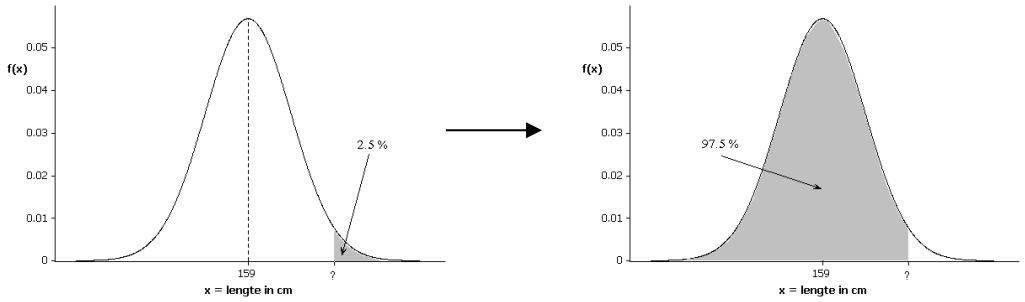


In je rapport vermeld je dat er 20 % meisjes van 13 zijn die niet meer dan 153 cm groot zijn. Je werkt hier immers met een studie waarbij de "continue" lengten achteraf afgerond werden tot op de centimeter. Nadat al je berekeningen zijn uitgevoerd mag jij ook achteraf afronden bij je eindconclusie.

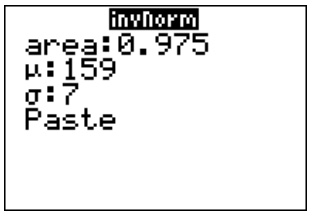
Als je wil behoren tot de 2.5 % grootste meisjes, hoe groot moet je dan zijn?

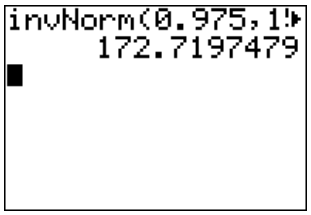
Schrijf eerst de vraag in de juiste notatie. Hier zoek je x waarvoor $P(X \geq x) = 0.025$. Dat is ook te schrijven als $P(x \leq X) = P(x \leq X < +\infty) = 0.025$ zodat je op zoek moet gaan naar een **rechterstaart** $[x ; +\infty[$ waarin je met kans 2.5 % terechtkomt.

Je hebt hier een probleem, want invNorm(vertelt je alleen wat de linkerstaart is. Gelukkig ken je een basiseigenschap van een dichtheidsfunctie nog. De totale oppervlakte onder de curve is gelijk aan 1 . Bovendien werk je met een continu kansmodel waarbij $P(X \geq x) = P(X > x)$ zodat $P(X \geq x) = P(X > x) = 1 - P(X \leq x)$. Als je dus op zoek moet gaan naar de x–waarde waarvoor $P(X \geq x) = 0.025$ dan moet voor die x ook gelden dat $P(X \leq x) = 0.975$. Nu kan je x vinden want je hebt de oorspronkelijke vraag kunnen veranderen in: "als de oppervlakte boven de **rechterstaart** 0.025 moet zijn dan moet de oppervlakte boven de **linkerstaart** 0.975 zijn". Als je weet waar de linkerstaart eindigt dan weet je ook waar de rechterstaart begint.



Ga nu te werk zoals zopas.





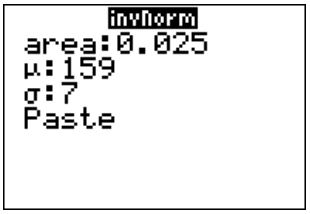
Druk [2nd] [DISTR] en 3:invNorm(. Vul in en loop naar Paste en druk 2 keer [ENTER]. Het antwoord is 172.72 zodat $P(X \le 172.72) = 0.975$. Hieruit volgt dat $P(X \ge 172.72) = 0.025$.

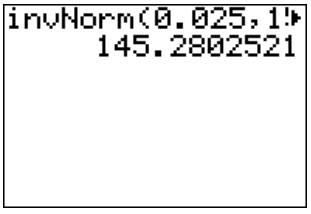
Als je (afgerond) minstens 173 cm bent dan ben je wel uitzonderlijk groot voor je leeftijd. Zo zijn er maar 2 à 3 meisjes per 100.

De waarde van x die je vindt bij een gegeven kans van de vorm $P(X \le x)$ of van de vorm $P(X \ge x)$ noem je een **kritisch punt**. Grafisch is dat dus het eindpunt van een linkerstaart of het beginpunt van een rechterstaart.

Soms werk je met 2 staarten tegelijkertijd. Je wil bijvoorbeeld de 5 % "extreme" meisjes zoeken, waarbij je symmetrisch te werk gaat, namelijk 2.5 % "extreem klein" samen met 2.5 % "extreem groot".

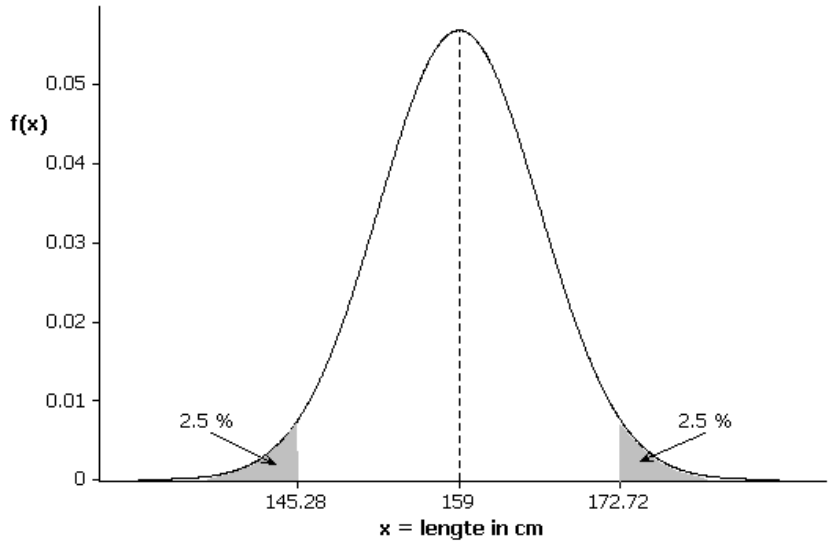
De 2.5 % extreem grote meisjes heb je al gevonden. Die lengten begonnen vanaf 172.72 cm. Om nu de 2.5 % extreem kleine te vinden kan je 2 dingen doen.





Je kan nog eens invNorm(gebruiken met je GRM. Als resultaat vind je dan dat $P(X \le 145.28) = 0.025$ zodat meisjes die niet meer dan (afgerond) 145 cm groot zijn tot de kleinste 2.5 % behoren.

Je kan ook gebruik maken van de symmetrie van elke normale curve rond haar gemiddelde. Maak daarbij een figuur. Dat is handig om je niet te vergissen.



Figuur 6

Als je de grootste 2.5 % lengten kent en je weet dat die vanaf 172.72 cm beginnen dan kan je ook zeggen dat die beginnen vanaf "13.72 cm boven het gemiddelde". Ga dan vanuit het gemiddelde ook 13.72 cm naar beneden. Dan vind je 159 – 13.72 = 145.28 cm en dat is inderdaad het symmetrische kritische punt voor de linkerstaart waar de 2.5 % kleinste lengten liggen.

3.3. De invloed van μ en σ

Je hebt reeds gezien dat het gemiddelde μ en de standaardafwijking σ van een normale populatie de plaats en de spitsheid van de normale curve beïnvloeden.

Opdracht 6

Geef een kort antwoord op de volgende vragen.

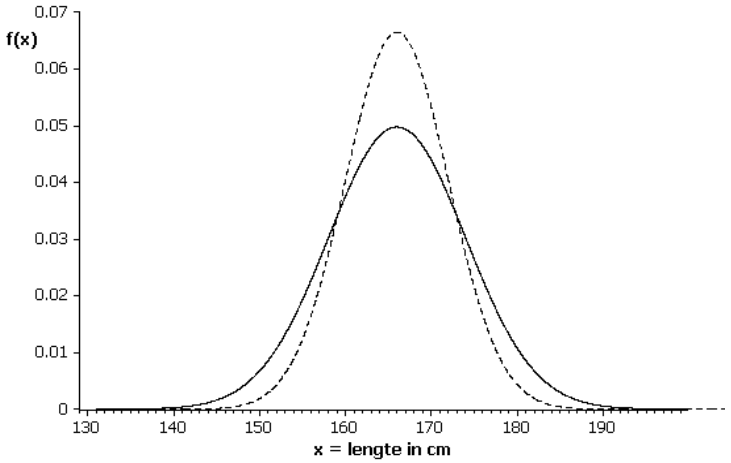
- Wanneer lijkt de curve qua vorm "bevroren" en zie je alleen maar een verschuiving?

- Wanneer blijft de curve "ter plaatse" maar wordt zij spitser of breder?

- Wat gebeurt er als μ en σ beide veranderen?

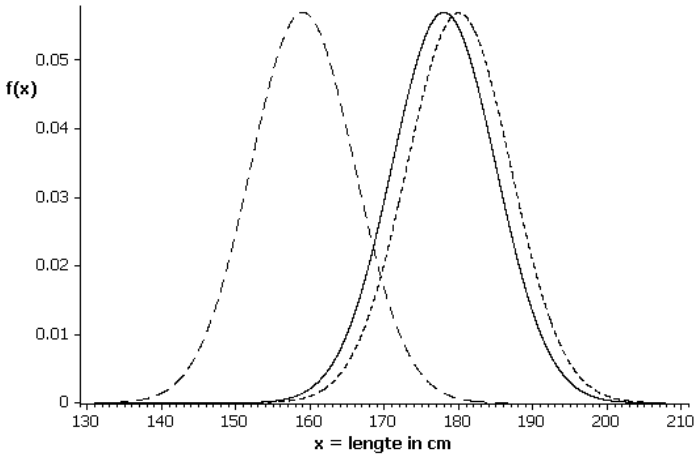
Opdracht 7

Hieronder zie je enkele voorbeelden waarbij jij moet aangeven welke curve bij welke populatie past. Motiveer steeds je antwoord.



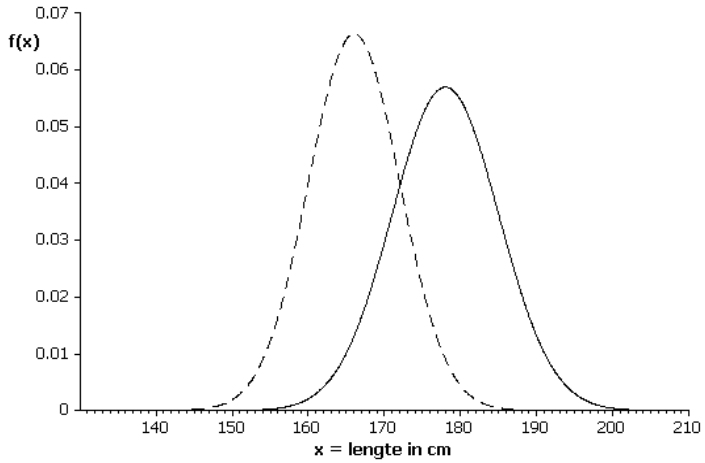
Figuur 7

Figuur 7. Lengte van Vlaamse jongeren (in cm)		
	μ	σ
jongens van 14 jaar	166	8
meisjes van 17 jaar	166	6



Figuur 8

Figuur 8. Lengte van Vlaamse jongeren (in cm)		
	μ	σ
meisjes van 13 jaar	159	7
jongens van 17 jaar	178	7
jongens van 19 jaar	180	7



Figuur 9

Figuur 9. Lengte van Vlaamse jongeren (in cm)		
	μ	σ
meisjes van 17 jaar	166	6
jongens van 17 jaar	178	7

4. Een nieuwe meetlat: z–scores

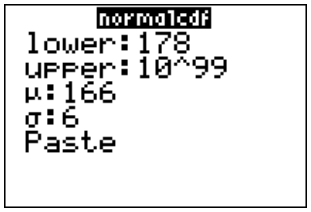
In figuur 7 heb je gezien dat meisjes van 17 en jongens van 14 gemiddeld 166 cm groot zijn.

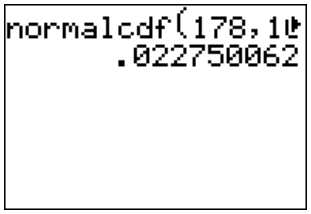
Katia heeft het een tijdje moeilijk gehad om zich goed in haar vel te voelen. Zij stak er altijd met kop en schouders bovenuit. Nu is zij het gewoon, maar binnen haar leeftijdsgroep is zij helemaal niet gewoon. Zij is 1 meter 78 cm en dat is veel meer dan de gemiddelde lengte van meisjes van 17. Dat is zomaar eventjes 12 cm meer.

Is Katia echt uitzonderlijk groot? Hoeveel percent van de meisjes van 17 zit in die topgroep van meisjes die minstens 12 cm groter zijn dan het gemiddelde?

Dat kan je eenvoudig vinden met je GRM. "Minstens 12 cm groter dan het gemiddelde" betekent dat je de oppervlakte boven een rechterstaart moet zoeken vanaf 178 tot $+\infty$. De curve die je hier gebruikt is de normale dichtheid met gemiddelde $\mu = 166\,\text{cm}$ en standaardafwijking $\sigma = 6\,\text{cm}$ want voor meisjes van 17 is de lengte X verdeeld als $X \sim N(166;6)$.

Druk [2nd] [DISTR] en 2: normalcdf(. Vul in, loop naar Paste en druk 2 keer [ENTER]. Het antwoord is 0.02 zodat $P(X \geq 178) = 0.02$.





In die populatie zijn er slechts 2 % meisjes die minstens 178 cm groot zijn. Katia hoort bij de top 2 % van haar leeftijdsgroep. Dat is echt uitzonderlijk.

Roel is 14 jaar. Hij is 1 meter 78 cm en hij is heel wat groter dan veel van zijn leeftijdsgenoten. De gemiddelde lengte van jongens van 14 is immers 166 cm. Roel is de broer van Katia en hij heeft van haar gehoord dat zij bij de top 2 % van haar leeftijdsgroep hoort.

De conclusie van Roel is snel en eenvoudig. Hij is ook 1 meter 78 cm en in zijn leeftijdsgroep is de gemiddelde lengte ook 166 cm. Hij steekt daar 12 cm bovenuit, juist zoals dat bij Katia het geval is. Dus hoort hij ook bij de top 2 % van zijn leeftijdsgroep, juist zoals bij Katia.

Is dat waar?

Je kan dat eenvoudig controleren. Waar ligt de top 2 % voor jongens van 14? Hun lengte is verdeeld als $X \sim N(166;8)$. Je moet dus op zoek naar een gepaste rechterstaart $[\, x; +\infty \,[$ waar je met kans 0.02 in terechtkomt. Wat is x ?

Je weet dat je deze vraag kan oplossen met invNorm(. Je weet ook dat dit commando je enkel zegt wat een **linkerstaart** is als jij de kans geeft om in die staart te vallen (dat is de oppervlakte boven die linkerstaart).

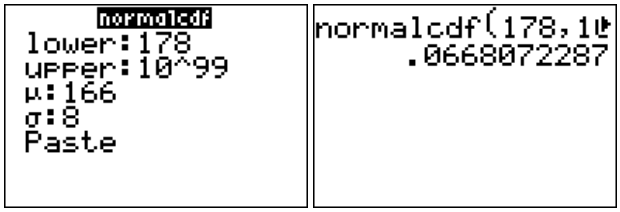
Met kans 2 % in $[\, x; +\infty \,[$ terechtkomen is hetzelfde als met kans 98 % in $]\, -\infty; x\,]$ terechtkomen. Zo stap je over van rechterstaart op linkerstaart.

Druk [2nd] [DISTR] en 3:invNorm(. Vul in, loop naar Paste en druk 2 keer [ENTER]. Het antwoord is 182.43 zodat $P(X \leq 182.43) = 0.98$. Hieruit volgt dat $P(X \geq 182.43) = 0.02$.

Tegen Roel zeg je dat hij (afgerond) minstens 182 cm moet zijn om tot de top 2 % te horen.

Tot welke topgroep hoort Roel dan wel? Wat is de oppervlakte boven $[178; +\infty[$ voor zijn leeftijdsgroep?

Druk [2nd] [DISTR] en 2: normalcdf(. Vul in, loop naar Paste en druk 2 keer [ENTER]. Het antwoord is 0.07 zodat $P(X \geq 178) = 0.07$. Er zijn dus 7 op 100 jongens die minstens zo groot zijn als Roel. Hij hoort zelfs niet bij de top 5 %.



Opdracht 8

Heb jij gezien wat hier gebeurt? Hoe kan je dat verklaren? Kon je zoiets vermoeden door naar de curven van figuur 7 te kijken? Wat vertellen zij over rechterstaarten? Kijk eens naar de oppervlakte boven $[178; +\infty[$. Wat zie je voor meisjes van 17 en wat gebeurt er bij jongens van 14?

- -

Hoeveel je afwijkt van het gemiddelde zegt blijkbaar nog niet hoe extreem je bent in je eigen leeftijdsgroep.

Bij jongens van 14 is de standaardafwijking groter dan bij meisjes van 17. Dat betekent dat er bij de jongens een grotere spreiding is rond het gemiddelde. Er zijn dus nogal wat jongens die heel wat groter en heel wat kleiner zijn dan 166 cm. Binnen zo'n populatie is een jongen van 178 cm geen uitzondering.

Om te weten hoe extreem je bent binnen je leeftijdsgroep heb je niet genoeg aan je afstand tot het gemiddelde. Dat zie je in het voorbeeld van Katia en Roel. De spreiding binnen je eigen groep speelt ook een rol en misschien is het een goed idee om die spreiding als een maat te gebruiken bij het meten van verschillen. Zo kom je aan een nieuwe "meetlat". Die meetlat zegt hoeveel standaardafwijkingen je van het gemiddelde ligt.

Katia is 178 cm. Voor haar leeftijdsgroep is het gemiddelde $\mu = 166\,\text{cm}$ en de standaardafwijking is $\sigma = 6\,\text{cm}$. Hoeveel standaardafwijkingen ligt zij boven het gemiddelde? Of met andere woorden, "hoeveel keer 6 cm" is het verschil tussen 178 cm en 166 cm? Dat is $\frac{178\,\text{cm} - 166\,\text{cm}}{6\,\text{cm}} = 2$.

Het getal dat je zo bekomt wordt een z–score genoemd. Het is eenheidsloos en het is een maatstaf voor de afwijking tot het populatiegemiddelde μ wanneer je als meetlat de spreiding van diezelfde populatie neemt (de standaardafwijking σ).

> Een z–score is niets anders dan een getal dat aangeeft hoeveel standaardafwijkingen een populatiewaarde x verwijderd is van het populatiegemiddelde μ (hierbij moet je het teken mee in rekening nemen en krijg je negatieve z–scores voor x–waarden die kleiner zijn dan μ).

Roel is ook 178 cm. Voor zijn leeftijdsgroep is het gemiddelde ook $\mu = 166\,\text{cm}$ maar de standaardafwijking is $\sigma = 8\,\text{cm}$. In zijn groep zit meer spreiding. Een maatstaf voor zijn afwijking tot het gemiddelde wordt gegeven door zijn z–score. Die is $\frac{178\,\text{cm} - 166\,\text{cm}}{8\,\text{cm}} = 1.5$. Dat is heel wat minder dan de z-score van Katia. Binnen zijn groep is de lengte van Roel minder extreem.

5. Het standaard normale kansmodel

5.1. Een transformatie

Katia behoort tot de top 2 % van haar leeftijdsgroep omdat zij, in verhouding tot de rest van haar groep, veel groter is dan het gemiddelde. Het gemiddelde μ en de standaardafwijking σ spelen beide een rol, en het is de z–score die met beide tegelijkertijd rekening houdt. Dat heb je gezien. Voor Katia is de z–score gelijk aan 2. Dat betekent dat zij 2 standaardafwijkingen boven het gemiddelde zit.

Je hebt gezien dat jongens van 14 minstens 182 cm moeten zijn om tot de grootste 2 % te horen. Steven is 14 jaar en hij is 182 cm. Wat is zijn z–score? Dat is heel eenvoudig als je weet dat de lengte van jongens van 14 verdeeld is als $X \sim N(166;8)$. De z–score van Steven is $\frac{182\,\text{cm} - 166\,\text{cm}}{8\,\text{cm}} = 2$.

> Overstappen op z–scores brengt je tot een gestandaardiseerde manier om met kansmodellen te werken **die normaal verdeeld zijn**. Om tot de top 2 % te horen moet je z–score minstens gelijk zijn aan 2, wat het gemiddelde μ en de standaardafwijking σ van je leeftijdsgroep ook moge wezen.

Om z–scores wat nader te bekijken neem je als voorbeeld de populatie $X \sim N(166;6)$ van de meisjes van 17 en je noteert de lengte van 8 verschillende meisjes. Die zijn (geordend): 157 cm, 160 cm, 163 cm, 166 cm, 167 cm, 169 cm, 172 cm en 178 cm.

Als je de z–score van die meisjes wil kennen, dan moet je 2 dingen doen. Je moet van hun lengte de gemiddelde lengte μ aftrekken en die resultaten moet je dan delen door de standaardafwijking σ.

Als je de eenheid (cm) even achterwege laat, dan zie je dat de oorspronkelijke getallen verschuiven en rond nul gaan liggen. Inderdaad, trek 166 af van {157; 160; 163; 166; 167; 169; 172; 178} en je krijgt {–9; –6; –3; 0; 1; 3; 6; 12}. Eigenlijk is dit nogal logisch. Je hebt te maken met een populatie waarvan de waarden rond 166 liggen. Trek daar nu overal 166 van af en je hebt een nieuwe populatie waarvan de waarden rond nul liggen.

Door getallen alleen maar te verschuiven (bijvoorbeeld over een afstand van 166) verander je wel het gemiddelde (dat verschuift mee) maar niet de spreiding rond dat gemiddelde. Voor die verschoven populatie is σ nog altijd gelijk aan 6. Maar nu ga je alle getallen ook nog delen door 6. Dat verandert hun spreiding wel en die wordt nu gelijk aan 1. Zo krijg je uiteindelijk een populatie met gemiddelde $\mu = 0$ en standaardafwijking $\sigma = 1$. Dat is de populatie van de z–scores.

De 8 getrokken lengten uit de oorspronkelijke populatie zijn nu veranderd in 8 z–scores uit de nieuwe populatie. In dit voorbeeld moet je {–9; –6; –3; 0; 1; 3; 6; 12} nog delen door 6 en dat wordt {–1.5; –1; –0.5; 0; 0.17; 0.5; 1; 2}. Dit zijn de z–scores van die 8 lengten.

De nieuwe populatie moet je nog een naam geven. Als je de uitkomsten (de z–scores) noteert met een kleine letter z dan noteer je het kansmodel met een hoofdletter Z. Dit kansmodel is het **standaard normale** model $Z \sim N(0;1)$.

De lengte van Katia is 178 cm. Dat is een uitkomst uit het populatiemodel $X \sim N(166;6)$ en je kan haar lengte met een algemene notatie aanduiden door een kleine letter x. Als je de z–score van Katia noteert met z dan kan je $\frac{178\text{ cm} - 166\text{ cm}}{6\text{ cm}} = 2$ ook schrijven als

$$\frac{x-\mu}{\sigma} = z$$

Dit is, in algemene notatie, de manier waarop je overstapt van x–waarden uit een willekeurig normaal kansmodel $X \sim N(\mu;\sigma)$ naar de corresponderende z–waarden van het standaard normale kansmodel $Z \sim N(0;1)$. In modelnotatie is dit:

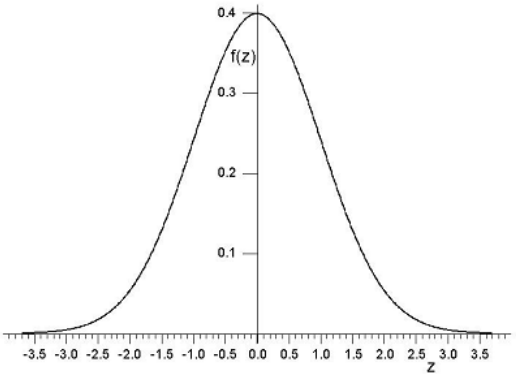
$$\frac{X-\mu}{\sigma} = Z$$

In woorden betekent dit het volgende:

> Start met gelijk welk normaal kansmodel $X \sim N(\mu;\sigma)$. Trek daar het gemiddelde μ van af en deel het resultaat daarna door de standaardafwijking σ. Het nieuwe model dat je zo krijgt is altijd het standaard normale kansmodel Z waarbij het gemiddelde altijd 0 is en de standaardafwijking altijd 1.

Als je $\mu = 0$ en $\sigma = 1$ invult in de algemene formule $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ voor $-\infty < x < +\infty$ dan krijg je de dichtheidsfunctie van het standaard normale kansmodel. Die ziet eruit als:

$$f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}} \text{ voor } -\infty < z < +\infty$$



De standaard normale dichtheidsfunctie

Vroeger, toen er geen GRM's en PC's waren, moest je tabellen gebruiken om kansen uit te rekenen. Je weet dat er oneindig veel verschillende normale kansmodellen zijn (bij elke andere μ en bij elke andere σ hoort een ander model) maar men kon toch niet oneindig veel tabellen maken! De kennis dat je met een eenvoudige transformatie kan overstappen van gelijk welk normaal model naar het standaard normale model maakt het mogelijk om alle vragen over normale modellen te berekenen met behulp van slechts één tabel, namelijk de tabel van de standaard normale.

Nu hoef je geen tabellen meer te gebruiken. Dit betekent niet dat je niet moet weten wat een z-score betekent en hoe je op een gestandaardiseerde manier moet leren denken bij al die verschillende normale modellen. Kijk maar eens goed naar de volgende figuur 10. Waarom moeten meisjes van 17 minstens 178 cm zijn en jongens van 14 minstens 182 cm om tot de top 2 % te behoren? En waarom kan je zowel bij die meisjes als bij die jongens identiek hetzelfde zeggen als je eerst transformeert naar z–scores? Om tot de top 2 % te behoren moet de z–score minstens gelijk zijn aan 2, voor beide groepen.

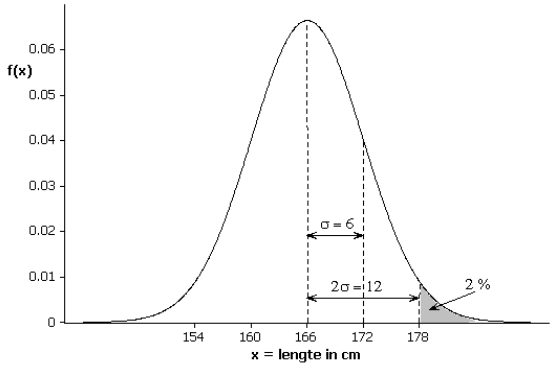
In formulevorm kan je dat als volgt schrijven (herken je de gebruikte transformaties?).

Voor de meisjes van 17 met $X \sim N(166;6)$ is

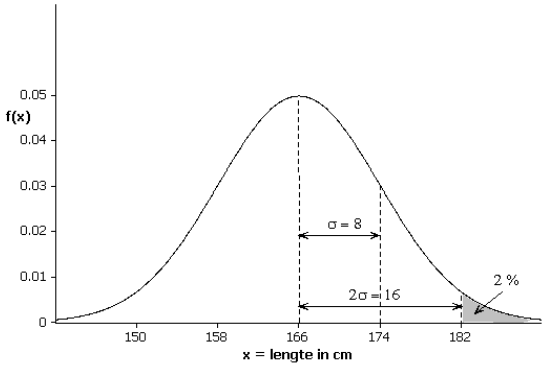
$$P(X \geq 178) = P\left(\frac{X-166}{6} \geq \frac{178-166}{6}\right) = P(Z \geq 2) = 0.02$$

Voor de jongens van 14 met $X \sim N(166;8)$ is

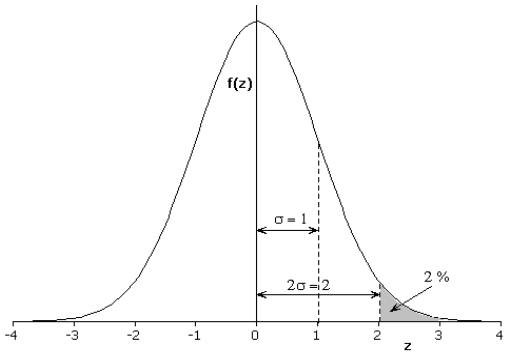
$$P(X \geq 182) = P\left(\frac{X-166}{8} \geq \frac{182-166}{8}\right) = P(Z \geq 2) = 0.02$$



meisjes van 17
$X \sim N(166;6)$



jongens van 14
$X \sim N(166;8)$



standaard normale
$Z \sim N(0;1)$

Figuur 10

5.2. Kansuitspraken

Door bij elk normaal model een eigen meetlat te gebruiken (namelijk de standaardafwijking σ van dat model) kom je tot gestandaardiseerde uitspraken voor alle normale modellen. In figuur 10 zie je zo'n uitspraak. Die zegt: "bij gelijk welk normaal model heb je kans 0.02 om terecht te komen in een rechterstaart die loopt van $\mu + 2\sigma$ tot $+\infty$".

Je kan dat ook vlotter zeggen: "bij elk normaal model is er 2 % kans om getallen uit te komen die minstens twee standaardafwijkingen groter zijn dan het gemiddelde".

In formulevorm schrijf je dat als: $P(X \geq \mu + 2\sigma) = 0.02$ voor elke $X \sim N(\mu ; \sigma)$.

Als je van X wil overstappen op Z dan gebruik je de gekende transformatie. Inderdaad

$$P(X \geq \mu + 2\sigma) = P(X - \mu \geq 2\sigma) = P\left(\frac{X-\mu}{\sigma} \geq 2\right) = P(Z \geq 2) = 0.02$$

Je kan natuurlijk ook omgekeerd te werk gaan. Zoek eerst eigenschappen van de standaard normale Z en pas die dan toe op een willekeurige normale $X \sim N(\mu ; \sigma)$. Je gebruikt dan dezelfde transformatie $\frac{X-\mu}{\sigma} = Z$ maar in de andere richting, van Z naar X. Doe dit nu in de volgende opdracht.

Opdracht 9

Zoek eens binnen welk gebied rond het centrum (dus rond nul) een standaard normale met kans 95 % terechtkomt. Gebruik je GRM en schrijf je resultaat als een kansuitspraak in de juiste notatie.

Begrijp je de standaard normale echt heel goed? Schrijf dan 1.96 eens anders op, namelijk als $0 + (1.96) \cdot 1$ want dat is $\mu + 1.96\,\sigma$ voor $Z \sim N(0 ; 1)$.

In woorden kan je voor $P(-1.96 \leq Z \leq 1.96) = 0.95$ zeggen dat een standaard normale in het interval [–1.96 ; 1.96] valt met kans 95 %. Maar het is veel interessanter om die 1.96 te herschrijven en op te merken dat $P(-1.96 \leq Z \leq 1.96) = P(\mu - 1.96\,\sigma \leq Z \leq \mu + 1.96\,\sigma) = 0.95$. Nu kan je zeggen dat een standaard normale met 95 % kans niet verder dan 1.96 standaardafwijkingen van zijn gemiddelde valt.

Als je dat op deze manier zegt, dan heb je niet enkel een kansuitspraak voor een standaard normale maar voor alle normale kansmodellen. Kijk maar naar de meisjes van 17 en de jongens van 17.

> Kansuitspraak:
>
> "*gelijk welk normaal kansmodel valt met 95 % kans niet verder dan 1.96 standaardafwijkingen van zijn gemiddelde*".

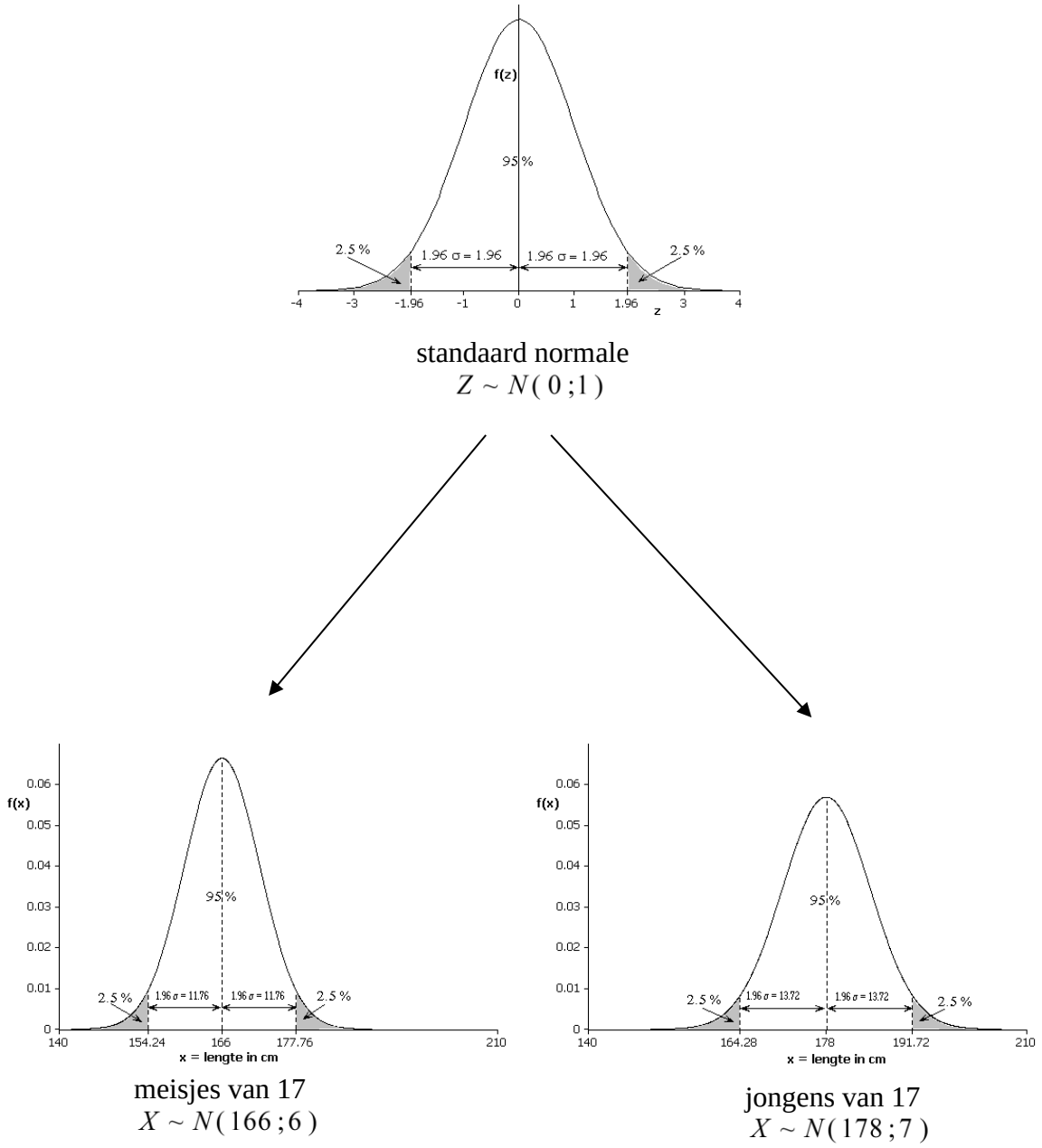
Je kan vooraf voorspellen wat er zal gebeuren en dat daarna controleren met je GRM.

Opdracht 10

- De lengte van 17-jarige meisjes gedraagt zich als $X \sim N(166;6)$.
  Hier is $\mu = 166$ en $\sigma = 6$ zodat $1.96\,\sigma = 11.76$.
  Volgens de kansuitspraak hebben 95 % van die meisjes een lengte tussen (afgerond) 154 cm en 178 cm want $P(\mu - 1.96\,\sigma \le X \le \mu + 1.96\,\sigma) = P(154.24 \le X \le 177.76) = 0.95$. Is dat waar? Controleer dit met je GRM.

- De lengte van 17-jarige jongens gedraagt zich als $X \sim N(178;7)$.
  Hier is $\mu = 178$ en $\sigma = 7$ zodat $1.96\,\sigma = 13.72$.
  Volgens de kansuitspraak hebben 95 % van die jongens een lengte tussen (afgerond) 164 cm en 192 cm want $P(\mu - 1.96\,\sigma \le X \le \mu + 1.96\,\sigma) = P(164.28 \le X \le 191.72) = 0.95$. Is dat waar? Controleer dit met je GRM.

Onderstaande figuur illustreert dit voorbeeld. Zij toont hoe je uit een eigenschap van de standaard normale kan overstappen op gelijk welke andere normale. Zorg ervoor dat je die figuur goed begrijpt. Kijk ook naar de ligging en de vorm van de curven.



standaard normale
$Z \sim N(0;1)$

meisjes van 17
$X \sim N(166;6)$

jongens van 17
$X \sim N(178;7)$

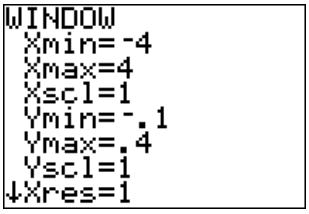
Figuur 11

Opdracht 11

Vul in (toon je redenering en berekening):

- "Elk normaal kansmodel valt met kans …… niet verder dan 1.645 standaardafwijkingen van zijn gemiddelde".

  Toon dit ook grafisch voor de situatie van het standaard normale kansmodel. Gebruik je GRM. Pas de vensterinstellingen ([WINDOW]) aan zoals aangegeven. Druk dan [2nd][DISTR] loop naar Draw en druk 1:ShadeNorm. In welk interval kom je dan terecht en waar wordt de kans getoond?



- "Elk normaal kansmodel valt met kans 68 % niet verder dan (afgerond) …… standaardafwijkingen van zijn gemiddelde".
  Los dit op met je GRM en zeg hoe je dit doet.

  Je kan nu, zonder gebruik van GRM, onmiddellijk zeggen dat 68 % van de 17-jarige jongens tussen …… cm en ……cm groot is.