De normale verdeling

Hilde Eggermont
(redactie Uitwiskeling)

Dag van de Wiskunde
15 november 2003

Inhoud

1. Inleiding
2. De start: histogrammen beschrijven met een dichtheidsfunctie
3. Relatieve frequenties vinden m.b.v. de normale dichtheidsfunctie
4. Vuistregels
5. Terugzoeken
6. Normale dichtheidsfuncties en de standaardnormale verdeling
7. Niet alle gegevens zijn normaal verdeeld!
8. Relatieve frequentiedichtheid
9. Normale verdelingen vergelijken
10. Toepassing: de machine goed instellen
11. Historische noot
12. Commando's i.v.m. de normale verdeling op de TI-83: samenvatting

1. Inleiding

De eindtermen voor de derde graad ASO vermelden de normale verdeling. De leerlingen die vanaf 1 september 2004 de derde graad aanvatten, zullen moeten kennismaken met de normale verdeling, ook diegenen die voor een studierichting met 3 of 4 uur wiskunde kiezen. Voor veel wiskundeleerkrachten is dat een relatief onbekend onderwerp. Daarom willen we in deze tekst tonen hoe dit onderwerp behandeld kan worden in een 3- of 4-uursrichting uit het ASO. De normale verdeling wordt hierbij niet bekeken vanuit de kansrekening, maar vanuit de beschrijvende statistiek: als een wiskundig model voor klokvormige frequentieverdelingen.

De tekst die volgt, is gebaseerd op een tekst geschreven door Johan Deprez, Jan Roels en Hilde Eggermont en die verschenen is in het nummer 18/1 (december 2001) van het tijdschrift Uitwiskeling.

Eindtermen en leerplannen

De eindtermen voor de derde graad ASO schrijven voor alle leerlingen het onderwerp 'de normale verdeling' voor. Deze eindtermen zullen vanaf september 2004 ingevoerd worden. We voegen hier ter informatie de eindtermen toe:

- De leerlingen kunnen in betekenisvolle situaties gebruik maken van een normale verdeling als continu model bij data met een klokvormige frequentieverdeling en het gemiddelde en de standaardafwijking van de gegeven data gebruiken als schatting voor het gemiddelde en de standaardafwijking van deze normale verdeling.
- De leerlingen kunnen het gemiddelde en de standaardafwijking van een normale verdeling grafisch interpreteren.
- De leerlingen kunnen grafisch het verband leggen tussen een normale verdeling en de standaardnormale verdeling.
- De leerlingen kunnen bij een normale verdeling de relatieve frequentie interpreteren van een verzameling gegevens met waarden tussen twee gegeven grenzen, met waarden groter dan een gegeven grens en met waarden kleiner dan een gegeven grens als oppervlakte van een gepast gebied.

De studie van de normale verdeling is binnen de eindtermen een onderdeel van de beschrijvende statistiek. In de tweede graad maken de leerlingen kennis met grafische voorstellingen van statistische gegevens en centrum- en spreidingsmaten. De centrum- en spreidingsmaten vatten de data samen in getallen. Soms vertoont een hele grote set gegevens een regelmatig patroon dat beschreven kan worden door een functie. De normale dichtheidsfunctie is een voorbeeld van zo'n functie.

Alle ASO-leerlingen

Eindtermen zijn doelen die alle leerlingen moeten bereiken. De aanpak die we in deze tekst voorstellen (geïnspireerd op [2], [6] en [7]), houdt daar rekening mee. We werken het onderwerp uit zoals het in een 3-uursrichting behandeld zou kunnen worden. In richtingen met 6 uur wiskunde per week zal er meer kansrekenen en statistiek op het programma staan dan wat de eindtermen voorschrijven. We zijn er echter van overtuigd dat onze aanpak ook bij een zwaarder pakket een goede invalshoek kan zijn. Voor TSO en KSO is een beperkte kennismaking met de normale verdeling voorzien. We hopen dat onze aanpak ook voor deze richtingen inspirerend kan werken.

Ook een nieuw onderwerp voor de leerkrachten

Voor de meeste leerkrachten is dit een nieuw onderwerp. We hebben daarom geprobeerd de nascholing zo uit te werken dat de lezer geen voorkennis over dit onderwerp nodig heeft.

Grafisch rekentoestel (of computer)

Lezers die vertrouwd zijn met het onderwerp weten dat men vroeger om relatieve frequenties (of kansen) bij een normale verdeling te bepalen gebruik moest maken van tabellen en omrekenformules. Dit is niet meer nodig. Men kan die nu eenvoudig met de grafische rekenmachine (of computer) bepalen. Daarnaast kunnen we met de grafische rekenmachine enkele belangrijke aspecten van de normale verdeling goed grafisch illustreren. Je zult merken dat we in deze nascholing heel veel gebruik maken van een grafische rekenmachine. Wij gebruiken een TI-83, maar met een andere grafische rekenmachine, een computer of met figuren op papier of transparant kan het natuurlijk ook. We hebben veel schermafdrukken opgenomen zodat je de tekst ook kunt lezen als je je rekenmachine niet bij de hand hebt. Maar het is natuurlijk wel veel leuker als je je rekenmachine er bij neemt. We veronderstellen dat je de basisvaardigheden voor het gebruik van de TI-83 voor functies en beschrijvende statistiek onder de knie hebt (meer uitleg daarover kan je vinden in Uitwiskeling 15/3, in de handleiding van je rekenmachine, [3], …). Alles wat verder gaat dan deze basisvaardigheden wordt in de tekst uitgelegd. Achteraan (paragraaf 12) vind je ook een overzichtje van de commando's op de TI-83 die verband houden met de normale verdeling.

Wat volgt

Eerst geven we aan hoe de normale verdeling in de klas kan aangebracht worden en hoe de normale verdeling gebruikt kan worden om statistische vragen op te lossen. In paragraaf 9 en 10 geven we enkele toepassingen. We besluiten de tekst met een historische noot.

2. De start: histogrammen beschrijven met een dichtheidsfunctie

Een histogram en een grafiek

Het inleidende voorbeeld hebben we overgenomen uit [2]. In 1947 werd in opdracht van N.V. Magazijn 'De Bijenkorf' een statistisch onderzoek verricht naar de lichaamsafmetingen van de Nederlandse vrouwen. Het doel van het onderzoek was beter passende damesconfectiekleding te kunnen maken. Voor dit onderzoek werden bij 5000[1] willekeurig gekozen volwassen, vrouwelijke klanten vijftien lichamelijke kenmerken gemeten. De resultaten voor wat de lichaamslengte betreft, vind je in de onderstaande tabel.

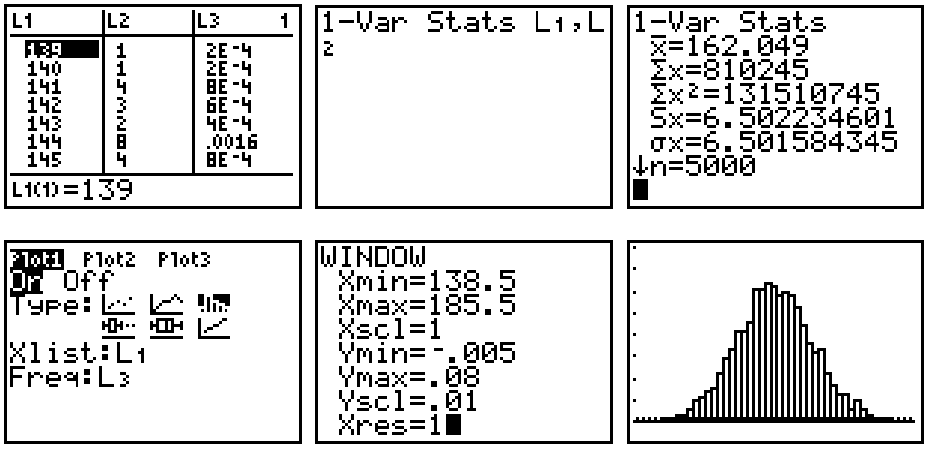
In de tweede graad hebben de leerlingen geleerd dat ze gegevens overzichtelijker kunnen maken door gebruik te maken van kengetallen zoals gemiddelde, mediaan, standaardafwijking, … en/of door de gegevens grafisch voor te stellen. M.b.v. de rekenmachine berekenen we kengetallen van de bovenstaande gegevens en maken we een histogram (we zetten verticaal de *relatieve* frequenties uit). Je vindt de schermafdrukken onder de tabel.

lengte (in cm)	frequentie	relatieve frequentie	lengte (in cm)	frequentie	relatieve frequentie
[138,5; 139,5[	1	0,0002	[161,5; 162,5[	313	0,0626
[139,5; 140,5[	1	0,0002	[162,5; 163,5[	290	0,0580
[140,5; 141,5[	4	0,0008	[163,5; 164,5[	294	0,0588
[141,5; 142,5[	3	0,0006	[164,5; 165,5[	291	0,0582
[142,5; 143,5[	2	0,0004	[165,5; 166,5[	261	0,0522
[143,5; 144,5[	8	0,0016	[166,5; 167,5[	222	0,0444
[144,5; 145,5[	4	0,0008	[167,5; 168,5[	184	0,0368
[145,5; 146,5[	17	0,0034	[168,5; 169,5[	157	0,0314
[146,5; 147,5[	18	0,0036	[169,5; 170,5[	167	0,0334
[147,5; 148,5[	32	0,0064	[170,5; 171,5[	109	0,0218
[148,5; 149,5[	51	0,0102	[171,5; 172,5[	86	0,0172
[149,5; 150,5[	54	0,0108	[172,5; 173,5[	65	0,0130
[150,5; 151,5[	71	0,0142	[173,5; 174,5[	62	0,0124
[151,5; 152,5[	78	0,0156	[174,5; 175,5[	29	0,0058
[152,5; 153,5[	115	0,0230	[175,5; 176,5[	49	0,0098
[153,5; 154,5[	149	0,0298	[176,5; 177,5[	28	0,0056
[154,5; 155,5[	170	0,0340	[177,5; 178,5[	17	0,0034
[155,5; 156,5[	208	0,0416	[178,5; 179,5[	5	0,0010
[156,5; 157,5[	208	0,0416	[179,5; 180,5[	10	0,0020
[157,5; 158,5[	231	0,0462	[180,5; 181,5[	6	0,0012
[158,5; 159,5[	301	0,0602	[181,5; 182,5[	3	0,0006
[159,5; 160,5[	302	0,0604	[182,5; 183,5[	1	0,0002
[160,5; 161,5[	321	0,0642	[183,5; 184,5[	2	0,0004

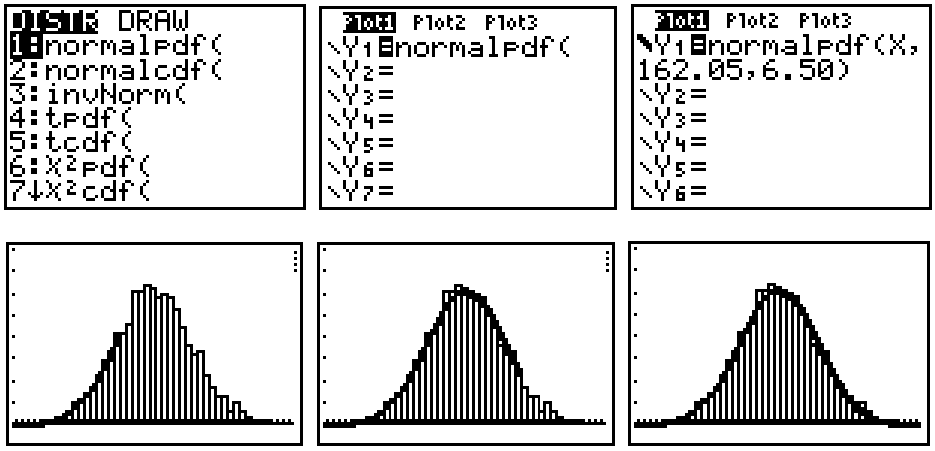
Tabel: Lengte van 5000 vrouwen

In de figuur hieronder zie je dat de klassenmiddens opgeslagen zijn in de lijst L_1, de frequenties in de lijst L_2 en de relatieve frequenties in de lijst L_3.

[1] We doen de werkelijkheid hier een klein beetje geweld aan. We hebben één meetresultaat geëlimineerd omdat er één klasse teveel was voor de TI-83. Dit rekentoestel kan met hoogstens 47 klassen werken.

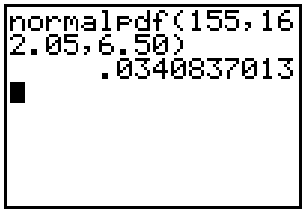


Je kan het gemiddelde en de standaardafwijking van een set gegevens opvatten als getallen die een samenvatting geven van de gegevens. De betekenis van deze getallen is echter krachtiger dan je op het eerste gezicht zou verwachten. Als je de functie 'normalpdf' van de rekenmachine gebruikt in combinatie met het gemiddelde en de standaardafwijking van de gegevens (zie de eerste drie schermafdrukken), krijg je iets moois te zien. (Om aan het eerste scherm te komen, moet je op [2nd] [DISTR] drukken.)



We stellen vast dat de grafiek van de functie het histogram verbazingwekkend goed weergeeft. Als we de grafiek van de functie zien, herkennen we dadelijk de globale vorm van het histogram. De hoogte van de staven is voor de meeste klassen ongeveer gelijk aan de functiewaarde van het klassemidden.

Dit betekent dat de functie de relatieve frequentietabel min of meer kan vervangen. Veronderstel nu eens dat we alleen over de functie en niet over de tabel zouden beschikken. Als we dan bijvoorbeeld willen weten hoeveel vrouwen 155 cm lang zijn, kunnen we de functiewaarde van 155 berekenen:



Omdat we hier wel over de tabel beschikken, kunnen we nagaan dat het resultaat goed overeenstemt met de echte relatieve frequentie (0.034).

Op basis van wat we hierboven vastgesteld hebben, kunnen we zeggen dat de functie een zeer compacte (maar enigszins geïdealiseerde) beschrijving vormt van de gegevens over de lengte van de vrouwen. Deze functie is een *wiskundig model* voor de frequentieverdeling van de gegevens over de lengte van de volwassen vrouwen uit 1947: het is een *eenvoudige* functie die de relatieve frequenties *bij benadering* weergeeft. De leerlingen hebben voordien al geregeld gebruik gemaakt van functies als wiskundig model

voor situaties uit de realiteit (eerste- en tweedegraadsfuncties, exponentiële functies, …). Nu gebruiken we dus een functie als een wiskundig model voor een frequentietabel.

De functie

Nu gaan we iets dieper in op de functie die we hierboven gebruikten. De functie 'normalpdf' van de rekenmachine staat voor het volgende voorschrift:

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot 6{,}50} 2{,}71828^{-\frac{(x-162{,}05)^2}{2 \cdot 6{,}50^2}}.$$

(We gebruikten 2,71828 als decimale benadering voor het getal *e* omdat niet alle leerlingen dat getal kennen.) Het voorschrift ziet er op het eerste gezicht nogal indrukwekkend uit omdat er veel bewerkingen in voorkomen. Daarom is het handig dat de functie onder een eenvoudige naam beschikbaar is op de rekenmachine. Toch staat er uiteindelijk niets in het voorschrift dat voor leerlingen uit een 3-uurscursus onbegrijpelijk is. Je kunt het volledig uitgeschreven voorschrift dus wel eens laten zien aan alle leerlingen (zonder er daarom achteraf verder mee te werken). Het is per slot van rekening merkwaardig dat je de relatieve frequenties (statistische gegevens) bij benadering kunt berekenen met een formule! De leerlingen hebben gemiddelde en standaardafwijking leren kennen als *getallen* die een *samenvatting* geven van de frequentieverdeling. Nu krijgen we een *functie* (*grafiek, formule,* …) die een samenvatting geeft van de frequentieverdeling.

We onderzoeken nu de benaming die de rekenmachine gebruikt. De staart 'pdf' staat voor 'probability density function', in het Nederlands 'kansdichtheidsfunctie'. Het gebruik van het woord 'kans' in de benaming kun je als volgt begrijpen: we hebben vastgesteld dat zo'n dichtheidsfunctie gebruikt wordt om relatieve frequenties te vinden en kansen zijn 'geïdealiseerde relatieve frequenties'. De functie wordt zowel gebruikt binnen de kansrekening als binnen de beschrijvende statistiek. Wij houden het in deze nascholing echter bij beschrijvende statistiek en dus bij relatieve frequenties. Het woord 'dichtheid' kunnen we pas goed verklaren in paragraaf 8. In deze nascholing zullen we de term *dichtheidsfunctie* gebruiken, zonder verwijzing naar kansen, omdat wij de functie gebruiken in het kader van de beschrijvende statistiek. Het woord 'normal' verwijst naar 'normaal'. We hebben in het voorgaande voorbeeld immers met een dichtheidsfunctie van een speciale vorm gewerkt, namelijk een *normale dichtheidsfunctie*. We zeggen ook dat de gegevens *normaal verdeeld* zijn. Een andere naam die voor de grafiek gebruikt wordt, is Gausskromme (zie ook paragraaf 11). Soms wordt ook de benaming 'verdelingsfunctie' gebruikt i.p.v. 'dichtheidsfunctie'. Het is beter om dat niet te doen omdat dit woord een andere betekenis heeft in de kansrekening. Zie hiervoor paragraaf 12. Van heel wat gegevens is geweten dat ze normaal verdeeld zijn (lengte van volwassenen, lengte van allerlei beenderen, IQ, meetfouten, werkelijke inhoud van een machinaal gevulde fles melk, …). Daar ligt de oorsprong van het gebruik van het woord 'normaal'. Anderzijds zijn er nog heel wat meer gegevens die juist niet normaal verdeeld zijn (inkomen, zwangerschapsduur, leeftijd bij overlijden, …; we gaan daar dieper op in in paragraaf 7). In die zin is het gebruik van het woord 'normaal' toch wat eigenaardig. Deze andere gegevens zijn immers niet zeldzaam of 'abnormaal'!

Er zijn oneindig veel van die normale dichtheidsfuncties. De twee getallen die in het voorschrift voorkomen, dienen om de juiste normale dichtheidsfunctie te selecteren. Ze zijn het gemiddelde en de standaardafwijking van de gegevens. We spreken daarom van *de* normale dichtheidsfunctie met gemiddelde 162,05 en standaardafwijking 6,50.

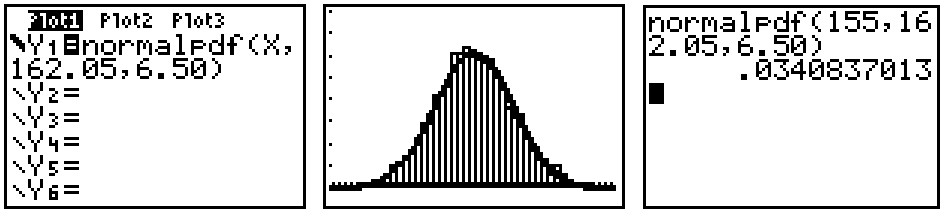
Toepassing: lengte van volwassen mannen

Er wordt gegeven dat de lengte (in cm) van 10 000 volwassen mannen normaal verdeeld is met een gemiddelde van 178,4 en een standaardafwijking van 7,4. We beschikken niet over verdere gegevens en

hebben dus geen frequentietabel. Schat m.b.v. een normale dichtheidsfunctie hoeveel van deze mannen (afgerond) 168 cm lang zijn. (*Antwoord: ongeveer 201 mannen.*)

Samenvatting

De *normale dichtheidsfunctie* met gemiddelde 162,05 en standaardafwijking 6,50 is een *wiskundig model* voor de frequentieverdeling van de gegevens over de lengte van de 5000 volwassen vrouwen uit 1947. De grafiek van deze functie geeft het histogram heel goed weer. We kunnen deze functie ook gebruiken om de relatieve frequenties bij benadering te berekenen.

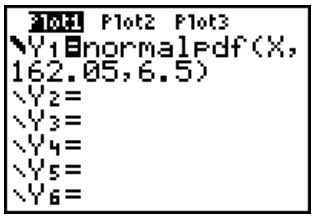


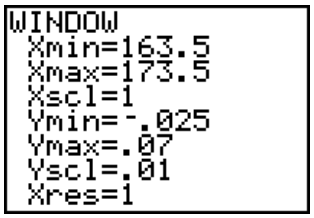
3. Relatieve frequenties vinden m.b.v. de normale dichtheidsfunctie

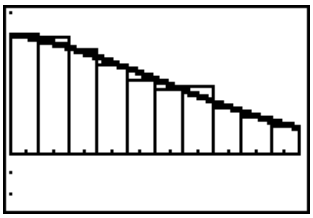
De functie *normalpdf*(*x*,162.05,6.5) is een *model* voor het histogram. Uit dat model moeten we dezelfde informatie kunnen halen als uit het histogram. In vele gevallen zullen we immers niet over het histogram beschikken, maar alleen over het model. Een eerste vraag waar we een antwoord op zoeken, is *hoe je (bij benadering) de relatieve frequenties van een groter gebied kunt vinden m.b.v. de normale dichtheidsfunctie*. We onderzoeken dit aan de hand van de volgende vraag:

Hoeveel procent van de vrouwen van de steekproef is tussen de 165,5 *en* 171,5 *cm lang?*

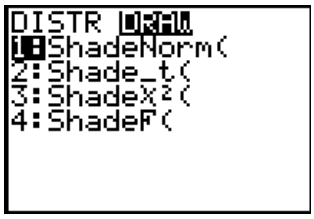
Om dit te vinden aan de hand van het histogram tel je de hoogten van de staven bij 166 tot en met 171 op. Je vindt dan 0,22 of 22%. Omdat de staven breedte 1 hebben, is de hoogte van een staaf gelijk aan zijn oppervlakte. De gezochte relatieve frequentie is dan de totale oppervlakte van de staven. In het model komt dit overeen met *de oppervlakte onder de grafiek* tussen de verticale rechten $x = 165{,}5$ en $x = 171{,}5$. Door naar de oppervlakte te kijken heb je een beter zicht op welke stukjes te veel en welke te weinig gerekend worden bij het overstappen van het histogram op de dichtheidsfunctie.

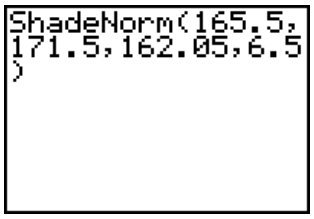


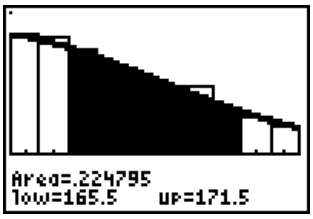




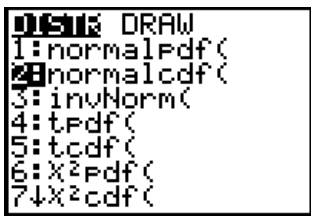
Op de grafische rekenmachine is een toets voorzien die de oppervlakte onder de dichtheidsfunctie tussen twee grenzen tekent en berekent. Je vindt die ook bij [2nd][DISTR]. Je kiest nu voor DRAW en vervolgens 1:ShadeNorm(. Na het ingeven van de gewenste parameters krijg je een figuur met de ingekleurde en berekende oppervlakte. Je vindt voor de oppervlakte onder de normale dichtheidsfunctie 0,2248.

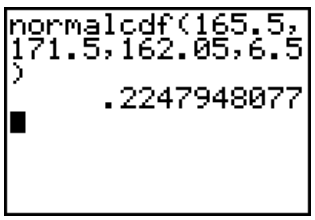






Je kunt ook de oppervlakte laten berekenen zonder er een figuur bij te maken. Dit kan door bij [2nd][DISTR] te kiezen voor 2:normalcdf(. Het achtervoegsel ‘cdf’ staat voor ‘cumulative distribution function’, of in het Nederlands, ‘cumulatieve verdelingsfunctie’. ‘Cumulatief’ wil zeggen dat je waarden samenvoegt of optelt. Dit is wat we doen door klassen samen te nemen.





Je vindt opnieuw voor de oppervlakte 0,2248. Bijgevolg hebben 22,48% van de vrouwen een lengte tussen de 165,5 en 171,5 cm.

Toepassing: lengte van volwassen mannen (bis)

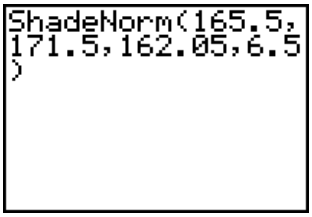
We hernemen de toepassing van de vorige paragraaf en berekenen de frequentie van een grotere klasse.

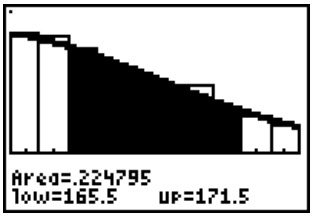
Er wordt gegeven dat de lengte (in cm) van 10 000 volwassen mannen normaal verdeeld is met een gemiddelde van 178,4 en een standaardafwijking van 7,4. Schat hoeveel van deze mannen een lengte tussen 162 en 173 cm hebben.

(Bereken eerst Normalcdf(162;173;178,4;7,4). Als antwoord vind je dan ongeveer 2194 mannen.)

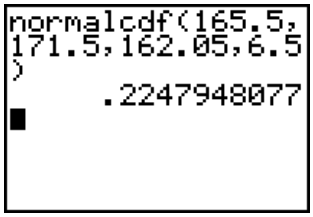
Samenvatting

De relatieve frequentie van een klasse van een normaal verdeelde variabele is de *oppervlakte van het gebied onder de normale dichtheidsfunctie* tussen de grenzen van de klasse. Je kunt dit op twee manieren met je grafisch rekentoestel berekenen: ofwel met een figuur erbij, ofwel rechtstreeks.





of

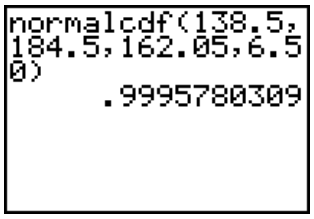


4. Vuistregels

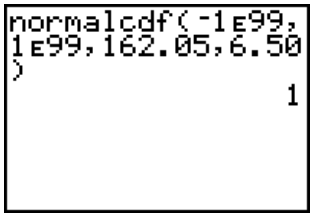
Omdat we nog vaak met de cumulatieve verdelingsfunctie zullen werken, is het handig om enkele veel voorkomende waarden te kennen.

De totale oppervlakte onder de kromme

Alle vrouwen van de steekproef hebben een lengte tussen 138,5 en 184,5 cm. De totale relatieve frequentie is dus 1. Als je dit berekent met het model, krijg je een (kleine) afwijking:



Dat kleine verschil is te verklaren doordat we *niet alle* oppervlakte onder de kromme berekend hebben. Aan de twee uiteinden wordt de functie weliswaar zéér klein, maar nooit helemaal nul. Als je werkelijk de totale oppervlakte onder de modelkromme wilt berekenen, moet je alle mogelijke x-waarden bekijken, d.w.z. van $-\infty$ tot $+\infty$. Omdat op de TI-83 geen symbool voor ∞ staat, voeren we de hoogste macht van 10 in die voor het toestel mogelijk is, namelijk 10^{99}. Zo vinden we inderdaad 1 als totale oppervlakte onder de kromme.



Toepassing: lengte van volwassen mannen (ter)

We kunnen nu ook de relatieve frequentie van een gebied links of rechts van een gegeven lengte bepalen.

Hoeveel procent van die mannen van de vorige toepassing zijn kleiner dan 165,5 cm?

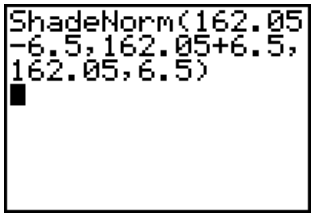
(3,5%)

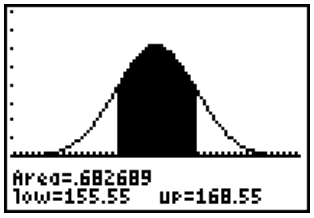
Hoeveel procent van die mannen zijn groter dan 165,5 cm?

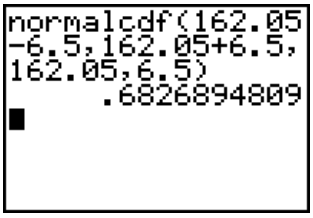
(1 – 0,035 = 0,965 of 96,5 %)

Het σ- en 2σ-gebied

Bij de Bijenkorf-steekproef van de vrouwen was het gemiddelde 162,05. De standaardafwijking is een gemiddelde afwijking. Je kunt de waarden die binnen deze afwijking vallen als 'normaal' beschouwen. We onderzoeken nu bij hoeveel vrouwen de lengte niet te sterk afwijkt van dit gemiddelde. Concreter gezegd berekenen we nu aan de hand van het model bij hoeveel vrouwen de lengte niet meer dan één standaardafwijking verschilt van het gemiddelde.

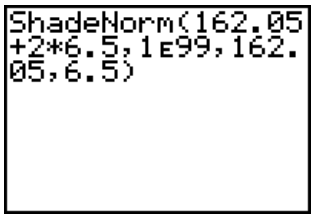


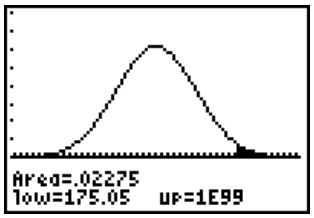


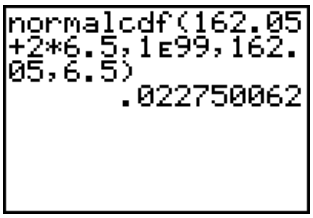


Dit betekent dat 68% van de vrouwen tot deze middengroep behoren. Men zegt ook wel dat 68% van de data in het σ-gebied ligt. Ook bij een ander gemiddelde en standaardafwijking ligt 68% van de data in het σ-gebied.

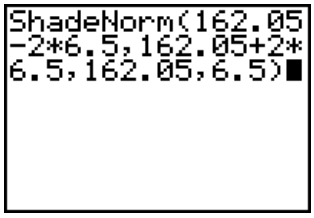
Op dezelfde manier kunnen we nagaan hoeveel vrouwen 'uitzonderlijk lang' zijn. We bedoelen hiermee dat ze meer dan twee standaardafwijkingen langer zijn dan het gemiddelde. Het resultaat is af te lezen van het schermpje hieronder.

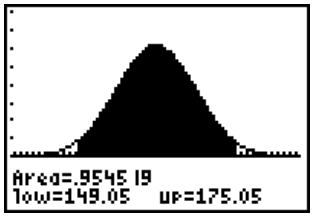


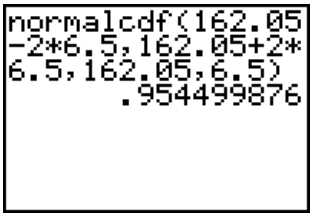




Bijgevolg is 2,3% van de vrouwen uitzonderlijk lang. Omwille van de symmetrie kunnen we zeggen dat ook 2,3% van de vrouwen uitzonderlijk klein is. Dit komt neer op een andere vuistregel uit de beschrijvende statistiek die zegt dat ongeveer 95% van de waarnemingen niet meer dan twee standaardafwijkingen afwijkt van het gemiddelde. Men noemt dit het 2σ-gebied. We kunnen ook rechtstreeks narekenen dat dit gebied ongeveer 95% van de data bevat.







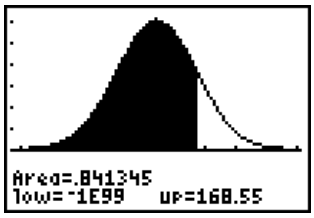
Samenvatting

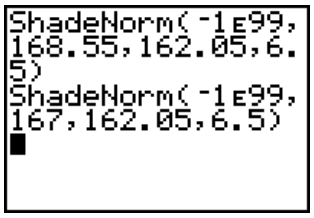
- De totale oppervlakte onder de normale dichtheidsfunctie is 1.
- Als de gegevens normaal verdeeld zijn, ligt 68 % van de data in het σ-gebied (zie figuur links).
- Als de gegevens normaal verdeeld zijn, ligt 95 % van de data in het 2σ-gebied (zie figuur rechts).

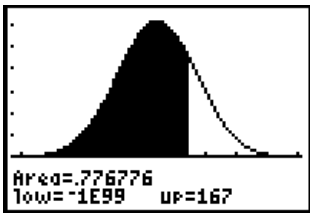
5. Terugzoeken

Voor een volgende opgave vertrekken we van een percentage (een oppervlakte) en zoeken we de bijbehorende grenswaarde. Als voorbeeld zoeken we hoe groot een Nederlandse vrouw anno 1947 moest zijn opdat 75% van de vrouwen kleiner zou zijn dan zij. Dit komt overeen met het derde kwartiel.

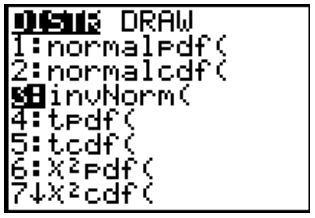
Een goede start is de grafiek van de normale dichtheidsfunctie. Je kunt dan een gokje wagen en schatten hoe ver je moet gaan om driekwart van de totale oppervlakte te arceren. De vuistregels kunnen een handje helpen bij dat afschatten. Als eerste gokje probeerde ik 162,05 + 6,5. De tweede gok zat al aardig in de buurt.

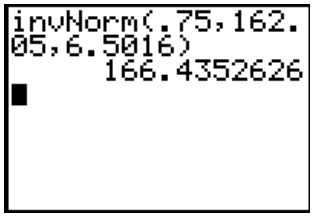




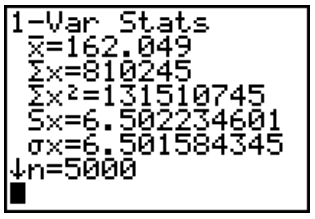


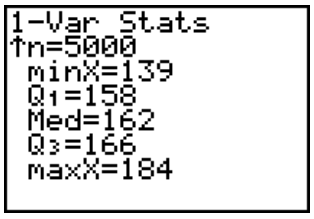
Het is uiteraard heel omslachtig om zo tot het goede antwoord te komen. Gelukkig is er een instructie voorzien op de grafische rekenmachine waarmee je dit kunt berekenen. Je vindt die onder het menu [DISTR] en vervolgens 3:invNorm(. We geven de juiste parameters in (het percentage, het gemiddelde en de standaardafwijking) en vinden zo het antwoord op de vraag waar het derde kwartiel zich bevindt.





Anno 1947 moest een vrouw dus 166,4 cm lang zijn om langer te zijn dan driekwart van de vrouwen. We kunnen dit vergelijken met het derde kwartiel Q_3 van het histogram. Het rekentoestel berekende dit reeds helemaal in het begin, maar toen bleef deze waarde verborgen. Door in dat scherm verder te scrollen, kunnen we Q_3 aflezen. Deze waarde (166) ligt heel dicht bij de waarde die we berekenden in het model.





Merk op dat de functie *invNorm* niet zomaar de inverse is van de functie *normalcdf*. Bij de verdelingsfunctie *normalcdf* geef je twee grenswaarden in en krijg je de overeenkomstige oppervlakte. Bij de functie *invNorm* geef je een oppervlakte in en krijg je één grenswaarde. Het gebied met de gegeven oppervlakte is dan het stuk links van de verticale lijn door de berekende grenswaarde.

Een toepassing: de lengte van komkommers

In de volgende opgave wordt dit alles nog eens ingeoefend in een andere context. De opgave is geïnspireerd op [2].

Komkommertijd

Op een groenteveiling worden in een bepaalde periode van de zomer te veel komkommers aangevoerd. Het zijn er zo veel dat er een overschot van 25% van de aanvoer is. Om de komkommerprijs niet te laten instorten besluit de directie van de veiling, in overleg met de kwekers, de 25% kleinste komkommers niet op de markt te brengen. Uit een steekproef leidt men af dat lengte van de komkommers klokvormig verdeeld is met een gemiddelde lengte van 40 cm en een standaardafwijking van 6 cm.

1. Laat je rekenmachine de grafiek tekenen van de normale dichtheidsfunctie die de lengte van de komkommers beschrijft. Teken deze grafiek over in je schrift.
2. Welk percentage van de komkommers zal langer zijn dan 50 cm? Duid de overeenkomstige oppervlakte aan op je grafiek.

   (4,8%)
3. De 25% kleinste komkommers zullen niet geveild worden. Hoelang moet een komkommer dan minstens zijn om op de markt te komen?

   (Minstens 36 cm.)
4. De 10% langste komkommers krijgen het etiket “jumbo-komkommer”. Vanaf welke lengte is een komkommer een jumbo?

   (Vanaf 47,7 cm.)

6. Normale dichtheidsfuncties en de standaardnormale verdeling

De leerlingen hebben ondertussen het nut ervaren van normale dichtheidsfuncties. Ze hebben geleerd om het gemiddelde en de standaardafwijking te gebruiken als parameters in deze normale dichtheidsfuncties. Nu vestigen we meer nadrukkelijk de aandacht op de invloed van de parameters op de grafiek. Aan de hand van de volgende werktekst laten we de leerlingen enkele normale dichtheidsfuncties onderzoeken.

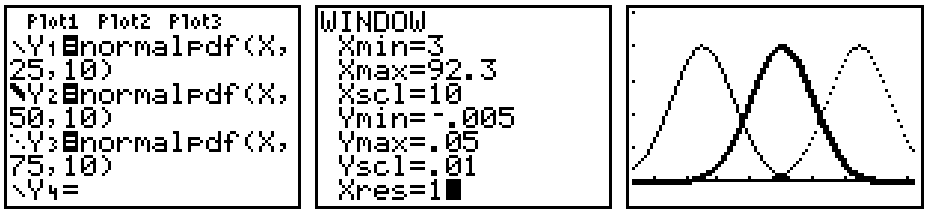
Normale dichtheidsfuncties en de standaardnormale dichtheidsfunctie

1. Zoek een goed tekenvenster voor de normale dichtheidsfunctie met gemiddelde 50 en standaardafwijking 10.

   (Je krijgt een mooie afbeelding als je er voor zorgt dat de x-waarde 50 in het midden van het tekenvenster ligt. Een vensterbreedte van 50 eenheden geeft een goed resultaat. Een maximale y-waarde van 0,05 eenheden volstaat.)

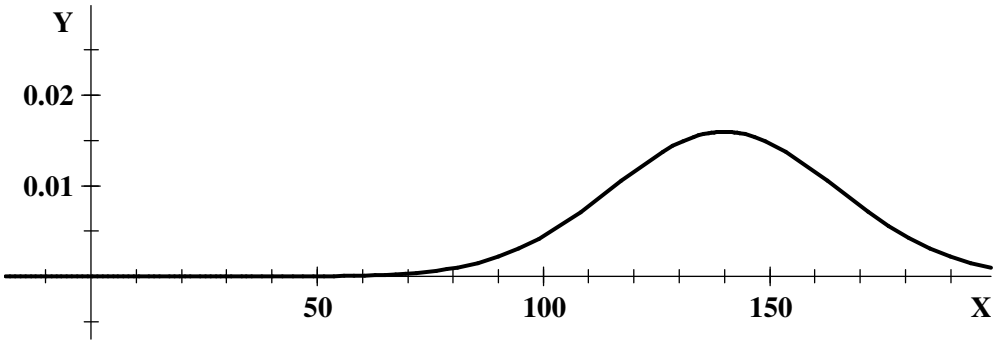
2. Stel de normale dichtheidsfuncties met gemiddelde 25, 50 en 75 en met standaardafwijking 10 op één tekening voor. Welk verband bestaat er tussen deze drie grafieken?

   (De onderstaande schermafdrukken tonen een goed tekenvenster en de bijbehorende grafieken.



   De drie grafieken zijn gewoon horizontaal verschoven ten opzichte van elkaar.)

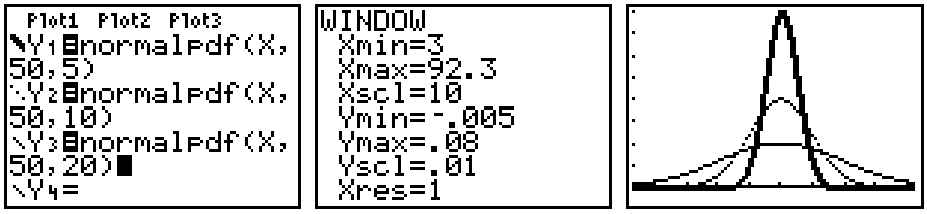
3. Leg uit hoe je het gemiddelde van een normale dichtheidsfunctie kunt aflezen op de grafiek. Wat is het gemiddelde van de normale dichtheidsfunctie die je hieronder ziet?



   (Het gemiddelde is de x-coördinaat van het hoogste punt. Hier is dat 140. De rechte $x = 140$ *is de symmetrie-as van de grafiek.)*

4. Stel de normale dichtheidsfuncties met gemiddelde 50 en standaardafwijking 5, 10 en 20 op één tekening voor. Het is nu moeilijker om precies te beschrijven hoe je van de ene naar de andere grafiek overgaat. Probeer toch maar eens.

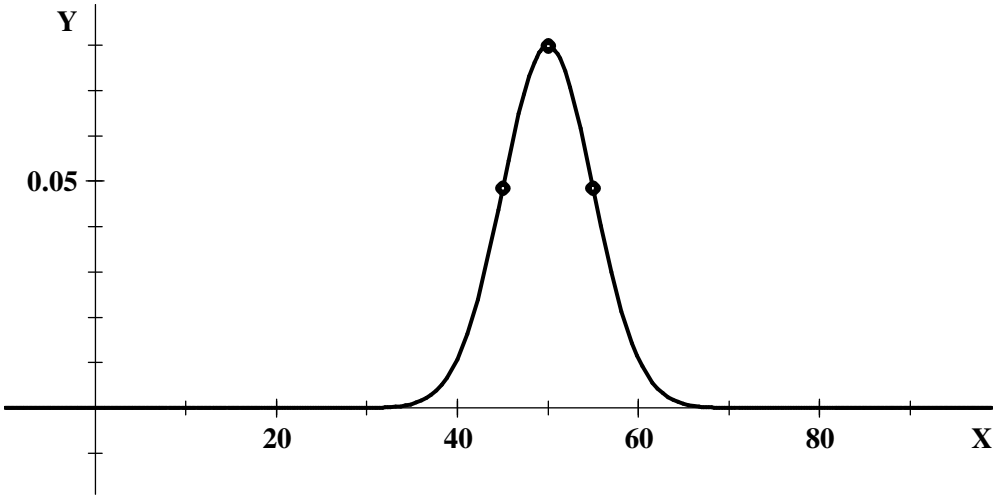
   (De onderstaande schermafdrukken tonen een goed tekenvenster en de bijbehorende grafieken.

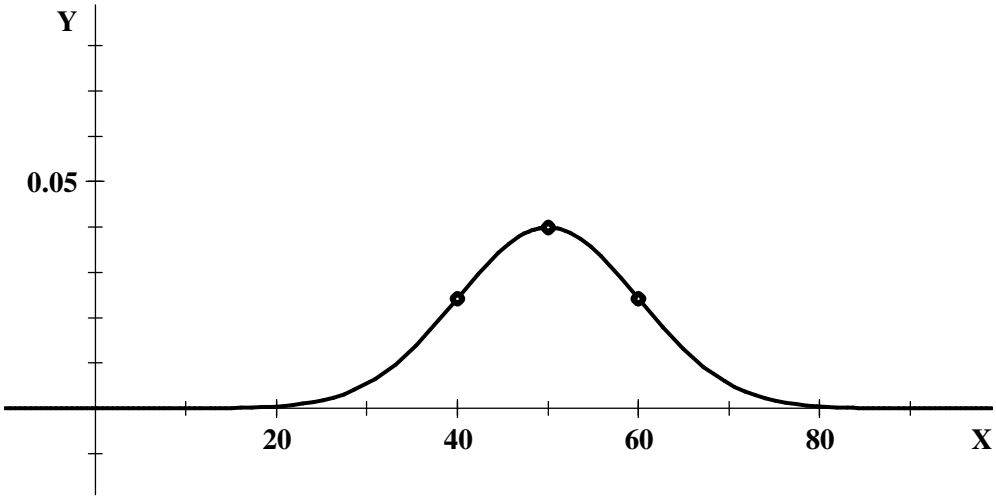


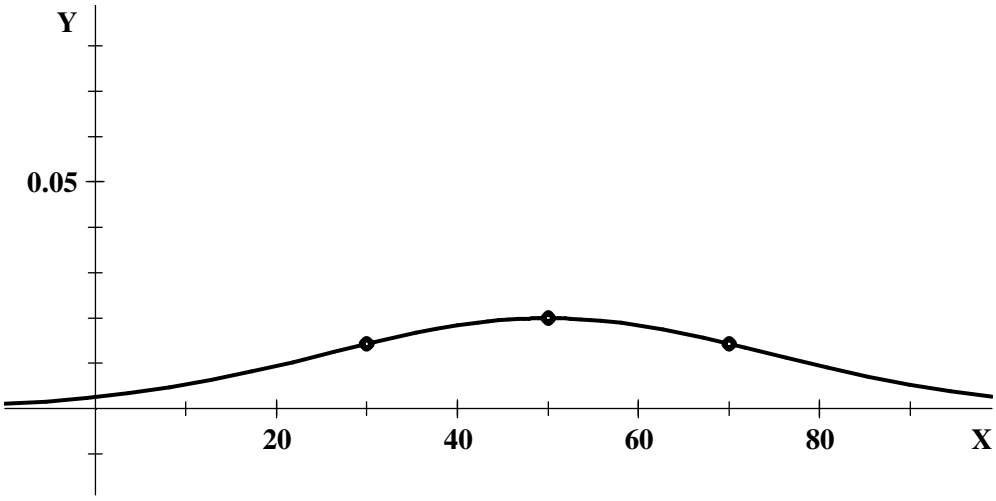
Als de standaardafwijking groter wordt, gebeuren er twee dingen: de grafiek wordt 'breder' en 'platter'. In meer wiskundige termen: de grafiek wordt in de horizontale richting uitgerekt (ten opzichte van de symmetrie-as) en in de verticale richting samengedrukt (ten opzichte van de horizontale as). Op de figuur is duidelijk te zien dat de hoogte van de top omgekeerd evenredig is met de standaardafwijking. Het is niet moeilijk om dit af te leiden uit het voorschrift.)

We hebben er vroeger de aandacht op gevestigd dat de totale oppervlakte onder de kromme gelijk is aan 1. Dat verklaart waarom we zowel horizontaal als verticaal moeten uitrekken/samendrukken. Horizontaal uitrekken vergroot immers de oppervlakte onder de grafiek. Om dat te compenseren moeten we verticaal samendrukken.

5. De onderstaande figuren tonen meer nauwkeurige (computer)grafieken van de drie dichtheidsfuncties uit de vorige opgave. Op elk van de drie grafieken zijn drie punten aangeduid. Probeer in woorden uit te leggen wat er speciaal is aan deze punten en geef telkens de *x*-coördinaat van deze punten. Welk verband bestaat er tussen het voorschrift van de functies (zie vraag 4) en de *x*-coördinaten van de punten?







(Het middelste punt is telkens het hoogste punt van de grafiek. Leerkrachten weten dat de andere twee punten buigpunten van de grafiek zijn. Niet alle leerlingen uit een 3-uurscursus kennen echter dit begrip. Daarom zochten we naar andere omschrijvingen voor deze twee speciale punten. We vonden er verschillende, waaruit je kan kiezen naargelang van de voorkennis van je leerlingen en je eigen smaak. De omschrijving die voor de leerlingen wellicht het duidelijkst is, maakt gebruik van de holle zijde van de grafiek. Het meest linkse punt markeert de overgang tussen een stuk grafiek met holle kant naar boven en een stuk grafiek met holle kant naar onder. Bij het meest rechtse punt is het net andersom. Eigenlijk is dat de definitie van een buigpunt. De tweede omschrijving maakt gebruik van een andere meetkundige eigenschap van een buigpunt: de twee punten zijn de punten waarin de raaklijn aan de grafiek het steilste is. De laatste mogelijkheid is helemaal anders. Die is gebaseerd op de eerste vuistregel uit paragraaf 4: tussen de twee aangeduide punten ligt ongeveer 68% van de oppervlakte. Het is ook leuk om te onthouden dat de hoogte waarop deze punten gelegen zijn ongeveer 60% bedraagt van de hoogte van de top. Voor de eerste grafiek zijn de gevraagde x-coördinaten: 45, 50 en 55; voor de tweede: 40, 50 en 60 en voor de derde: 30, 50 en 70. De x-coördinaat van het hoogste punt is het gemiddelde. Om de x-coördinaten van de andere punten te vinden, moet je de standaardafwijking optellen bij/aftrekken van het gemiddelde.)

6. Probeer nu uit te leggen hoe je de standaardafwijking van een normale dichtheidsfunctie kunt aflezen op de grafiek. Wat is de standaardafwijking van de normale dichtheidsfunctie uit de figuur bij opdracht 3?

   (Je moet letten op de buigpunten van de grafiek (of …) of gebruik maken van de eerste vuistregel. De x-coördinaten van deze punten worden gegeven door het gemiddelde vermeerderd/verminderd met de standaardafwijking. Als je het gemiddelde kent, kan je hiermee de standaardafwijking bepalen. In de figuur is de x-coördinaat van het hoogste punt 140. Het is niet gemakkelijk om de buigpunten exact te localiseren. De x-coördinaten zijn (ongeveer) 115 en 165. Je vindt dus (ongeveer) 25 voor de standaardafwijking.)

In de vorige opgaven hebben we je gevraagd te letten op de *verschillen* tussen de grafieken. Er zijn echter ook opvallende *overeenkomsten*: alle grafieken liggen boven de horizontale as, ze zijn symmetrisch ten opzichte van een verticale rechte door het hoogste punt, … Het voornaamste gemeenschappelijke kenmerk is dat de grafieken van alle normale dichtheidsfuncties *klokvormig* zijn.

De grafieken van alle normale dichtheidsfuncties lijken dus heel goed op elkaar. Bovendien kan je de ene grafiek afleiden uit een andere door verschuiven en horizontaal/verticaal uitrekken/samendrukken. Daarom wordt één van de normale dichtheidsfuncties de *standaardnormale dichtheidsfunctie* genoemd, namelijk de normale dichtheidsfunctie met gemiddelde 0 en standaardafwijking 1. Alle andere normale dichtheidsfuncties kunnen hiervan afgeleid worden. Als je de standaardnormale dichtheidsfunctie met je rekenmachine tekent, hoef

je geen parameters op te geven. In plaats van 'normalpdf(X,0,1)' kan je dus gewoon 'normalpdf(X)' ingeven.

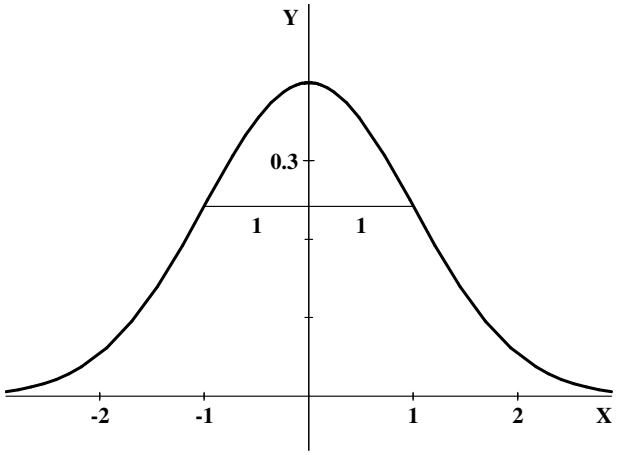
7. Teken deze dichtheidsfunctie met je rekenmachine.
8. Leg uit hoe je de grafiek van de normale dichtheidsfunctie met gemiddelde μ en standaardafwijking σ kunt afleiden uit de grafiek van de standaardnormale dichtheidsfunctie.

   (Je moet de grafiek de goede vorm geven door horizontaal/verticaal uitrekken/samendrukken. Dit wordt aangegeven door de standaardafwijking σ. Je moet de grafiek ook horizontaal verschuiven zo dat het hoogste punt het gemiddelde μ als x-coördinaat krijgt.)

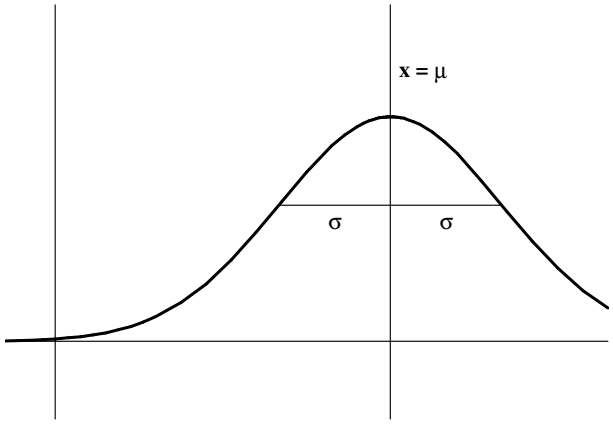
Als je gebruik maakt van het volledig uitgeschreven voorschrift (bijvoorbeeld in een 6- of 8-urencursus), kan je natuurlijk verder gaan dan wat in de bovenstaande opgaven gevraagd wordt. Je kan dan bijvoorbeeld vragen om de vaststellingen ook te verklaren m.b.v. de vergelijking van de functie of je kan het buigpunt exact localiseren m.b.v. de tweede afgeleide.

De tekst in het onderstaande kader vat samen wat we uit de werktekst geleerd hebben.

Er zijn oneindig veel normale dichtheidsfuncties. De *standaardnormale dichtheidsfunctie* heeft gemiddelde 0 en standaardafwijking 1. De grafieken van de andere normale dichtheidsfuncties kunnen afgeleid worden uit de grafiek van de standaardnormale dichtheidsfunctie door die horizontaal te verschuiven en tegelijk horizontaal en verticaal uit te rekken / samen te drukken. Alle normale dichtheidsfuncties hebben een *klokvormige grafiek*. Het gemiddelde μ is de *x*-coördinaat van het hoogste punt van de grafiek. De punten met *x*-coördinaat $\mu \pm \sigma$ markeren de overgang tussen een gedeelte van de grafiek met holle zijde naar boven en een gedeelte met holle zijde naar onder. Een grote waarde voor σ betekent dat de grafiek 'plat' en 'breed' is en een kleine waarde voor σ betekent dat de grafiek 'hoog' en 'smal' is.



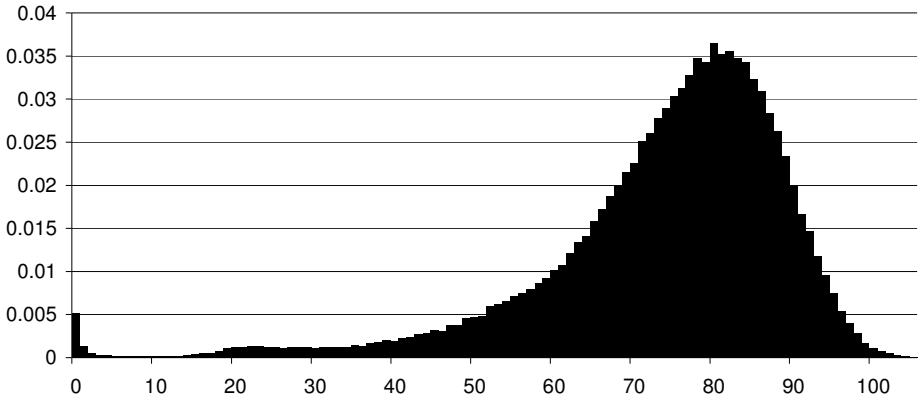
standaardnormale dichtheidsfunctie



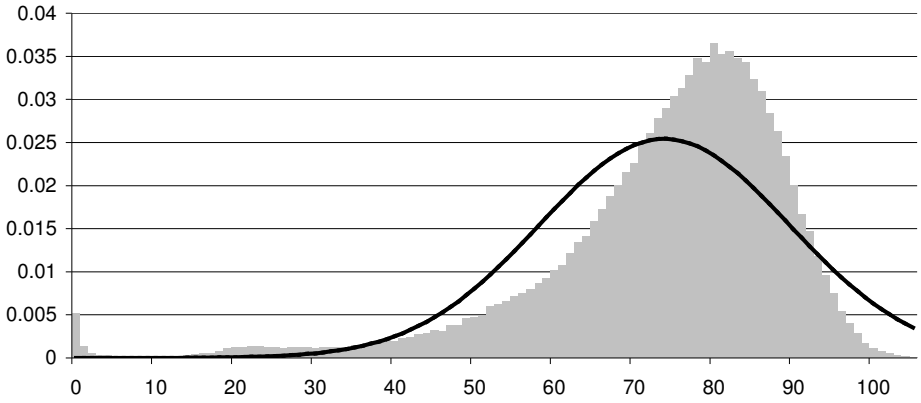
normale dichtheidsfunctie met gemiddelde μ en standaardafwijking σ

7. Niet alle gegevens zijn normaal verdeeld!

Het onderstaande histogram geeft informatie over de leeftijd van Belgische mannen bij overlijden (gebaseerd op de sterftetafel 1995-1997 op p. 43 uit Uitwiskeling 15/3).



We proberen ook hier om het histogram te modelleren m.b.v. een normale dichtheidsfunctie. Het gemiddelde is 73,80 en de standaardafwijking is 15,69. De volgende figuur toont dat het duidelijk mis gaat.



We hadden dit vooraf kunnen verwachten: het histogram is hier helemaal niet symmetrisch. Het heeft een 'staart naar links' en daarom zeggen we dat de gegevens linksscheef verdeeld zijn. Je kunt de staart naar links hier als volgt verklaren. Er zijn meer mannen die sterven vóór de modale leeftijd bij overlijden (ongeveer 80) dan omgekeerd. Bovendien kan het verschil met de modale leeftijd grotere waarden aannemen als je jonger sterft dan als je ouder sterft (20 jaar langer leven is reeds zeldzaam, 40 jaar minder lang leven is veel minder zeldzaam).

We merken in dit voorbeeld dat niet alle gegevens normaal verdeeld zijn. Nog heel wat andere gegevens zijn links- of rechtsscheef en kunnen daarom niet goed door een normale verdeling weergegeven worden. Zo zou het ook niet gelukt zijn als we gewerkt hadden met het gewicht i.p.v. de lengte van volwassen vrouwen. Het histogram van de gewichten heeft een 'staart naar rechts' (zie bv. [2, p. 9]). Ook inkomens zijn traditioneel sterk rechtsscheef verdeeld (leerlingen kunnen dit zelf uitleggen). Zwangerschapsduur is dan weer een ander voorbeeld van een linksscheve verdeling (gegevens vind je bijvoorbeeld in [5, p. 30]).

Gebrek aan symmetrie is overigens niet de enige reden waarom het mis kan lopen. Het histogram van het aantal ogen bij heel veel worpen met een dobbelsteen is bijvoorbeeld wel symmetrisch maar het wordt het best gemodelleerd door een horizontaal lijnstuk en niet door een 'klokvormige' kromme. Ook de som van het aantal ogen bij heel veel worpen met twee dobbelstenen zal symmetrisch zijn, maar zal niet de vereiste 'klokvorm' vertonen. Het histogram wordt dan het best gemodelleerd door een lijnstuk met helling $\frac{1}{6}$ gevolgd door een lijnstuk met helling $-\frac{1}{6}$.

Sommige gegevens geven wel aanleiding tot een 'bekende' dichtheidsfunctie die echter niet normaal is. Veronderstel bijvoorbeeld dat we aan een tankstation telkens de tijd noteren die verstrijkt tussen het aankomen van een klant en de volgende. Deze gegevens kunnen benaderd worden door een exponentiële verdeling. Het aantal tweelingen dat jaarlijks geboren wordt in een bepaalde materniteit volgt ongeveer een Poissonverdeling. Het aantal ogen bij het gooien met één dobbelsteen geeft aanleiding tot een uniforme verdeling. Sommige van deze andere dichtheidsfuncties vind je ook op de rekenmachine.

Omdat gegevens niet automatisch normaal verdeeld zijn, is het nodig om dat te onderzoeken. Je zult met je leerlingen dus een aantal oefeningen moeten maken waarin gevraagd wordt of het gebruik van een normale dichtheidsfunctie al dan niet gerechtvaardigd is. We hebben er in paragraaf 2 al op gewezen dat van sommige gegevens geweten is dat ze tot een normale verdeling aanleiding geven (lengte van volwassenen, IQ, meetfouten, …). In andere gevallen kan je het histogram beoordelen. In een aantal gevallen is er een eenvoudige verklaring te geven voor het feit dat de gegevens niet normaal verdeeld zijn. Bij zwangerschapsduur is het bijvoorbeeld niet moeilijk om in te zien dat de gegevens linksscheef verdeeld zullen zijn. De meeste zwangerschappen duren ongeveer 40 weken. Veel langer dan 40 weken kan een zwangerschap niet duren omdat die anders kunstmatig afgebroken wordt. Er zijn daarentegen wel zwangerschappen die veel korter duren dan 40 weken. Ook je leerlingen kunnen dergelijke verklaringen vinden. Het is dus de moeite om er hen naar te vragen.

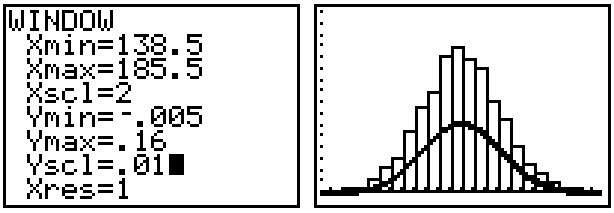
Een meer formele manier om na te gaan of gegevens normaal verdeeld zijn, bestaat er in gebruik te maken van zogenaamd normaal waarschijnlijkheidspapier (zie bv. [2]). Ook dat kan ondertussen m.b.v. de rekenmachine. Het valt echter buiten het bestek van deze tekst om dit verder uit te diepen.

Het onderstaande kadertje vat samen wat we in deze paragraaf geleerd hebben.

> Niet alle gegevens zijn normaal verdeeld. Het histogram is dan bijvoorbeeld onvoldoende symmetrisch of wel symmetrisch, maar niet klokvormig. Vóór we een normale dichtheidsfunctie gebruiken, moeten we dus nagaan of de gegevens wel (bij benadering) normaal verdeeld zijn. We doen dit door het te beoordelen m.b.v. het histogram. Van sommige gegevens is ook algemeen bekend dat ze normaal verdeeld zijn (lengte van volwassenen, IQ, meetfouten, …).

8. Relatieve frequentiedichtheid

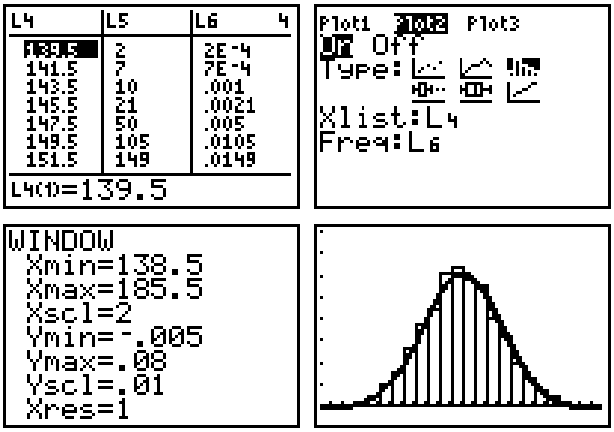
We keren terug naar het voorbeeld met de lengte van 5000 volwassen vrouwen uit 1947. We stelden bij dit voorbeeld vast dat het histogram heel goed weergegeven kon worden door een normale dichtheidsfunctie. We werkten met klassenbreedte 1 en veranderen de klassenbreedte nu in 2. Op de rekenmachine gaat dat heel eenvoudig: in het WINDOW-scherm geven we de variabel Xscl de waarde 2 i.p.v. 1 (we passen ook Ymax aan). De onderstaande schermafdruk toont dat de dichtheidsfunctie nu niet langer overeenstemt met het histogram.



Het is niet moeilijk om te begrijpen wat er gebeurt. Neem bijvoorbeeld de klasse [154,5; 156,5[. De (relatieve) frequentie van zo'n nieuwe klasse is de som van de (relatieve) frequenties van de klassen [154,5; 155,5[en [155,5; 156,5[. De hoogte van de rechthoek in het nieuwe histogram is de som van de hoogten van twee rechthoeken uit het vorige histogram. De rechthoeken zijn dus ongeveer twee keer zo lang geworden (dat is de reden waarom we Ymax moesten aanpassen). De dichtheidsfunctie is echter niet veranderd.

De oplossing bestaat er in de relatieve frequenties te delen door de klassenbreedte. Zo krijg je de relatieve frequentie*dichtheden* van de klassen. De dichtheidsfunctie is in feite geen wiskundig model voor de relatieve frequenties maar voor de relatieve frequentiedichtheden (ze heet niet voor niets een *dichtheids*functie!).

We tekenen het histogram van de relatieve frequentiedichtheden nu m.b.v. de rekenmachine. We moeten de nieuwe klassenmiddens en hun frequenties eerst ingeven. Dat doen we in de lijsten L_4 en L_5. In L_6 vind je de relatieve frequentiedichtheden. De schermafdruk rechts onderaan toont ons dat histogram en dichtheidsfunctie nu weer met elkaar overeenstemmen.



De *hoogten* van de rechthoeken in het histogram zijn relatieve frequentiedichtheden. De *oppervlakten* van de rechthoeken zijn nog altijd relatieve frequenties (relatieve frequentiedichtheid maal klassenbreedte!). In de voorgaande paragrafen waren de hoogten en de oppervlakten aan elkaar gelijk omdat de klassenbreedte steeds 1 was.

In de voorbeelden uit de vorige paragrafen hebben we steeds gewerkt met klassen van breedte 1 en daarom trad het probleem niet eerder op. De voorbeelden uit de vorige paragrafen zijn dus wiskundig volledig correct behandeld. We hebben ons alleen moeten houden aan de beperking van klassenbreedte 1. Voor leerlingen uit een 3-uurscursus zouden we ons ook aan die beperking houden. Het begrip relatieve frequentiedichtheid zouden we bij deze leerlingen dus niet introduceren. Uiteindelijk moet bij deze leerlingen de klemtoon liggen op het aanbrengen van de normale dichtheidsfunctie zonder te technisch te

worden. Voor leerlingen uit een 6-urencursus (en voor jullie, leerkrachten) liggen de zaken anders. Daarom hebben we deze paragraaf in onze tekst opgenomen.

Een dichtheidsfunctie is in feite een wiskundig model voor de relatieve frequentie*dichtheden* en niet voor de relatieve frequenties. Als de klassenbreedte gelijk is aan 1, is de relatieve frequentiedichtheid gelijk aan de relatieve frequentie en kunnen dichtheidsfuncties dus toch gebruikt worden als model voor relatieve frequenties.

9. Normale verdelingen vergelijken

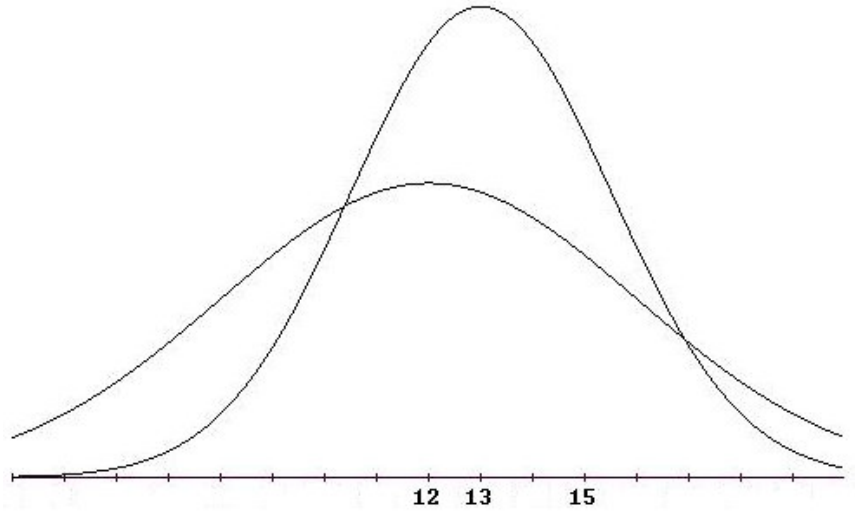
De normale verdeling (evenals andere verdelingen) kan gezien worden als een wiskundig model voor statistische gegevens. Die worden samengevat in één grafiek. Door verschillende statistische gegevens op dezelfde manier te modelleren, kunnen ze vergeleken worden. De vorm van de grafieken geeft onmiddellijk heel wat informatie over de verdeling van de data. In deze paragraaf gaan we daar dieper op in.

In de volgende opgave maken de leerlingen kennis met twee methoden om data van twee reeksen normaal verdeelde gegevens te vergelijken. De cijfers zijn spijtig genoeg fictief, voorlopig krijgen we de juiste cijfers niet te pakken.

Examenresultaten vergelijken

Aan het ingangsexamen geneeskunde en tandheelkunde hebben 450 kandidaten meegedaan. Dit examen bestaat uit twee delen: een wetenschappelijk deel (wiskunde, chemie, fysica en biologie) en een niet-wetenschappelijk deel. Uit onderzoek is gebleken dat de resultaten van beide delen normaal verdeeld zijn. Stel dat voor het wetenschappelijke deel het gemiddelde 12 was en de standaardafwijking 4 en dat voor het andere deel het gemiddelde 13 was met standaardafwijking 2,5. Hans heeft een 15 op beide delen. Welk deel heeft hij (in vergelijking met de andere kandidaten) het beste afgelegd?

1. Schets de normale verdeling bij beide resultaten.
2. Duid op beide tekeningen de resultaten van Hans aan.



3. Hoe zou je de resultaten op de twee delen kunnen vergelijken?

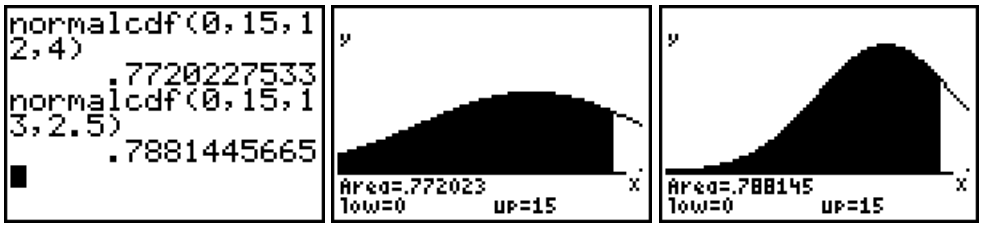
   (Mogelijke antwoorden zijn: door de afstanden van de resultaten van Hans tot de gemiddelden te vergelijken, door deze afstanden uit te drukken in aantal maal de bijbehorende standaardafwijking, door voor beide proeven na te gaan hoeveel procent van de deelnemers minder goed presteerden dan Hans, …)
4. Je kan bv. kijken naar de afwijking van het resultaat van Hans t.o.v. het gemiddelde.
   a. Welke test heeft Hans volgens dit criterium het beste gemaakt?
   b. Wat vind je van deze manier om de resultaten te vergelijken? Waar hou je bv. geen rekening mee?

(Met het feit dat de resultaten bij wetenschappen veel meer gespreid zijn.)

5. Een mogelijkheid om ook met de standaardafwijking rekening te houden bij het vergelijken van beide resultaten is de afstand tussen het resultaat van Hans en het gemiddelde te vergelijken met de standaardafwijking. Doe dit.

   (Voor het wetenschappelijk deel: $15 - 12 = 3$ *en dit is* $\frac{3}{4}$ *van de standaardafwijking; voor het andere deel:* $15 - 13 = 2$ *en dit is* $\frac{2}{2{,}5} = \frac{4}{5}$ *van de standaardafwijking. Zo bekeken heeft Hans het niet-wetenschappelijke deel iets beter gedaan dan het wetenschappelijke deel: zijn resultaat voor dit deel wijkt een groter deel van de standaardafwijking af van het gemiddelde.)*

6. Je kan ook bij beide verdelingen het percentage leerlingen berekenen dat een lagere score dan Hans heeft. Doe dit.



7. Voor welk onderdeel heeft Hans het beste gepresteerd volgens het criterium uit vraag 6?

   (Hans heeft ook nu voor het niet-wetenschappelijke deel een heel klein beetje beter gepresteerd: bij het wetenschappelijke deel heeft volgens het model ongeveer 77% van de deelnemers minder goed gepresteerd dan Hans, voor het andere deel net geen 79%.)

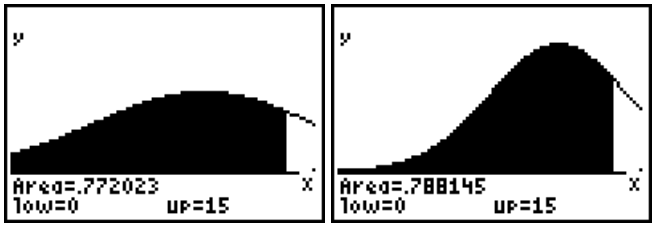
In de bovenstaande opgave vergeleken we op twee manieren twee normaal verdeelde variabelen.

- In vraag 5 vergeleken we de afwijkingen t.o.v. het gemiddelde met de standaardafwijking. De getallen die je zo berekende, noemen we de *z-scores*:

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

  Voor het wetenschappelijke deel is de *z-score* van het resultaat van Hans dus $z = \frac{15 - 12}{4} = \frac{3}{4} = 0{,}75$.

- Een andere manier is het aandeel kleinere resultaten te vergelijken. Dit deden we in opgave 6. Dit komt neer op de vergelijking van de oppervlaktes onder beide normale verdelingen links van de te vergelijken resultaten.



Bij normaal verdeelde variabelen en het bijbehorende wiskundige model zal je bij beide criteria steeds tot dezelfde conclusie komen. (Als je de criteria toepast op de gegevens zelf (en dus niet het wiskundige model), kom je uiteraard niet altijd tot dezelfde conclusie.)

We geven nog een tweede oefening hierbij. De gegevens over de lengte van 18-jarige Belgische mannen in 1950 en in 2000 zijn afkomstig van het Belgische leger. We danken hiervoor Med. Maj. Goossens, C Med Medische Basisselectie (Centrum voor Medische expertise, Militair Hospitaal Brussel).

Ben ik groter dan mijn grootvader?

Zoals je weet worden mensen almaar langer. Hiermee bedoelt men niet dat ieder van ons steeds blijft groeien maar dat de gemiddelde lengte, bv. van alle Belgen, gedurende de laatste eeuw(en) toegenomen is. In 1950 was een 18-jarige van 1m80 echt groot terwijl er nu in je klas vermoedelijk wel enkele jongens groter dan 1m80 zijn. We willen in deze opgave toch proberen individuele lengten nu en vroeger te vergelijken.

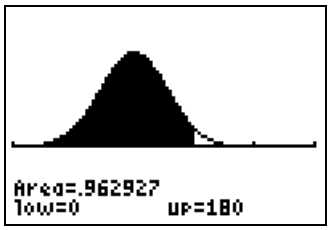
Jeroens grootvader was 18 jaar in 1950 en hij was 1m80 groot. In 2000 was Jeroen 18 jaar en 1m90 groot. Uiteraard is Jeroen groter dan zijn grootvader. Maar hoe zit dat in vergelijking met de rest van de bevolking?

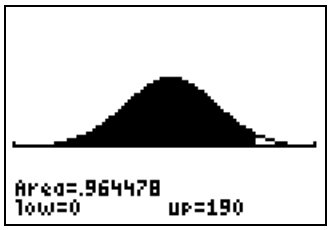
Het Belgische leger houdt veel statistieken bij van de militairen (en vroeger dus van bijna alle ongeveer 18-jarige jongens). De lengte van jongens (en meisjes!) op een bepaalde leeftijd is normaal verdeeld. In 1950 was de gemiddelde lengte van 18-jarige jongens 170,0 cm en de standaardafwijking 5,6. In 2000 was het gemiddelde 176,1 cm en de standaardafwijking 7,7.

1. Vergelijk beide lengten met behulp van de *z*-scores. Duid je resultaten aan op schetsen van de normale verdelingen.

$$\left(\frac{180-170}{5,6}=1,7857 \; en \; \frac{190-176,1}{7,7}=1,8052\right)$$

2. Vergelijk beide lengten ook door het percentage kleinere personen te berekenen. Duid ook nu je resultaten aan op schetsen van de normale verdelingen.





Uiteraard is het niet zo belangrijk of Jeroen groter is dan zijn grootvader in vergelijking tot hun respectievelijke leeftijdsgenoten. Maar de gemiddelde lengte verschilt ook van land tot land. O.a. voor de kledingindustrie is het belangrijk om te weten hoe deze lengten van land tot land verdeeld zijn om de ontwerpen en de te produceren maten daaraan aan te passen. Dit alles biedt voldoende stof voor een groter *statistisch project*. Op de site www.tallpages.com/nl/index vind je hiervoor bij 'Statistieken' o.a. gegevens over de gemiddelde lengte en de standaardafwijking in een aantal landen. Deze site zou ook als start van de studie van de normale verdeling gebruikt kunnen worden. Heel wat aspecten van de normale verdeling worden op een heel eenvoudige manier in het kader van de lengte van mensen aangehaald.

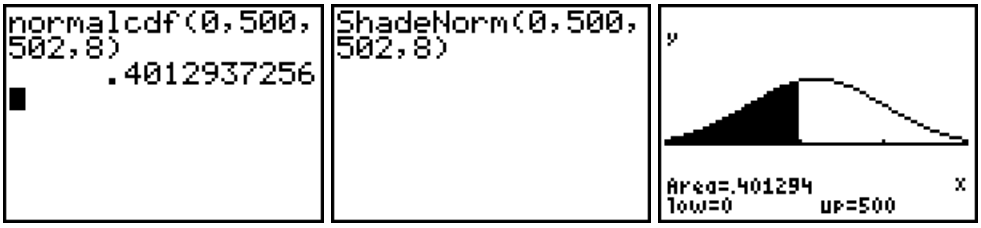
10. Toepassing: de machine goed instellen

De volgende toepassing gaat verder dan wat de eindtermen voorschrijven. We zien ze dan ook eerder als uitbreiding.

Dozen erwten vullen

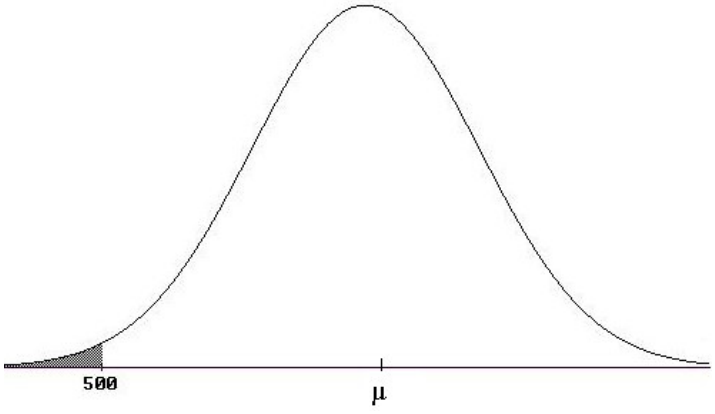
Sommige klanten van een warenhuis hebben het vermoeden dat de dozen erwten van 500 gram van een bepaald merk te weinig wegen. Iemand beweert zelfs dat een vijfde van de pakken minder dan 500 gram weegt. Een verbruikersorganisatie neemt een steekproef van 1000 pakken. Het gemiddeld gewicht blijkt 502 gram te zijn. De standaardafwijking bedraagt 8 gram. We mogen aannemen dat het gewicht van de dozen erwten normaal verdeeld is.

1. Maak een grafiek van de normale verdeling van het gewicht van de dozen erwten.
2. Hoeveel procent van de pakken uit de steekproef heeft een gewicht van minder dan 500 gram?
   (Via een berekening en grafisch zien we dat dit een groot deel is: 40%.)



De ondernemer die de erwten verpakt, wil geen nieuwe klacht. Hij wil dat hoogstens 1% van de pakken te weinig weegt. Hij weet dat de vulmachine een gewicht aflevert dat normaal verdeeld is met een standaardafwijking van 8 gram. Het gewicht dat aangeduid wordt als vulgewicht is ook ongeveer het gemiddelde gewicht van de gevulde dozen. De vraag is op welk gewicht de ondernemer de machine moet afstellen opdat slechts 1% van de dozen een gewicht zou hebben beneden de 500 gram.

We moeten dus μ zoeken zo dat de gearceerde oppervlakte in de onderstaande figuur 0,01 is.

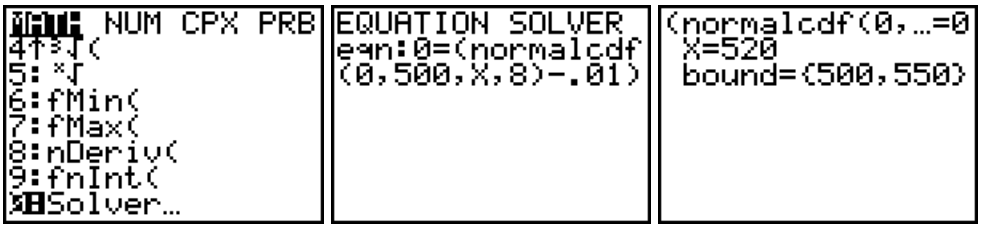


Hiervoor is er geen commando op het rekentoestel. Bij de commando's voor de normale verdeling heb je steeds het gemiddelde nodig. Met de 'solver' op je rekentoestel kunnen we μ wel vinden.

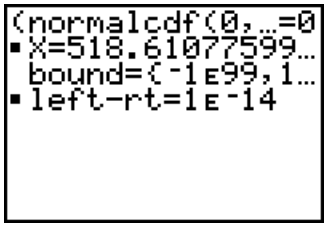
3. Schrijf de voorwaarde waaraan μ moet voldoen eens op met de notaties van je rekentoestel.
   (normalcdf(0, 500, μ, 8) = 0.01)

Om deze vergelijking door de rekenmachine te laten oplossen, kies je '0:Solver' onder de toets [MATH]. Deze opdracht kan enkel vergelijkingen met rechterlid 0 oplossen. Voer de juiste vergelijking in (kies *X* i.p.v. μ als onbekende) en druk vervolgens op [ENTER]. In het schermpje

dat dan verschijnt, vul je een gok voor de oplossing in bij *X* en pas je bij 'bound' de grenzen waarbinnen het rekentoestel naar een oplossing moet zoeken eventueel aan. (Hier weten we bv. dat de oplossing die we zoeken zeker groter is dan 500 en als bovengrens kunnen we bv. 550 nemen.)



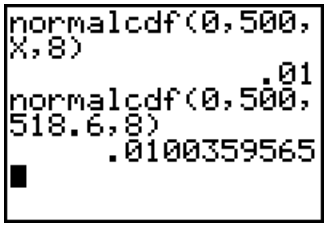
Zet de cursor vervolgens terug bij het getal dat je bij *X* invulde en druk op [ALPHA] [SOLVE].



Zo vinden we de oplossing van de vergelijking en dus het gevraagde gemiddelde.

4. Ga na dat bij dit gemiddelde slecht 1% van de dozen een gewicht heeft beneden de 500 gram.

Merk op dat als je dit gemiddelde afrondt, je uiteraard niet juist 0,01 vindt. Je kan met het niet-afgeronde getal verder werken omdat het rekentoestel dit automatisch opslaat in *X*. Je ziet dit in het onderstaande schermpje.



Op dezelfde manier kunnen problemen waarbij wel het gemiddelde, maar niet de standaardafwijking bekend is, opgelost worden. We maakten in deze werktekst gebruik van de mogelijkheden van een grafisch rekentoestel om vergelijkingen (numeriek) op te lossen. Op deze manier kunnen we zulke problemen aanpakken zonder de transformatieformules voor de omzetting naar de standaardnormale verdeling uitvoerig te behandelen. De essentie kan zo gaan naar de statistiek, het interpreteren van het statistisch probleem.

11. Historische noot

De ontdekking van de normale verdeling wordt toegeschreven aan *Abraham de Moivre* (1667-1754). Deze Fransman, die later uitweek naar Engeland en bevriend was met Newton, voorzag in zijn levensonderhoud door voor gegoede burgers hun winstkansen bij kansspelen te berekenen. In 1718 publiceerde hij zijn werk 'Doctrine of Changes' over kansberekeningen bij kansspelen. Later, in 1733, publiceerde hij een artikel waarin hij de binomiale verdeling benaderde door een vloeiende kromme. Deze kromme werd later de normale verdeling genoemd. Het artikel bevat ook reeds de vuistregels: 2/3 van de waarnemingen wijkt niet meer dan één standaardafwijking af van het gemiddelde en 95% niet meer dan twee standaardafwijkingen.

Het artikel van de Moivre bleef onopgemerkt tot Karl Pearson het in 1924 herontdekte. Intussen werd de normale verdeling wel herontdekt door Laplace en Gauss. *Pierre Simon Laplace* (1749-1827) gebruikte de normale verdeling in 1783 om de *verdeling van meetfouten* te beschrijven. Later (in 1809) gebruikte *Carl Friedrich Gauss* (1777-1855) de normale verdeling ook om gegevens uit de astronomie te analyseren. Ook hij kwam tot de bevinding dat meetfouten normaal verdeeld zijn.

De eerste persoon die de normale verdeling toepaste op sociale gegevens was de Belg *Adolph Quetelet* (1796-1874). Hij verzamelde gegevens over de borstomvang van Schotse soldaten en de lengte van Franse soldaten. Hij constateerde dat beide normaal verdeeld waren. Aanvankelijk hield Quetelet zich bezig met sterrenkunde. Via de meetfoutentheorie maakte hij kennis met de normale verdeling. Hij was ervan overtuigd dat niet alleen meetfouten ontstonden als gevolg van het toeval, maar dat heel wat aspecten van het menselijk leven ook bepaald worden door het toeval. Quetelet definieerde de "gemiddelde mens", een begrip dat op heel wat weerstand stuitte. Hij verzamelde en analyseerde in opdracht van de staat o.a. misdaadcijfers en sterftecijfers. Hij kan beschouwd worden als de vader van de sociale wetenschappen.

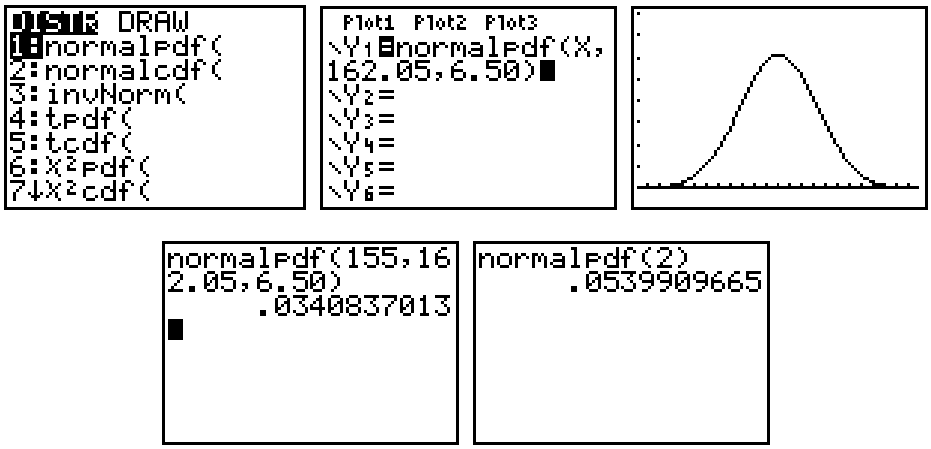
Francis Galton (1822-1911), een neef van Charles Darwin, was geen wiskundige maar een wetenschapper. Hij was niet zoals Quetelet geïnteresseerd in het gemiddelde, maar juist in de afwijkingen van het gemiddelde. Hij onderbouwde zijn onderzoek over erfelijkheid door nieuwe statistische begrippen toe te passen. Hij was niet de eerste die dit deed, maar zijn werk had wellicht het meeste invloed. Hij construeerde het naar hem genoemde bord om tijdens lezingen te laten zien hoe een reeks opeenvolgende toevallige gebeurtenissen (naar links of naar rechts vallen) tot een normale verdeling leiden. Hij introduceerde de begrippen regressie en correlatiecoëfficiënt. Hij was een man uit de praktijk die deze begrippen vorm en inhoud gaf. Hij was echter geen theoreticus die deze begrippen in een groter geheel kon plaatsen. Dit laatste was het werk van Karl Pearson.

Ten slotte is ook de bijdrage van *Florence Nightingale* (1820-1910) een vermelding waard. Zij is het best gekend om haar verplegend werk tijdens de Krimoorlog. Anderzijds was zij ook een uitstekende wiskundige die sterk beïnvloed was door het werk van Quetelet. Zij verzamelde gegevens over de doodsoorzaak van de soldaten. Hieruit bleek dat er meer soldaten stierven aan infecties opgedaan in de ziekenboeg dan aan hun verwondingen van op het slagveld. Zo kon zij, gebruik makend van haar statistische gegevens en voorstellingen, de autoriteiten ervan overtuigen te investeren in hygiënische hervormingen in de militaire hospitalen.

12. Commando's i.v.m. de normale verdeling op de TI-83: samenvatting

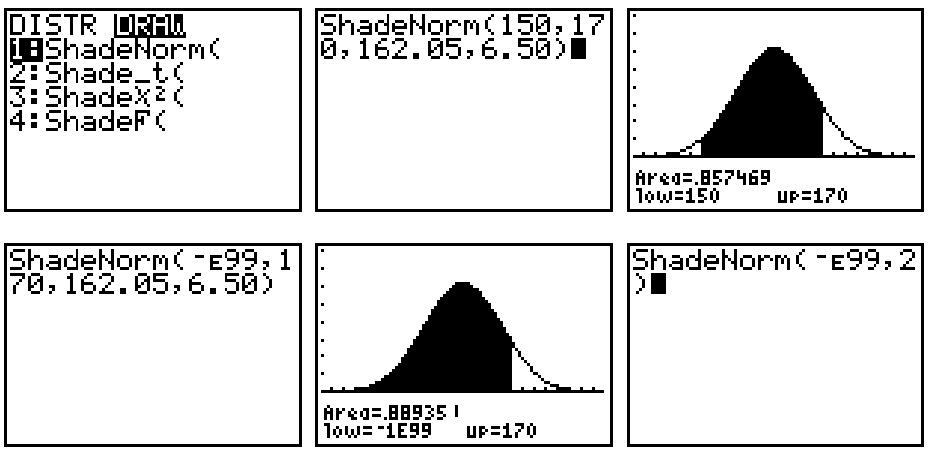
normalpdf

Voluit is de naam van deze functie: normal probability density function, en in het Nederlands: normale kansdichtheidsfunctie. Je kunt deze functie plakken in het basisscherm (voor berekeningen) en in het [Y=]-scherm (om de grafiek te tekenen) via [2nd] [DISTR]. In het [Y=]-scherm moet je na het opengaande haakje eerst de veranderlijke *X* toevoegen en vervolgens het gemiddelde en de standaardafwijking. In het basisscherm moet je in de plaats van de veranderlijke *X* een getal invullen. Als je met de standaardnormale dichtheidsfunctie werkt, hoef je het gemiddelde en de standaardafwijking niet in te voeren.



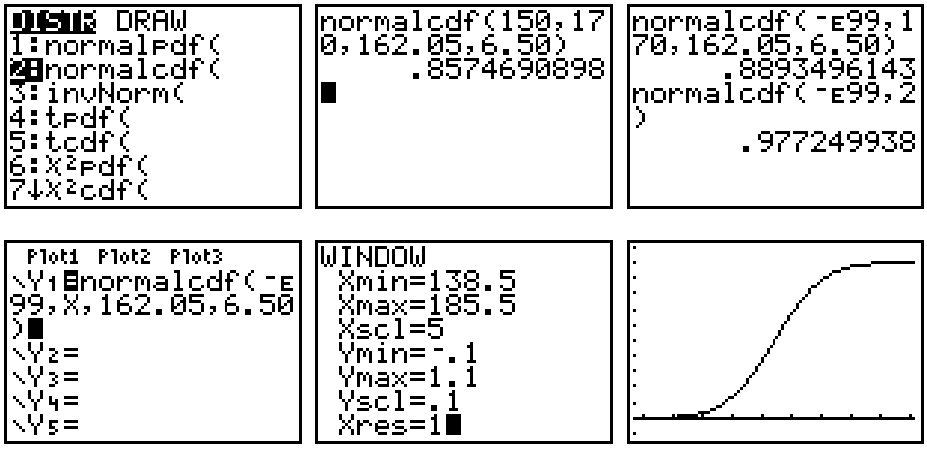
shadeNorm

Het commando shadeNorm gebruikten we in deze tekst enkel in het basisscherm. Je kunt dit commando plakken in het basisscherm via [2nd] [DISTR] [DRAW]. Er zijn vier argumenten: eerst twee getallen (de grenzen) en vervolgens het gemiddelde en de standaardafwijking. Als je het commando uitvoert, wordt een grafiek gemaakt (in het tekenvenster dat op dat ogenblik van kracht is; vergeet het dus niet aan te passen!) van de normale dichtheidsfunctie met de opgegeven parameters en de oppervlakte onder de dichtheidsfunctie tussen de opgegeven twee getallen wordt berekend en gearceerd. Vaak heb je min oneindig als ondergrens nodig. Dan moet je voor de ondergrens -10^{99} invullen. Voor de standaardnormale hoef je gemiddelde en standaardafwijking niet in te vullen.



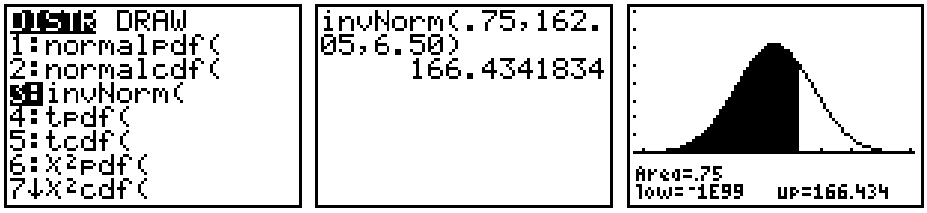
normalcdf

Voluit is de naam van deze functie: normal cumulative density function. We hebben ze alleen in het basisscherm gebruikt. Je kunt deze functie plakken vanuit [2nd] [DISTR]. Na het opengaande haakje moeten twee getallen volgen en verder het gemiddelde en de standaardafwijking. Deze functie berekent de oppervlakte onder de grafiek van de normale dichtheidsfunctie tussen de twee opgegeven grenzen. De functie heeft dus dezelfde werking als het commando shadeNorm, maar tekent er geen figuur bij. Als je deze functie in het [Y=]-scherm gebruikt en min oneindig als ondergrens invult, krijg je de verdelingsfunctie van de normale verdeling, maar die hebben we in deze tekst niet gebruikt.



invNorm

Deze functie is de inverse van de verdelingsfunctie (normalcdf of shadeNorm met ondergrens min oneindig). Voor een getal tussen 0 en 1 berekent ze met andere woorden de grens waarvoor de oppervlakte links ervan gelijk is aan het opgegeven getal. Dat kan je vaststellen in de rechtse twee schermafdrukken. Je kan de naam van deze functie plakken vanuit [2nd] [DISTR]. Na het opengaande haakje moet een getal komen dat tussen 0 en 1 ligt, gevolgd door het gemiddelde en de standaardafwijking. Vul je niets in voor het gemiddelde en de standaardafwijking, dan wordt de standaardnormale verdeling gebruikt.



Bibliografie

[1] A. Bakker, *Historical and didactical phenomenolgy of the average values*, Histoire et épistémologie dans l'éducation mathématique, Proceedings I (Louvain-la-Neuve-Leuven), 2001.

[2] J. de Langhe, M. Kindt, *De normale verdeling*, Educaboek (Culemborg), 1986, ISBN 90-11-010558.

[3] M. Doorman, P. Drijvers, *De TI-83 en TI-83+, kennismaken en toepassen*, Wolters-Noordhoff (Groningen), 2001, ISBN 90-01-832-88-1.

[4] G. Herweyers en K. Stulens, *Statistiek met een grafisch rekentoestel*, Acco (Leuven), 2000.

[5] M.J.B.T. Janssens, G. Nieuwenhuis, *Opgaven bij statistiek in de economie*, Academic Service (Schoonhoven), 1993, ISBN 90-5261-279.

[6] D.S. Moore, G.P. McCabe, *Statistiek in de praktijk*, Academic Service (Schoonhoven), 1994, ISBN 90-395-0576-4.

[7] H. Staal e.a., *Pascal, Wiskunde voor de tweede fase*, Thieme (Zutphen), 1998.

Enkele sites:

http://www.tallpages.com/nl/index

http://www.tld.jcu.edu.au/hist/stats/

http://www.mrs.umn.edu/~sungurea/introstat/history/w98/Quetelet.html

http://www.mugu.com/galton/statistician.html

http://www-groups.dcs.st-andrews.ac.uk/~history/Mathematicians/

Een aantal van de lijsten voor de TI83 grafische rekenmachine kun je downloaden van de site van Uitwiskeling:

www.uitwiskeling.be