Normaal-waarschijnlijkheidspapier

Als een variabele normaal verdeeld is, zal het juist uitzetten van de gegevens op normaal waarschijnlijkheidspapier een rechte lijn opleveren. Vanuit die rechte lijn kunnen dan het gemiddelde μ en de standaardafwijking σ worden bepaald.

Voorbeeld:

In de tabel staan de gewichten van 1000 willekeurig gekozen vrouwen in de leeftijdsklasse van 20 tot 25 jaar.

Onderzoek of de gegevens normaal verdeeld zijn.

Indien het een normale verdeling is, bepaal dan het gemiddelde en de standaardafwijking bij de gegevens.

Gewicht	frequentie
35-<45	25
45-<55	100
55-<65	200
65-<75	325
75-<85	250
85-<95	75
95-<105	25
totaal	1000

Om dit onderzoek uit te voeren moet je de volgende stappen zetten.

- De relatieve cumulatieve frequentie bepalen.
- Klassenmiddens bepalen.
- Gevonden waarden uitzetten op normaal waarschijnlijkheidspapier.
- Nagaan of de punten op een rechte lijn liggen.
- Is het een rechte lijn dan :
  1. aflezen bij 50% voor het gemiddelde.
  2. aflezen bij 16 of 84 % om de standaardafwijking te bepalen.

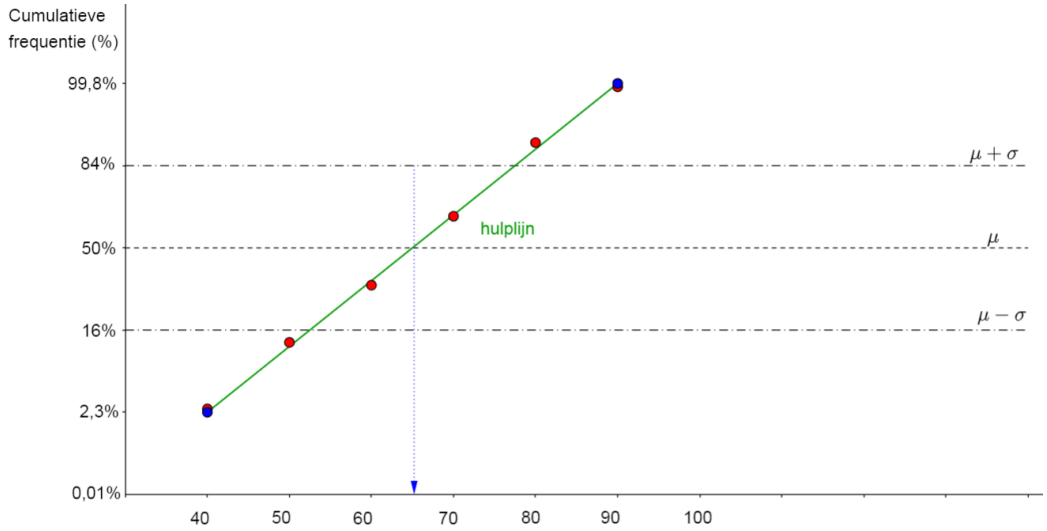
Gewicht	**Klassenmidden**	**frequentie**	**cum. Freq.**	**Rel cum. Freq.**
Kg	**Kg**	**aantal**	**aantal**	**%**
35-<45	40	25	25	2,5
45-<55	50	100	125	12,5
55-<65	60	200	325	32,5
65-<75	70	325	650	65
75-<85	80	250	900	90
85-<95	90	75	975	97,5
95-<105	100	25	1000	100
totaal		1000		

De gegevens zijn verwerkt en zijn uitgezet in de tabel hierboven. Enkel de waarden in de twee gekleurde kolommen zijn nu nog van belang.

Het klassenmidden zet je uit op de horizontale as en de relatieve cumulatieve frequentie op de verticale as bij het normaal waarschijnlijkheidspapier.

Nadat de punten zijn uitgezet op normaal-waarschijnlijkheidspapier onderzoek je of ze op een lijn liggen.

De punten uit het voorbeeld liggen bij benadering op een rechte lijn. Je mag nu de conclusie trekken dat de gegevens normaal verdeeld zijn.



Het snijpunt van die rechte lijn *(hulplijn)* met de horizontale lijn door 50% geeft het gemiddelde.

In dit geval is dat $\mu = 65\ kg$. Aflezen bij 84% geeft $78\ kg$. Hieruit volgt dat de standaardafwijking gelijk is aan: $\sigma = (78 - 65) = 13\ kg$.

Op basis van deze gegevens kun je nu met behulp van de vuistregels bij de normale verdeling uitspraken doen over de gehele populatie.

Al met al is het bepalen of de punten op een rechte lijn liggen een erg persoonlijke inschatting. Daarnaast is de nauwkeurigheid van aflezen van de waarden bij 50% en 84% ook erg afhankelijk van de persoon die het onderzoek uitvoert.

Als je dit onderzoek met de TI-84 zou kunnen uitvoeren zou dat heel wat meer houvast bieden en in ieder geval bij iedere onderzoeker tot dezelfde uitkomsten lijden.

Dat kan als je het programma hieronder op je TI-84 installeert.

Programma NWPAPIER

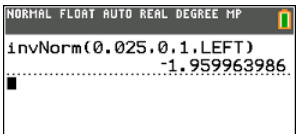
```
Disp "ONDERZOEK OF GEGEVENS"
Disp "NORMAAL VERDEELD ZIJN"
Disp "PLOT OP NORMAAL"
Disp "WAARSCHIJNLIJKHEIDS"
Disp "PAPIER"
Disp "ZET GEGEVENS IN L1 EN L2"          In L1 zet je de grootheid die je gemeten hebt
Pause                                    en in L2 de frequentie daarvan.
If min(L2)<0                             Negatieve frequentie kan niet; foutmelding.
Then
Goto Z
Else
PlotsOff                                 Zet plots uit.
FnOff                                    Zet functies uit.
AxesOn                                   Zet assen aan.
6→Xres                                   Door dit te doen kun je veel sneller plotten.
cumSum(L2)→L3                            Cumulatieve frequentie in L3.
L3/sum(L2)→L4                            Relatieve cumulatieve frequentie in L4.
3+seq(invNorm(L4(T)),T,1,dim(L4))→L5     Zie uitleg 1.
L5→L6                                    Kopieer L5 naar L6.
6→L6(dim(L5))                            Zie uitleg 2.
Plot1(Scatter,L1,L6)                     Plot de inhoud van L1 en L6.
min(L1)-1→Xmin                           Zet x-window goed.
max(L1)+1→Xmax
-0.2→Ymin                                Zet y-window goed.( zie ook uitleg 2)
6.2→Ymax
Full                                     Plot op volle weergave.
DispGraph
Pause                                    Vanaf hier worden de gegevens bepaald voor
L1→LX                                    de lineaire regressive. Zie ook uitleg 3.
L6→LY
dim(LX)→D
sum(LX)→M
sum(LY)→N
(LX)²→ LXK
LX*LY→LXY
sum(LXK)→O
sum(LXY)→P
M/D→E
N/D→F
P/D→G
E*F→H
stdDev(LX)→I
stdDev(LY)→J
(D*P-M*N)/(D*O-M²)→A                     Bepaal helling van de regressie lijn.
(F-A*E)→B                                Bepaal startwaarde van de lijn.
```

(G-H)/(I*J)→R	Bepaal de correlatiecoëfficiënt bij de gegevens.
(3-B)/A→U	x-waarde van gemiddelde : μ.
(3+invNorm(0.84)-B)/A→V	x-waarde bij 84% : $(\mu + \sigma)$.
V-U→W	Verschil is gelijk aan de standaardafwijking.
"AX+B"→Y_1	Plot de regressielijn door de punten.
DispGraph	
Text(140,160,"R=",R)	
If R≥0.85	Als correlatiecoëfficiënt $R \geq 0{,}85$ dan
Then	is dit zichtbaar.
Text(5,2,"NORMAAL VERDEELD")	
Text(20,2,"u=",U)	
Text(35,2,"σx=",W)	
Else	Slechte correlatie dan dit zichtbaar.
Text(5,2,"NIET NORMAAL")	
Text(20,2,"VERDEELD")	
Stop	
Lbl Z	Foutmelding.
Disp "LIJST 2 HEEFT NEGATIEVE"	
Disp "WAARDEN. DAT MAG NIET"	
Stop	

De uitleg is erg ingewikkeld. Mag ook worden overgeslagen!

Uitleg 1: Waarom : **3+seq(invNorm(L_4(T)),T,1,dim(L_4)→L_5 ?**

De TI-84 kan niet het commando invNorm(..) op de gegevens in een lijst uitvoeren, dan krijg je een foutmelding. Om dat te omzeilen is er gekozen voor het commando: seq(..) Dan worden de waarden uit lijst4 één voor één met invNorm(.) berekend en opgeslagen in lijst 5.

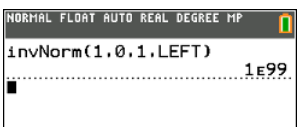


De inversenormaal functie werkt in dit programma met de standaardnormale verdeling die als kenmerken heeft $\mu = 0\ en\ \sigma = 1$. Een waarde als 2,5 % geeft dan als antwoord: $-1{,}9599..$

Je krijgt zo bij de omzetting van de inhoud van L4 getallen die lopen van -3 tot $+3$ want bij de standaardnormale verdeling met $\mu = 0\ en\ \sigma = 1$ geldt: $(\mu - 3\sigma) = -3\ en\ (\mu + 3\sigma) = 3$.
Tussen de grenzen -3 en 3 zit $99{,}7\%$ van alle waarnemingen. Voor deze plot is dat voldoende nauwkeurig.

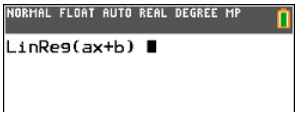
Het commando **Plot1(Scatter,L_1,L_5)** zou een plot genereren met de x-as storend midden in beeld. Om dat tegen te gaan wordt bij iedere uitkomst 3 opgeteld zodat de waarden in lijst 5 lopen van $0 \longrightarrow 6$

Uitleg 2: waarom **6→L_6(dim(L_5))** ?



In lijst 4 is het laatste getal altijd 100%. Bij omzetting via invNorm wordt dat in lijst 5 omgezet naar 10^{99} en geen 3 (zie screenshot).
Dit hele grote getal is een waarde die je niet wilt meenemen in je plot en overige berekeningen. Die moet vervangen worden door het getal 6. Dat doe je hier. Zoals gezegd introduceer je dan een kleine maar acceptabele foutmarge.

Uitleg 3: Waarom dit hele blok met berekeningen?



De TI-84 heeft een **LinReg (ax+b)** optie. Die optie werkt prima maar de gegevens die je wilt weten, de waarde van a, b en de regressie coëfficiënt r worden in een register gezet waar je vanuit een programma niet bij kunt. Je zal , als je die gegevens wilt gebruiken, ze zelf moet gaan genereren. Dat doe je in dit hele blok met berekeningen.

Wat moet je allemaal doen ?

Als je de theorie hierachter wil weten volg dan de link:
https://nl.wikipedia.org/wiki/Regressie-analyse

Hier wordt slechts summier aangegeven wat er is gedaan.

$L1 = Lx$	$L6 = Ly$	$(Lx)^2$	$Lx \cdot Ly$
..	..	..	..
..	..	..	..
$\sum Lx$	$\sum Ly$	$\sum(Lx)^2$	$\sum Lx \cdot Ly$

De lijsten L1 en L6 worden gekopieerd naar een andere lijst om de herkenbaarheid in de berekeningen te vergroten.

Bepaal vervolgens:

Lengte van lijst Lx, dat is de variabele D

Gemiddelden: $\overline{Lx} = \frac{\sum Lx}{D}$ en $\overline{Ly} = \frac{\sum Ly}{D}$

Helling van de regressielijn: $a = \frac{D \cdot \sum Lx \cdot Ly - \sum Lx \cdot y}{D \cdot \sum (Lx)^2 - (\sum Lx)^2}$

De startwaarde van de regressielijn: $b = \overline{Ly} - a \cdot \overline{Lx}$

De correlatiecoëfficiënt r via:

Bepaal daarvoor de Standaardafwijking van lijsten Lx en Ly, respectievelijk *StwX* en *StwY*

$r = \frac{\frac{\sum Lx \cdot Ly}{D} - \overline{Lx} \cdot \overline{Ly}}{Stwx \cdot Stwy}$

De mate van samenhang tussen de punten is af te lezen aan de waarde van de correlatiecoëfficiënt.

0,7 - 0,85	sterk
0,85 - 0,95	zeer sterk
> 0,95	uitzonderlijk sterk

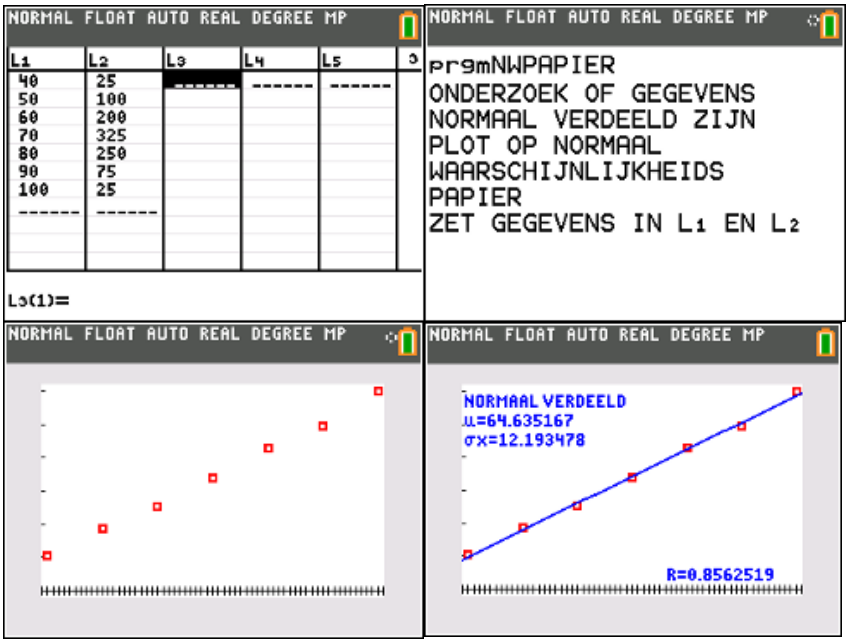
In het programma is de waarde voor R op $0{,}85$ gezet. Ligt de correlatiecoëfficiënt boven deze grens, dan wel normaal verdeeld. Ligt die eronder dan niet normaal verdeeld.

Deze grens is arbitrair vastgesteld door de auteur van het programma dus het staat U vrij hierin andere keuzes te maken.

VB1:

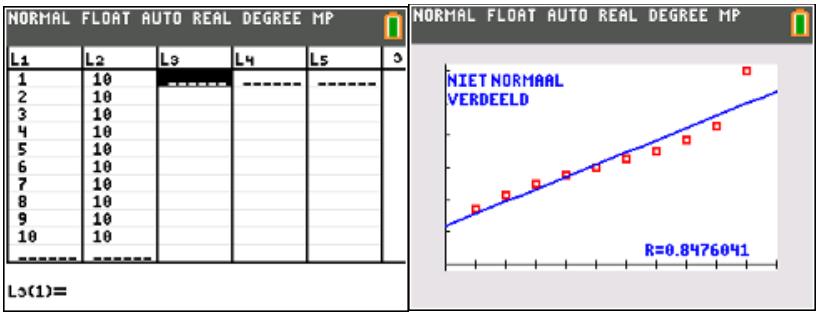
Het voorbeeld van het begin nog een keer maar nu met de TI84.

Gewicht	frequentie
35-<45	25
45-<55	100
55-<65	200
65-<75	325
75-<85	250
85-<95	75
95-<105	25
totaal	1000

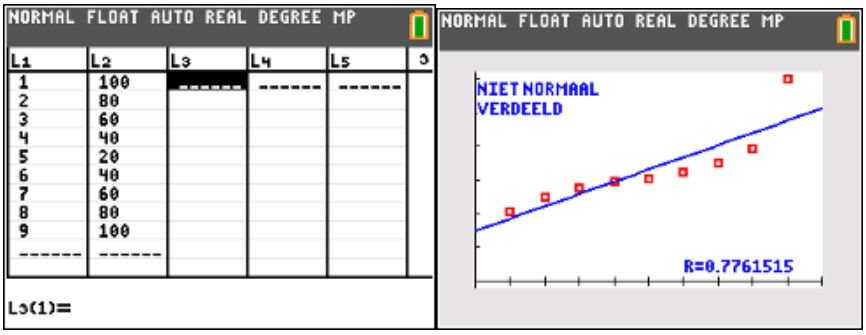


Na invoer van de gegevens is het programma gestart.
1-e scherm geeft de punten weer, tweede scherm de punten met de regressielijn en de gegevens die horen bij de normale verdeling.

VB2: Niet normaal verdeeld.



VB3: Niet normaal verdeeld.



VB 4:

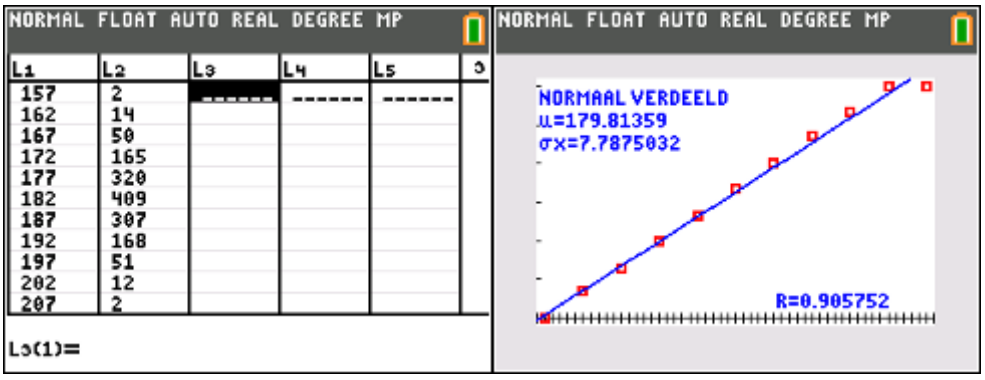
Van een groep jongens is de lengte gemeten.

Er wordt aangenomen dat de variabele lengte normaal verdeeld is.

Onderzoek of deze aanname klopt.

Is dat het geval, geef dan het gemiddelde en de standaarddeviatie van die normale verdeling.

Lengte (cm)	Frequentie
155-159	2
160-164	14
165-169	50
170-174	165
175-179	320
180-184	409
185-189	307
190-194	168
195-199	51
200-204	12
205-209	2
Totaal	1500

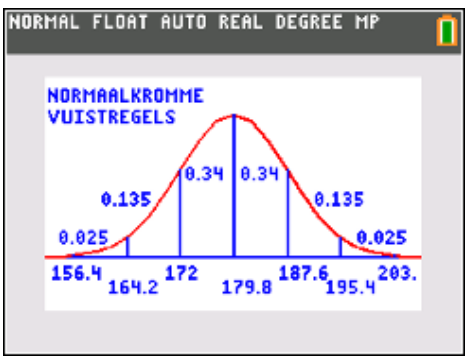


In lijst 1 zijn de klassenmiddens ingevoerd.

Het blijkt dat de gegevens normaal verdeeld zijn met

$\mu = 179{,}8\ cm$ en $\sigma = 7{,}8\ cm$.

Op basis van de gegevens kan je nu bepalen dat 2,5 % van de jongens kleiner is dan 164,2 cm of dat 16% groter is dan 187,6 cm etc, etc.



(Dit screenshot is afkomstig uit het TI84 programma VUISTREG van ELJ.)