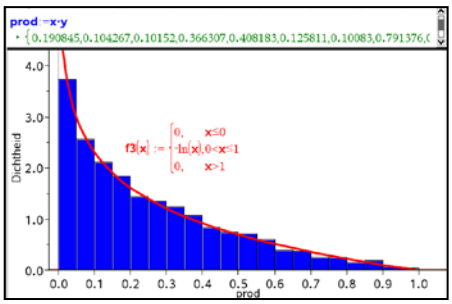
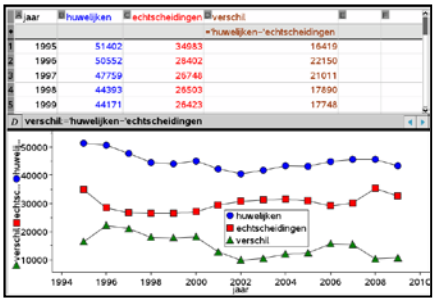
Cahiers T³ Europe
Vlaanderen nr. 28

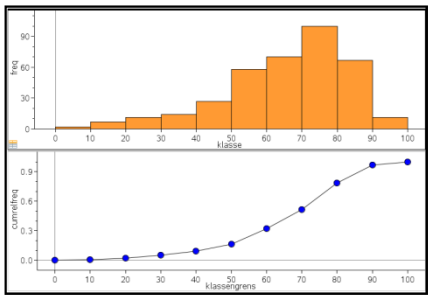
Statistiek

Dynamisch visualiseren en simuleren met TI-Nspire 3.0

Guido Herweyers







KATHOLIEKE UNIVERSITEIT LEUVEN

Inhoudsopgave

Inleiding 1

Deel 1: beschrijvende statistiek 3

1) Kwalitatieve niet gegroepeerde data 3
2) Kwalitatieve gegroepeerde data 6
3) Kantitatieve niet gegroepeerde data 10
4) Kwantitatieve gegroepeerde data 20
5) Spreidingsdagrammen en regressie 22

Deel 2: simuleren met TI-Nspire 25

1) Lukrake getallen genereren 25
2) Steekproeven met en zonder terugleggen 26
3) Twee dobbelstenen werpen 27
4) Kansverdeling van het steekproefgemiddelde 28
5) Steekproefvariabiliteit bij boxplots 32

Deel 3: kansverdelingen ontdekken 34

1) z-scores versus t-scores 34
2) Functies van één toevalsvariabele 35
3) Functies van twee toevalsvariabelen 37

Deel 4: statistische inferentie en simulaties 39

Referenties 46

Statistiek
Dynamisch visualiseren en simuleren met TI-Nspire 3.0

Guido Herweyers
KHBO Campus Oostende
guido.herweyers@khbo.be

Inleiding

"Meten is weten" is meer dan ooit actueel. Op woensdag 20 oktober 2010 vond de eerste Wereldstatistiekdag ('World Statistics Day') plaats, op initiatief van de Verenigde Naties. Bij deze gelegenheid publiceerde het **NIS** (Nationaal Instituut voor Statistiek) [1] een speciale uitgebreide versie van de kerncijfers 2009, waarbij België in een Europees perspectief wordt geplaatst [2].

Technologie evolueert en biedt nieuwe mogelijkheden voor het onderwijs. Dit is zeker zo voor statistiek. Concrete data kunnen snel worden gevisualiseerd op verschillende wijzen, simulaties van steekproeven en kansexperimenten kunnen worden onderzocht. Hiermee kan men statistische begrippen in een vroeger stadium invoeren, waarbij de data op de voorgrond treden en de formules op de achtergrond. Het is de bedoeling om met dit cahier via concrete voorbeelden een mogelijke aanpak te illustreren. De statistische basisbegrippen worden hier niet gedefinieerd, de nodige achtergrond hiervoor vindt men in de talrijke boeken over statistiek.

Deel 1 is een kennismaking met de mogelijkheden van TI-Nspire (softwareversie 3.0) [3] , waarbij voornamelijk de beschrijvende statistiek aan bod komt van de tweede graad en de derde graad met twee tot vier wekelijkse lestijden wiskunde. Er wordt ook gewerkt met data van het internet [1], [4].

Deel 2 gaat over simuleren met TI-Nspire. Hoe trekt men een enkelvoudige aselecte steekproef uit een populatie? Hoe komt de normale verdeling hierbij tevoorschijn als kansmodel voor het streekproefgemiddelde, bij grafische observatie van de variabiliteit ervan in meer dan 1000 steekproeven? Wat is de invloed van het kansmodel van de populatie en van de steekproefgrootte?

Deel 3 laat zien hoe men nieuwe kansverdelingen kan ontdekken via simulatie, uitgaande van de continue uniforme kansverdeling op het interval [0,1] . Ook de t-verdeling komt aan bod.

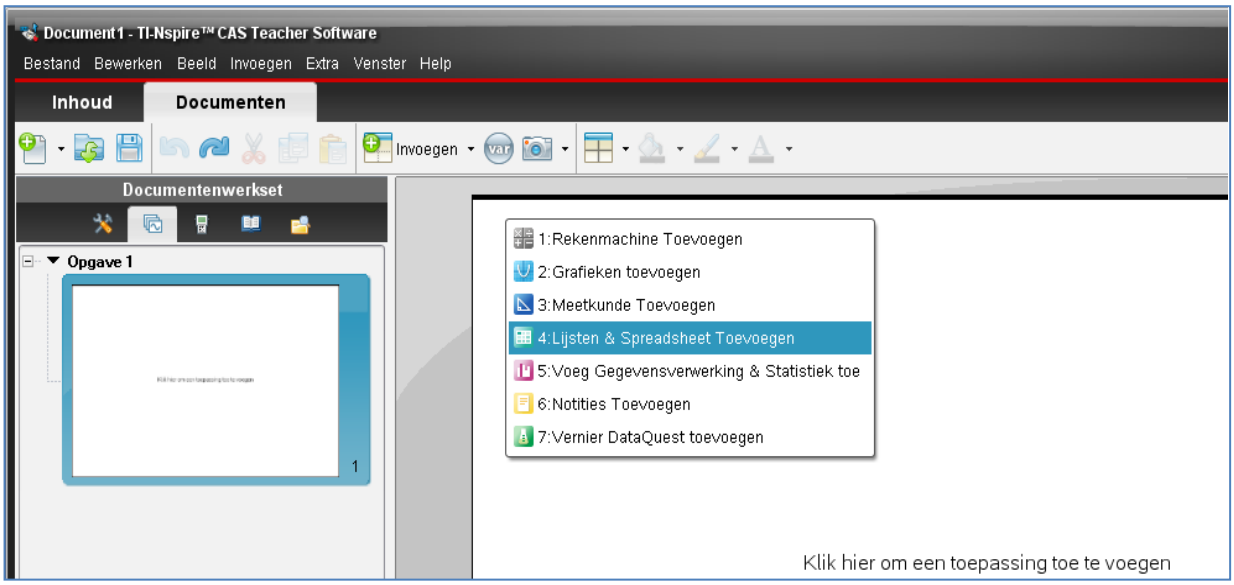
Deel 4 illustreert een mogelijke aanpak van toetsen van hypothesen, ondersteund door simulaties.

Traditioneel komt ook dit cahier op de website van T^3 Vlaanderen, www.t3vlaanderen.be , samen met het TI-Nspire bestand dat de data bevat van alle voorbeelden in dit cahier.

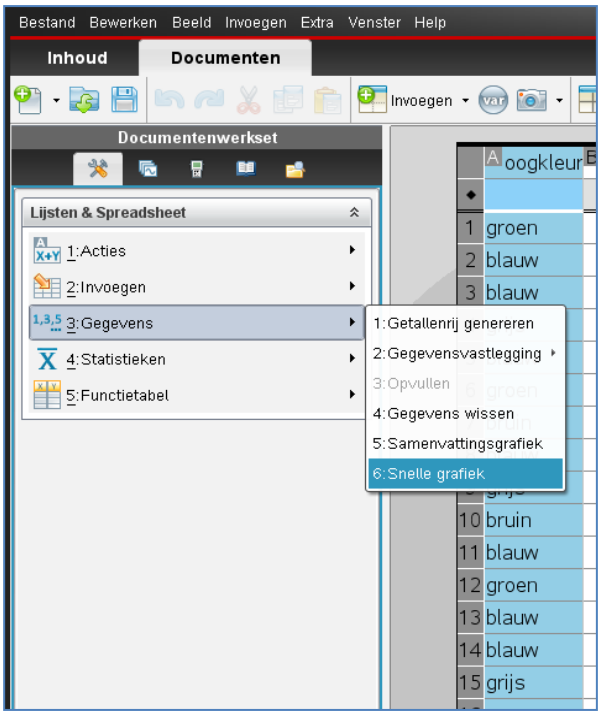
Deel 1: beschrijvende statistiek

1) Kwalitatieve niet gegroepeerde data

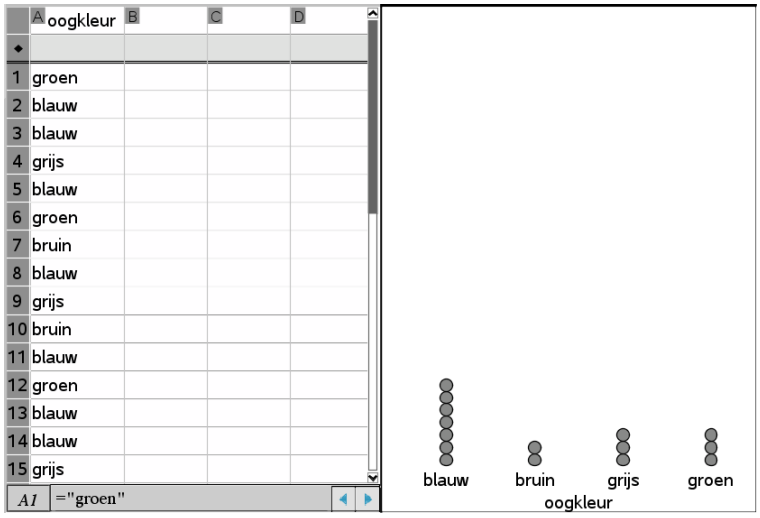
Open TI-Nspire en sluit het welkomscherm, voeg een pagina met een **lijsten en spreadsheet toepassing** toe aan het bestand (er zijn 7 toepassingen ter beschikking):



Voorbeeld 1:
Gegeven de oogkleur van 15 personen: groen, blauw, blauw, grijs, blauw, groen, bruin, blauw, grijs, bruin, blauw, groen, blauw, blauw, grijs.
→ Geef de eerste kolom A bovenaan de naam **oogkleur**.
→ Noteer de kleuren (start met cel A1) telkens tussen aanhalingstekens (tekstgegevens).

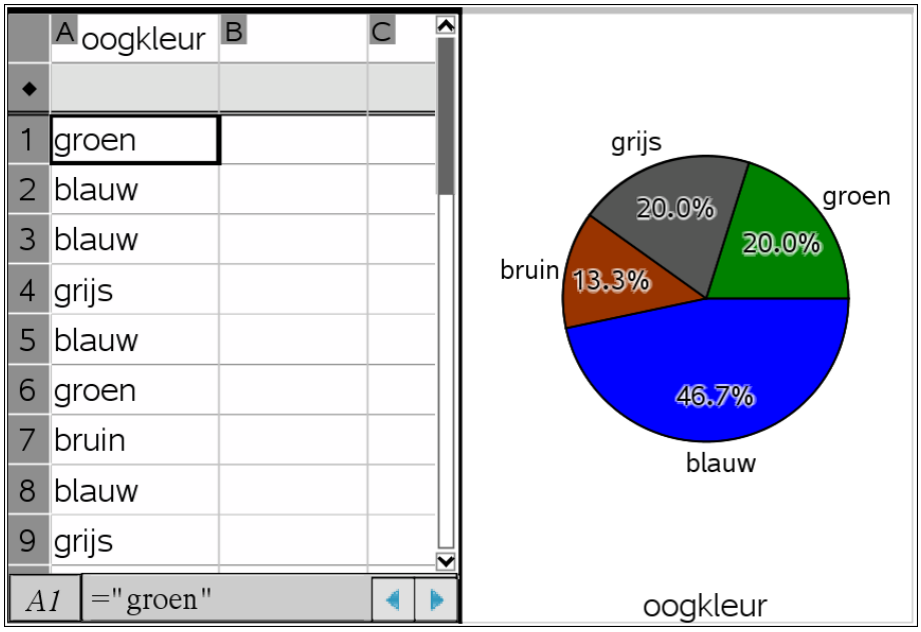


→ Selecteer de hele kolom (linkermuisklik bovenaan op de letter A), kies links het eerste menu  (documenttools**)** in de **documentenwerkset** ; dit toont het menu van de actieve (spreadsheet)toepassing. Kies hierin de optie **gegevens** met daaronder **snelle grafiek**. Het venster wordt gesplitst met rechts een **puntendiagram** of **dot plot**.

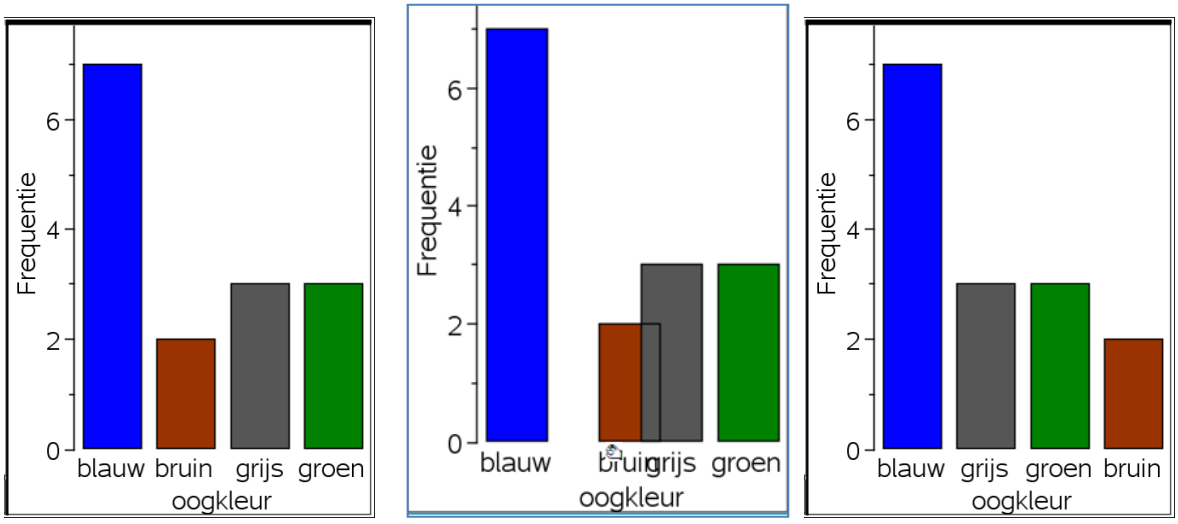


→ Selecteer het rechtervenster (linkermuisklik op dat venster). Wijzig de grafiek in een cirkeldiagram (of een taartdiagram) via het menu Plot-type onder  .
Met een rechtermuisklik op de cirkel kan men o.a. de kleuren aanpassen en de procentuele verdeling laten verschijnen (via **alle labels tonen**).

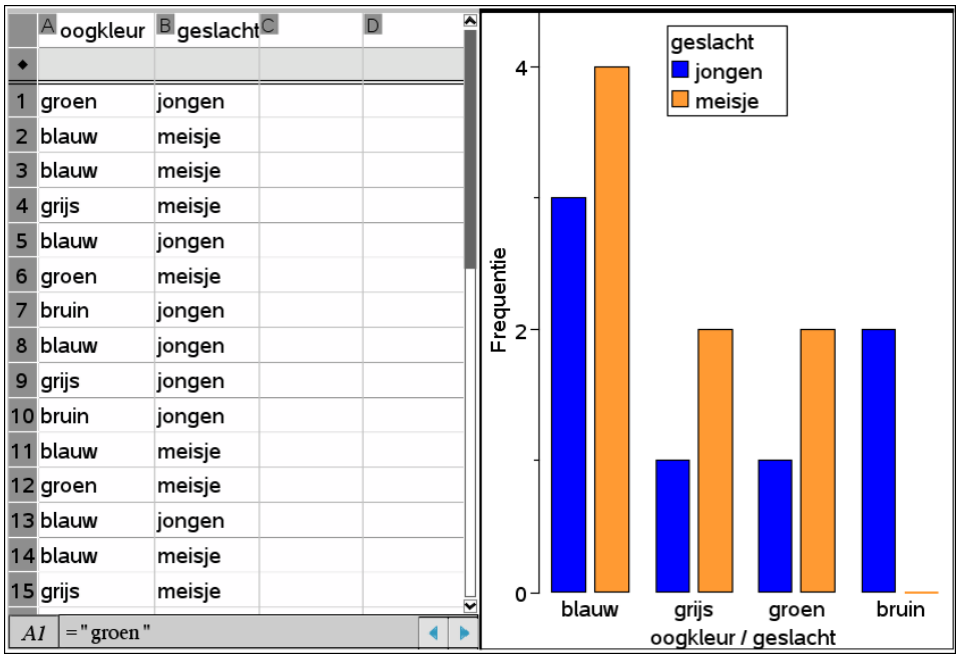
Tip: het **contextmenu** dat verschijnt bij een rechtermuisklik is handig om snel een actie ter plaatse uit te voeren. Een deelvenster kan men verwijderen door het eerst te selecteren (linkermuisklik) en vervolgens de toetsen **Ctrl** en **K** samen in te duwen (de rand van het venster vibreert), vervolgens op **Delete** drukken.



→ Wijzig de grafiek in een staafdiagram (rechtermuisklik in het rechtervenster).
Verschuif de categorie bruin naar rechts (linkermuisklik op het woord bruin, blijven induwen en verslepen).



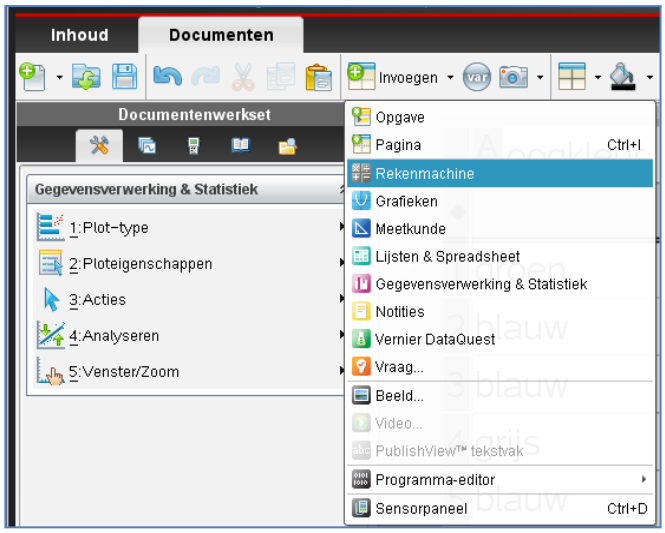
→ Noteer naast de oogkleur het geslacht van de personen en splits de categorie van de oogkleuren volgens de tweede categorie geslacht (rechterklik op de X-variabele oogkleur, optie **categorieën splitsen per variabele**)



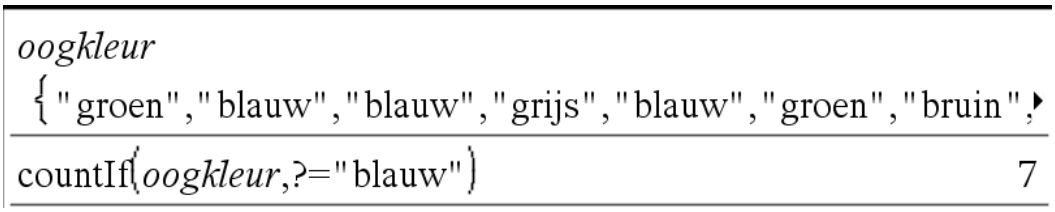
	A oogkleur	B geslacht	C	D
◆				
1	groen	jongen		
2	blauw	meisje		
3	blauw	meisje		
4	grijs	meisje		
5	blauw	jongen		
6	groen	meisje		
7	bruin	jongen		
8	blauw	jongen		
9	grijs	jongen		
10	bruin	jongen		
11	blauw	meisje		
12	groen	meisje		
13	blauw	jongen		
14	blauw	meisje		
15	grijs	meisje		

A1 = "groen"

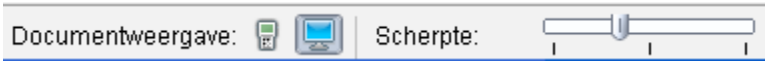
→ Voeg een **rekenmachine pagina** toe aan het bestand via de optie **invoegen** van het algemene **documentenmenu** bovenaan.



→Typ **oogkleur** gevolgd door **enter**; de variabele oogkleur is een **lijst** van tekstgegevens. Het commando **countif(oogkleur, ? = ”blauw”)** geeft het aantal gegevens “blauw” in de lijst.



Tip: de tekstgrootte kan men onderaan op het scherm aanpassen met de schuifregelaar

scherpte: 

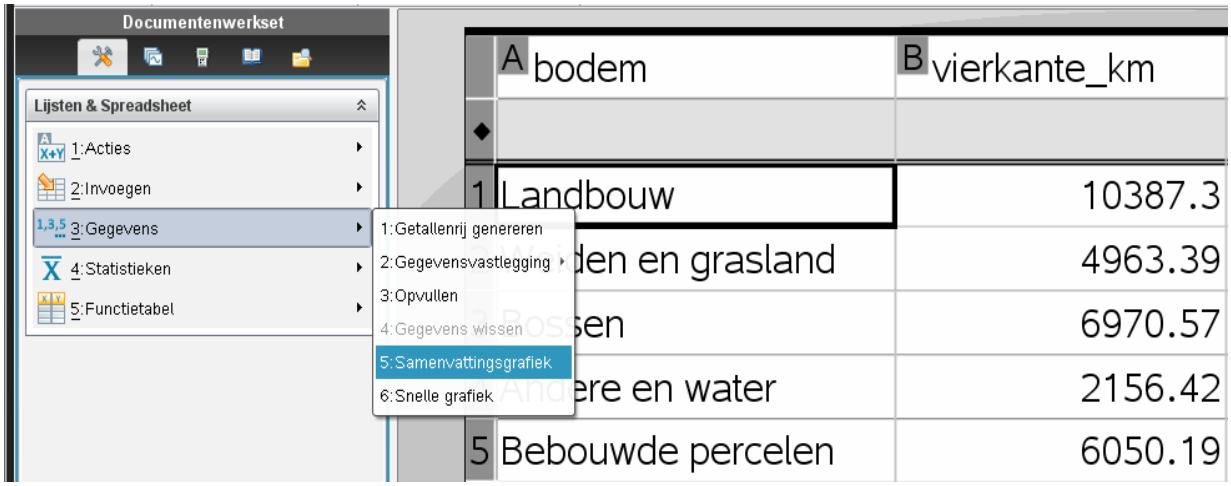
2) Kwalitatieve gegroepeerde data

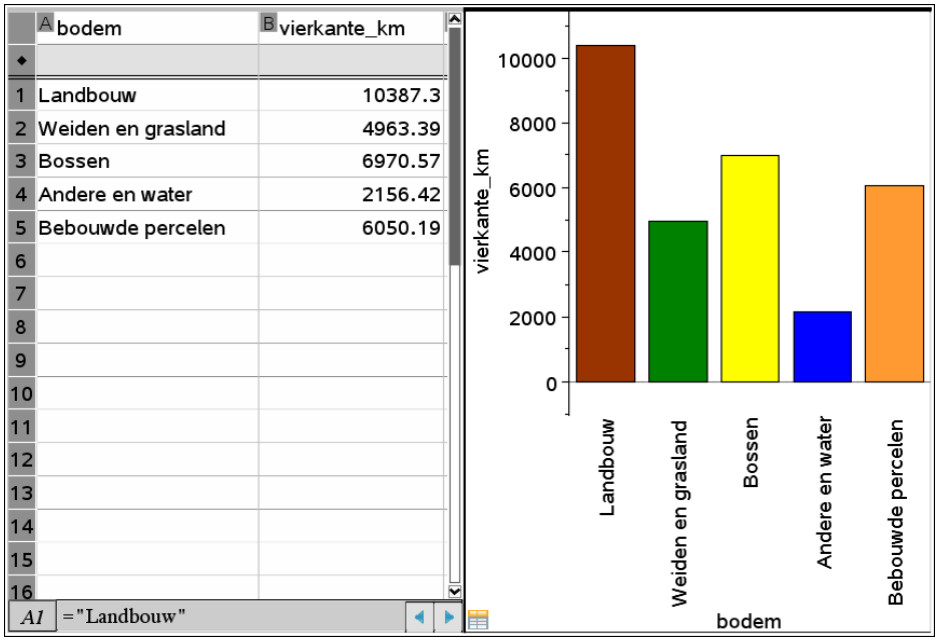
Voorbeeld 2:
Op de website van het NIS staat de verdeling van het bodemgebruik in België (2009):

Bodemgebruik België 2009 (in km²) bron: NIS	
Landbouw	10387,34
Permanente weiden en grasland	4963,39
Bossen	6970,57
Andere en water	2156,42
Bebouwde percelen	6050,19

→ Voeg een nieuwe spreadsheet pagina toe aan het bestand en vul de gegevens in.

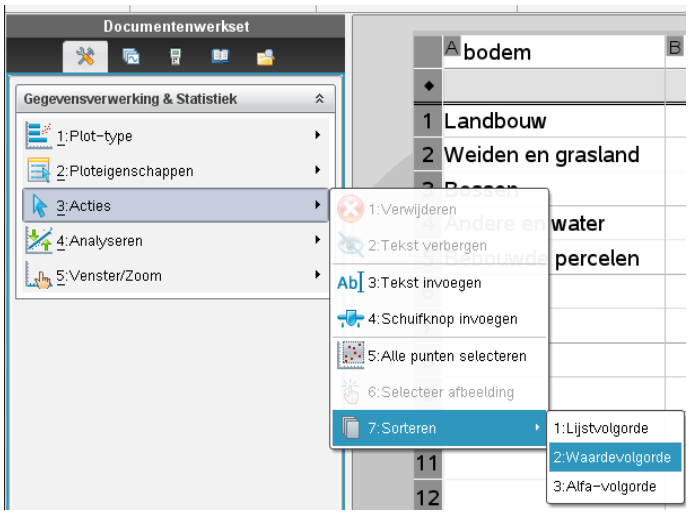
→ Kies **samenvattingsgrafiek** uit het menu gegevens van de spreadsheet toepassing, vervolgens bodem als X-lijst en vierkante_km als samenvattingslijst.

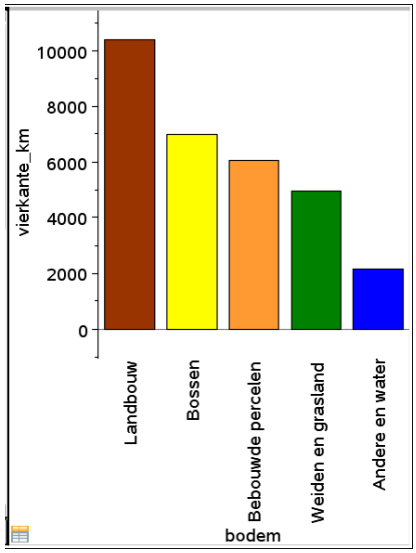




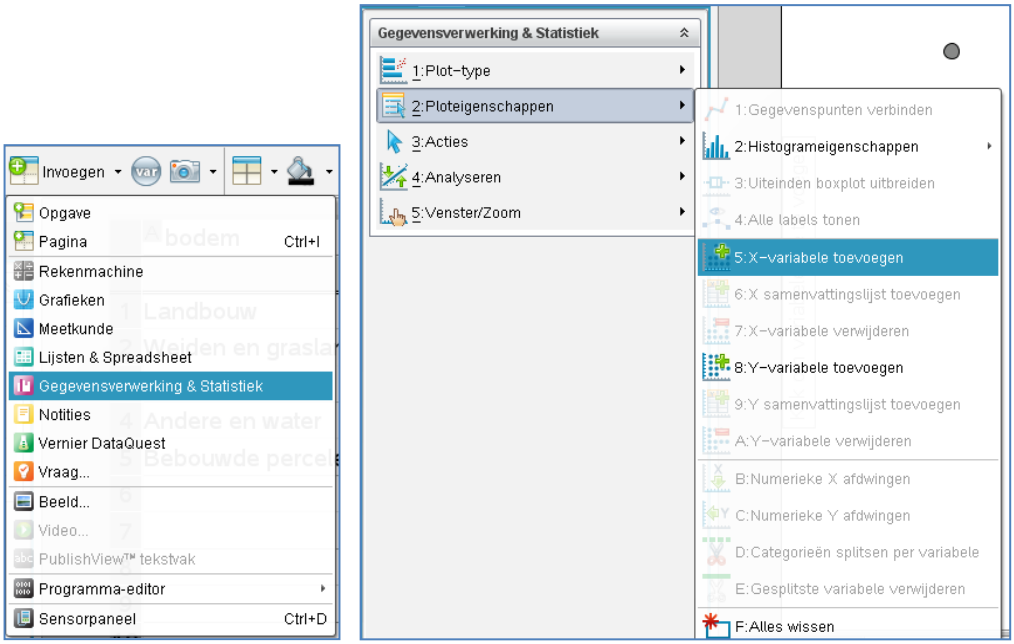
→ Selecteer het rechtervenster. Wijzig het staafdiagram volgens aflopende waarden via het menu **acties →sorteren→waardevolgorde**. Dit kan ook door de staven te verslepen.

Maak er een cirkeldiagram van met het menu Plot-type (ofwel met een rechtermuisklik op een staaf).

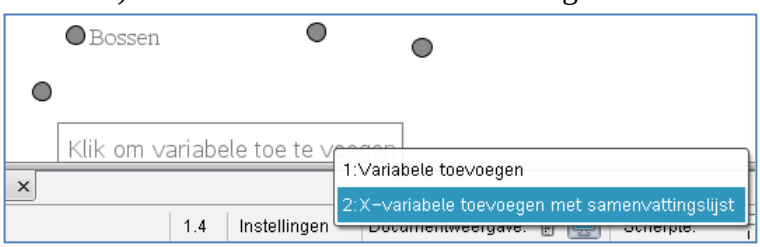




De statistische grafieken verschijnen in een **gegevensverwerking en statistiek toepassing**. Men kan deze toepassing ook eerst invoegen op een aparte pagina en vervolgens via het menu ploteigenschappen eerst een **X-variabele toevoegen** (kies bodem) en daarna een **Ysamenvattingslijst toevoegen** (kies vierkante_km) .

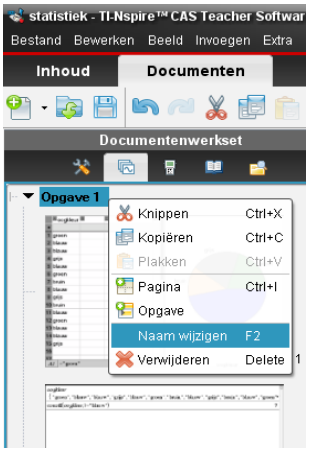


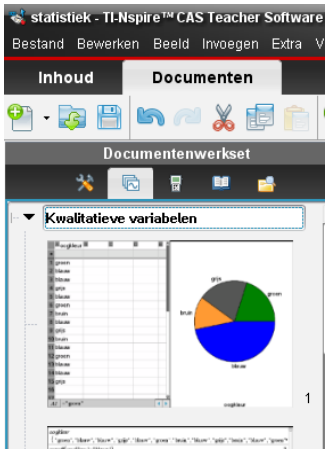
Het gaat echter sneller met een rechtermuisklik op “klik om variabele toe te voegen” (onderaan in het venster) en de keuze “X-variabele toevoegen met samenvattingslijst”.



→ Sla het bestand op onder de naam statistiek (menu **bestand**, **opslaan als**).

Open het menu paginasorteerder  van de documentenwerkset.
Wijzig de standaardnaam “opgave 1” in “kwalitatieve variabelen” door rechts te klikken op de naam opgave 1.





Een bestand kan verschillende opgaven bevatten, elke opgave kan verschillende pagina’s bevatten (selecteerbaar in de paginasorteerder ), elke pagina kan bestaan uit één tot 4 toepassingen (er zijn 7 verschillende toepassingen ter beschikking). Elke toepassing heeft een eigen menu, ter beschikking onder  in de documentenwerkset.

Binnen één opgave heeft een variabele dezelfde waarde op elke pagina. De variabele *x* in de ene opgave heeft echter niets te maken met de variabele *x* in een andere opgave.

Voorbeeld 3:

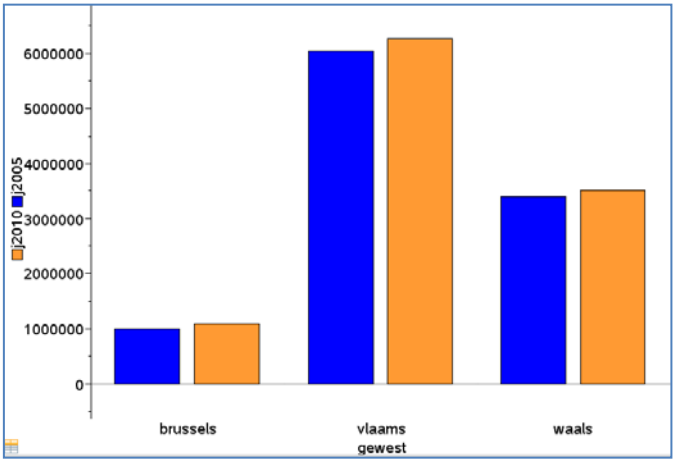
België bestaat uit drie gewesten: het Vlaamse gewest, het Brussels hoofdstedelijk gewest en het Waalse gewest. Hier volgt een tabel van de wettelijke bevolking (op 1januari van het vermelde jaar, bron NIS).

Gewest	2005	2006	2007	2008	2009	2010
Brussels Hoofdstedelijk Gewest	1.006.749	1.018.804	1.031.215	1.048.491	1.068.532	1.089.538
Vlaams Gewest	6.043.161	6.078.600	6.117.440	6.161.600	6.208.877	6.251.983
Waals Gewest	3.395.942	3.413.978	3.435.879	3.456.775	3.475.671	3.498.384
waarvan Duitstalige gemeenschap	72.512	73.119	73.675	74.169	74.540	75.222

→Maak een staafdiagram van de gewesten voor de jaren 2005 en 2010.

Werkwijze: men kan verschillende Y samenvattingslijsten toevoegen bij een gegeven X-variabele.

	A gewest	B j2005	C j2006	D j2007	E j2008	F j2009	G j2010
◆							
1	brussels	1006749	1018804	1031215	1048491	1068532	1089538
2	vlaams	6043161	6078600	6117440	6161600	6208877	6251983
3	waals	3395942	3413978	3435879	3456775	3475671	3498384



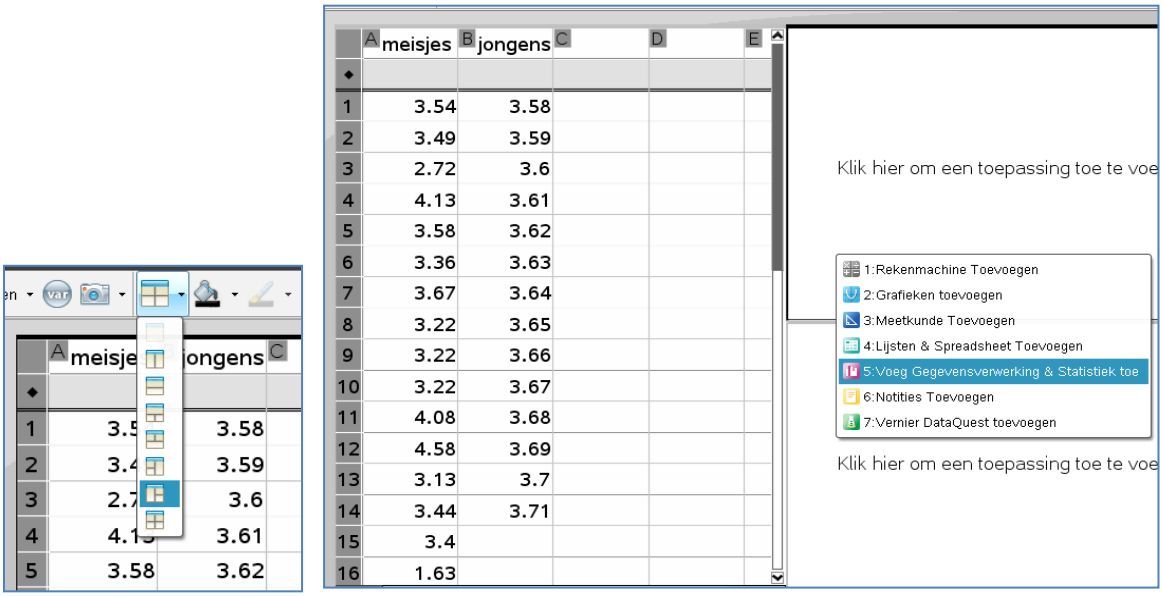
3) kwantitatieve niet gegroepeerde data

Voorbeeld 4:

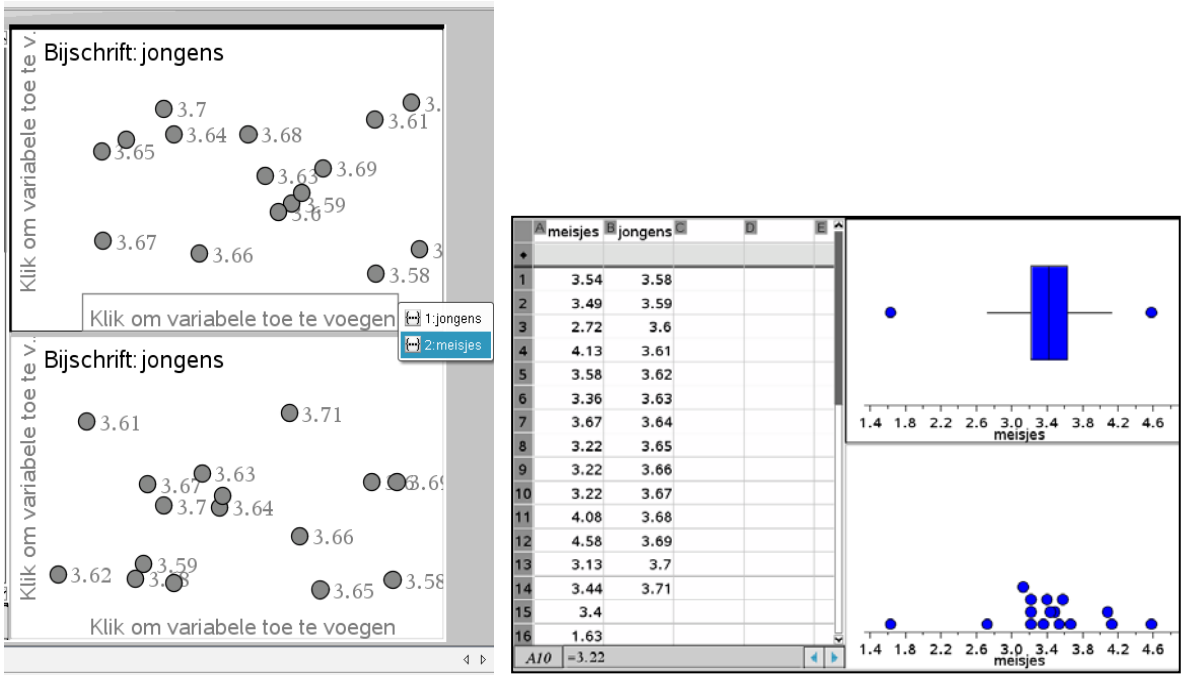
Gegeven de volgende geboortegewichten van 16 meisjes en 14 jongens (in kg). Data uit een Excelbestand kunnen met copy en paste rechtstreeks naar een TI-Nspire spreadsheet pagina worden overgebracht, het decimaalteken moet hiertoe in Excel eerst worden ingesteld op een punt i.p.v. een komma.

Meisjes	Jongens
3.54	3.58
3.49	3.59
2.72	3.60
4.13	3.61
3.58	3.62
3.36	3.63
3.67	3.64
3.22	3.65
3.22	3.66
3.22	3.67
4.08	3.68
4.58	3.69
3.13	3.70
3.44	3.71
3.40	
1.63	

→ Begin aan een nieuwe opgave binnen hetzelfde bestand (documentenmenu **invoegen opgave**), wijzig de naam “opgave 2” in “kwantitatieve variabelen”, voeg een **lijsten & spreadsheet** pagina toe, noteer de data in kolommen A en B, geef die kolommen de naam meisjes en jongens. Splits de pagina in drie zoals aangegeven (documentenmenu pagina-indeling). Voeg twee keer een toepassing **gegevensverwerking en statistiek** toe in de lege vensterdelen, met deze toepassing worden statistische gegevens grafisch voorgesteld.

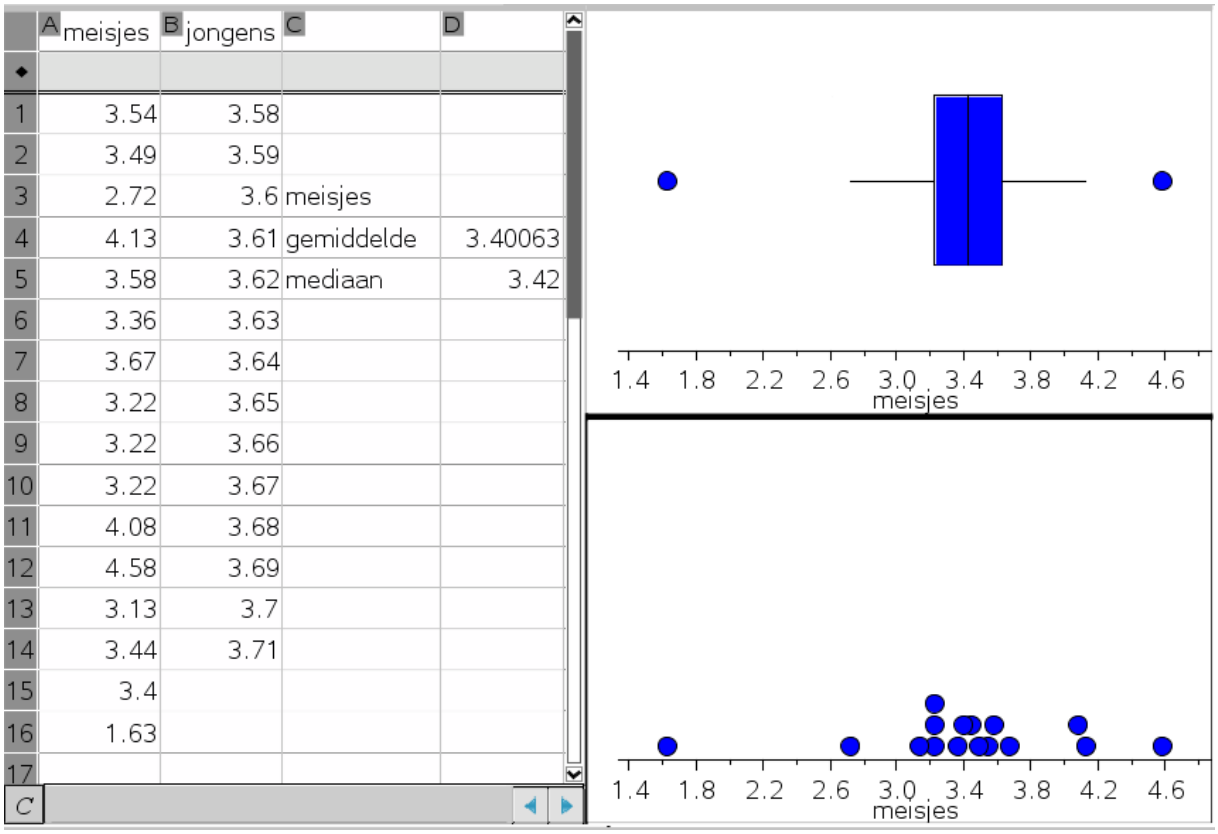


→ Klik onderaan om de variabele meisjes toe te voegen, doe dit voor beide vensters rechts. Standaard verschijnt een dot plot van de gegevens, wijzig het bovenste venster in een boxplot (rechtermuisklik in dat venster).



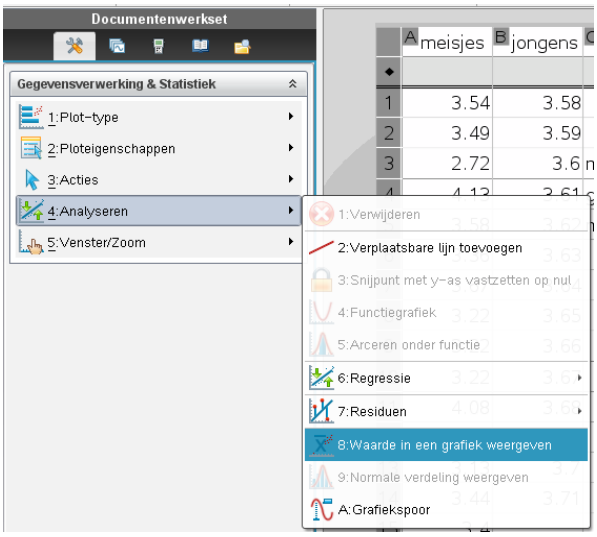
→ Typ tussen aanhalingstekens “meisjes” in cel C3 van de spreadsheet , “gemiddelde” in cel C4 en “mediaan” in C5.

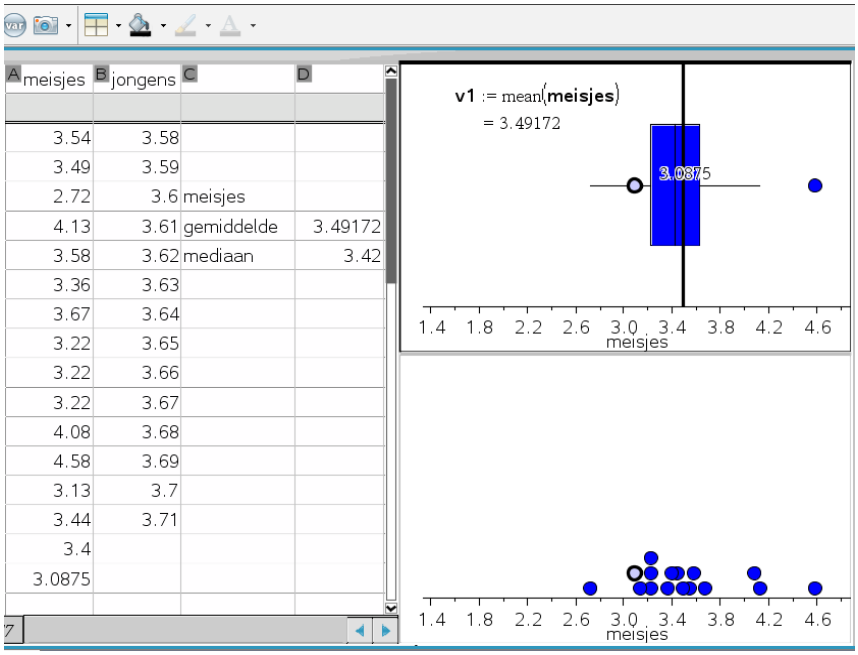
Typ **=mean(meisjes)** in cel D4 gevolgd door **enter**; het gemiddelde verschijnt.
Typ **=median(meisjes)** in cel D5 gevolgd door **enter**; de mediaan verschijnt.



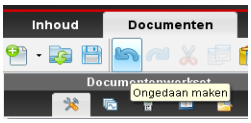
	A meisjes	B jongens	C	D
1	3.54	3.58		
2	3.49	3.59		
3	2.72	3.6	meisjes	
4	4.13	3.61	gemiddelde	3.40063
5	3.58	3.62	mediaan	3.42
6	3.36	3.63		
7	3.67	3.64		
8	3.22	3.65		
9	3.22	3.66		
10	3.22	3.67		
11	4.08	3.68		
12	4.58	3.69		
13	3.13	3.7		
14	3.44	3.71		
15	3.4			
16	1.63			
17				

→Klik in het boxplotvenster, kies het toepassingsmenu **analyseren**, **waarde in een grafiek weergeven**, vul v1:= **mean(meisjes)** in. Een verticale lijn ter hoogte van het gemiddelde verschijnt. Vervolgens kan men een uitschieter vastpakken en verplaatsen (linkermuistoets blijven indrukken op de uitschieter). Merk op hoe de data mee wijzigen, samen met het gemiddelde. Bestudeer de invloed van uitschieters op het gemiddelde.



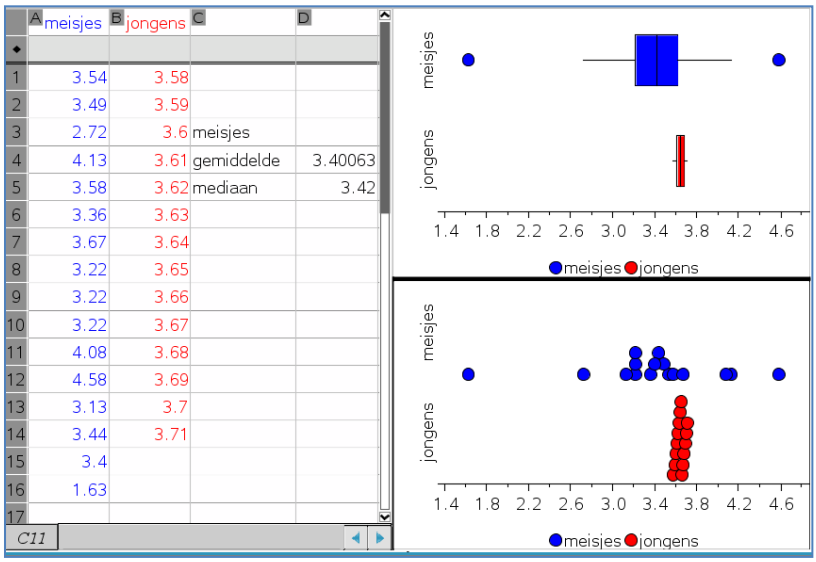


Tip: men kan de oorspronkelijke data terug verkrijgen met het documentenmenu “ongedaan maken” .

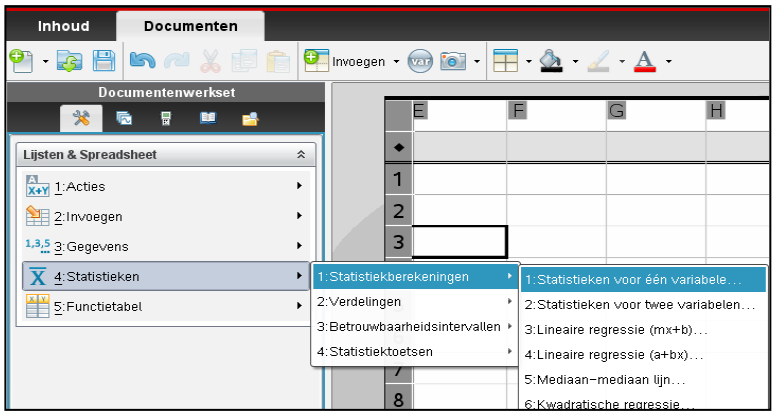


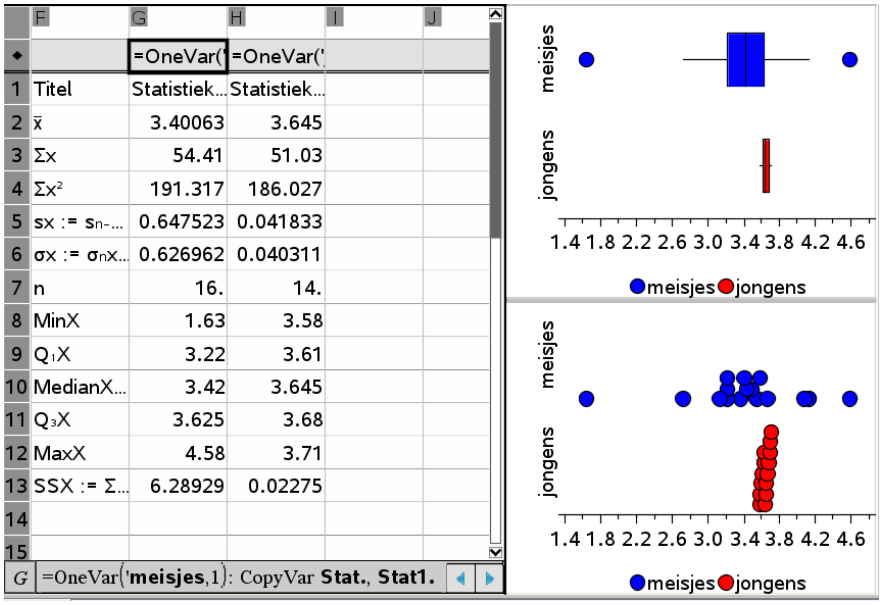
Wijziging van data kunnen (gelukkig) ook verboden worden door de variabele eerst te vergrendelen, open hiertoe een rekenmachine pagina en typ het bevel **lock(meisjes)** gevolgd door **enter**. Het terug vrijgeven van de data gebeurt met het bevel **unlock(meisjes)**.

→ Herstel de oorspronkelijke data met het menu “ongedaan maken”. Wis de verticale lijn voor het gemiddelde (selecteer de lijn en druk **delete**). Kies, met een rechtermuisklik op de naam meisjes van de horizontale as, voor **X-variabele toevoegen**, klik vervolgens op de variabele jongens. Doe dit zowel voor de boxplots als voor de puntenplots. Pas de kleuren aan. Vergelijk de data meisjes versus jongens voor deze steekproeven.

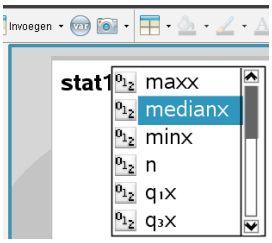


→ Bepaal nu de statistieken van de data.
Ga in de spreadsheet naar kolom E en selecteer (klik) daar een cel, kies vervolgens het menu statistieken, statistiekberekeningen, statistieken voor één variabele, 2 lijsten.





→ Wijzig opnieuw de data (door verslepen) en observeer de wijziging van de statistieken.
→ Open een rekenmachine toepassing, typ **stat1.** en selecteer vervolgens een statistiek uit de lijst die dan verschijnt, gevolgd door enter. Analoog met **stat2.** voor de statistieken van de tweede lijst, **stat1.results** levert een matrix met een overzicht van de statistieken van de meisjes. TI-Nspire slaat automatisch alle statistische berekeningen op in stat1, stat2, stat3, ….



stat1.MedianX		3.42
stat2.MedianX		3.645
stat1.results	"Titel"	"Statistieken voor één variabele"
	"$\bar{x}$"	3.40063
	"Σx"	54.41
	"Σx^2"	191.317
	"$sx := s_{n-1}x$"	0.647523
	"$\sigma x := \sigma_n x$"	0.626962
	"n"	16.
	"MinX"	1.63
	"Q_1X"	3.22
	"MedianX"	3.42
	"Q_3X"	3.625
	"MaxX"	4.58
	"$SSX := \Sigma(x-\bar{x})^2$"	6.28929

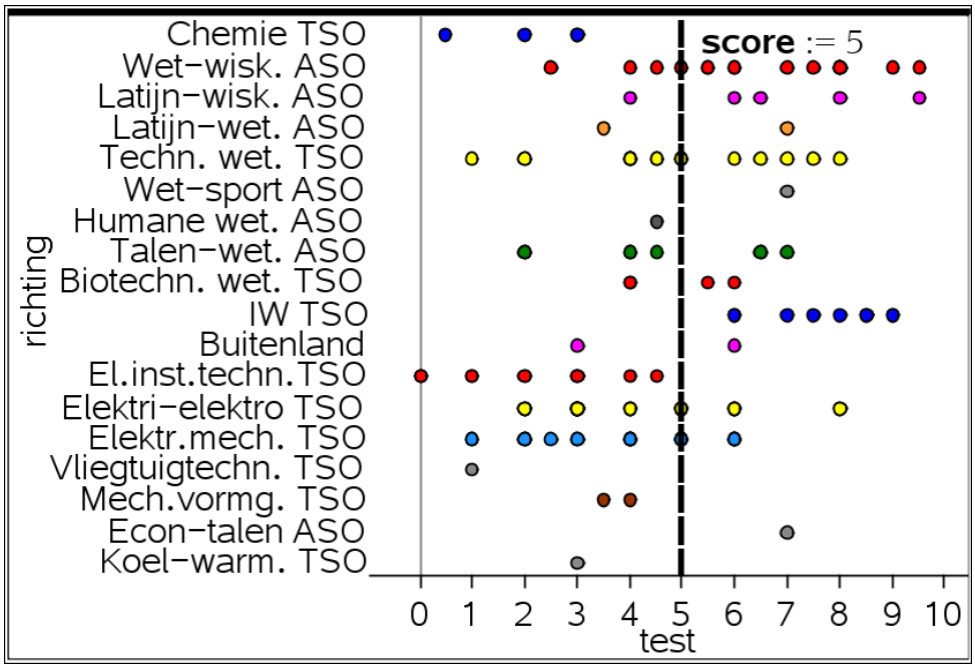
Statistische berekeningen kan men overigens ook in een rekenmachine toepassing uitvoeren via het menu statistieken.

Voorbeeld 5:

In dit voorbeeld worden kwantitatieve niet gegroepeerde data opgesplitst per categorie.
De populatie is het eerste jaar professionele bachelor aan de KHBO Campus Oostende academiejaar 2008-2009 (137 studenten). Hier volgt een overzicht van de resultaten van een test wiskunde tijdens het eerste semester (score op 10), samen met de richting die de student heeft gevolgd in het secundair onderwijs. De resultaten komen in de lijsten test en richting:

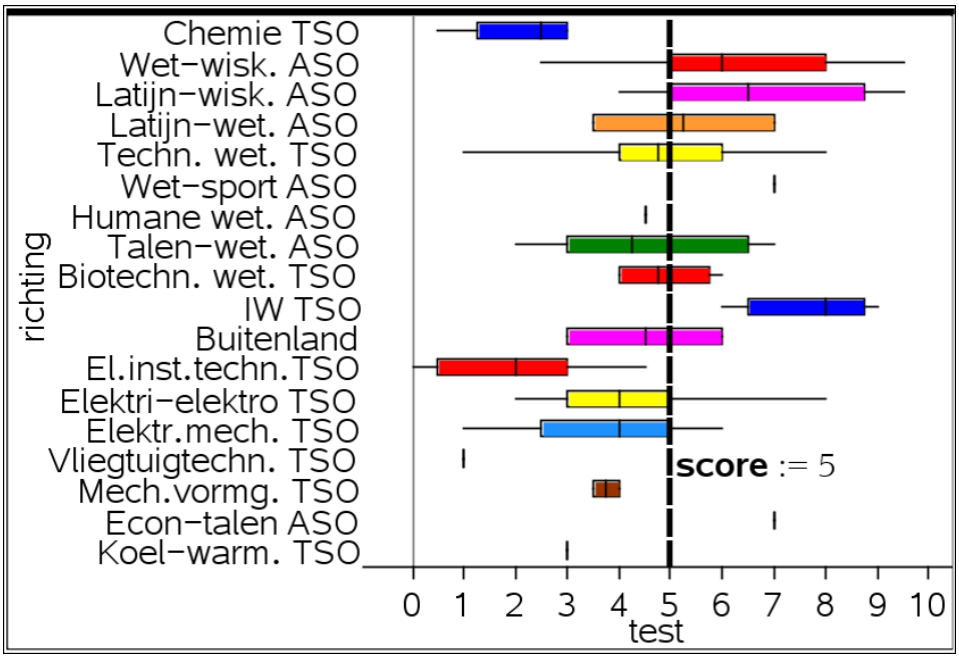
	A test	B richting
◆		
1	3	Chemie TSO
2	5	Wet-wisk. ASO
3	5.5	Wet-wisk. ASO
4	5	Wet-wisk. ASO
5	6	Latijn-wisk. ASO
6	7	Latijn-wet. ASO
7	2	Techn. wet. TSO
8	2	Chemie TSO
9	7	Wet-sport ASO
10	4	Wet-wisk. ASO
11	4.5	Humane wet. ASO
12	4.5	Techn. wet. TSO
13	7	Talen-wet. ASO
14	4.5	Wet-wisk. ASO
B1	="Chemie TSO"	

Voeg een nieuwe pagina gegevensverwerking en statistiek toe.
Zet de variabele test uit op de X-as en de variabele richting op de Y-as. Dit levert de puntenplots van de testresultaten opgesplitst per richting, voeg er de lijn ter hoogte van score 5 aan toe (menu **analyseren →waarde in een grafiek weergeven** , dan nog de naam van de waarde wijzigen in score).

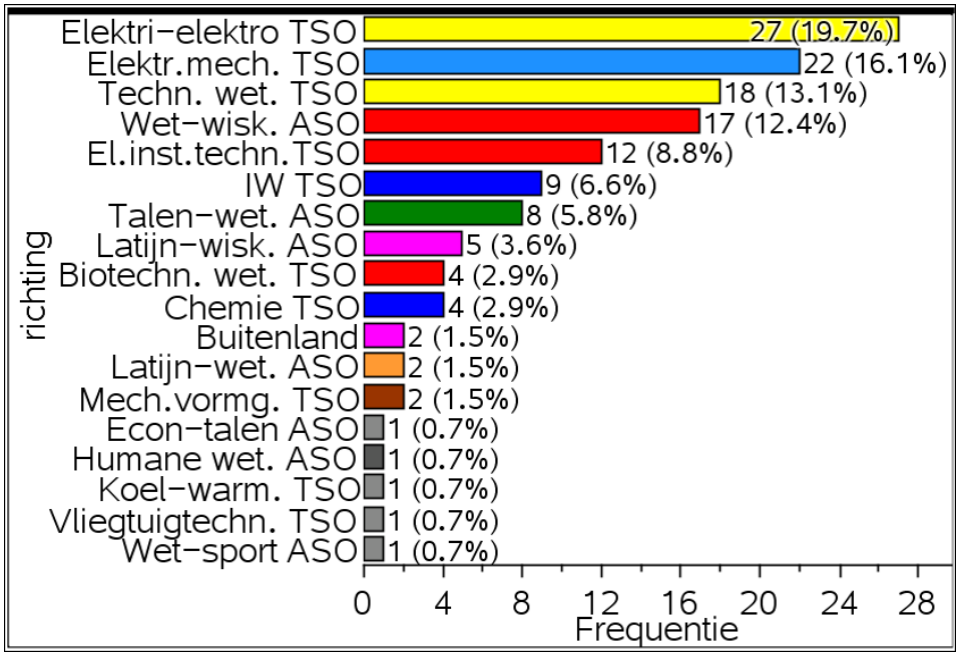


De puntenplots geven een duidelijk beeld van de individuele resultaten en de richtingen die laag of hoog scoren.

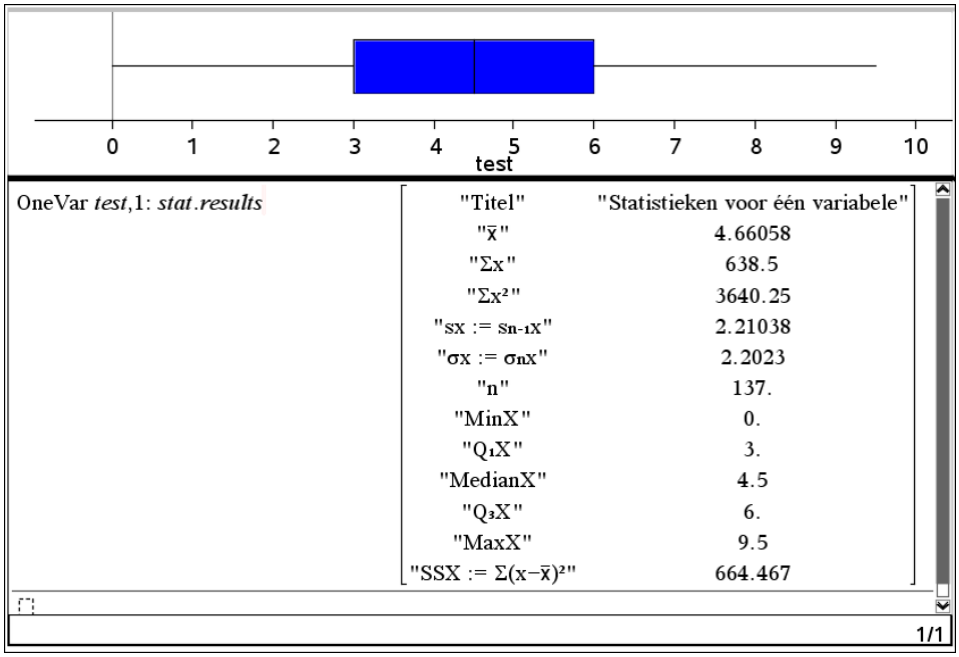
→ Wijzig de puntenplots in boxplots, er zijn 6 richtingen met een mediaan hoger dan 5 en 12 richtingen met een mediaan kleiner dan 5 :



→Verwijder de variabele test op de X-as (rechterklik op test en **X-variabele verwijderen**), wijzig het puntendiagram in een staafdiagram, sorteer dit volgens waardevolgorde en toon alle labels om een overzicht te hebben van de studentenaantallen per richting (telkens met een rechterklik in het venster).



→ Open een nieuwe pagina en splits ze horizontaal in een toepassing gegevensverwerking en statistiek en een toepassing rekenmachine. Vat alle testresultaten samen met een boxplot en de statistieken van de data.



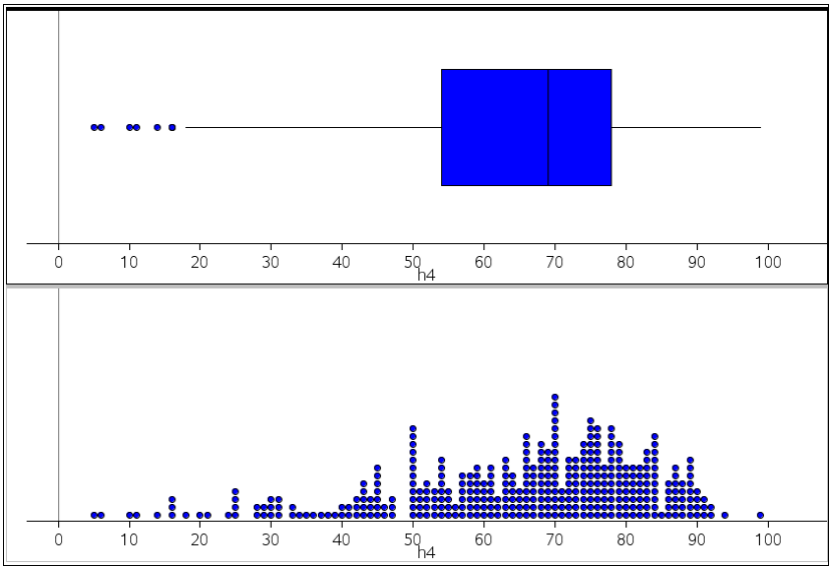
Voorbeeld 6:

Tijdens de zomervakantie worden de stranden aan de Belgische kust bewaakt door redders aan zee. Heel wat studenten voelen zich aangetrokken tot deze avontuurlijke vakantiejob. De opleiding tot redder aan zee is echter niet te onderschatten; naast een uitgebreide theoretische cursus moet men ook slagen in een aantal zware zwemproeven.

De opleiding wordt jaarlijks georganiseerd. Eerst moeten de studenten het theoretisch examen afleggen. De cursus bestaat uit zeven hoofdstukken en deze worden apart gequoteerd, elk op 100 punten. Om te slagen voor het theoretisch deel moet men minstens 50% behalen voor elk van de zeven hoofdstukken. Enkel de studenten die geslaagd zijn voor theorie mogen nadien deelnemen aan het praktisch examen (zwemproeven, knopenleer, eerste hulp bij ongevallen). De resultaten van het theoretisch examen verschijnen jaarlijks op een website [4].

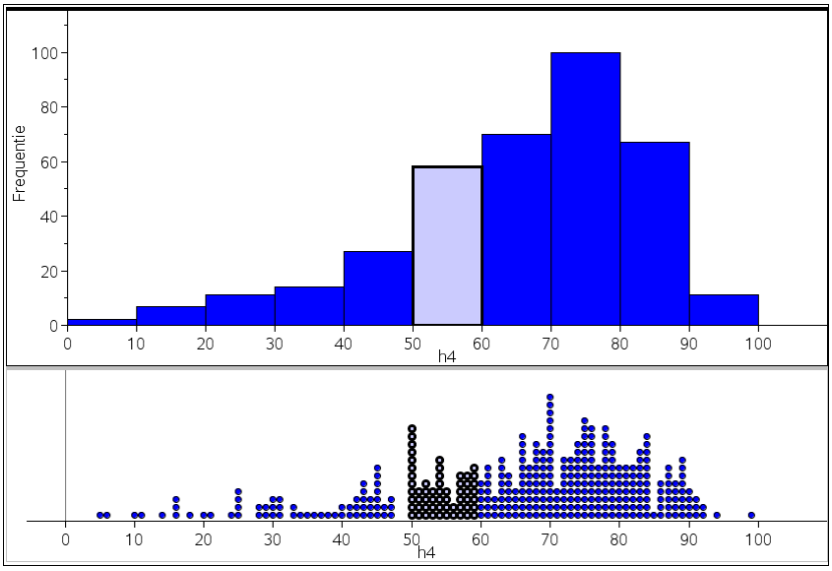
Er waren 367 deelnemers voor de opleiding in 2010-2011, elke deelnemer krijgt een nummer toegewezen tussen 1 en 367, de resultaten vindt men volgens oplopend nummer in het bestand "statistiek.tns".

→ Open een rekenmachine pagina en vergrendel de data met het commando
lock h1,h2,h3,h4,h5,h6,h7
→ Open een nieuwe pagina met twee toepassingen gegevensverwerking en statistiek.
Maak een puntenplot en een boxplot van de resultaten voor hoofdstuk 4 (variabele h4), de data zijn links scheef verdeeld.



→ Wijzig de boxplot in een histogram, vergroot het histogramvenster (horizontale venstergrens verslepen) en experimenteer met de klassenbreedte (beweeg de cursor naar een opstaande rechthoekszijde tot een dubbele pijl verschijnt, druk vervolgens de linkermuistoets in en versleep die zijde). De klassenbreedte kan ook manueel worden ingesteld (rechtermuisklik in het histogramvenster) met **Klasse-instellingen**. Beweeg de cursor over het histogram om de frequenties af te lezen. Het histogram voor de relatieve frequenties vindt men met een rechtsklik op het histogramvenster en de keuze **schaal**, **percentage**.

Bij selectie van een klasse worden de corresponderende data getoond in de puntenplot. De selectie wordt ongedaan gemaakt door een klik buiten het histogram.



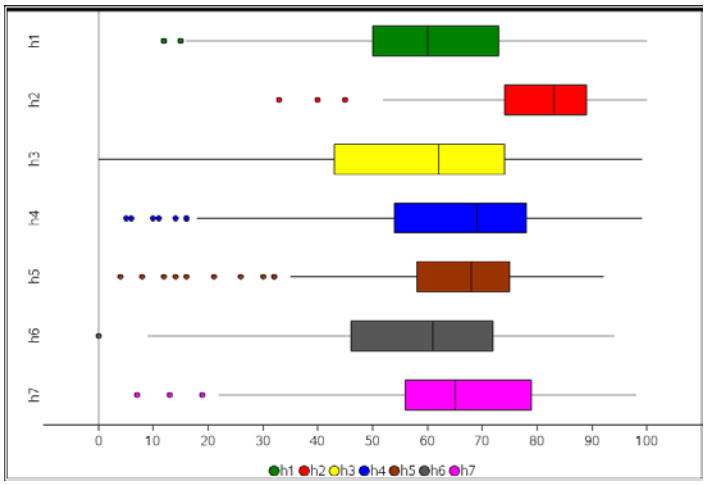
→ De klassenfrequenties kunnen worden berekend in een rekenmachine toepassing met het commando countif . Het symbool ≤ vindt men met het menu hulpprogramma's  van de documtenwerkset onder symbolen.
Een lijst van frequenties horende bij de klassen [0,10[, [10,20[, [20,30[, …… [100,110[wordt verkregen met **seq (countif (h4 , 10k ≤ ? < 10(k+1)) , k , 0 , 10)**
De lijst wordt ook rechtstreeks in een kolom van een spreadsheet toepassing gegenereerd door

= seq (countif (h4 , 5k ≤ ? < 5(k+1)) , k , 1 , 10)

te typen in de grijze cel onder de naam van de kolom (d.i. de formulecel voor de hele kolom).

Lock *h1,h2,h3,h4,h5,h6,h7*	*Gereed*
countIf(*h4*,20≤?<30)	11
seq(countIf(*h4*,10·*k*≤?<10·(*k*+1)),*k*,0,10)	{2,7,11,14,27,58,70,100,67,11,0}

→ Vergelijk alle hoofdstukken in één venster via hun boxplots en bespreek het verschil.



→ Bepaal de statistieken van de drie eerste hoofdstukken, leg het verband met hun boxplots.

OneVar 3,*h1,h2,h3*: *stat.results*

"Titel"	"Statistieken voor één variabele"	"_____"	"_____"
"$\bar{x}$"	59.9019	81.0198	57.8392
"Σx"	21984.	28681.	21227.
"Σx^2"	1.43359E6	2.36305E6	1.38082E6
"s_X := $s_{n-1}X$"	17.8566	10.5536	20.4505
"σ_X := σ_{nX}"	17.8323	10.5387	20.4226
"n"	367.	354.	367.
"MinX"	12.	33.	0.
"Q_1X"	50.	74.	43.
"MedianX"	60.	82.	62.
"Q_3X"	73.	89.	74.
"MaxX"	100.	100.	99.
"SSX := $\Sigma(x-\bar{x})^2$"	116702.	39316.9	153070.

De hoofdstukken h2 en h7 bevatten lege cellen, voor de studenten die niet hebben deelgenomen aan het examen van die hoofdstukken.

Lege plaatsen in een lijst (of een kolomcel van een spreadsheet) worden ingevoerd met een onderstrepingsteken _ (zonder aanhalingstekens) of met het woord **void**

dim(*h2*)	367
mean(*h2*)	28681/354
sum(*h2*)/367	28681/367
isVoid(*h2*)	{false,false,false,false,false,false,false,false,false,false,false,false,false,false,fals▸
countIf(isVoid(*h2*),?=false)	354
countIf(isVoid(*h2*),?=true)	13
h2	{90,97,78,83,90,85,85,92,68,90,86,63,93,78,97,95,88,79,89,93,95,79,66,89,74,7▸

4) Kwantitatieve gegroepeerde data

Voorbeeld 7:

Hier volgen de frequenties voor hoofdstuk 4 van de redders voor de klassen

[0,10[, [10,20[, [20,30[, [30,40[, [40,50[, [50,60[, [60,70[, [70,80[, [80,90[, [90,100[:

2 , 7 , 11 , 14 , 27 , 58 , 70 , 100 , 67 , 11

→ Maak een histogram uitgaande van deze frequentietabel, bepaal tevens de relatieve en de cumulatieve relatieve frequenties. Teken een cumulatieve relatieve frequentiepolygoon.

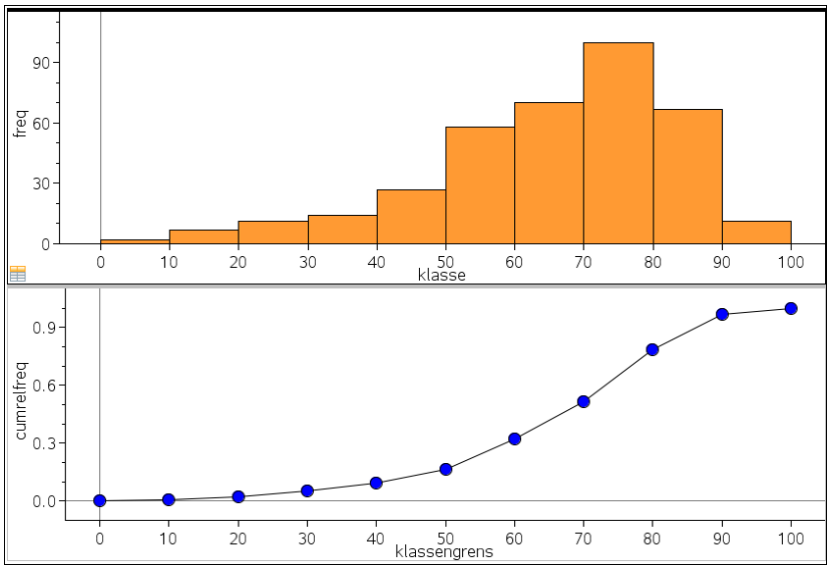
	A klassengrens	B klasse	C freq	D relfreq	E cumrelfreq
◆	=seq(10*k,k,0,10)			=round(freq/(sum(freq)),3)	
1	0	5	2	0.005	0
2	10	15	7	0.019	0.005
3	20	25	11	0.03	0.024
4	30	35	14	0.038	0.054
5	40	45	27	0.074	0.092
6	50	55	58	0.158	0.166
7	60	65	70	0.191	0.324
8	70	75	100	0.272	0.515
9	80	85	67	0.183	0.787
10	90	95	11	0.03	0.97
11	100				1.

De tweede kolom van de spreadsheet bevat de lijst “klasse” der klassenmiddens.
Om deze lijst te maken kan men als volgt te werk gaan: klik op cel B1 en typ

= (a1+a2)/2

gevolgd door enter (zoals in Excel start een formule met een gelijkheidsteken).
Klik vervolgens op cel B1 en beweeg de cursor naar de rechterbenedenhoek tot er een plusteken verschijnt, hou dan de linkermuistoets ingedrukt en beweeg de cursor naar beneden om de formule uit te breiden.

In kolom E komen de cumulatieve relatieve frequenties: typ eerst 0 in cel E1, in cel E2 komt de formule **= e1+ d1**, breid die formule vervolgens uit naar beneden.



Om de polygoon te tekenen: zet horizontaal de variabele klassengrens en verticaal de variabele cumrelfreq uit, verbind vervolgens de punten (keuze via rechtermuisklik).

5) Spreidingsdiagrammen en regressie

Voorbeeld 8:

Hier volgt een tabel van het percentage van de huishoudens in België met toegang tot het internet thuis (bron NIS)

jaar	2005	2006	2007	2008	2009
percentage	50	54	60	64	67

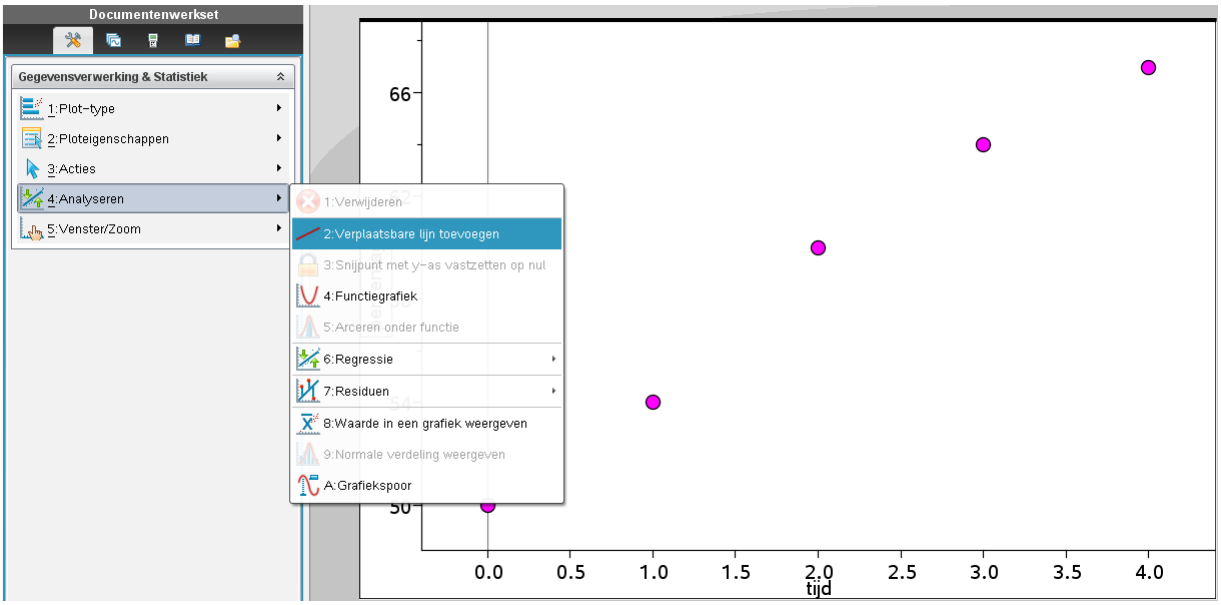
→ Maak een spreidingsdiagram van deze data.

Open een lijsten & spreadsheet pagina en voer eerst de gegevens in (tijdstip 0 stemt overeen met jaar 2005).

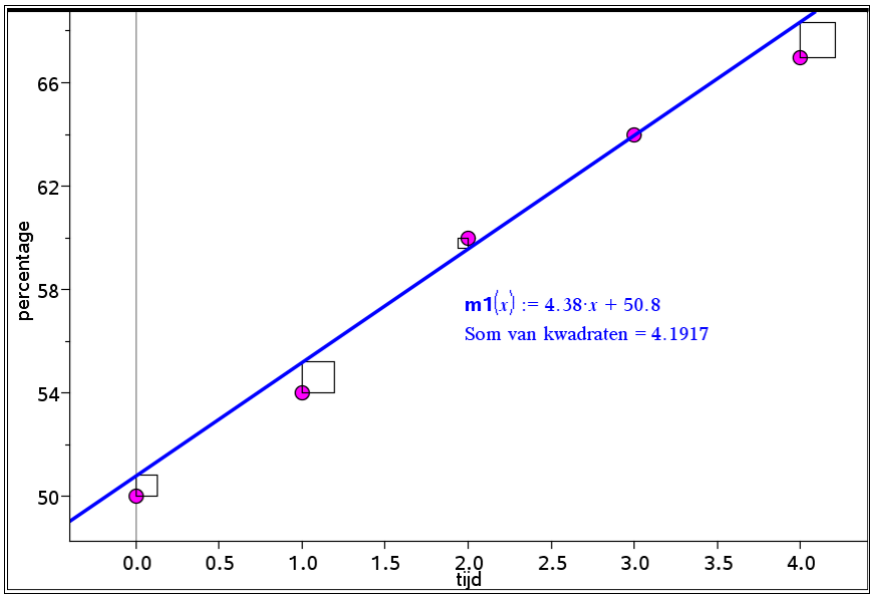
	A tijd	B percentage
◆		
1	0	50
2	1	54
3	2	60
4	3	64
5	4	67

Zet de tijd uit op de horizontale as en het percentage op de verticale as in een gegevensverwerking en statistiek pagina.

Voeg via het menu **analyseren** een **verplaatsbare lijn** toe

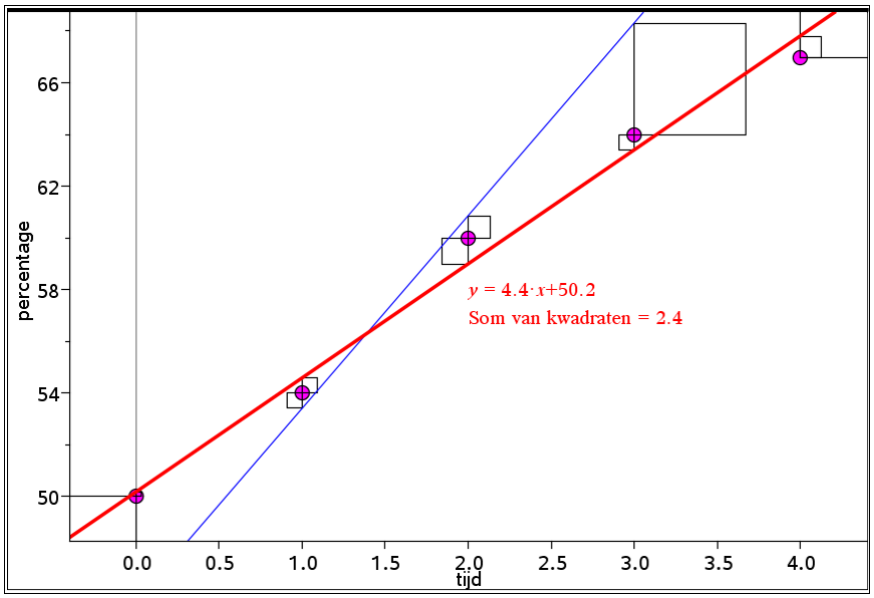


Kies vervolgens het menu **analyseren→residuen→residukwadraten weergeven**

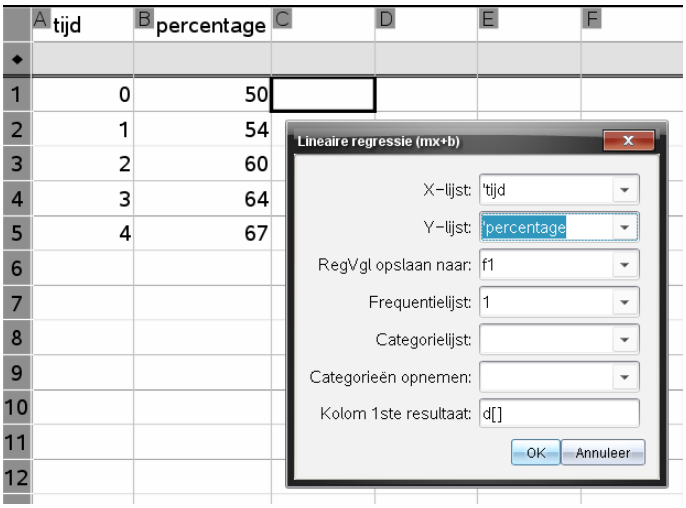


Verplaats de lijn (rotatie door aanwijzen van de lijn op een uiteinde en linkermuistoets blijven induwen, translatie door aanwijzen van de lijn halverwege). Ga zo experimenteel op zoek naar de rechte met de kleinste som van de residukwadraten (d.i. kleinste kwadraten rechte).

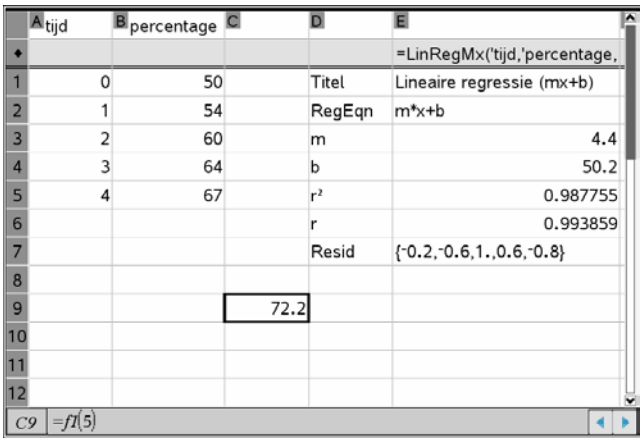
Vergelijk tenslotte uw "beste rechte" met de kleinste kwadraten rechte via het menu **analyseren→regressie→lin.regressie weergeven (mx+b)** en vervolgens **analyseren→residuen→residukwadraten weergeven**



Keer terug naar de gegevensverwerking en statistiek pagina (met de paginasorteerder of indrukken van de toetsen **Ctrl** en ←).
Klik in een cel van kolom C en kies het menu
statistieken→statistiekberekeningen→lineaire regressie (mx+b)



De determinatiecoëfficiënt r^2=98,8% is prima. De regressievergelijking wordt bewaard in de functie f1, voorspel het percentage in 2010 met de formule **=f1(5)** in cel C9.



Voorbeeld 9:

Een grafiek van een tijdreeks is een spreidingsdiagram met de meetwaarden van een kwantitatieve variabele op de verticale as in functie van de tijd op de horizontale as, waarbij de opeenvolgende punten worden verbonden door lijnstukken. Op die wijze kan men de trend of de evolutie van een variabele in de tijd bestuderen.

Onderstaande tabellen geven de evolutie van het totaal aantal huwelijken en echtscheidingen in België voor de jaren 1995 tot 2009.

Aantal huwelijken per gemeente, 1995 - 2009

Refnis	ADMINISTRATIEVE EENHEDEN	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Totaal		51.402	50.552	47.759	44.393	44.171	45.123	42.110	40.434	41.777	43.296	43.141	44.813	45.561	45.613	43.303

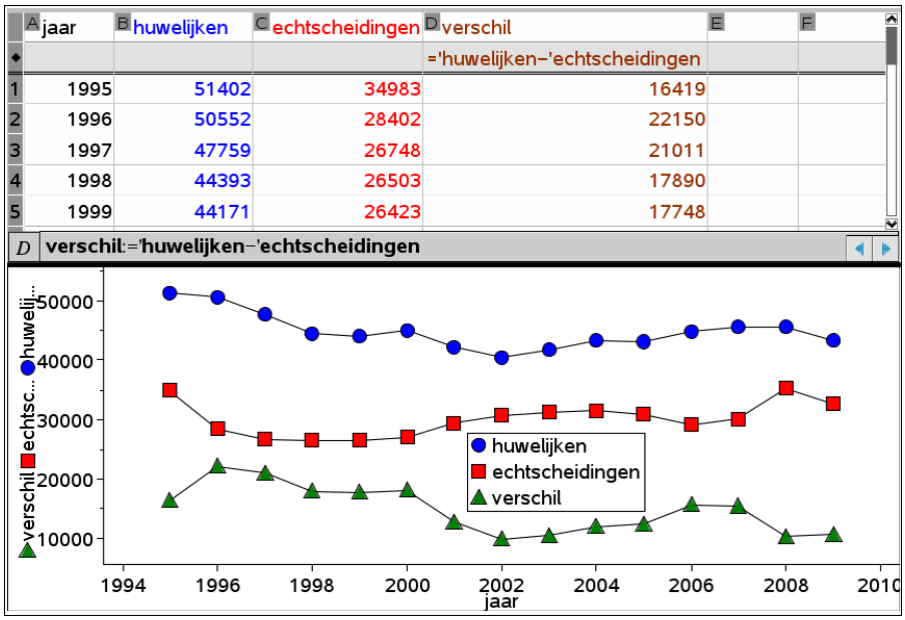
Bron (verplichte vermelding): Algemene Directie Statistiek en Economische Informatie - Thematische Directie Samenleving.

Aantal echtscheidingen per gemeente, 1995 - 2009

Refnis	ADMINISTRATIEVE EENHEDEN	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Totaal		34.983	28.402	26.748	26.503	26.423	27.002	29.314	30.628	31.355	31.405	30.840	29.189	30.081	35.366	32.606

Bron (verplichte vermelding): Algemene Directie Statistiek en Economische Informatie - Thematische Directie Samenleving.

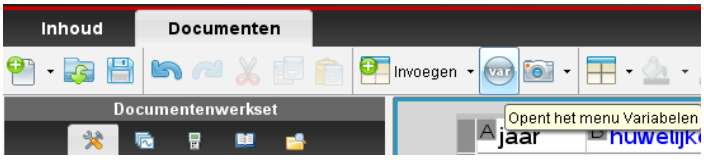
→ Stel de tijdreeksen grafisch voor en zet ook het verschil van de data uit.



Na het invoeren van de naam **verschil** voor kolom D verschijnt automatisch **verschil:=** door te klikken op de grijze formulecel onder de naam.
Vul de formule verder aan met

verschil:= huwelijken – echtscheidingen

Tip: de namen van de variabelen kan men opvragen in een lijst via de toets  onder het documentenmenu, hiermee kan men de namen snel invoegen in een formule.



Deel 2: simuleren met TI-Nspire

1) Lukrake getallen genereren

Open een rekenmachine pagina. Onder het menu **kansen, willekeurig**, vindt men de commando's om lukrake getallen te genereren.

Typ eerst het commando **randseed** gevolgd door een spatie en een natuurlijk getal, om een nieuwe reeks van lukrake getallen te beginnen.

Voorbeelden om lukrake getallen te genereren (cfr. TI-84 Plus commando's):

→ **rand( )** : een lukraak "reëel" getal tussen 0 en 1
zo geeft $3 + 5 \cdot \text{rand}()$ een lukraak getal uit het interval [3 , 8]

→ **rand(10)** : een lijst van 10 getallen tussen 0 en 1.

→ **randint(1,6)** : een natuurlijk getal tussen 1 en 6 (een dobbelsteen werpen).

→ **randint(1,6,20)** : een lijst van 20 natuurlijke getallen tussen 1 en 6.

→ **randbin(10,0.5)** : het aantal keer kop (succes) bij 10 keer werpen van een correct muntstuk (met kans op succes 0.5)

→ **randbin(10, 0.5, 20)** : een lijst van 20 resultaten van dergelijke binomiaalexperimenten.

→ **randnorm(175, 10)** : een lukraak getal uit een normale verdeling met gemiddelde 175 en standaardafwijking 10.

→ **randnorm(175, 10, 50)** : een steekproef (lijst) van 50 getallen uit een normale verdeling met gemiddelde 175 en standaardafwijking 10.

RandSeed 834	*Gereed*
randInt(1,6,10)	{5,5,6,3,5,5,5,5,1,5}
randInt(1,6,10)	{2,3,1,3,4,2,6,6,1,3}
randBin(10,0.5,20)	{3,7,6,4,6,3,5,9,8,4,4,4,6,6,6,3,9,4,7,5}
3+5·rand(4)	{6.78668,3.18189,3.97924,5.4059}

2) Steekproeven met en zonder terugleggen

Een steekproef wordt met het commando **randsamp** getrokken uit een lijst van numerieke of categorische data. De syntax is:

zonder terugleggen: **randsamp(lijstnaam, steekproefgrootte, 1)**
met terugleggen: **randsamp(lijstnaam, steekproefgrootte)**

Een lukrake steekproef trekken zonder terugleggen:
randsamp(lijstnaam, steekproefgrootte , 1)
Voorbeeld: een lottotrekking
lotto:=seq(k,k,1,42)
randSamp(**lotto**,6,1) ▸ {41,14,36,26,17,31}

Een lukrake steekproef met terugleggen:
randsamp(lijstnaam, steekproefgrootte)
Voorbeeld: 10 keer het roulette–rad draaien
roulette:=seq(k,k,0,36)
▸ {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36}
randSamp(**roulette**,10) ▸ {35,15,10,29,20,11,27,10,24,16}

Voor de spelregels van roulette en de bijhorende winstkansen zie [5].

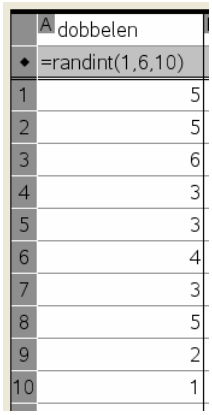
Het bovenstaande scherm werd gemaakt in een **toepassing notities**.

De toepassing notities is zeer nuttig:

- Men kan tekst met wiskundige uitdrukkingen (zoals in een rekenmachinetoepassing) combineren. Een wiskundige uitdrukking wordt geschreven in een math box of wiskunde-vak , dit kan men activeren door de toetsen **Ctrl** en **M** gelijktijdig in te drukken. Men kan de uitwerking van een uitdrukking verbergen: rechterklik op het wiskunde-vak, kies kenmerken wiskunde-vak.
- De toepassing notities is ook dynamisch (in tegenstelling tot de statische rekenmachine toepassing): wijziging van een definitie veroorzaakt een onmiddellijke herberekening van alle uitdrukkingen in de notitiepagina.
- Een simulatie (een uitdrukking waarin **rand**… optreedt) wordt automatisch opnieuw uitgevoerd door op de definitie van de uitdrukking te klikken en op enter te duwen (of gelijktijdig **Ctrl enter** voor snelle opeenvolgende herhalingen).

De toepassing notities is dus een (elementaire) tekstverwerker samen met een dynamische rekenmachine toepassing en een omgeving waarin simulaties snel kunnen worden herhaald.

Tip: als men in een spreadsheet toepassing een kolom (dit is één lijst) definieert met een simulatie-commando, dan wordt die simulatie ook daar vernieuwd door in de spreadsheet te klikken en de toetsen **Ctrl** samen met **R** in te drukken.

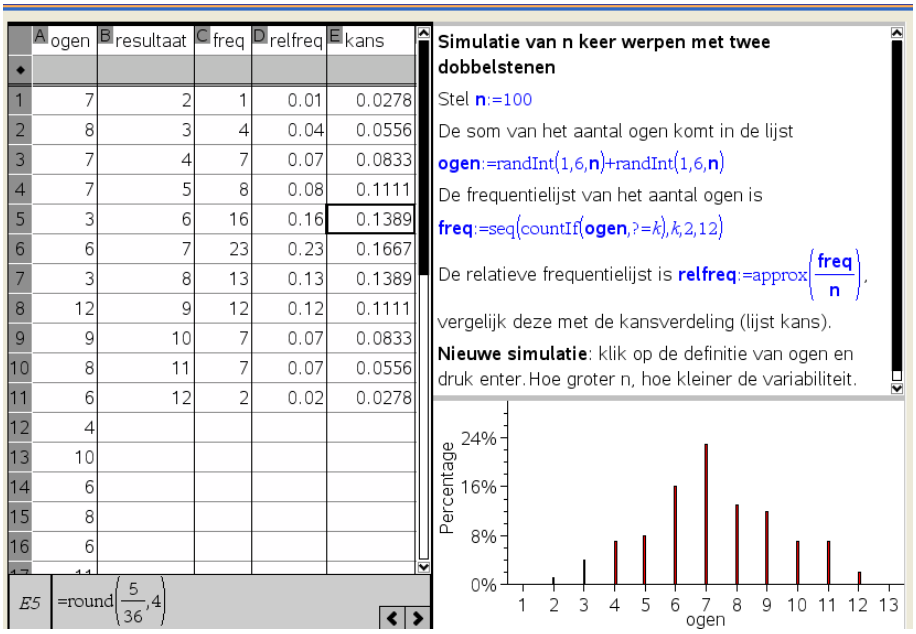


	A dobbelen
◆	=randint(1,6,10)
1	5
2	5
3	6
4	3
5	3
6	4
7	3
8	5
9	2
10	1

3) Twee dobbelstenen werpen

De pagina wordt opgesplitst in drie toepassingen: lijsten en spreadsheet, notities, gegevensverwerking en statistiek. De steekproefgrootte **n** en de simulatie **ogen** worden gedefinieerd bij de notities. De spreadsheet en het staafdiagram wijzigen dynamisch mee bij vernieuwing van de simulatie en bij keuze van een andere waarde voor **n**.

Onderzoek de invloed van de steekproefgrootte op de steekproefvariabiliteit en de “convergentie” naar het kansmodel voor twee dobbelstenen.



Het is sapristi jammer dat de geschreven tekst in dit cahier statisch is …

4) Kansverdeling van het steekproefgemiddelde

Hier volgt het principe om een benadering te vinden voor het kansmodel van het steekproefgemiddelde.

- simuleer een groot aantal steekproeftrekkingen uit een populatie met een discrete of continue kansverdeling.
- noteer bij elke steekproef telkens het steekproefgemiddelde in een groeiende lijst, via automatische gegevensvastlegging in een spreadsheet toepassing.
- een puntenplot van de lijst laat zien of het kansmodel voor het steekproefgemiddelde discreet of continu zal zijn.
- observeer grafisch de convergentie van de procentuele verdeling van het staafdiagram voor een discreet model of de convergentie van het histogram op dichtheidsschaal (met voldoend kleine klassenbreedte) voor een continu model naar een stabiele situatie.
- de “evenwichtssituatie” toont een benadering van de kansverdeling van het steekproefgemiddelde.

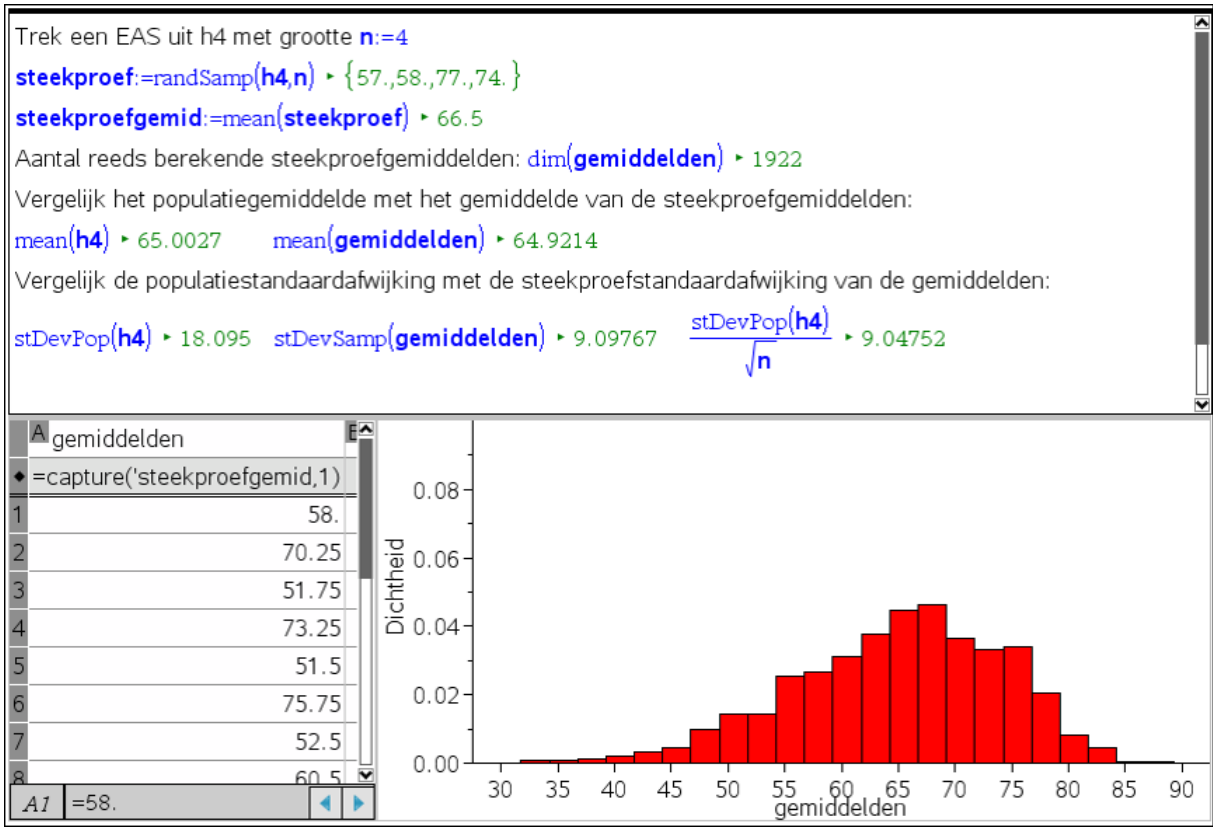
Hierbij is het didactisch aangewezen om een steekproef te trekken uit een concrete gegeven populatie van getallen, zoals de resultaten van hoofdstuk 4 voor de redders, een populatie van 367 data met een links scheve verdeling (zie pagina 18).

Dit is immers voor de student concreter dan een trekking uit een denkbeeldige populatie, waarbij de data gegenereerd worden door een proces dat aangestuurd wordt door het kansmodel van de populatie (bijvoorbeeld de uniforme kansverdeling bij het werpen van één dobbelsteen).

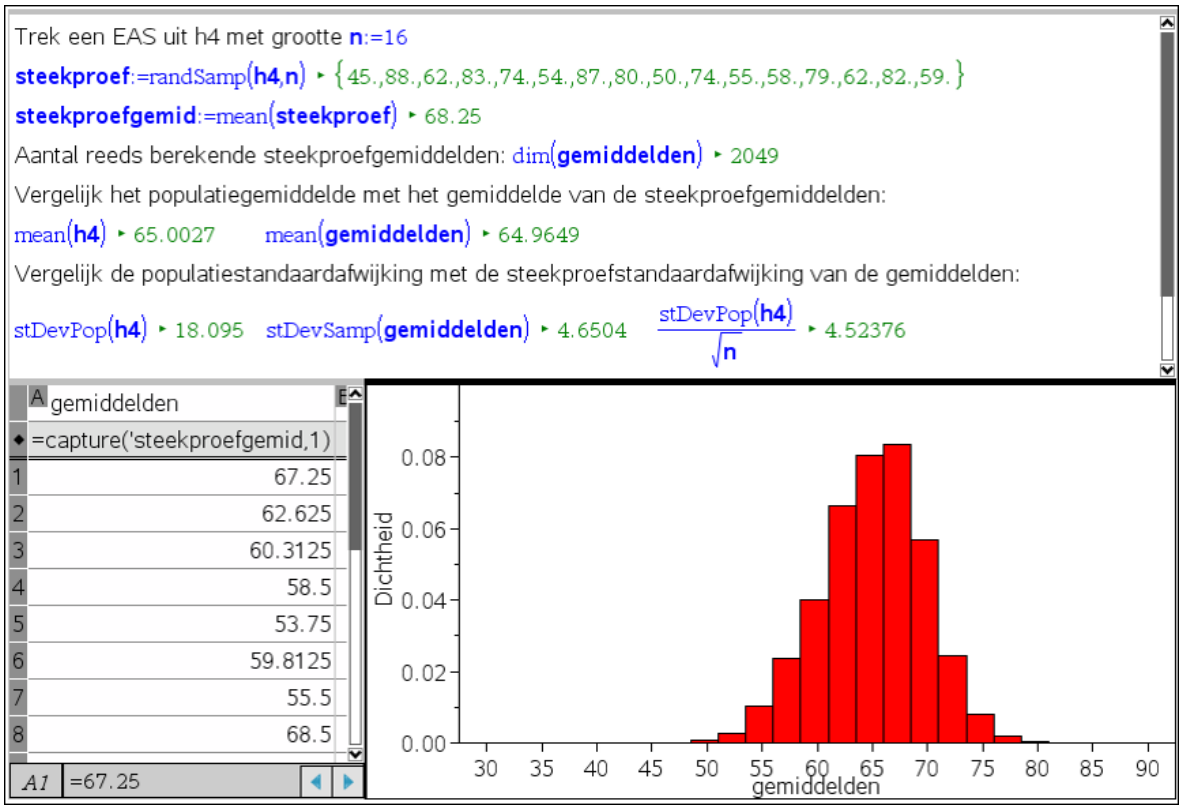
Voer de gegevens in zoals hieronder staat aangegeven in de notitietoepassing.
Soms is het nodig om een commando, bijvoorbeeld **mean(h4)**, uit te voeren met **Ctrl enter** i.p.v. met enter alleen; hierdoor verschijnt het resultaat decimaal i.p.v. exact in breukvorm.

Het gemiddelde van elke steekproef wordt automatisch toegevoegd aan de lijst gemiddelden, die gedefinieerd wordt in een spreadsheet toepassing: klik hiertoe op de formulecel en gebruik het menu **gegevens→gegevensvastlegging→automatisch** , wijzig tenslotte **capture(var,1)** in **capture(steekproefgemid,1)**.

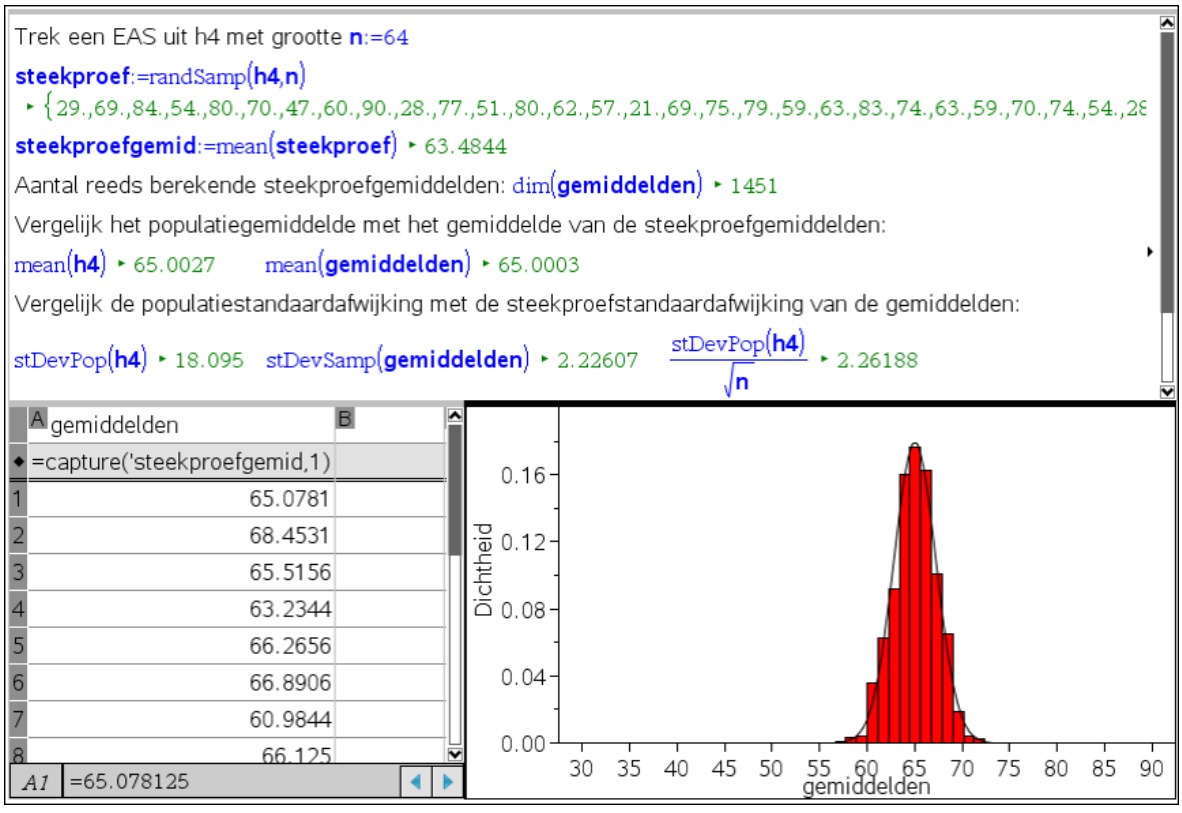
Steekproefsimulaties worden verkregen door te klikken op het commando **steekproef:=randsamp(h4,n)** (de rode math box verschijnt) en door telkens op **enter** te drukken. De steekproefgemiddelden variëren van steekproef tot steekproef (een illustratie van de steekproefvariabiliteit). Snelle herhaaldelijke steekproefsimulaties verkrijgt men door te klikken op het commando **steekproef:=randsamp(h4,n)** en vervolgens de toetsen **Ctrl enter** samen te blijven indrukken. Observeer hoe het dichtheidshistogram evolueert naar een "stabiele" toestand.



Bij steekproefgrootte 4 is de kansverdeling van $\overline{X}$ minder scheef dan die van de populatie. Simulatie van 1922 steekproefgemiddelden. Hun gemiddelde 64,92 is nagenoeg gelijk aan het populatiegemiddelde 65,00. De spreiding is nagenoeg gehalveerd!



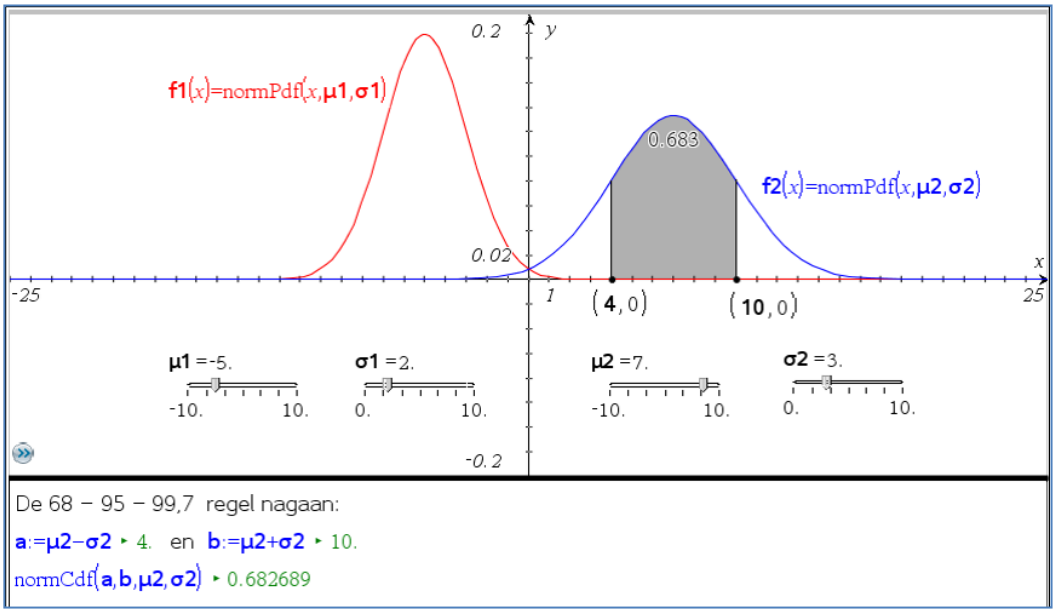
Steekproefgrootte 16, simulatie van 2049 steekproefgemiddelden



Steekproefgrootte 64, illustratie van de centrale limietstelling!

De lezer wordt uitgenodigd om te experimenteren met steekproeftrekkingen uit andere (discrete of continue) populaties! Dit zal leiden tot het vermoeden dat het kansmodel van de populatie der steekproefgemiddelden, bij voldoend grote steekproefgrootte, kan benaderd worden door een klokvormig model: de normale verdeling.
Dit vermoeden wordt bevestigd door de centrale limietstelling.

Het klokvormig kansmodel van een normaal verdeelde populatie ligt volledig vast door het populatiegemiddelde μ en de populatiestandaardafwijking σ. De invloed van deze parameters op de grafiek van de dichtheidsfunctie wordt geïllustreerd met schuifbalken :



Werkwijze om de bovenstaande pagina te realiseren:

- Splits de pagina in een grafiektoepassing en een notitietoepassing.
- Klik op de grafiektoepassing en voer de vier schuifbalken in met het menu **acties →schuifknop invoegen**. De Griekse letters staan bij de symbolen onder de hulpprogramma's van de documentenwerkset.
- Definieer de functies f1(x) en f2(x) onderaan op de invoerregel, deze verschijnt door te klikken op .
- Definieer 2 punten op de x-as met het menu **punten en lijnen → punt op** .
  Druk daarna op de **Esc** toets om uit het puntenmenu te gaan.
- Bepaal de coördinaten van de punten met het menu **acties →coördinaten en vergelijkingen**.
- Bepaal de oppervlakte onder de curve tussen de twee punten met het menu **grafiek analyseren → integraal** , wijs hierbij de punten als grenzen aan.
- Klik op de notitietoepassing en definieer de variabelen a en b zoals aangegeven.
- Klik op de grafiektoepassing, beweeg de cursor naar de x-coördinaat van het eerste punt (het woord "tekst" verschijnt). Met een rechtermuisklik op die plaats en het menu **variabelen →koppelen aan**, kan men de x-coördinaat koppelen aan variabele a (de x-coördinaat van het punt wordt gelijk aan a). Ga analoog te werk voor het tweede punt.

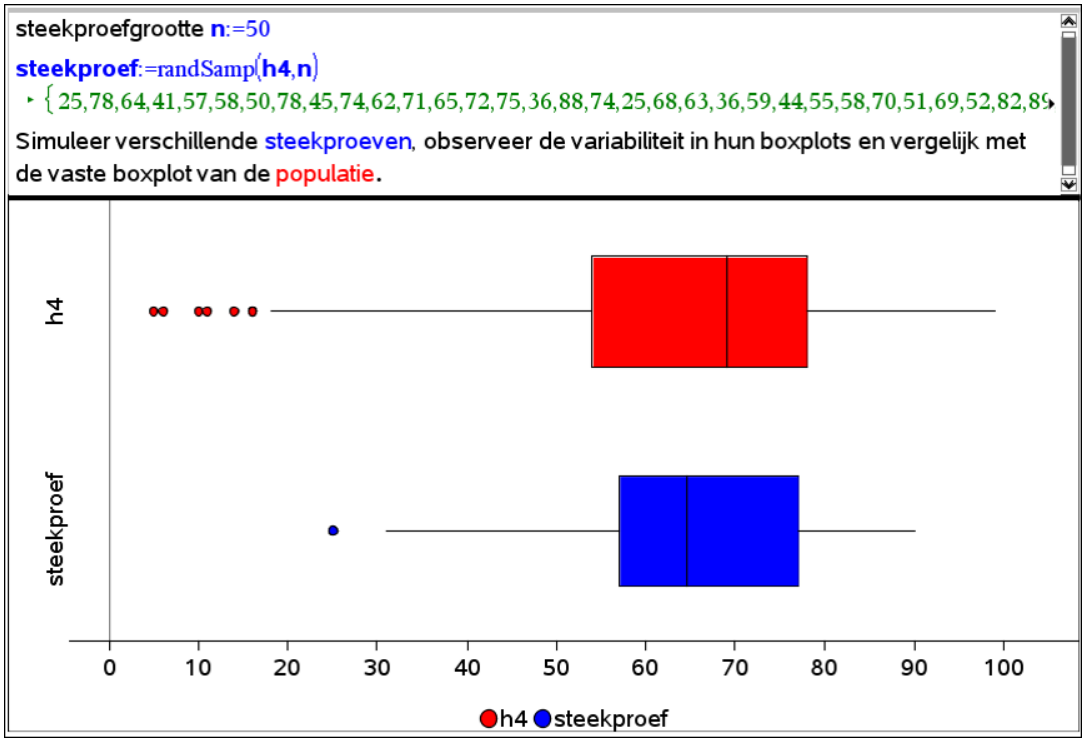
Tip: onder het menu **statistieken → verdelingen** (in een toepassing rekenmachine of spreadsheet) of het menu **berekeningen →statistieken → verdelingen** (in een toepassing notities) vindt men de verschillende kansverdelingen.
Als $X \sim N(\mu,\sigma)$, dan is
→ normpdf(x,μ,σ) = f(x) , met f de normale dichtheidsfunctie
→ normcdf(a,b,μ,σ)= P(a<X<b)
→ invnorm(a,μ,σ)= x met a = P(X<x)

5) Steekproefvariabiliteit bij boxplots

Het loont zeker de moeite om de steekproefvariabiliteit grafisch te observeren bij de boxplots van de opeenvolgende steekproeven.

Bij een kleine steekproefgrootte zullen de boxplots erg variëren en ze kunnen erg afwijken van de boxplot van de populatie.
Naarmate de steekproefgrootte toeneemt, schommelen de boxplots van de opeenvolgende steekproefsimulaties minder en minder, deze door het toeval verkregen boxplots zullen ook beter aansluiten bij de boxplot van de populatie.

Beschouw opnieuw als populatie de resultaten van hoofdstuk 4 voor de redders (367 data).
Hieronder worden lukrake steekproeven van grootte *n* (met terugleggen) gesimuleerd .



Deel 3: kansverdelingen ontdekken

Dit deel illustreert hoe men nieuwe kansverdelingen kan ontdekken via simulatie.

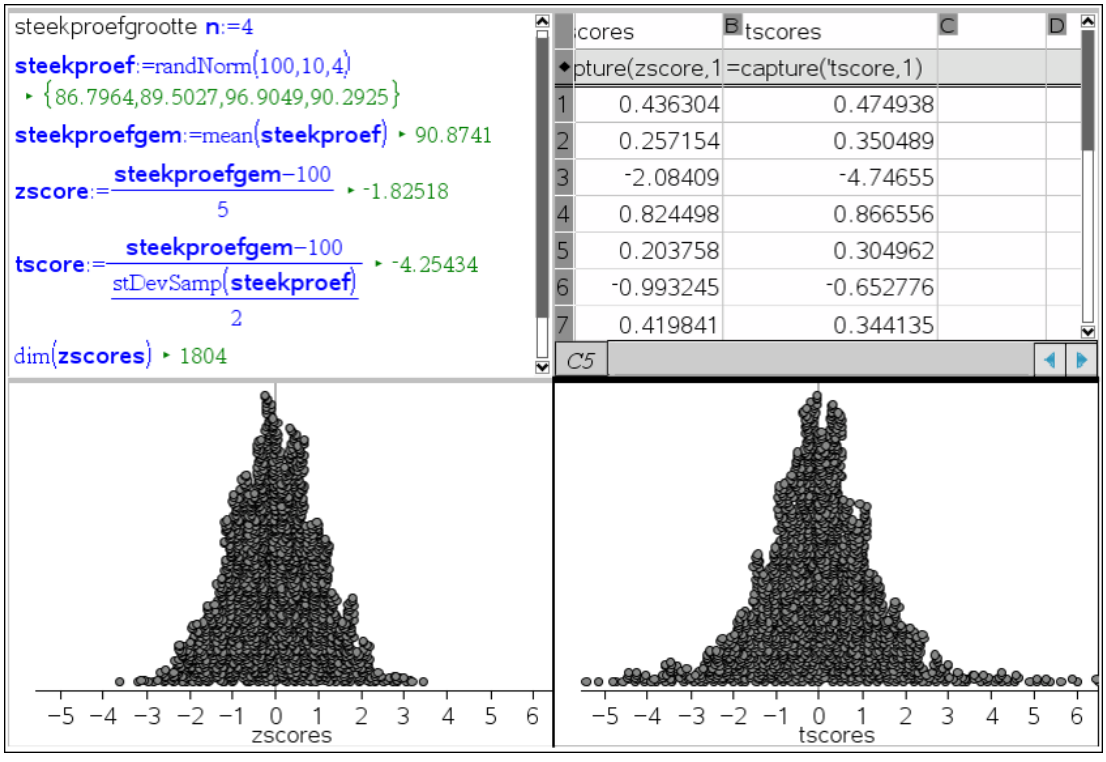
1) z-scores versus t-scores

Stel $X \sim N(\mu,\sigma)$ en trek uit deze populatie een steekproef met grootte *n*, dan geldt voor het steekproefgemiddelde: $\overline{X} \sim N\left(\mu,\frac{\sigma}{\sqrt{n}}\right)$.

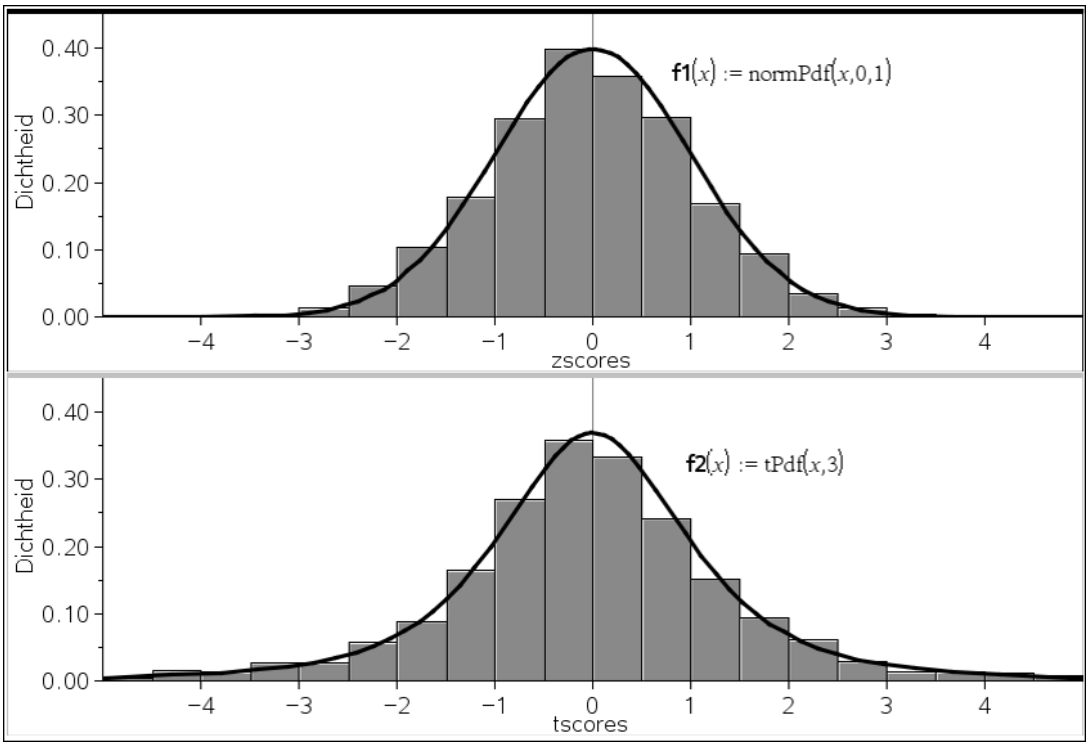
Standaardiseren levert $Z=\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$.

Als de populatiestandaardafwijking σ niet gekend is, dan wordt ze geschat met de standaardafwijking s. Zo ontstaat de nieuwe toevalsvariabele $T=\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$.

We vergelijken de "z-scores" $z=\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ met de "t-scores" $t=\frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$ via hun dichtheidshistogrammen in een concrete simulatie met populatiegemiddelde $\mu=100$, populatiestandaardafwijking $\sigma=10$, steekproefgrootte $n=4$.



De t-scores vertonen een grotere spreiding dan de z-scores.



De t-scores zijn verdeeld volgens een t-verdeling met 3 vrijheidsgraden ($3 = n - 1$), wel klokvormig maar niet normaal verdeeld!

2) Functies van één toevalsvariabele

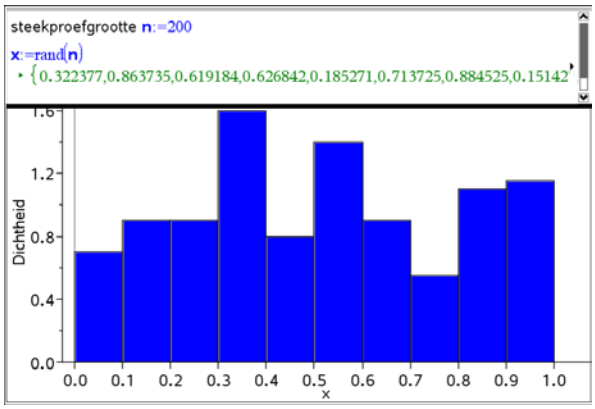
Kies lukraak een reëel getal *X* in het interval [0,1].
De toevalsvariabele *X* (hoofdletter!) heeft een uniforme kansverdeling met als dichtheidsfunctie

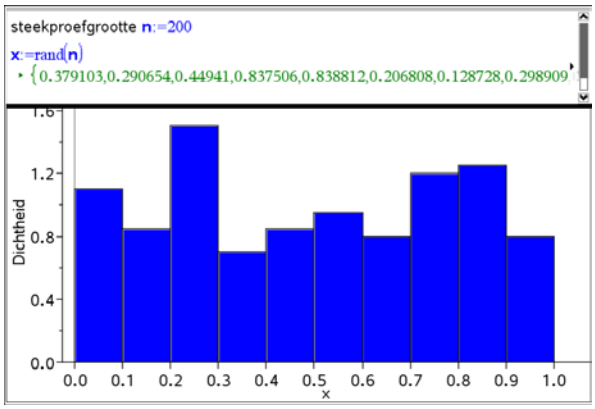
$$f(x) = \begin{cases} 1 \text{ als } 0 \le x \le 1 \\ 0 \text{ elders} \end{cases}$$

Telkens er een lukraak getal wordt gekozen tussen 0 en 1, neemt de toevalsvariabele *X* één of andere concrete getalwaarde *x* (kleine letter!) aan, aangestuurd volgens het kansmechanisme van de dichtheidsfunctie *f*.

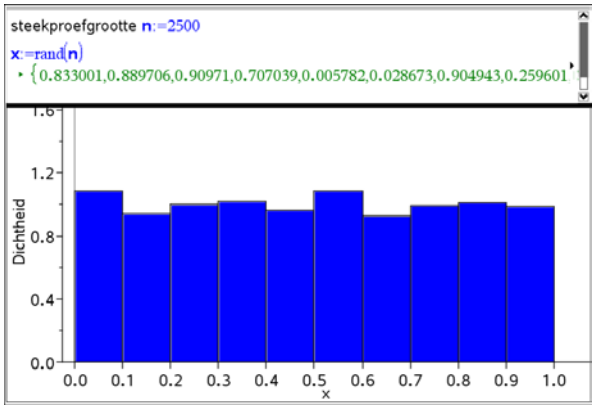
Door zeer vaak een lukraak getal te kiezen tussen 0 en 1 en deze zo verkregen getallen voor te stellen met een dichtheidshistogram, zal dit histogram meer en meer stabiliseren en "convergeren" naar de dichtheidsfunctie (kansen zijn relatieve frequenties op de lange duur: wet van de grote aantallen).

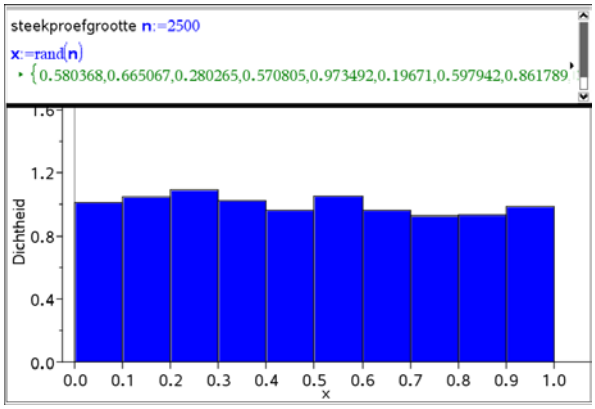
Bij 200 simulaties is er nog veel variatie in het dichtheidshistogram:





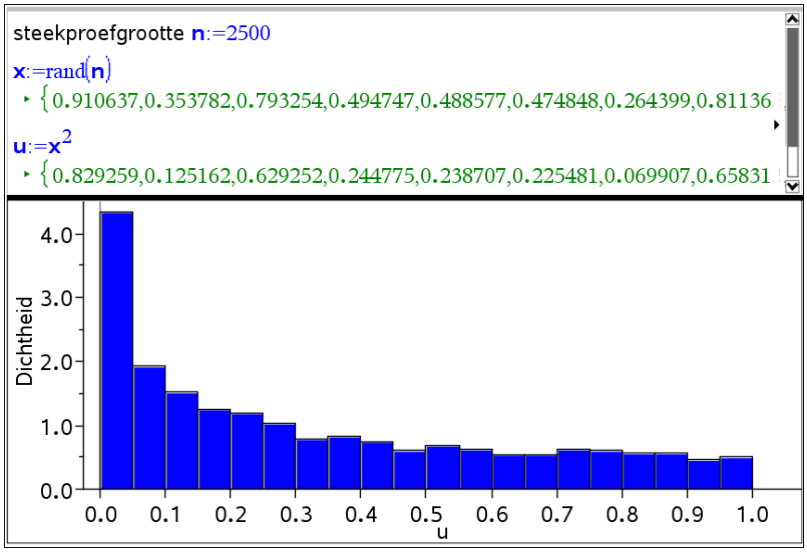
Bij 2500 simulaties is de variatie al veel kleiner: het dichtheidshistogram laat al meer de rechthoekvorm van de dichtheidsfunctie zien (bij een niet te kleine klassenbreedte).





Op deze wijze kan men door simulatie een goed idee krijgen van de kansverdeling van een toevalsvariabele.

Met de variabele X kan men nieuwe variabelen definiëren, bijvoorbeeld de toevalsvariabele $U = X^2$. De kansverdeling (de dichtheidfunctie) van U wordt bij benadering bepaald door het dichtheidshistogram van een groot aantal gegenereerde data $u = x^2$:



De dichtheidsfunctie van U is de afgeleide van de verdelingsfunctie F_U van U:
voor x tussen 0 en 1 geldt:

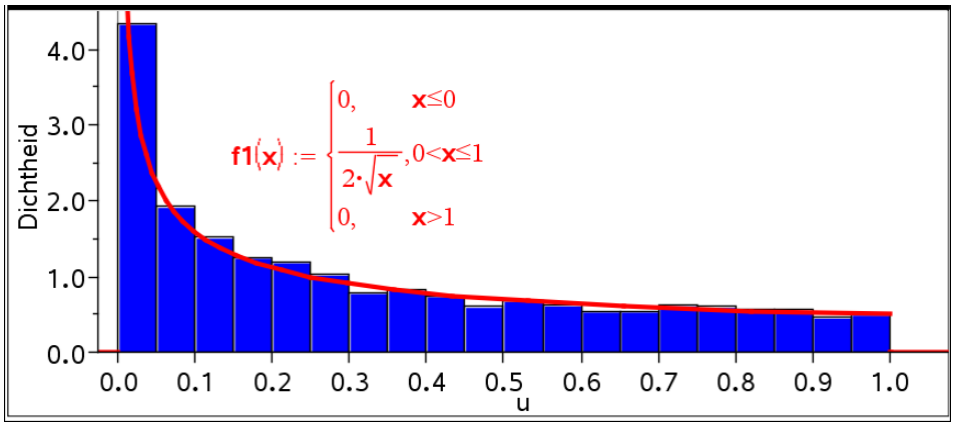
$$F_U(x) = P(U \le x) = P\left(X^2 \le x\right) = P\left(-\sqrt{x} \le X \le \sqrt{x}\right) = \int_0^{\sqrt{x}} 1 \cdot dt = \sqrt{x}$$

zodat de dichtheidsfunctie $f_U(x) = F_U'(x)$ van U gegeven wordt door:

$$f_U(x) = F_U'(x) = \begin{cases} \dfrac{1}{2\sqrt{x}} & \text{als } 0 < x \le 1 \\ 0 & \text{elders} \end{cases}$$

Maak een grafiek van de dichtheidsfunctie (menu → **analyseren→functiegrafiek**, het meervoudig voorschrift via het menu →**wiskundetemplates**→).

Het dichtheidshistogram sluit aan bij de dichtheidsfunctie:



Reken na in de notitietoepassing dat de (oneigenlijke) integraal van de dichtheidsfunctie over het interval [0 , 1] wel degelijk 1 oplevert:

$$\int_0^1 \frac{1}{2\cdot\sqrt{x}}\, dx = 1$$

2) Functies van twee toevalsvariabelen

Kies lukraak een reëel getal X en een reëel getal Y in het interval [0 , 1].
De toevalsvariabelen X en Y hebben elk een uniforme kansverdeling op [0 , 1].

Met deze twee variabelen kan men nieuwe toevalsvariabelen definiëren, zoals:
$S = X + Y$, $P = X \cdot Y$, $M = \max(X, Y)$, $K = \min(X, Y)$, ….

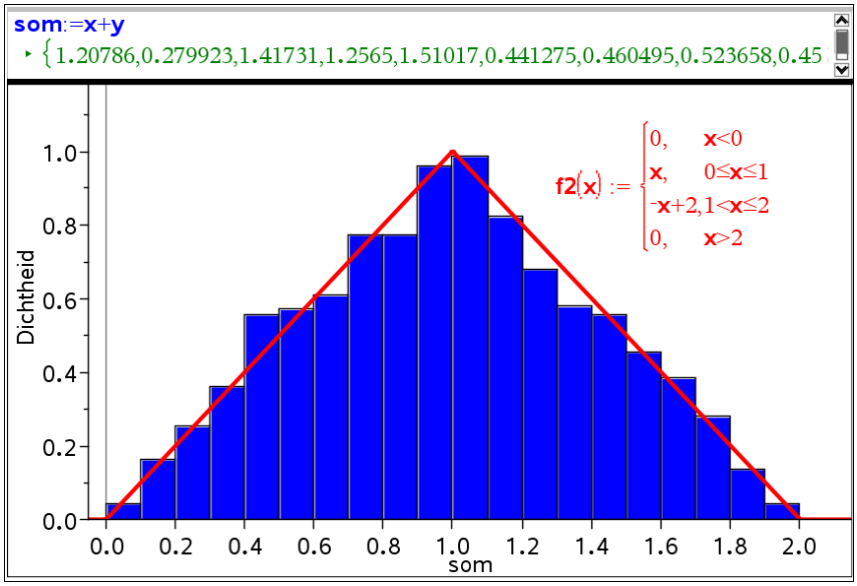
Wat is de kansverdeling voor deze nieuwe variabelen?

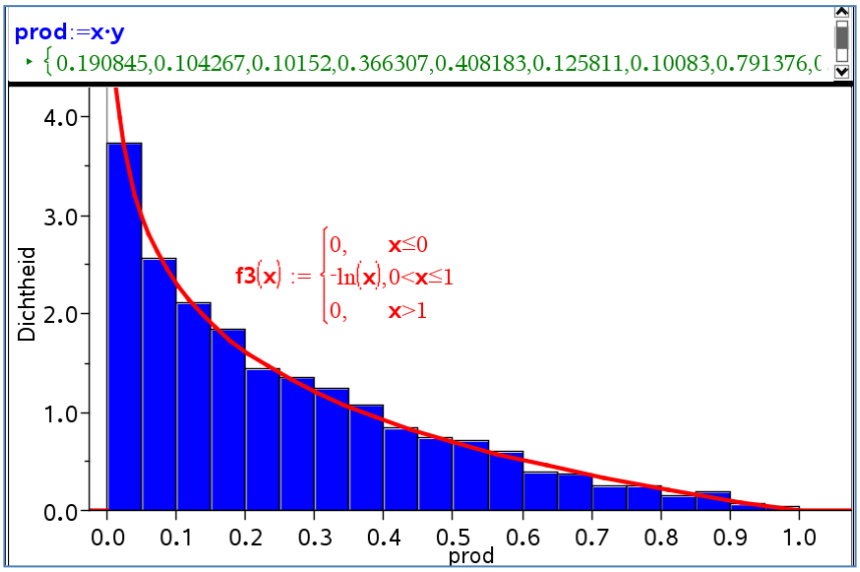
Neem als voorbeeld de toevalsvariabele $S = X + Y$. Na lukrake keuze van twee getallen tussen 0 en 1 levert dit een concrete som $s = x + y$. Herhaal dit experiment zeer vaak.
De verkregen getallen *x* en *y* worden aangestuurd volgens het uniforme kansmodel op [0,1] en de sommen wijzigen voortdurend, met waarden tussen 0 en 2. Het dichtheidshistogram van de gegenereerde sommen zal op de lange duur meer en meer aansluiten bij de dichtheidsfunctie of het kansmodel van S.

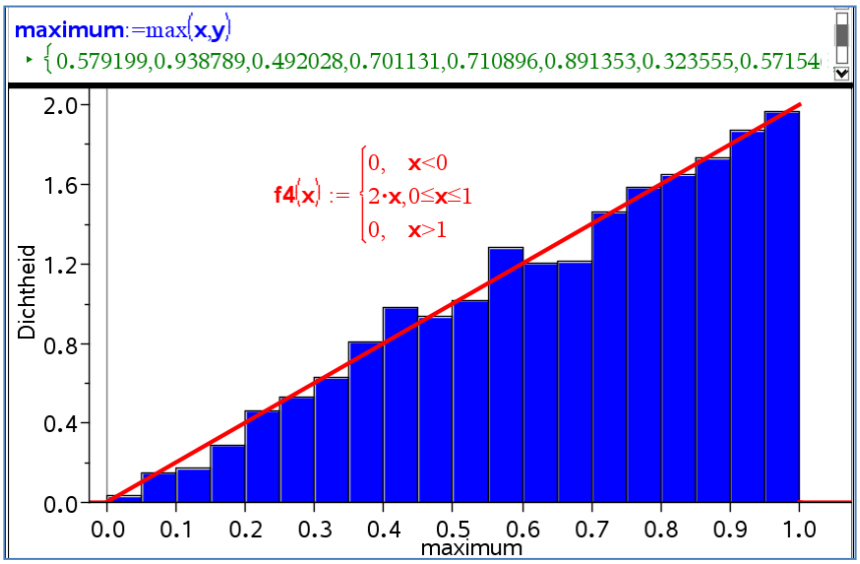
Ga analoog te werk voor de andere variabelen *P*, *M* en *K*.

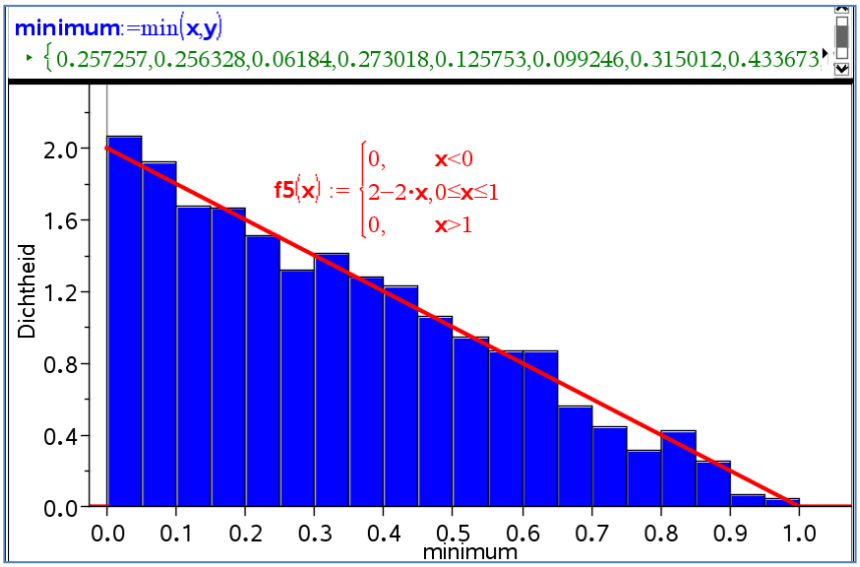
```
steekproefgrootte n:=3
x:=rand(n):y:=rand(n)  (twee commando's gescheiden door een dubbelpunt worden samen uitgevoerd na enter)
x ▸ {0.669211,0.208168,0.612423}
y ▸ {0.309759,0.130504,0.837721}
som:=x+y ▸ {0.97897,0.338672,1.45014}
prod:=x·y ▸ {0.207294,0.027167,0.51304}
maximum:=max(x,y) ▸ {0.669211,0.208168,0.837721}
minimum:=min(x,y) ▸ {0.309759,0.130504,0.612423}
```

Hier volgt het verkregen dichtheidshistogram bij steekproefgrootte *n* = 2500 , samen met een grafiek van de dichtheidsfunctie (ga zelf het voorschrift na):




prod:=x·y
{0.190845,0.104267,0.10152,0.366307,0.408183,0.125811,0.10083,0.791376,(
f3(x) := 0, x≤0
-ln(x), 0<x≤1
0, x>1
Dichtheid
prod
0.0
1.0
2.0
3.0
4.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


maximum:=max(x,y)
{0.579199,0.938789,0.492028,0.701131,0.710896,0.891353,0.323555,0.57154
f4(x) := 0, x<0
2·x, 0≤x≤1
0, x>1
Dichtheid
maximum
0.0
0.4
0.8
1.2
1.6
2.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


minimum:=min(x,y)
{0.257257,0.256328,0.06184,0.273018,0.125753,0.099246,0.315012,0.433673
f5(x) := 0, x<0
2−2·x, 0≤x≤1
0, x>1
Dichtheid
minimum
0.0
0.4
0.8
1.2
1.6
2.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Oefening:

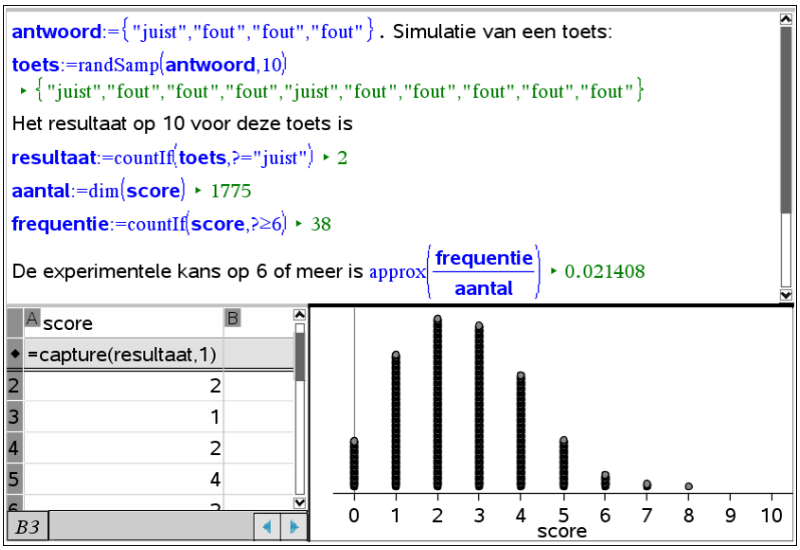
Beschouw de toevalsvariabele $K = \min(X, Y)$ van het laatste voorbeeld.
Simuleer een steekproef met grootte 2500 en bepaal het steekproefgemiddelde $\overline{x}$ (**mean**) , dit is een schatting voor de theoretische verwachtingswaarde $E(K) = \mu = \int_0^1 x \cdot (2-2x)\,dx$.
Bepaal tevens de steekproefstandaardafwijking *s* (**stdevsamp**) van de steekproef , dit is een schatting voor de theoretische standaardafwijking $\sigma = \sqrt{\mathrm{Var}(K)} = \sqrt{\int_0^1 (x-\mu)^2 \cdot (2-2x)\,dx}$.
Bereken μ en σ . Ga na of het steekproefgemiddelde en de steekproefstandaardafwijking inderdaad in de buurt liggen van μ en σ . Doe dit voor enkele steekproeven met grootte 2500. Herhaal dit voor enkele steekproeven met grootte 10. Conclusie?

Deel 4: statistische inferentie en simulaties

Voor de statistische achtergrond van dit onderwerp zie [6] [7] [8] [9] [10] [11].

Voorbeeld 1: is de student eerlijk?

Een toets bestaat uit 10 meerkeuzevragen met elk vier mogelijke antwoorden waarvan er één juist is. Een juist antwoord levert één punt op, een fout antwoord 0.
Een student haalt 6 op 10 voor die toets en beweert lukraak gegokt te hebben voor elke vraag. Gelooft u deze student?
Door deze toets dikwijls te simuleren en de score telkens te bewaren via automatische gegevensvastlegging (capture) , kan men onderzoeken hoe vaak een score 6 of hoger optreedt:



De experimentele kans voor een score van 6 of meer, na een simulatie van 1775 toetsen, is slechts 2,1%. Men kan dus veilig beweren dat de student liegt.

Stel X het juiste aantal antwoorden, bij lukraak gokken voor elke vraag, dan is X binomiaal verdeeld met parameters $n = 10$ en succeskans $p = \frac{1}{4}$.

Ter controle wordt de theoretische kansverdeling vergeleken met de experimentele:

	A	B theoretisch	C experimenteel	D	E	F
•		=binompdf(10,1/4)	=seq(approx(coun			
1	0	0.056314	0.067606			
2	1	0.187712	0.194366			
3	2	0.281568	0.247324			
4	3	0.250282	0.23662			
5	4	0.145998	0.164507			
6	5	0.058399	0.068169			
7	6	0.016222	0.017465			
8	7	0.00309	0.00338			
9	8	0.000386	0.000563			
10	9	0.000029	0.			
11	10	9.53674E-7	0.			

c **experimenteel**:=seq(approx(countif(**score**,?=k) / dim(**score**)),k,0,10)

De hypothesetest verloopt als volgt:

H_0: $p = \frac{1}{4}$ (de student heeft gegokt).

H_1: $p > \frac{1}{4}$ (de student heeft gestudeerd en niet gegokt, stel dat elke vraag dezelfde succeskans heeft).

De testvariabele is het aantal juiste antwoorden X, met $X \sim B\left(10, \frac{1}{4}\right)$ (redeneer in de veronderstelling dat H_0 waar is).

De geobserveerde waarde van de testvariabele is $x = 6$. De overschrijdingskans of p-waarde is $P(X \geq 6) = P(6 \leq X \leq 10) =$ binomcdf $\left(10, \frac{1}{4}, 6, 10\right) = 1{,}97\%$.

binomCdf$\left(10, \frac{1}{4}, 6, 10\right)$	0.019728
1−binomCdf$\left(10, \frac{1}{4}, 5\right)$	0.019728

Deze kans is klein, zodat H_0 wordt verworpen.

Voorbeeld 2: is de dobbelsteen correct?

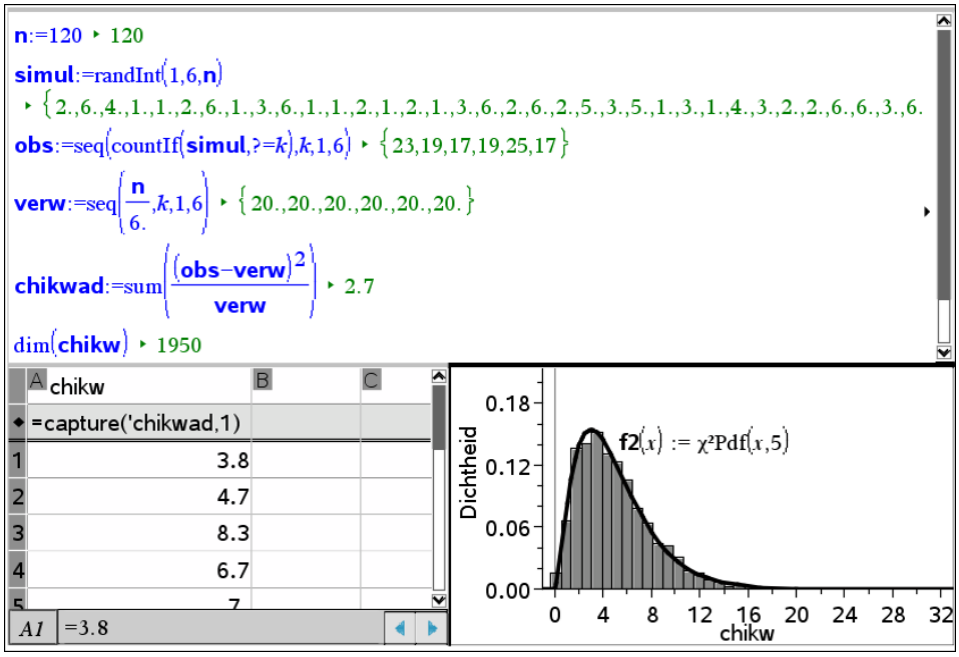
Bij 120 keer werpen van een eerlijke dobbelsteen verwacht men dat het aantal ogen 1, 2, 3, 4, 5, 6 elk ongeveer 20 keer zal optreden. In welke mate kunnen hiervan afwijkingen optreden? Een grootheid die een maat is voor de afwijking van de verwachte waarden is de

chi-kwadraat-maat: $\chi^2 = \sum \frac{(\text{geobserveerde waarde} - \text{verwachte waarde})^2}{\text{verwachte waarde}} = \sum_i \frac{(O_i - e_i)^2}{e_i}$

Als deze toevalsvariabele een kleine waarde oplevert, dan is er niets aan de hand met de dobbelsteen. Als men echter een grote chi-kwadraat-maat observeert, dan is het mogelijk dat de echte kansverdeling van de dobbelsteen niet overeenstemt met de verwachte discrete uniforme verdeling.

Om een idee te verkrijgen van de kansverdeling van de toevalsvariabele χ^2 volstaat het een groot aantal simulaties uit te voeren en de geobserveerde chi-kwadraat-waarden uit te zetten met een dichtheidshistogram, dat op de duur stabiliseert.

De verdeling is rechts scheef. Zoals blijkt uit de onderstaande grafische voorstelling levert de χ^2-verdeling met 5 vrijheidsgraden een goed model voor de kansverdeling.



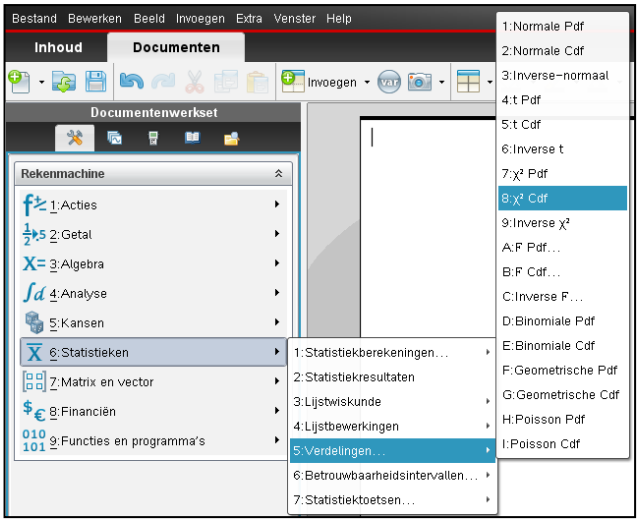
Werp nu zelf 120 keer een dobbelsteen om te testen of deze correct is. De resultaten zijn:

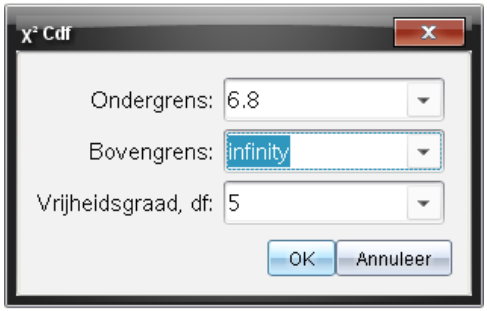
Aantal ogen	1	2	3	4	5	6
geobserveerd	17	12	23	18	25	15
verwacht	20	20	20	20	20	20

De geobserveerde χ^2-waarde wordt:

$$\chi^2 = \frac{(17-20)^2}{20} + \frac{(12-20)^2}{20} + \frac{(23-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(25-20)^2}{20} + \frac{(15-20)^2}{20} = 6.8$$

Is dit resultaat verdacht? De p-waarde is $P(\chi^2 \geq 6.8) = \text{chi2cdf}(6.8, \infty, 5) = 23.6\%$





$\chi^2\text{Cdf}(6.8, \infty, 5) = 0.235945$

De kans dat de χ^2-statistiek een waarde aanneemt die minstens even groot is als de geobserveerde waarde 6.8 is 23.6%, dit is niet uitzonderlijk.

De nulhypothese kan dus niet worden verworpen in deze "goodness-of-fit test":

H_0: de dobbelsteen heeft een uniforme discrete verdeling

versus

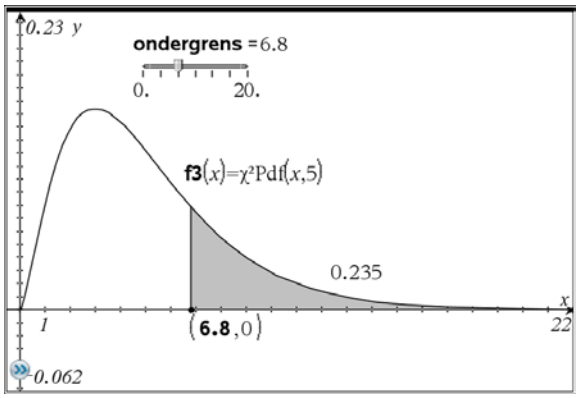
H_1: de kansverdeling van de dobbelsteen is niet uniform

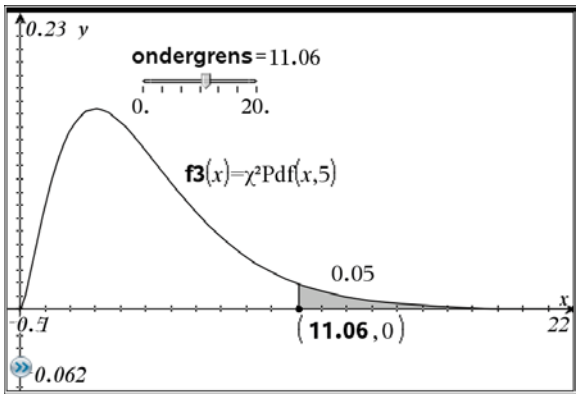
Vanaf welke geobserveerde χ^2-waarde zal men de nulhypothese verwerpen, met een significantieniveau $\alpha = 5\%$?

Deze kritieke waarde wordt verkregen met $\text{invchi2}(0.95,5) \simeq 11.07$:

$\text{inv}\chi^2(0.95,5) = 11.0705$

De kritieke waarde kan men ook grafisch dynamisch bepalen met een schuifknop:

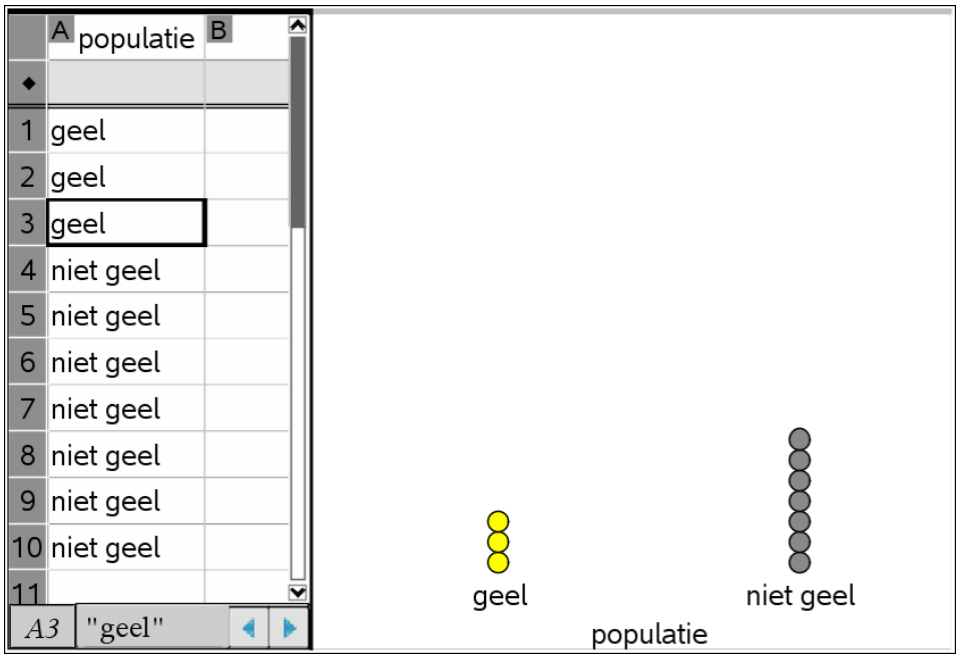


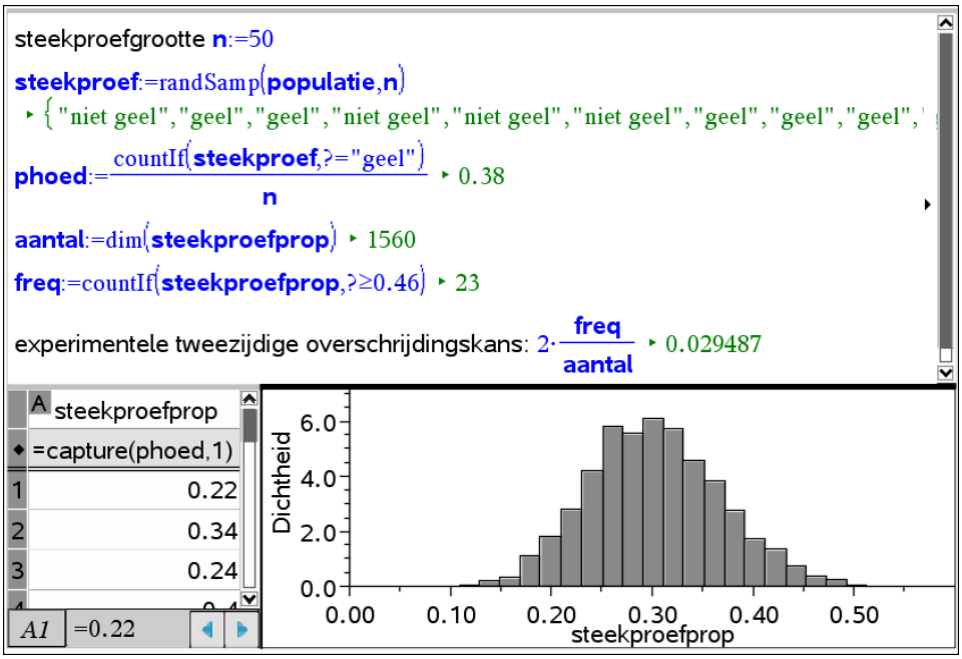


Voorbeeld 3: test op een proportie

Een snoepjesfabrikant beweert dat 30% van zijn snoepjes geel zijn. Je wenst dit na te gaan.
Onderzoek een pakje met 50 snoepjes, daarvan blijken er 23 geel te zijn of 46%!
Mag men hiermee besluiten dat de uitspraak van de fabrikant niet klopt?
Stel p de populatieproportie.
De hypothesetest is H_0: $p = 0.3$ versus H_1: $p \neq 0.3$ (tweezijdige test).

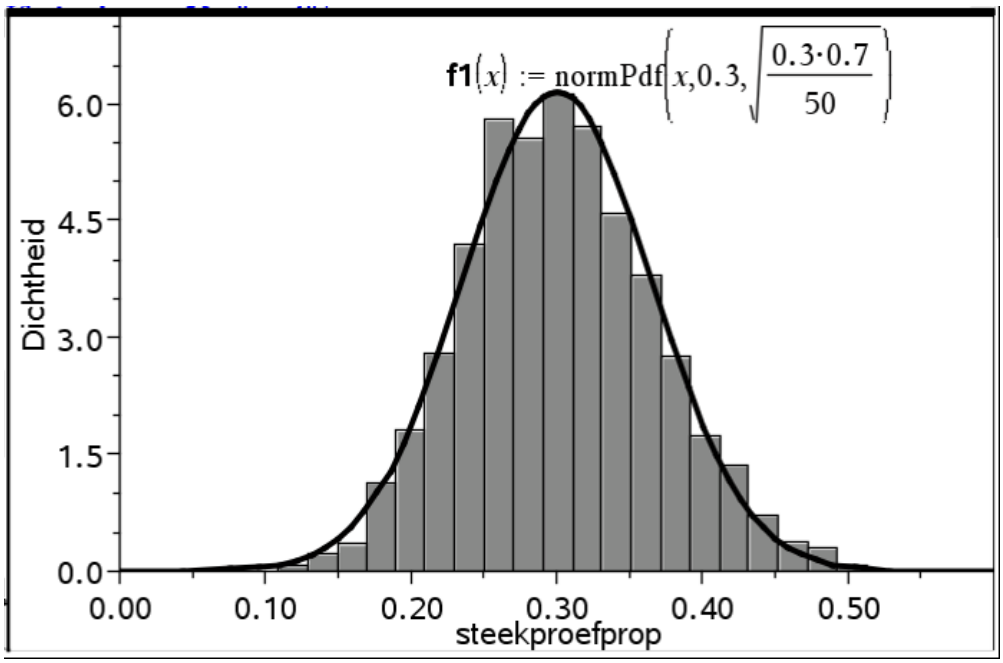
Simulatie:
veronderstel dat H_0 waar is, neem lukrake steekproeven van 50 snoepjes uit een populatie met 30% gele snoepjes.
Ga na in welke mate de steekproefproporties $\hat{p}$, verkregen volgens een toevalsproces, kunnen afwijken van de (vaste) populatieproportie p.





De experimentele tweezijdige overschrijdingskans (bij 1560 steekproeven) is slechts 2,9%, zodat men H_0 kan verwerpen met significantieniveau $\alpha = 5\%$.

Alhoewel de kansverdeling van de steekproefproporties discreet is (dit valt op wanneer men de klassenbreedte kleiner maakt dan 0.02) , kan men ze goed benaderen door een normale verdeling (kies een dichtheidshistogram met voldoend grote klassenbreedte) met gemiddelde $\mu = 0.3$ en standaardafwijking $\sqrt{\frac{0.3 \cdot 0.7}{50}} \simeq 0.065$:



De tweezijdige overschrijdingskans, met deze normale verdeling als kansmodel voor de steekproefproportie $\hat{P}$, wordt $2 \cdot P(\hat{P} \geq 0.46) \simeq 1.4\%$ (zonder continuïteitscorrectie!)

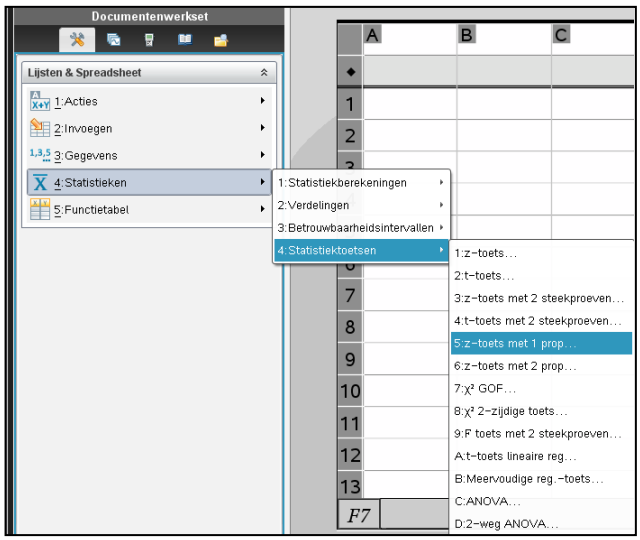
$$\sqrt{\frac{0.3 \cdot 0.7}{50}} \qquad 0.064807$$

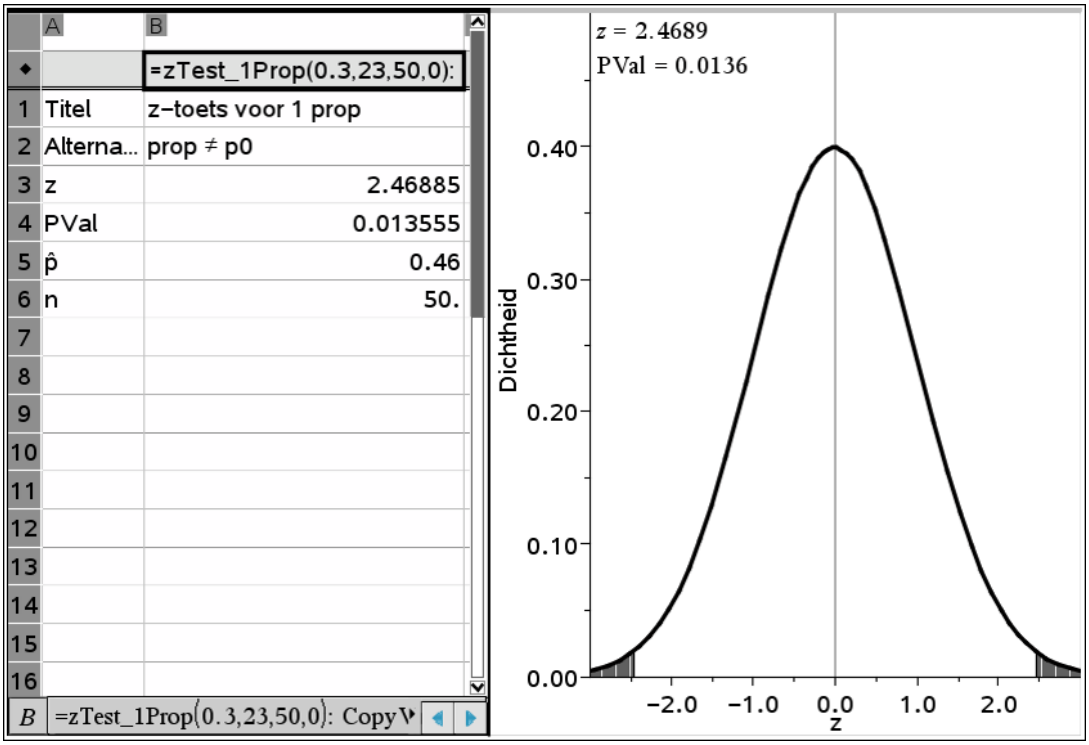
$$2 \cdot \text{normCdf}\left(0.46, 1, 0.3, \sqrt{\frac{0.3 \cdot 0.7}{50}}\right) \qquad 0.013555$$

De z-toets met 1 proportie werkt met de standaard normaal verdeelde testvariabele

$Z = \dfrac{\hat{P} - 0.3}{\sqrt{\dfrac{0.3 \cdot 0.7}{50}}}$, de geobserveerde waarde is $z = \dfrac{0.46 - 0.3}{\sqrt{\dfrac{0.3 \cdot 0.7}{50}}} \simeq 2.47$.

De overschrijdingskans of p-waarde is $2 \cdot P(Z \geq 2.47) \simeq 1.4\%$





Referenties

Websites

[1] Het **NIS** (Nationaal Instituut voor Statistiek) is de grootste statistische overheidsorganisatie in België: http://statbel.fgov.be/nl

[2] Kerncijfers 2009, België in een Europees perspectief http://economie.fgov.be/nl/modules/pressrelease/statistieken/generale/world_statistics_day.jsp

[3] Informatie over TI-Nspire: www.education.ti.com

[4] Data redders aan zee: www.redderaanzee.wobra.be

[5] Roulette, spelregels en winstkansen: http://www.casino-gids.be/artikels/spelregels/beginners/roulette.php

[6] G. Herweyers, *Cahier 8: Betrouwbaarheidsintervallen en het testen van hypothesen*, ter beschikking op www.t3vlaanderen.be

Boeken

[7] J. Beirlant, G. Dierckx, M. Hubert, *Statistiek en Wetenschap*, Acco, Leuven, 2005.

[8] M.H. DeGroot, M.J.Schervish, *Probability and Statistics*, Pearson International Edition, 2010.

[9] D.S. Moore, G.P. McCabe, *Statistiek in de Praktijk*, Academic Service, Schoonhoven, 2006.

[10] D.S.Yates, D.S. Moore, G.P. McCabe, *The practice of Statistics, TI-83 Graphing Calculator Enhanced*, W.H. Freeman and Company, New York, 1999.

[11] R. E. Walpole, R. H. Myers, S. L. Myers, K. Ye, *Probability & Statistics for Engineers and Scientists*, Pearson International Edition, 2011.

Technologie evolueert en biedt nieuwe mogelijkheden voor het onderwijs. Dit is zeker zo voor statistiek. Concrete data kunnen snel worden gevisualiseerd op verschillende wijzen, simulaties van steekproeven en kansexperimenten kunnen worden onderzocht. Hiermee kan men statistische begrippen in een vroeger stadium invoeren, waarbij de data op de voorgrond treden en de formules op de achtergrond. Het is de bedoeling om met dit cahier via concrete voorbeelden een mogelijke aanpak te illustreren.

GUIDO HERWEYERS doceert wiskunde en statistiek aan het Departement Industriële Wetenschappen en Technologie van de Katholieke Hogeschool Brugge-Oostende en is wetenschappelijk medewerker aan de K.U.Leuven.

Mei 2011