STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Steekproefmodellen en normaal verdeelde steekproefgrootheden

5. Het steekproefgemiddelde

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Het steekproefgemiddelde

1. Een concreet voorbeeld

Start met een populatie X waarvan het kansmodel in tabel 1 staat (de rode dobbelsteen).

x	1	3	6
$P(X=x)$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Kansmodel van de populatie X

Tabel 1

Als jij uit deze populatie X een steekproef van grootte $n=2$ trekt dan zou je bijvoorbeeld als eerste resultaat een 3 kunnen vinden (dan is voor jou $x_1=3$) en als tweede resultaat een 1 (dan is voor jou $x_2=1$). Van die 2 gevonden getallen kan je het gemiddelde berekenen. Een gemiddelde van steekproefgetallen noem je een steekproefgemiddelde. Dat wordt zoals elk gemiddelde van getallen voorgesteld door $\bar{x}$. Voor jou is de uitkomst van je steekproefgemiddelde gelijk aan $\bar{x}=\frac{x_1+x_2}{2}=\frac{3+1}{2}=2$. Maar je had natuurlijk ook iets anders kunnen vinden. Als je 2 keer een zes had gevonden dan zou jij $\bar{x}=6$ gehad hebben.

De vraag die nu komt verwacht je waarschijnlijk.

Als je uit die populatie een steekproef van grootte *n=2* **zou** trekken en je **zou** het steekproefgemiddelde berekenen, wat **zou** je dan vinden?

Je weet dat een antwoord op zo'n vraag altijd gegeven wordt in de vorm van een kansmodel. Je moet daarvoor alle mogelijke uitkomsten opschrijven, samen met hun kansen. Een uitkomst bij het berekenen van het steekproefgemiddelde wordt genoteerd als $\bar{x}$. Het kansmodel stel je voor door $\bar{X}$ (hoofdletter) waarbij $\bar{X}=\frac{X_1+X_2}{2}$.

2. Een kansmodel voor het steekproefgemiddelde

Om te weten wat alle mogelijke uitkomsten $\bar{x}$ zijn moet je weten welke waarden x_1 en x_2 kunnen aannemen want $\bar{x}=\frac{x_1+x_2}{2}$. Je kan dat afleiden uit het kansmodel van de steekproef. Dat ken je al. Het staat in onderstaande tabel.

Opdracht 1

In tabel 2 is een kolom toegevoegd om $\bar{x} = \frac{x_1 + x_2}{2}$ te berekenen. Vul de juiste waarden in.

(x_1 , x_2) uitkomst steekproef	$P(X_1 = x_1 , X_2 = x_2)$ kans van deze uitkomst	$\bar{x}$ uitkomst van het steekproefgemiddelde
(1 , 1)	$P(X_1 = 1 , X_2 = 1) = \frac{3}{6} \times \frac{3}{6} = \frac{9}{36}$	$\bar{x} = \frac{x_1 + x_2}{2} =$
(1 , 3)	$P(X_1 = 1 , X_2 = 3) = \frac{3}{6} \times \frac{2}{6} = \frac{6}{36}$	$\bar{x} = \frac{x_1 + x_2}{2} =$
(1 , 6)	$P(X_1 = 1 , X_2 = 6) = \frac{3}{6} \times \frac{1}{6} = \frac{3}{36}$	$\bar{x} = \frac{x_1 + x_2}{2} =$
(3 , 1)	$P(X_1 = 3 , X_2 = 1) = \frac{2}{6} \times \frac{3}{6} = \frac{6}{36}$	$\bar{x} = \frac{x_1 + x_2}{2} =$
(3 , 3)	$P(X_1 = 3 , X_2 = 3) = \frac{2}{6} \times \frac{2}{6} = \frac{4}{36}$	$\bar{x} = \frac{x_1 + x_2}{2} =$
(3 , 6)	$P(X_1 = 3 , X_2 = 6) = \frac{2}{6} \times \frac{1}{6} = \frac{2}{36}$	$\bar{x} = \frac{x_1 + x_2}{2} =$
(6 , 1)	$P(X_1 = 6 , X_2 = 1) = \frac{1}{6} \times \frac{3}{6} = \frac{3}{36}$	$\bar{x} = \frac{x_1 + x_2}{2} =$
(6 , 3)	$P(X_1 = 6 , X_2 = 3) = \frac{1}{6} \times \frac{2}{6} = \frac{2}{36}$	$\bar{x} = \frac{x_1 + x_2}{2} =$
(6 , 6)	$P(X_1 = 6 , X_2 = 6) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$	$\bar{x} = \frac{x_1 + x_2}{2} =$

Tabel 2

In tabel 2 zie je alle mogelijke uitkomsten van de steekproef samen met het corresponderende steekproefgemiddelde. Sommige waarden hiervan komen meerdere keren voor.

Opdracht 2

Gebruik tabel 2 om voor het steekproefgemiddelde $\bar{X}$ een nieuwe tabel (tabel 3 hieronder) te maken waarin alle **verschillende** uitkomsten één keer voorkomen, samen met hun kansen. Die tabel is het kansmodel voor het steekproefgemiddelde $\bar{X}$.

$\bar{x}$	
$P(\bar{X} = \bar{x})$	

Kansmodel voor het steekproefgemiddelde $\bar{X} = \frac{X_1 + X_2}{2}$

Tabel 3

In deze tekst blijf je verder werken met dit eenvoudige voorbeeld (het gemiddelde van een steekproef van grootte $n = 2$ uit de rode dobbelsteen). Daarna mag je aannemen (en het is juist !) dat de eigenschappen die je hier op een voorbeeld ontdekt algemeen geldig zijn.

Zoals bij de studie van elk kansmodel ga je op zoek naar 3 eigenschappen: het centrum, de spreiding en de globale vorm.

3. Het gemiddelde van het steekproefgemiddelde

Voor elk discreet kansmodel is het gemiddelde (of de verwachtingswaarde) gelijk aan de "gewogen" som van "uitkomsten maal hun kansen". Dat wist je al.

Kijk nu eens wat het gemiddelde is van de populatie X waaruit je trekt en het gemiddelde van het steekproefgemiddelde $\bar{X}$ waarvan het model in tabel 3 staat.

Voor de populatie X heb je vroeger berekend dat $E(X) = 1 \cdot \frac{3}{6} + 3 \cdot \frac{2}{6} + 6 \cdot \frac{1}{6} = \frac{15}{6} = 2.5$ en omdat het hier over de populatie gaat schrijf je dit als $\mu = 2.5$.

Voor het steekproefgemiddelde $\bar{X}$ gebruik je tabel 3. Dan krijg je

$$E(\bar{X}) = 1 \cdot \frac{9}{36} + 2 \cdot \frac{12}{36} + 3 \cdot \frac{4}{36} + 3.5 \cdot \frac{6}{36} + 4.5 \cdot \frac{4}{36} + 6 \cdot \frac{1}{36} = \frac{90}{36} = 2.5$$

Dat je hier twee keer als resultaat 2.5 krijgt is geen toeval.

Je kan hier als volgt over nadenken. Trek een steekproef van grootte $n = 2$ en bereken het gemiddelde. Dat levert je een eerste resultaat. Trek uit dezelfde populatie terug een steekproef van grootte $n = 2$ en bereken het gemiddelde. Je hebt dan een tweede gemiddelde dat waarschijnlijk verschilt van het eerste. Blijf nu maar steekproeven van grootte $n = 2$ trekken en bereken telkens het steekproefgemiddelde. Je krijgt dan heel veel gemiddelden. Als je daarvan het gemiddelde berekent dan valt dat (in de long run) exact samen met het gemiddelde van de populatie waaruit je trekt.

Je hebt hier een algemene eigenschap ontdekt die geldig is voor elk steekproefgemiddelde $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Het is altijd waar dat $E(\bar{X}) = \mu$ zowel bij het trekken uit een discrete populatie als uit een continue. Zo'n belangrijke eigenschap kan je best ook goed in woorden leren formuleren.

Voor elke steekproef $(X_1, X_2, \ldots X_n)$ uit eender welke populatie X geldt :

het gemiddelde van het steekproefgemiddelde is gelijk aan het populatiegemiddelde

$$E(\bar{X}) = \mu$$

4. De standaardfout van het steekproefgemiddelde

De standaardafwijking van een kansmodel is een maat voor de spreiding rond het gemiddelde.

Voor een discreet kansmodel heb je geleerd hoe je de standaardafwijking berekent. Je begint met de variantie. Daarvoor maak je voor elke uitkomst het verschil tussen die uitkomst en het modelgemiddelde. Dat verschil kwadrateer je en dan maak je de "gewogen" som van "gekwadrateerde verschillen maal hun kansen". Dat is de variantie.

Voor de populatie X heb je vroeger berekend dat de variantie gelijk is aan

$$\text{var}(X) = (1-2.5)^2 \cdot \frac{3}{6} + (3-2.5)^2 \cdot \frac{2}{6} + (6-2.5)^2 \cdot \frac{1}{6} = \frac{19.5}{6} = 3.25$$

De standaardafwijking $sd(X)$ van de populatie noteer je als σ. Hier is $\sigma = \sqrt{3.25}$

Voor het steekproefgemiddelde $\bar{X}$ pas je dezelfde formule toe waarbij je ermee rekening houdt dat $E(\bar{X}) = 2.5$. Je hebt dan dat

$$\text{var}(\bar{X}) = (1-2.5)^2 \cdot \frac{9}{36} + (2-2.5)^2 \cdot \frac{12}{36} + (3-2.5)^2 \cdot \frac{4}{36} + (3.5-2.5)^2 \cdot \frac{6}{36} +$$

$$(4.5-2.5)^2 \cdot \frac{4}{36} + (6-2.5)^2 \cdot \frac{1}{36} = \frac{58.5}{36} = 1.625 = \frac{3.25}{2}$$

De standaardafwijking van het steekproefgemiddelde ziet er uit als $sd(\bar{X}) = \sqrt{\frac{3.25}{2}} = \frac{\sqrt{3.25}}{\sqrt{2}}$.

Je ziet hier een breuk waar in de teller de standaardafwijking van de populatie staat en in de noemer de vierkantswortel uit de steekproefgrootte. Die was hier immers $n = 2$. De eigenschap die je in dit voorbeeld ziet geldt terug algemeen.

De standaardafwijking van het steekproefgemiddelde $\bar{X}$ heb je genoteerd als $sd(\bar{X})$. Daar is niets mis mee, want $sd(...)$ is de klassieke notatie voor de standaardafwijking van elk kansmodel. Maar in de statistiek is het de gewoonte om een speciale naam te gebruiken voor de standaardafwijking van grootheden die opgebouwd zijn met elementen van een steekproef. Die naam is **standaardfout**. De afkorting die hierbij hoort is $se(...)$ want standaardfout is in het Engels *standard error*.

Om de spreiding van het steekproefgemiddelde $\bar{X}$ aan te geven spreek je dus in het vervolg over "*de standaardfout van het steekproefgemiddelde*" en je noteert dat als $se(\bar{X})$.

Voor elke steekproef $(X_1, X_2, ... X_n)$ uit eender welke populatie X geldt :

de standaardfout van het steekproefgemiddelde is gelijk aan
de standaardafwijking van de populatie gedeeld door de wortel uit de steekproefgrootte

$$se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

5. De vorm van het kansmodel

Voor het kansmodel van het steekproefgemiddelde $\bar{X}$ ken je nu al twee eigenschappen die altijd waar zijn: $E(\bar{X}) = \mu$ en $se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ **.** Die eigenschappen hangen **NIET** af van de populatie waaruit je trekt en zij hangen **NIET** af van de grootte van de steekproef.
Je hebt nu al informatie over "centrum" en "spreiding". Maar hoe zit het met de globale vorm?

Hoe het kansmodel van $\bar{X}$ er qua vorm uitziet hangt **WEL** af van de populatie waaruit je trekt en hangt **WEL** af van de grootte van de steekproef.

Om een idee te krijgen over de globale vorm kan je wat experimenteren met je GRM.
- Kijk in de volgende paragrafen naar wat er gebeurt met $\bar{X}$ als je *uit een discrete populatie* trekt waarbij je een steekproef van grootte $n = 2$ of $n = 30$ of $n = 100$ neemt.
- Bestudeer daarna de vorm van het kansmodel van $\bar{X}$ als je *uit een continue populatie* trekt.

5.1. Trekken uit een discrete populatie

Met een simulatie kan je de echte kansen van een kansmodel alleen maar benaderen. Als je wil weten welke uitkomsten het steekproefgemiddelde $\bar{X}$ allemaal kan hebben en met welke kansen, dan kan je heel veel keren een steekproef trekken en telkens kijken waar het steekproefgemiddelde terechtkomt. Als je dan de relatieve frequenties van die uitkomsten berekent, heb je een **benaderend** idee over wat de bijhorende kans zou kunnen zijn.

Met je GRM ga je 250 keer een steekproef van grootte $n = 2$ trekken en telkens $\bar{x}$ berekenen. Voor die 250 gevonden gemiddelden $\bar{x}$ maak je een frequentietabel en teken je een staafdiagram. Dat is een ***benadering*** voor de echte kansverdeling van het steekproefgemiddelde $\bar{X}$. Let erop dat in dit voorbeeld $\bar{X}$ het gemiddelde voorstelt van een steekproef van grootte $n = 2$ die getrokken wordt uit de populatie X van tabel 1.

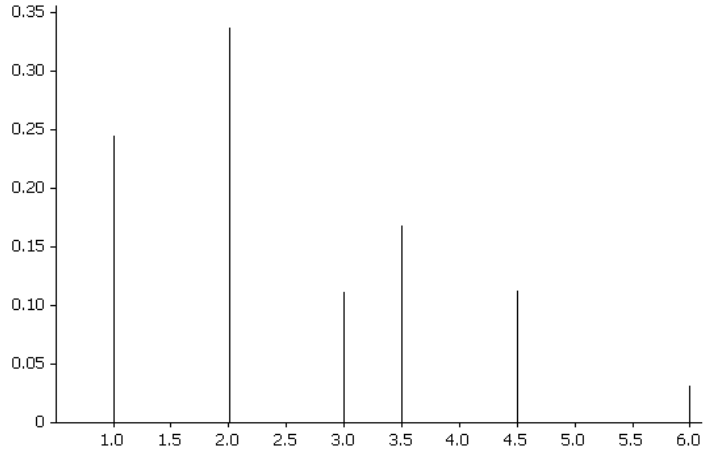
Opdracht 3

- Het juiste vaasmodel van tabel 1 zet je in je GRM door de getallen { 1, 1, 1, 3, 3, 6 } in lijst [L5] te tikken.
- Druk dan [PRGM], loop naar CLSVAAS en druk 2 keer [ENTER]. Geef aan dat je met een steekproef van grootte $n = 2$ wil werken en dat je het trekken van zo'n steekproef 250 keer wil herhalen. Het programma zal dan 250 keer $\bar{x}$ berekenen en die 250 steekproefgemiddelden in [L1] plaatsen. Voor het uitvoeren van dit programma heeft je GRM ongeveer anderhalve minuut nodig.
- Druk [PRGM], loop naar FREQDISC en druk 2 keer [ENTER]. Met dit programma stel je een frequentietabel op van al die gevonden steekproefgemiddelden.
- Druk [PRGM], loop naar STAAFDGR en druk 2 keer [ENTER] om een staafdiagram te tekenen. Uit tabel 3 weet je dat de kleinst mogelijke uitkomst 1 is en de grootst mogelijke 6. Om zeker te zijn dat je het volledige interval [1 ; 6] te zien krijgt kies je bij de vraag over de vensterinstellingen voor "zelf ingeven= 2". Stel dan Xmin= 1 en Xmax= 6 wanneer dit gevraagd wordt. Voor de hoogte van de staafjes kies je: "relatieve frequentie = 2" want je wil een benadering van de echte kans (je

relatieve frequenties zouden samenvallen met de echte kansen als je met een oneindig aantal herhalingen kan werken in plaats van met 250).

Herinner je dat jij hier geen oneindig aantal herhalingen nodig hebt om de echte kansen te kennen. Je hebt die zelf berekend in tabel 3.

Je gevonden staafdiagram is een benadering voor het echte kansmodel van $\bar{X}$. Het staafdiagram van het echte kansmodel is het staafdiagram dat bij tabel 3 hoort. Het ziet er als volgt uit.



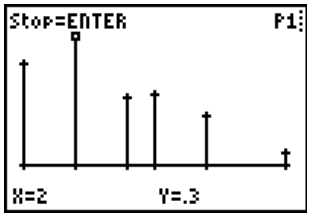
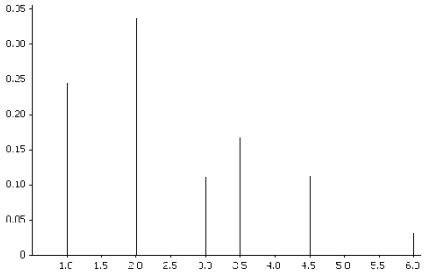
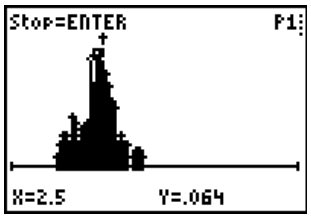
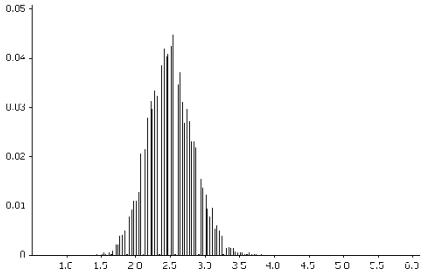
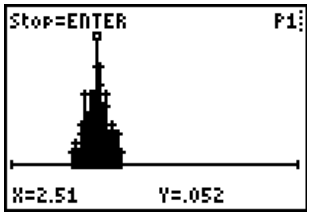
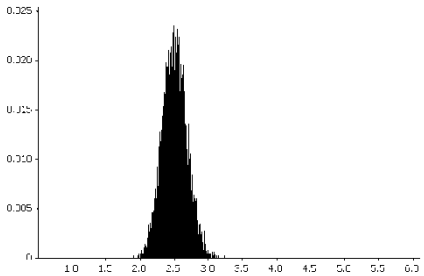
Het kansmodel van het steekproefgemiddelde $\bar{X}$ voor een steekproef van $n = 30$ en een steekproef van $n = 100$ kan je proberen te benaderen door simulatie. Je zal dan resultaten krijgen die er uitzien zoals de linker kolom in onderstaande figuur. In de rechter kolom zie je wat je krijgt als je met een computer telkens 5000 keer een steekproef van grootte n trekt.

Voor de globale vorm van het kansmodel van het steekproefgemiddelde $\bar{X}$ geldt:

als de steekproefgrootte n groter en groter wordt
- dan wordt de globale vorm meer en meer symmetrisch
- met een top in het midden
- en met staafjes die kleiner en kleiner worden naarmate je (links en rechts) verder verwijderd bent van de locatie waar de top ligt.

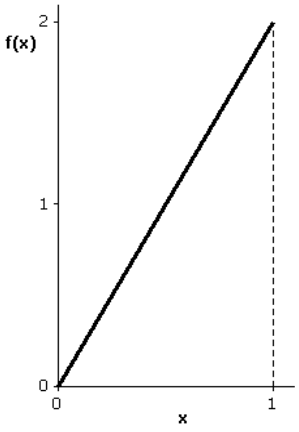
Op onderstaande figuur zie je deze eigenschappen duidelijk bij $n = 100$. Bij $n = 2$ is dat nog niet het geval: de steekproef is daar te klein en je hebt helemaal nog geen "klokvormige" figuur die symmetrisch is rond één top.

Het kansmodel van het steekproefgemiddelde $\bar{X}$
bij een steekproef van grootte n uit de populatie X van tabel 1.

	Benaderend kansmodel op basis van een beperkte simulatie van 250 herhalingen	Benaderend kansmodel op basis van een simulatie van 5000 herhalingen
$n = 2$		
$n = 30$		
$n = 100$		

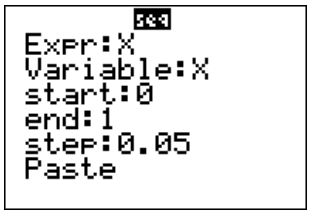
5.2. Trekken uit een continue populatie

Op eenzelfde manier kan je nu werken met steekproeven uit een continue populatie. Neem daarvoor de populatie X die als dichtheidsfunctie $f(x) = 2x$ voor $0 \le x \le 1$ heeft. Deze dichtheidsfunctie heb je vroeger al ontmoet. Kijk maar even hoe die eruit ziet. Uit deze continue populatie trek je een steekproef van grootte $n = 2$ en je herhaalt dat 250 keer. Dat doe je in de volgende opdracht.

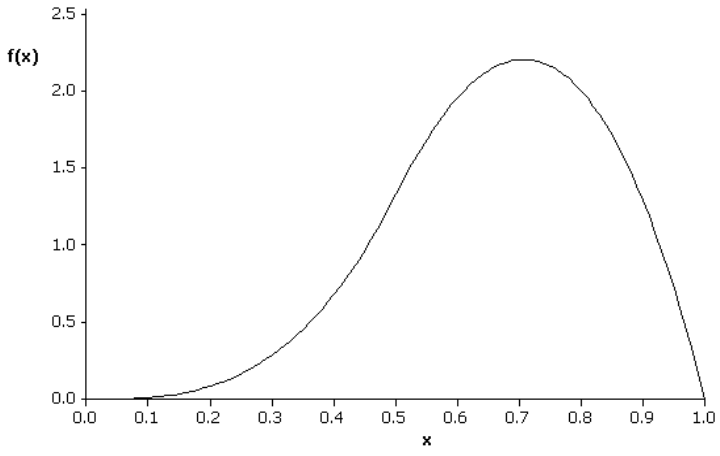


Opdracht 4

- Vooraf zet je in [L2] de klassengrenzen voor de op te stellen frequentietabel met klassenindeling. Aangezien de uitkomsten van de populatie X allemaal tussen 0 en 1 liggen zullen ook alle uitkomsten van $\bar{X}$ tussen 0 en 1 liggen, wat de steekproefgrootte ook weze. Je kan het interval [0 ; 1] bijvoorbeeld in 20 deelintervallen opdelen. Dat doe je als volgt: druk [2nd][LIST] loop naar OPS en kies 5:seq(.
  Vul in, loop naar Paste, druk [ENTER] en dan [STO▶] en [2nd][L2].

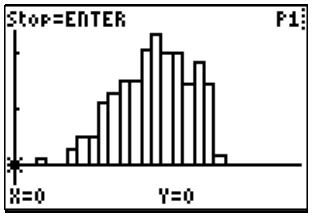
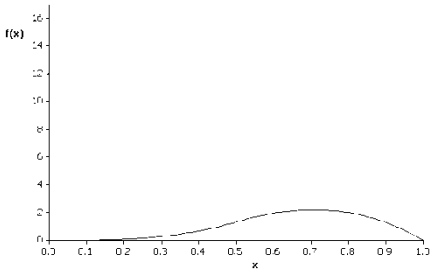
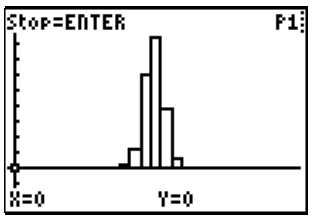
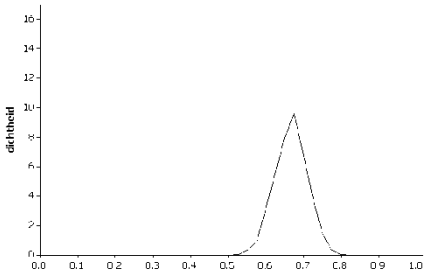
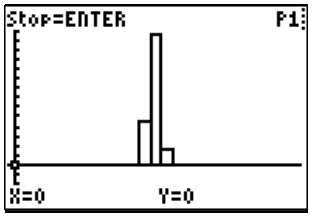
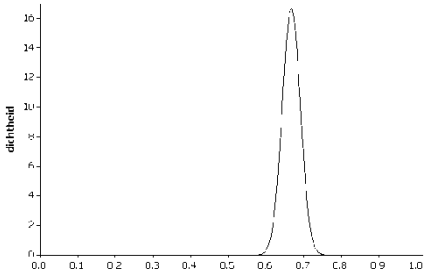


- Druk [PRGM], loop naar CLS2X en druk 2 keer [ENTER]. Zeg dat je uit de populatie X een steekproef van grootte $n = 2$ wil trekken en dat je dit 250 keer wil herhalen. Het programma zal telkens $\bar{x}$ berekenen en de uitkomst in [L1] plaatsen.
- Druk [PRGM], loop naar FREQCONT en druk 2 keer [ENTER] om een frequentietabel met klassenindeling op te stellen.
- Druk [PRGM], loop naar HISDICH en druk 2 keer [ENTER]. Kies 1=Histogram om een globale vorm te ontdekken. Het programma HISDICH tekent een histogram op de dichtheidsschaal (totale oppervlakte = 1) als benadering van de echte dichtheidsfunctie van $\bar{X}$. Let zowel op de plaats van de curve als op haar vorm.
- Je kan je figuur ook vergelijken met de exacte dichtheidsfunctie van $\bar{X} = \frac{1}{2}(X_1 + X_2)$. Zij is scheef naar links en ziet er als volgt uit.



Naarmate $\bar{X}$ het gemiddelde is van een grotere steekproef lijkt haar kansmodel meer en meer op een klokvormige curve, symmetrisch rond één top. Je kan dat in dit voorbeeld zien in de benaderende kansmodellen die met de computer werden opgesteld (rechterkolom van de onderstaande figuur).

Het kansmodel van het steekproefgemiddelde $\bar{X}$
bij een steekproef van grootte n
uit de populatie X met dichtheidsfunctie $f(x) = 2x$ voor $0 \le x \le 1$

	Benaderend kansmodel op basis van een beperkte simulatie van 250 herhalingen	Benaderend kansmodel op basis van een simulatie van 5000 herhalingen
$n = 2$		
$n = 30$		
$n = 100$		

5.3. Besluit: een klokvormige curve

Of je nu uit een discrete populatie trekt of uit een continue, voor de globale vorm van het kansmodel van het steekproefgemiddelde $\bar{X}$ is dat blijkbaar om het even. Als de steekproef maar groot genoeg is, dan krijg je meer en meer een klokvormige figuur die symmetrisch is rond één top. Die klokvormige figuur is, voor $n \to \infty$, niets anders dan de normale curve.

Bij een steekproef $(X_1, X_2, \ldots X_n)$ uit eender welke populatie X (discreet of continu)

geldt voor het steekproefgemiddelde $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- **altijd:** $E(\bar{X}) = \mu$
- **altijd:** $se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- **voor** n **groot:** de kans dat $\bar{X}$ in een willekeurig interval [a ; b] valt
  $\cong$ de kans dat een normaal kansmodel in dat interval [a ; b] valt.

Nota.
In heel wat gevallen is $n \geq 30$ *al voldoende om over "grote steekproef" te spreken en mag je de bovenstaande eigenschap toepassen.*