# Hoofdstuk 5

# Van beschrijvende naar verklarende statistiek

We hebben gezien in de *beschrijvende statistiek* hoe we data grafisch kunnen voorstellen en samenvatten door centrum- en spreidingsmaten als we beschikken over de data van een volledige populatie.

Indien we echter enkel beschikken over steekproefdata, wat meestal het geval is, dan wensen we niet alleen deze beschrijving van de data. We wensen nu ook uitspraken te doen over de volledige, voor ons onbekende, populatie.

Deze *statistische besluitvorming* omtrent de populatie wordt bestudeerd in de verklarende statistiek. In dergelijke uitspraken speelt het kansbegrip een belangrijke rol. Het begrip *stochastische veranderlijke* staat centraal in de verklarende statistiek.

## 5.1 Gemiddelde en variantie m.b.v. relatieve frequentie

Om het verband te begrijpen tussen de beschrijvende en de verklarende statistiek is het belangrijk de formules voor het rekenkundig gemiddelde en de standaard-afwijking van de data te schrijven in functie van de relatieve frequentie van de verschillende data.

Als voorbeeld beschouwen we de volgende data bestaande uit $n = 10$ numerieke gegevens :

| $x_i$ | 1 | 3 | 3 | 1 | 3 | 9 | 3 | 5 | 9 | 3 |
|---|---|---|---|---|---|---|---|---|---|---|

Herinner je dat we voor het (*rekenkundig) gemiddelde* de notatie $\mu$ gebruiken voor *het populatiegemiddelde* en $\overline{x}$ voor *een steekproefgemiddelde*.

Voor de bovenstaande lijst, als steekproef, vinden we $\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 4$.

Indien we van de bovenstaande data een frequentietabel construeren, noteren we de *verschillende* meetresultaten, van klein naar groot, samen met hun frequentie en de relatieve frequentie. Voor de bovenstaande data geeft dit, met $x_i$ $(i = 1,...,m)$ de $m$ *verschillende* data, de volgende tabel :

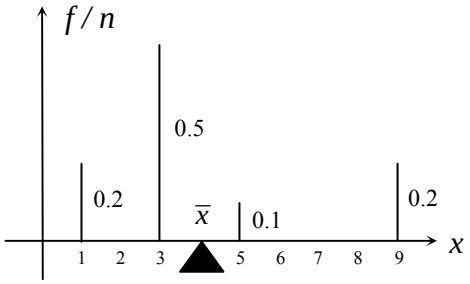
| $x_i$ | $f_i$ | $f_i / n$ |
|---|---|---|
| 1 | 2 | 0.2 |
| 3 | 5 | 0.5 |
| 5 | 1 | 0.1 |
| 9 | 2 | 0.2 |
| $\sum_i$ | $(n =)$ 10 | 1 |

We verkrijgen, op basis van deze tabel, voor het rekenkundig gemiddelde :

$$\overline{x} = \frac{\sum_{i=1}^{m} f_i . x_i}{n} = \sum_{i=1}^{m} \frac{f_i}{n} \cdot x_i \quad \text{met} \quad m \le n \text{ en } n \text{ het totale aantal gegevens.}$$

Uit de laatste formule volgt dat we het rekenkundig gemiddelde kunnen interpreteren als het *gewogen gemiddelde* van de *verschillende* getallen $x_i$, waarbij elke $x_i$ vermenigvuldigd wordt met het *gewicht* $f_i / n$. Dit gewicht is de *relatieve frequentie* van $x_i$. De som van de relatieve frequenties is steeds 1.

De frequentieverdeling of relatieve frequentieverdeling wordt aanschouwelijk voorgesteld in onderstaande grafiek of staafdiagram:




We kunnen $\overline{x}$ mechanisch als volgt interpreteren :

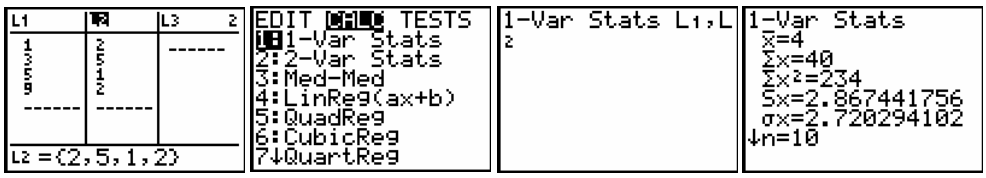
Plaats een gewichtsloze staaf op de $x$-as en massa's $f_i / n$ in de punten $x_i$. Nu is $\overline{x}$ het zwaartepunt van de massaverdeling. De staaf is in evenwicht indien we hem ondersteunen ter hoogte van $\overline{x}$.

Voor de *populatievariantie* $\sigma^2$ en de *steekproefvariantie* $s^2$ van $n$ getallen geldt :

| | zonder frequentietabel | met frequentietabel |
|---|---|---|
| populatievariantie | $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$ | $\sigma^2 = \dfrac{\sum_{i=1}^{M} f_i(x_i - \mu)^2}{N}$ |
| steekproefvariantie | $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ | $s^2 = \dfrac{\sum_{i=1}^{m} f_i(x_i - \bar{x})^2}{n-1}$ |

We plaatsen de data in de lijsten **L1** en **L2** van de **TI-83**, in de vorm van de frequentietabel, en berekenen vervolgens de kengetallen.

Reken manueel na dat $s^2 = 74/9 = 8.22$ .




## 5.2 Stochastische veranderlijken

De uitkomsten van een experiment hoeven geen getallen te zijn. Vaak zijn het zelfs niet de uitkomsten die ons interesseren maar bepaalde getallen geassocieerd met de uitkomsten.

Als voorbeeld beschouwen we het experiment dat bestaat uit het twee keer opwerpen van een muntstuk. De mogelijke uitkomsten zijn KK, KM, MK, MM. We zijn echter alleen benieuwd naar het aantal keer kop.

We stellen het aantal keer kop voor door $X$ .

We kunnen niet op voorhand zeggen welke waarde $X$ zal aannemen bij de start van het experiment. Dit hangt af van het toeval. We noemen $X$ een toevalsveranderlijke of stochastische veranderlijke of kortweg stochast. Stochasten noteren we steeds met hoofdletters.

Pas na uitvoeren van het experiment weten we hoeveel keer kop er tevoorschijn is gekomen. De stochast $X$ heeft een concrete waarde aangenomen. Deze waarde noteren we met de bijbehorende kleine letter $x$. Indien we het experiment nog eens herhalen, kunnen we een andere $x$ krijgen. Dit onderscheid tussen hoofdletters voor stochasten en kleine letters voor de concrete aangenomen waarden na het uitvoeren van het experiment is zeer belangrijk in de verklarende statistiek.

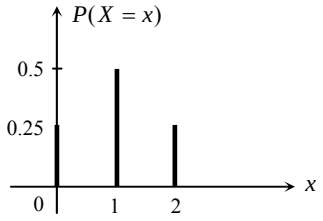
Zo spreken we over de kans P($X = x$), d.w.z. de kans dat $X$ de waarde $x$ aanneemt. De stochast $X$ is volledig gekenmerkt door de verschillende waarden die $X$ kan aannemen en de bijbehorende kans dat dit gebeurt. Dergelijke tabel noemen we de *kansverdeling* van $X$ :

| $x_i$ | 0 | 1 | 2 |
|---|---|---|---|
| $P(X = x_i)$ | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |

Hieronder vind je een grafische voorstelling van de kansverdeling :

Merk op dat $P(X = 1) = P(\{\text{KM, MK}\})$.

In dit voorbeeld is $X$ een *discrete stochast*. Dit is een stochast waarvan je de verschillende waarden die kunnen aangenomen worden, kan ordenen als een rij. Bovendien kan een discrete stochast oneindig veel waarden aannemen.




Een stochast is eigenlijk een functie van de uitkomstenverzameling van het experiment naar $\mathbb{R}$. In het vorige voorbeeld associeert $X$ met de uitkomsten KK, KM, MK, MM als beelden respectievelijk de getallen 2, 1, 1, 0. De verzameling van al de beelden levert ons een nieuwe uitkomstenverzameling $\{0,1,2\}$ waarin we geïnteresseerd zijn. De kansen op die uitkomsten 0,1,2 worden gegeven door de kansverdeling van $X$.

## 5.3 Verwachtingswaarde

De *verwachtingswaarde* of het *gemiddelde* van een discrete stochast $X$ wordt gedefinieerd door : $E(X) = \sum_i x_i \cdot P(X = x_i)$.

We sommeren over alle waarden die de stochast aanneemt.

Steeds geldt : $\sum_i P(X = x_i) = 1$.

Voor ons vorig voorbeeld vinden we : $E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$.

Wat is de betekenis van $E(X)$ ?

Herhaal het experiment "werp twee keer een muntstuk" heel vaak en noteer telkens het aantal keer kop. Het gemiddelde van al die getallen is gelijk aan het gewogen gemiddelde van de getallen 0,1 en 2 met als gewichten hun relatieve frequenties. Op de lange duur zullen die relatieve frequenties evolueren naar de kansen $P(X=0), P(X=1), P(X=2)$. Op de lange duur benadert het gemiddelde van die getallen bijgevolg $E(X)$.

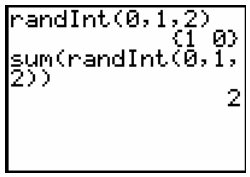
We noteren het gemiddelde van een stochast met $\mu$ ; $E(X) = \mu$.

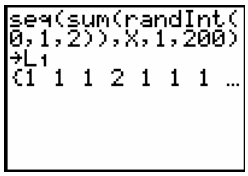
De Griekse letter $\mu$ hebben we vroeger ook reeds gebruikt als notatie voor het gemiddelde van een populatie van getallen. $E(X) = \mu$ is inderdaad te beschouwen als het gemiddelde van de populatie van alle getallen $x$ die we verkrijgen door het experiment heel vaak uit te voeren.

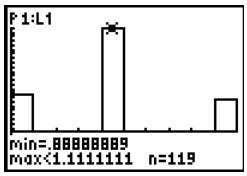
De kracht van de simulatie van een experiment met het rekentoestel bestaat hierin dat we de frequentieverdeling van de verkregen data snel kunnen genereren. Dit geeft, indien we de simulatie van het experiment maar vaak genoeg uitvoeren, meteen ook de relatieve frequentieverdeling van de data als goede benadering van de kansverdeling van de stochast.

Dit is soms de enige manier van werken om de kansverdeling van een stochast te achterhalen, indien de kansberekening te moeilijk wordt.

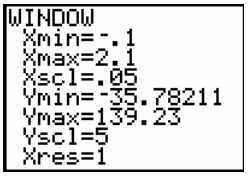
We simuleren het experiment "werp twee keer een muntstuk" 200 keer en noteren telkens het aantal keer kop. We coderen kop met 1 en munt met 0. De resultaten plaatsen we in lijst **L1**.

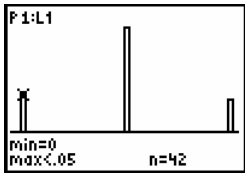


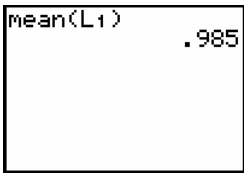




Definieer de **WINDOW**- en **MODE**-instellingen zoals hieronder om een staafdiagram te tekenen en bereken het gemiddelde van **L1**.

De simulatie leverde 42 keer 0, 119 keer 1 en 39 keer 2 als aantal keer kop. Het gemiddelde aantal is $0 \cdot \frac{42}{200} + 1 \cdot \frac{119}{200} + 2 \cdot \frac{39}{200} = 0.985$. Dit is een goede benadering of *schatting* van de theoretische waarde $E(X)$.

Ook de relatieve frequenties 0.21, 0.595, 0.195 geven reeds een ruw idee van de theoretische kansverdeling 0.25, 0.5, 0.25 van de stochast $X$.

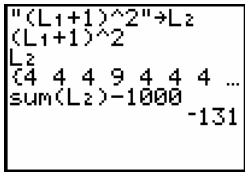
## Klasopgave

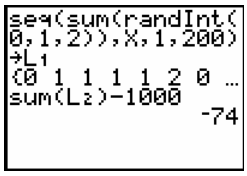
Combineer de simulaties van alle studenten. Wat verwacht je van de relatieve frequentieverdeling ?

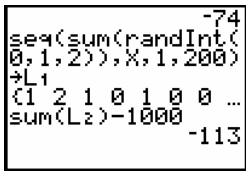
Met de stochast $X$ = "aantal keer kop bij twee keer werpen van een muntstuk" kunnen we nieuwe stochasten definiëren.

Stel bijvoorbeeld dat men jou in het casino het volgende kansspel voorstelt. Je werpt twee keer een muntstuk en telt het aantal keer kop. Tel hierbij 1 op en kwadrateer. Dit is het bedrag dat je krijgt in Euro. Als inzet moet je echter 5 Euro betalen per spel. Ga je dit spel een ganse avond spelen, in de veronderstelling dat je rijk genoeg bent om steeds je inzet te betalen ?

We simuleren dit met behulp van de lijst **L1** van voorgaande simulatie. De bedragen die we ontvangen komen terecht in lijst **L2 = "(L1 + 1)²"**. Vlug even narekenen wat we verliezen of winnen na 200 keer spelen.




De resultaten zijn onheilspellend. Bij de eerste simulatie verlies je 131 Euro. Een tweede en derde simulatie zijn snel uitgevoerd dankzij het koppelen van de formule in lijst **L2**. Ook hier verliezen we respectievelijk 74 en 113 Euro na telkens 200 keer spelen. Dit spel is af te raden.

Ook zonder simulatie kun je de gemiddelde winst per spel na lang spelen berekenen. De stochast $Y = (X+1)^2 - 5$ is de winst per spel.

Deze neemt de waarden $-4, -1$ en $4$ aan met respectievelijke kansen $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$.

Het gemiddelde van $Y$ is $E(Y) = -4 \cdot \frac{1}{4} + (-1) \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} = -0.5$ .
Dit is de gemiddelde "winst" na lang spelen.

Op de lange duur verlies je gemiddeld 0.5 Euro per spel. Na 200 keer spelen mag je een verlies verwachten in de buurt van 100 Euro. Als we de drie simulaties combineren, hebben we 113 + 131 + 74 = 318 Euro verlies na 600 keer spelen. Dit is aardig dicht bij de theoretische voorspelling van $600 \cdot 0.5 = 300$ .

Voor een eerlijk spel moet de gemiddelde winst per spel nul zijn en dit is hier niet het geval.

Voor een discrete stochast $X$ met waarden $x_i$ kan men bewijzen dat de verwachtingswaarde van een stochast $Y = g(X)$ gegeven wordt door :

$$E(g(X)) = \sum_i g(x_i) \cdot P(X = x_i)$$

## 5.4 Variantie

Een belangrijke illustratie van de laatste formule is de *variantie* van $X$ :

$Var(X) = E\left((X - \mu)^2\right)$ met $\mu = E(X)$ of
$Var(X) = \sum_i (x_i - \mu)^2 \cdot P(X = x_i)$

Voor $X$ = "aantal keer kop bij twee keer werpen van een muntstuk" geldt :

$$Var(X) = (0-1)^2 \cdot \frac{1}{4} + (1-1)^2 \cdot \frac{1}{2} + (2-1)^2 \cdot \frac{1}{4} = \frac{1}{2} .$$

We noteren de variantie van $X$ met $\sigma^2$ : $Var(X) = \sigma^2$ .

De *standaardafwijking*, $\sigma$ , is de positieve vierkantswortel van de variantie.

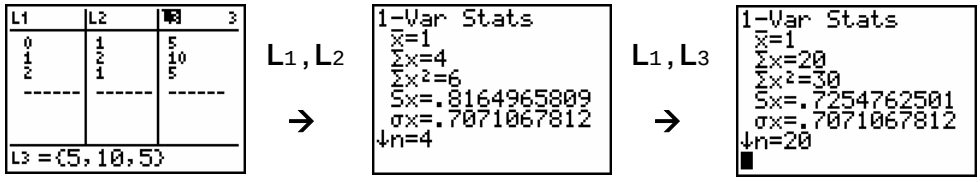
In ons voorbeeld geldt : $\sigma = \sqrt{0.5} = 0.707$ .

De steekproefvariantie van de getallen die we verkrijgen bij het heel vaak herhalen van het experiment, $s^2 = \frac{\sum_{i=1}^{m} f_i \left(x_i - \overline{x}\right)^2}{n-1}$ (1) , benadert het best $Var(X)$ .

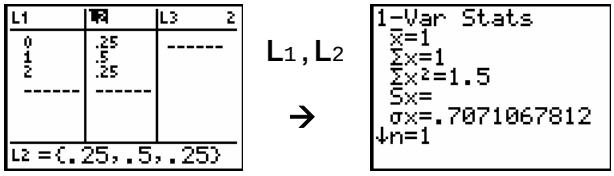
Je zou in (1) eerder $n$ verwachten als noemer als je de formule $Var(X) = \sum_i (x_i - \mu)^2 \cdot P(X = x_i)$ bekijkt en daar we op de lange duur bekomen dat $\frac{f_i}{n} \approx P(X = x_i)$. Maar vergeet niet dat we in de teller van (1) als benadering voor $\mu = E(X)$ het steekproefgemiddelde $\overline{x}$ gebruiken. In paragraaf 5.8 zien we waarom $s^2$ de beste schatting is van $Var(X) = \sigma^2$.

Het gemiddelde en de standaardafwijking (of de variantie) zijn de belangrijkste kenmerken van een discrete of continue (zie verder) stochast. Om het gemiddelde en de standaardafwijking van een discrete stochast te berekenen met het rekentoestel volstaat het een populatie van getallen in te voeren waarvan de relatieve frequentieverdeling samenvalt met de kansverdeling van de stochast.

Voor $X$ = "aantal keer kop bij twee keer werpen van een muntstuk" kan dit als volgt :




We kunnen ook de relatieve frequenties invoeren in **L2,** zie onderaan. Maar dan gebeurt er iets met Sx, kan je dat verklaren ? Waarom blijft σx wel juist ?




## 5.5 Lukrake trekking uit een populatie

Beschouw als populatie de schoenmaten van $n = 30$ volwassen mannen (zie ook 4.4) met onderstaande frequentietabel:

| $x_i$ | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 46 |
|---|---|---|---|---|---|---|---|---|
| $f_i$ | 2 | 4 | 7 | 5 | 6 | 2 | 3 | 1 |

Het populatiegemiddelde is $\mu = 41.1$, de populatiestandaardafwijking is $\sigma = 1.89$.

Schrijf die 30 schoenmaten op een kaartje en leg die kaarten in een doos. Vraag aan iemand, die de inhoud van de doos niet kent, om lukraak een getal te trekken uit deze populatie.

We noemen dit getal *X*. Dit is een stochast daar we niet op voorhand kunnen zeggen welk getal er zal gekozen worden. Elke kaart heeft dezelfde kans om getrokken te worden. De *kansverdeling* van $X$ valt samen met de *relatieve frequentieverdeling van de populatie* :

| $x_i$ | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 46 |
|---|---|---|---|---|---|---|---|---|
| $\frac{f_i}{n} = P(X = x_i)$ | $\frac{2}{30}$ | $\frac{4}{30}$ | $\frac{7}{30}$ | $\frac{5}{30}$ | $\frac{6}{30}$ | $\frac{2}{30}$ | $\frac{3}{30}$ | $\frac{1}{30}$ |

Bijgevolg zijn $E(X)$ en $Var(X)$ respectievelijk gelijk aan het gemiddelde $\mu$ en de variantie $\sigma^2$ van de gegeven populatie.

Voor een lukrake trekking $X$ uit een populatie van getallen geldt steeds :

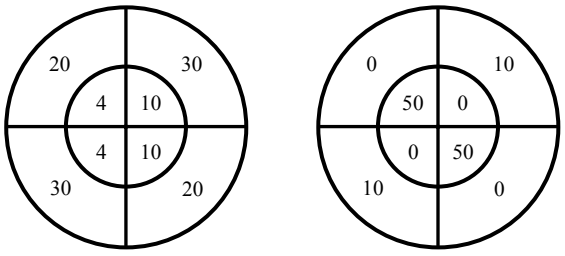
$E(X) =$ populatiegemiddelde $\mu$,
$Var(X) =$ populatievariantie $\sigma^2$.

Dit is een belangrijke vaststelling. We noemen *X* ook een *populatiestochast*.

## 5.6 Afhankelijke en onafhankelijke stochasten

Bij draaien van de hieronder afgebeelde raderen van fortuin komen we terecht in één van de vier kwadranten.

Stel $X$ het buitenste en $Y$ het binnenste getal.



Voor het rechtse rad geldt :

$E(X) = 0 \cdot \frac{1}{2} + 10 \cdot \frac{1}{2} = 5$ en $E(Y) = 0 \cdot \frac{1}{2} + 50 \cdot \frac{1}{2} = 25$ .

De verwachtingswaarden van $X + Y$ en $X \cdot Y$ zijn in dit geval :

$E(X + Y) = 50 \cdot \frac{1}{2} + 10 \cdot \frac{1}{2} = 30 = E(X) + E(Y)$

$E(X \cdot Y) = E(0) = 0 \neq E(X) \cdot E(Y)$

Reken na dat $Var(X + Y) \neq Var(X) + Var(Y)$ .

Voor het linkse rad, met dezelfde betekenis van $X$ en $Y$ , geldt :

$E(X) = 20 \cdot \frac{1}{2} + 30 \cdot \frac{1}{2} = 25$ en $E(Y) = 4 \cdot \frac{1}{2} + 10 \cdot \frac{1}{2} = 7$

$E(X + Y) = 40 \cdot \frac{1}{4} + 24 \cdot \frac{1}{4} + 34 \frac{1}{4} + 30 \cdot \frac{1}{4} = 32 = E(X) + E(Y)$

$E(X \cdot Y) = 300 \cdot \frac{1}{4} + 80 \cdot \frac{1}{4} + 120 \cdot \frac{1}{4} + 200 \cdot \frac{1}{4} = 175 = E(X) \cdot E(Y)$

Reken na dat in dit geval geldt dat $Var(X + Y) = Var(X) + Var(Y)$ .

Voor het linkse rad zegt men dat de stochasten $X$ en $Y$ *onafhankelijk* zijn. D.w.z. dat informatie over de ene stochast geen extra informatie geeft over de andere stochast.

Zo geldt voor het hele linkse rad dat de gebeurtenissen $Y = 4$ en $Y = 10$ als kans 1/2 hebben. Zegt men je dat de gebeurtenis $X = 20$ is opgetreden, hebben $Y = 4$ en $Y = 10$ (bekijk enkel het tweede en vierde kwadrant) nog steeds kans 1/2.

Bij het rechtse rad zijn $X$ en $Y$ *afhankelijk*. De gebeurtenis $Y = 0$ heeft kans 1/2. Maar als je weet dat $X = 10$ is opgetreden, heeft $Y = 0$ kans 1.

## 5.7 Eigenschappen van de operatoren *E* en *Var*

Stel $X$ , $Y$ stochasten bij eenzelfde experiment en $a$ een reëel getal. Er geldt :

$$\left.\begin{array}{l} E(X + Y) = E(X) + E(Y) \\ E(a \cdot X) = a \cdot E(X) \end{array}\right\}$$ de operator $E$ is *lineair*.

$$\left.\begin{array}{l} E(X \cdot Y) = E(X) \cdot E(Y) \\ Var(X + Y) = Var(X) + Var(Y) \end{array}\right\}$$ als $X$ en $Y$ onafhankelijk zijn.

Bovendien geldt :

$E(a) = a$ en $Var(a) = 0$

$Var(X + a) = Var(X)$ en $Var(a \cdot X) = a^2 \cdot Var(X)$

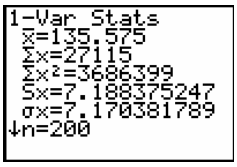
Deze eigenschappen zijn geldig voor discrete en continue (zie verder) stochasten.

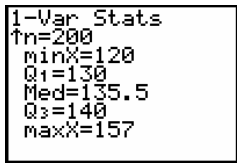
## 5.8 Het begrip steekproef in de verklarende statistiek

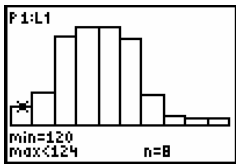
In de volgende tabel beschouwen we als populatie de lengte (in cm) van 200 kinderen van 10 jaar

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 120 | 123 | 151 | 142 | 137 | 128 | 133 | 142 | 136 | 137 |
| 144 | 126 | 135 | 142 | 135 | 134 | 140 | 149 | 137 | 129 |
| 128 | 140 | 129 | 137 | 141 | 137 | 135 | 129 | 133 | 139 |
| 132 | 125 | 124 | 132 | 129 | 139 | 132 | 145 | 140 | 138 |
| 137 | 133 | 137 | 138 | 131 | 137 | 131 | 127 | 134 | 134 |
| 150 | 140 | 144 | 137 | 133 | 139 | 130 | 141 | 136 | 124 |
| 130 | 135 | 124 | 122 | 136 | 132 | 133 | 133 | 142 | 127 |
| 142 | 130 | 135 | 125 | 136 | 132 | 153 | 145 | 131 | 131 |
| 134 | 145 | 139 | 132 | 136 | 143 | 138 | 141 | 141 | 141 |
| 136 | 148 | 128 | 137 | 134 | 138 | 130 | 145 | 135 | 141 |
| 131 | 143 | 146 | 132 | 127 | 129 | 133 | 142 | 157 | 133 |
| 139 | 128 | 123 | 140 | 140 | 152 | 136 | 125 | 130 | 153 |
| 130 | 126 | 129 | 157 | 144 | 142 | 128 | 138 | 142 | 135 |
| 141 | 139 | 132 | 135 | 145 | 134 | 140 | 136 | 138 | 143 |
| 122 | 141 | 122 | 132 | 136 | 129 | 138 | 130 | 129 | 135 |
| 134 | 141 | 133 | 128 | 121 | 131 | 137 | 140 | 133 | 135 |
| 138 | 132 | 140 | 145 | 128 | 140 | 134 | 128 | 146 | 132 |
| 131 | 142 | 133 | 137 | 126 | 128 | 129 | 124 | 137 | 127 |
| 139 | 141 | 157 | 146 | 128 | 136 | 130 | 141 | 129 | 143 |
| 137 | 143 | 139 | 141 | 121 | 131 | 128 | 133 | 136 | 146 |

We bepalen de statistische kengetallen en een histogram.

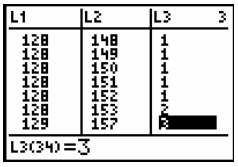



Stel $X$ een lukraak gekozen getal uit deze populatie. Daar de *kansverdeling* van $X$ samenvalt met de *relatieve frequentieverdeling van de gegeven populatie* geldt dat $E(X) =$ populatiegemiddelde $\mu$ en $Var(X) =$ populatievariantie $\sigma^2$.

In plaats van de discrete voorstelling te geven van de (relatieve)frequenties van de 34 *verschillende* data biedt een histogram een beter beeld. Bepaal een frequentietabel met het programma FREQTAB (zie opdracht 9 - paragraaf 4.8), met als resultaat het plaatje hiernaast.




We kiezen $n$ keer lukraak een getal, *dat we telkens terugleggen*, uit de populatie. Dit geeft een rij van *onafhankelijke* stochasten $X_1, X_2, \ldots, X_n$ met dezelfde (kans)verdeling als $X$. Zo een rij noemen we een steekproef van grootte $n$ uit een populatie met populatiestochast $X$.

Uitgaande van een steekproef berekent men vaak het *steekproefgemiddelde* $\overline{X}$ :

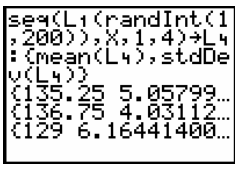
$$\overline{X} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

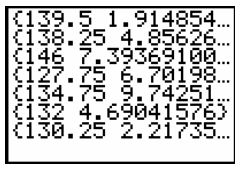
Dit is een stochast. Pas na het uitvoeren van een steekproef krijgen we één concrete getalwaarde $\overline{x} = \frac{x_1 + x_2 + \ldots x_n}{n}$, afhankelijk van het toeval. Als we nog een steekproef uitvoeren, krijgen we wellicht een andere getalwaarde $\overline{x}$.

Als je dat heel vaak doet, zal de (relatieve) frequentieverdeling van de zo verkregen gemiddelden langzaam maar zeker evolueren naar de kansverdeling van de stochast $\overline{X}$.

De exacte kansverdeling van $\overline{X}$ is de relatieve frequentieverdeling van de populatie van de gemiddelden van alle geordende *n*-tallen die we kunnen vormen met elementen uit de gegeven populatie.

We genereren enkele steekproeven van grootte 4 uit de populatie van de 200 lengtes die we in een lijst plaatsten. We berekenen telkens het steekproefgemiddelde en de steekproefstandaardafwijking.




We zien dat er veel variatie is in de resultaten. Theoretisch vinden we voor een steekproef van grootte $n$ uit een populatie met populatiestochast $X$ :

$$E(\overline{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n} \cdot \left(E\left(X_1\right) + \cdots + E\left(X_n\right)\right) = \frac{1}{n} \cdot n \cdot E\left(X\right) = E\left(X\right) = \mu$$

De verwachtingswaarde van het steekproefgemiddelde $\overline{X}$ is steeds gelijk aan het populatiegemiddelde $\mu$. Daarom noemt men $\overline{X}$ een *onvertekende schatter* van $\mu$. De concrete verkregen waarde $\overline{x}$ na het uitvoeren van een steekproef noemen we een *schatting* van $\mu$.

Zo verkregen we bij de bovenstaande steekproeven als schattingen ($\overline{x}$) voor $\mu = 135.575$ achtereenvolgens 135.25, 136.75, 129,….

Deze schattingen schommelen om en bij $E\left(\overline{X}\right) = \mu$. In het hoofdstuk over betrouwbaarheidsintervallen gaan we dieper in op de kwaliteit van zo een schatting.

De steekproefvariantie $S^2$ is per definitie : $S^2 = \dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$. Dit is ook een stochast. Men kan bewijzen dat $E\left(S^2\right) = Var(X) = \sigma^2$, m.a.w. dat $S^2$ een *onvertekende schatter* is van de populatievariantie $\sigma^2$.

Voor deze eigenschap is het noodzakelijk dat in de definitie van steekproef-variantie gedeeld wordt door $n-1$. Dit is de reden waarom we $s^2$ als *schatting* gebruiken voor $\sigma^2$ en bijgevolg schatten we $\sigma$ met $s$.

Bij bovenstaande steekproeven vonden we als *schattingen* ($s$) van $\sigma = 7.17$ achtereenvolgens de volgende steekproefstandaardafwijkingen : 5.06, 4.03, 6.16.

We kunnen aantonen dat voor de variantie van het steekproefgemiddelde $\overline{X}$ geldt :

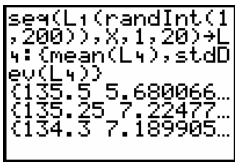
$$Var\left(\overline{X}\right) = \frac{Var(X)}{n} = \frac{\sigma^2}{n} \text{ of } \sigma_{\overline{X}}^2 = \frac{\sigma_X^2}{n}$$

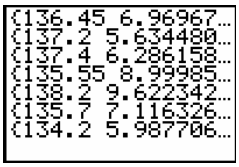
Dit is de populatievariantie gedeeld door de steekproefgrootte. Dit is een belangrijk resultaat.

De variantie van het steekproefgemiddelde wordt kleiner naarmate de steekproefomvang $n$ toeneemt. M.a.w. hoe groter de steekproef, hoe minder variatie er zal zijn in de verkregen steekproefgemiddelden $\overline{x}$ en hoe beter we $\mu$ kunnen schatten.

Ook $Var\left(S^2\right)$ wordt kleiner naarmate *n* groter wordt zodat we met $s^2$ ook beter $\sigma^2$ kunnen schatten. De steekproefgrootte heeft alvast invloed op de kwaliteit van een schatting.

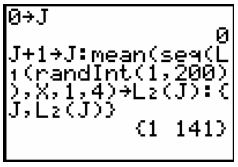
We illustreren dat we betere schattingen krijgen door enkele steekproeven te genereren van grootte 20 (telkens steekproeven met terugleggen).

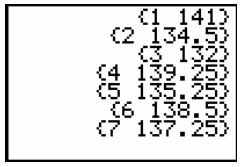


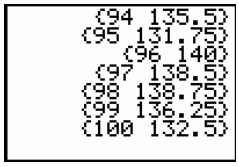




Er is inderdaad minder variatie in de resultaten en je krijgt betere schattingen voor $\mu$ en $\sigma$.

Tenslotte genereren we 100 steekproeven van grootte vier uit onze populatie van 200 getallen die zich in lijst **L1** bevinden. De steekproefgemiddelden komen in **L2.** Je ziet bijvoorbeeld dat de zesde steekproef als gemiddelde 138.5 heeft.

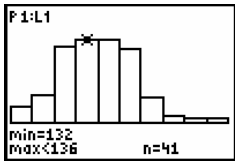


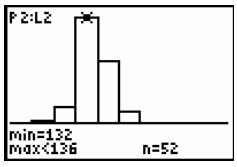



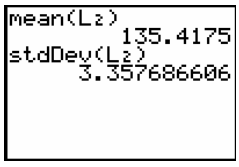




We vergelijken het histogram van de gegeven populatie in **L1** en dat van de 100 steekproefgemiddelden in **L2**.

Voor beide grafieken is **Xmin=120** , **Xmax=160** en **Xscl=4**.




We merken op dat er een kleinere spreiding is in de gemiddelden, zoals verwacht.

In ons voorbeeld vonden we 135.4175 als gemiddelde. Dit is een schatting van $\mu = 135.575$, het gemiddelde van de $200^4$ mogelijke steekproefgemiddelden van steekproeven van grootte vier uit onze populatie met 200 getallen !

In de simulatie was de steekproefstandaardafwijking van de 100 steekproef-gemiddelden gelijk aan 3.36.

Dit is een schatting van de theoretische $\sigma_{\bar{X}} = \sqrt{\frac{\sigma_X^2}{4}} = \frac{\sigma_X}{2} = \frac{7.17}{2} = 3.585$.

## 5.9 Steekproeven met en zonder terugleggen

In voorgaande paragraaf beschouwden we steekproeven met terugleggen. Een getrokken getal werd telkens teruggelegd vooraleer een nieuw getal werd getrokken. Dit garandeert dat elke $X_i$ van de steekproef dezelfde verdeling heeft als de populatiestochast *X*.

Bij steekproeven zonder terugleggen zijn de $X_i$'s echter niet onafhankelijk.

Hier geldt wel nog dat $E\left(\overline{X}\right) = \mu$ .

Deze intuïtief efficiëntere manier van werken resulteert dan ook in een kleinere standaardafwijking van $\overline{X}$ : $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$ , met *N* de populatiegrootte en *n* de steekproefgrootte. De formule voor $\sigma_{\overline{X}}$ wordt dus ingewikkelder.

Voor grote *N*, waarbij *n* gevoelig kleiner is dan *N*, wordt de reductiefactor $\sqrt{\frac{N-n}{N-1}}$ ongeveer 1.

In dit geval kunnen we de eenvoudige formule $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$ blijven gebruiken.

## 5.10 Opdrachten

1. Stel $X$ een stochast met $E(X)=\mu$. Toon aan dat $Var(X)=E\left(X^2\right)-\mu^2$.

2. Ziehier de verdeling van het aantal inwoners bij huisgezinnen in Amerika :

| aantal inwoners | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| fractie van de huisgezinnen | .25 | .32 | .17 | .15 | .07 | .03 | .01 |

   Kies lukraak een huisgezin en stel $X$ het aantal inwoners. De stochast *X* heeft een kansverdeling die gegeven wordt door de bovenstaande tabel.

   Bereken $E(X)$ en $Var(X)$.

3. Werp een dobbelsteen en stel *X* het aantal ogen. Bereken $E(X)$ en $Var(X)$. Bereken tevens $E\left(X^2\right)$.

   Simuleer honderd worpen met een dobbelsteen. Bereken het gemiddelde en de steekproefvariantie van het aantal ogen en vergelijk dit met de theoretische waarden $E(X)$ en $Var(X)$.

4. Werp twee dobbelstenen. Stel *X* het aantal ogen op de eerste dobbelsteen, *Y* het aantal ogen op de tweede dobbelsteen en $S=X+Y$ het totaal aantal ogen.

   Bereken $E(S)$ en $Var(S)$ :

   (a) met behulp van de resultaten van opdracht 3 en
   (b) uitgaande van de kansverdeling van $S$.

5. Stel $X$ en $Y$ onafhankelijke stochastische veranderlijken met $E(X)=2$, $E\left(X^2\right)=7$ en $E(Y)=-1$, $Var(Y)=5$.

   Bereken :

   (a) $E\left[(X-3)^2\right]$ (b) $Var(X)$ (c) $E(3X-2Y+8)$
   (d) $Var(2X-4Y)$ (e) $Var(Y+2X+10)$ (f) $Var(aX\pm bY)$

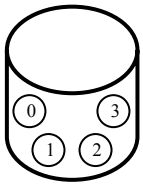
6. Beschouw het volgende kansspel.

   Werp twee dobbelstenen en tel het totaal aantal ogen. Indien dit groter is dan 7, krijg je 5 Euro. Zoniet betaal je 4 Euro. Is dit een eerlijk spel ?

   Stel $X$ de winst per spel. Voor een eerlijk spel moet $E(X)=0$, het spel is ongunstig voor de speler indien $E(X)<0$ en gunstig als $E(X)>0$.

   Simuleer 100 spelen en bereken nadien de verwachte winst.

7. Trek twee getallen uit onderstaande vaas. Noem $X$ het eerste getal en $Y$ het tweede getal.

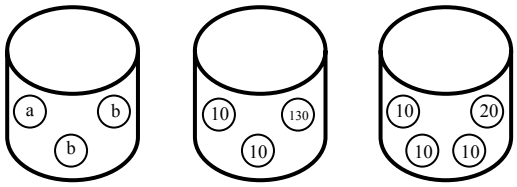



   Stel $M=\max(X,Y)$ en $S=X+Y$. Bepaal de kansverdeling van $M$ en $S$ en hun gemiddelde.

   (a) voor een trekking met terugleggen,
   (b) voor een trekking zonder terugleggen.

   Simuleer dit experiment 100 keer, bereken het gemiddelde en vergelijk met de theoretische waarde.

8. Beschouw de 3 onderstaande vazen. Een letter wordt uit de eerste vaas getrokken. Is dit de letter a, dan trekken we een getal uit de tweede vaas. Is het de letter b, dan trekken we een getal uit de derde vaas. Noem $X$ het getrokken getal. Teken een kansboom en bereken $E(X)$.




9. Uit Amerika komt het spel “chuck a luck” met dobbelstenen. De speler mag inzetten op één van de getallen 1,2,3,4,5,6. Vervolgens werpt hij drie dobbelstenen. Komt zijn getal 1,2 of 3 keer tevoorschijn, dan krijgt hij 1,2 of 3 keer zijn inzet met daarbij zijn inzet terug. Stel X de winst met als inzet 1 dollar. Bereken $E(X)$.

10. In een bepaalde wijk valt gedurende een maand een aantal straatlantaarns uit. Dit aantal $X$ heeft de volgende kansverdeling:

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $P(X = x)$ | 0.15 | 0.25 | 0.30 | 0.15 | 0.10 | 0.05 |

Een monteur gaat één keer per maand op controle in de wijk en vervangt de defecte lampen. De kosten hierbij zijn 25 Euro vast plus 5 Euro per vervangen lamp. Stel $K$ het bedrag te betalen aan de monteur.
Bereken $E(X)$ en $E(K)$.