De standaardafwijking

Prof. dr. Herman Callaert

Inhoudstafel

1 Motivatie

Teksten over statistiek zijn niet altijd duidelijk wanneer het over de standaardafwijking gaat. Zowel de formule als de notatie kan tot verwarring leiden. Die verwarring wordt soms nog versterkt door het gebruik van software (GRM of PC).

Deze tekst motiveert, vanuit verschillende invalshoeken, waarom je, bij de berekening van de standaardafwijking van een verzameling getallen, zo goed als altijd deelt door (n–1). Soms geven we een wiskundig bewijs en soms werken we intuïtief vanuit voorbeelden. Dit is geen lesmateriaal voor het secundair onderwijs. Het is een tekst voor geïnteresseerde leerkrachten, niet voor leerlingen.

2 Een groter kader: lineaire modellen

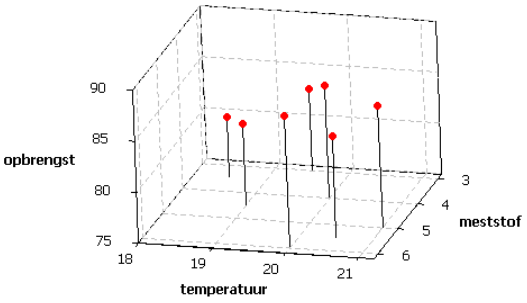
Sommen van kwadratische afwijkingen zijn bouwstenen om populatievarianties te schatten. Hoe dat werkt, illustreren we met eenvoudige modellen bij lineaire regressie. We tonen daarbij aan dat de “intuïtieve” reflex om een som te delen door het aantal termen helemaal niet gebruikelijk is bij het schatten van variabiliteit.

De methode der kleinste kwadraten klinkt vertrouwd bij het schatten van een “beste” vlak of een “beste” rechte. Waarschijnlijk denk je daar niet onmiddellijk aan bij het bepalen van een “beste” punt. Daarom bekijken we voorbeelden die van dimensie 3 over dimensie 2 naar dimensie 1 gaan.

In de tekst over regressie op http://www.uhasselt.be/lesmateriaal-statistiek kan je meer informatie vinden over de basisterminologie. Heel wat standaardwerken over (meervoudige) regressie bevatten wiskundige bewijzen van eigenschappen die we hieronder ter illustratie vermelden.

2.1 Drie dimensies, twee verklarende veranderlijken

In een studie over de teelt van maïs werd naast de opbrengst ook de hoeveelheid toegediende meststof en de temperatuur opgemeten. Het is de bedoeling om na te gaan hoe de opbrengst wijzigt in functie van meststof en temperatuur.



In deze studie zijn meststof en temperatuur verklarende veranderlijken en maïsopbrengst is de respons.

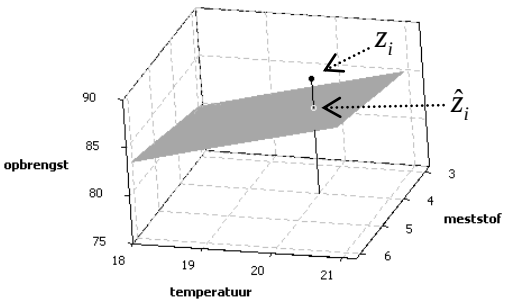
De resultaten waren als volgt:

x_i = meststof (kg/are)	3.0	3.5	4.0	4.5	5.0	5.5	6.0
y_i = temperatuur (°C)	19.5	18.5	20.0	19.0	21.0	20.5	20.0
z_i = opbrengst (kg/are)	83	81	86	83	87	85	88

In de tabel (en op de figuur) zie je dat bij een hoeveelheid meststof van 4 kg/are en bij een temperatuur van 20 °C een opbrengst van 86 kg/are is opgemeten. Als je een volgende keer werkt

met 4 kg/are meststof bij een temperatuur van 20 °C, dan verwacht je wellicht een opbrengst van dezelfde grootteorde, maar niet exact 86 kg/are. De opbrengst behandel je als een realisatie van "een onderliggend kansmodel voor opbrengst". Daarbij onderstel je dat, voor elke combinatie van meststof en temperatuur, de "kansmodellen voor de opbrengst" een gemiddelde hebben dat "in een vlak" ligt, waarbij de variabiliteit rond dat gemiddelde gekenmerkt wordt door een vaste maar niet gekende variantie σ^2.

Het "vlak van de gemiddelden" schat je op basis van je steekproef. Je bepaalt de vergelijking van een vlak $\hat{z}$ = ax + by + c zodanig dat $\sum_{i=1}^{n}\left(z_i - \hat{z}_i\right)^2$ minimaal is. Hierbij is z_i de opgemeten opbrengst die hoort bij de waarde $\left(x_i, y_i\right)$ = (meststof , temperatuur) en $\hat{z}_i$ is de verwachte opbrengst met waarde $ax_i + by_i + c$ (het punt in het vlak). De methode die je hier gebruikt, minimaliseert de som van de kwadratische afwijkingen: het is **de methode der kleinste kwadraten**.



Om een idee te hebben over de variabiliteit start je met een som waarvan de bouwstenen de kwadratische afwijkingen zijn van "opgemeten waarde " ten opzichte van "verwachte waarde": $\sum_{i=1}^{n}\left(z_i - \hat{z}_i\right)^2$. In softwarepakketten en (Engelstalige) tekstboeken wordt deze som genoteerd als SSE, waarbij SS staat voor "**S**um of **S**quares" en E voor "**E**rror".

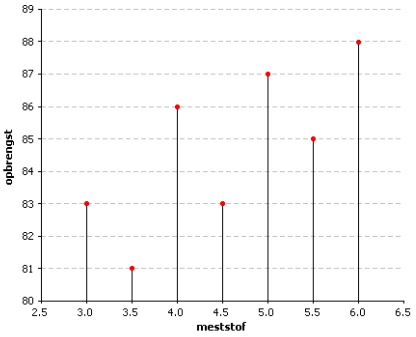
De kwadraatsom houdt rekening met alle observaties.
Als je nu eens dubbel zoveel observaties zou hebben die globaal eenzelfde variabiliteit rond het gemiddelde zouden vertonen, dan zou de nieuwe kwadraatsom zowat dubbel zo groot zijn. Om een goede schatting voor de (vaste) populatievariantie σ^2 te krijgen, moet er dus op een of andere manier gecompenseerd worden voor het aantal observaties. Men stapt dan over van SSE naar MSE (MS = **M**ean **S**quare). En hoewel je hier het woord "mean" (= gemiddelde) ontmoet, toch zal je **nergens** een tekst vinden waar de som der kwadraten gedeeld wordt door het aantal termen.

De kwadraatsom gepast standaardiseren betekent dat je hier moet delen door (n–3). Je werkt dan met $MSE = \frac{1}{(n-3)}\sum_{i=1}^{n}\left(z_i - \hat{z}_i\right)^2$. Als model geldt dat $E(MSE) = \sigma^2$: gemiddeld kom je exact op de populatievariantie σ^2 terecht.

Daarom werk je hier met $\frac{1}{(n-3)}\sum_{i=1}^{n}\left(z_i - \hat{z}_i\right)^2$.

2.2 Twee dimensies, één verklarende veranderlijke

In een studie over de teelt van maïs werd naast de opbrengst ook de hoeveelheid toegediende meststof opgemeten. Het is de bedoeling om na te gaan hoe de opbrengst wijzigt in functie van de meststof.



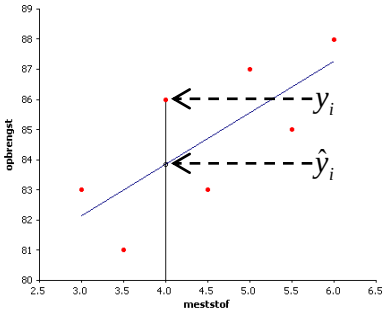
In deze studie is meststof de verklarende veranderlijke en maïsopbrengst is de respons.

De resultaten waren als volgt:

x_i = meststof (kg/are)	3.0	3.5	4.0	4.5	5.0	5.5	6.0
y_i = opbrengst (kg/are)	83	81	86	83	87	85	88

In de tabel (en op de figuur) zie je dat bij een hoeveelheid meststof van 4 kg/are een opbrengst van 86 kg/are is opgemeten. Als je een volgende keer werkt met 4 kg/are meststof, dan verwacht je een opbrengst van dezelfde grootteorde, maar niet exact 86 kg/are. De opbrengst behandel je als een realisatie van "een onderliggend kansmodel voor opbrengst". Daarbij onderstel je dat, voor elke niveau van toegediende meststof, de "kansmodellen voor de opbrengst" een gemiddelde hebben dat "op een rechte" ligt, waarbij de variabiliteit rond dat gemiddelde gekenmerkt wordt door een vaste maar niet gekende variantie σ^2.

De "rechte van de gemiddelden" schat je op basis van je steekproef. Je bepaalt de vergelijking van een rechte $\hat{y} = ax + b$ zodanig dat $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ minimaal is. Hierbij is y_i de <u>opgemeten opbrengst</u> die hoort bij de waarde x_i (meststof) en $\hat{y}_i$ is de <u>verwachte opbrengst</u> met waarde $ax_i + b$ (het punt op de rechte). De methode die je hier gebruikt, minimaliseert de som van de kwadratische afwijkingen: het is **de methode der kleinste kwadraten**.



Om een idee te hebben over de variabiliteit start je met een som waarvan de bouwstenen de kwadratische afwijkingen zijn van "opgemeten waarde " ten opzichte van "verwachte waarde": $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Ook hier noteer je deze som als SSE, met SS = "<u>**S**</u>um of <u>**S**</u>quares" en E = "<u>**E**</u>rror".

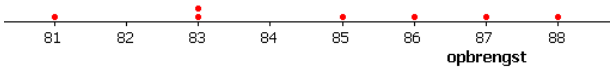
Je standaardiseert en stapt over van SSE naar MSE (MS= <u>**M**</u>ean <u>**S**</u>quare). En ook hier <u>zal je **nergens** een tekst vinden waar de som der kwadraten gedeeld wordt door het aantal termen</u>. Je werkt hier met $MSE = \frac{1}{(n-2)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Voor MSE (als model) geldt dat $E(MSE) = \sigma^2$. De kwadraatsom delen door (n–2) levert hier een grootheid die gemiddeld exact op σ^2 terechtkomt.

Daarom werk je hier met $\frac{1}{(n-2)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

2.3 Eén dimensie, geen verklarende veranderlijke

In een studie over de teelt van maïs werd, op eenzelfde perceel en onder dezelfde omstandigheden, meerdere keren de opbrengst opgemeten.

De resultaten waren als volgt:



x_i = opbrengst (kg/are)	83	81	86	83	87	85	88

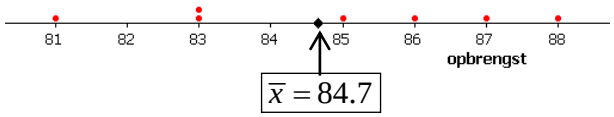
Dat je niet altijd dezelfde opbrengst vindt, is niet verwonderlijk. De opgemeten opbrengsten behandel je als realisaties van "een onderliggend kansmodel voor opbrengst" (= de populatie). Dat kansmodel heeft een (vast maar niet gekend) gemiddelde μ en een variabiliteit die gekenmerkt wordt door een (vaste maar niet gekende) variantie σ^2.

Bij twee verklarende veranderlijken heb je het vlak van de populatiegemiddelden geschat door een **vlak** dat volgens **de methode der kleinste kwadraten** het beste aansluit bij de meetpunten.

Bij één verklarende veranderlijke heb je de rechte van de populatiegemiddelden geschat door een **rechte** die volgens **de methode der kleinste kwadraten** het beste aansluit bij de meetpunten.

Als er geen verklarende veranderlijken in de studie zijn, dan schat je het populatiegemiddelde door een **punt** dat volgens **de methode der kleinste kwadraten** het beste aansluit bij de meetpunten.

Het punt a waarvoor $\sum_{i=1}^{n}(x_i - a)^2$ minimaal is wordt gegeven door $a = \bar{x}$. Het gemiddelde $\bar{x}$ is een schatting voor de verwachte opbrengst.



$\bar{x} = 84.7$

Om een idee te hebben over de variabiliteit start je met een som waarvan de bouwstenen de kwadratische afwijkingen zijn van "opgemeten waarde " ten opzichte van "verwachte waarde": $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Deze kwadraatsom moet je nu nog "standaardiseren". Ook hier deel je niet door het aantal termen maar wel door (n–1). Als model geldt immers voor $S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2$ dat $E(S^2) = \sigma^2$. De kwadraatsom delen door (n–1) levert een grootheid die gemiddeld exact op σ^2 terechtkomt.

Daarom werk je hier met $\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

3 Een wiskundig bewijs: $E(S^2)=\sigma^2$

Uit een populatie met vaste maar niet gekende variantie σ^2 trek je een steekproef en je berekent $s^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Uit diezelfde populatie trek je opnieuw een steekproef en je berekent terug s^2. En je vindt (zo goed als zeker) een andere waarde. Als je dit heel veel keren zou herhalen, waar kom je dan met al die s^2-waarden gemiddeld terecht? Om dit te beantwoorden moet je kijken waar het onderliggende model $S^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\bar{X})^2$ gemiddeld terechtkomt. Hieronder staat het bewijs dat $E(S^2)=\sigma^2$: de steekproefvariantie S^2 waarbij je deelt door *(n–1)* komt gemiddeld op de populatievariantie σ^2 terecht.

3.1 Inleidende eigenschappen van kansmodellen

Een uitgebreidere uitleg over de gebruikte begrippen en notaties kan je vinden in de teksten over kansmodellen op http://www.uhasselt.be/lesmateriaal-statistiek .

De populatie (als kansmodel) noteer je met een hoofdletter X en populatie-eigenschappen (populatieparameters) noteer je met een Griekse letter:

- het populatiegemiddelde $E(X)$ noteer je als μ
- de populatievariantie $var(X)$ noteer je als σ^2.

Een steekproef (als kansmodel) noteer je als $(X_1, X_2, \ldots X_n)$ waarbij

- $E(X_i)=\mu$ en $var(X_i)=\sigma^2$ voor elke i (1)
- de X_i's onafhankelijk zijn. (2)

De verwachtingswaarde E is een lineaire operator: $E\left(\sum_{i=1}^{n}a_i\,U_i\right)=\sum_{i=1}^{n}a_i\,E(U_i)$. (3)

De variantie *var* voldoet bij kansmodellen aan:

- $var(U)=E\left[\left(U-E(U)\right)^2\right]$ (4)
- $var\left(\sum_{i=1}^{n}a_i\,U_i\right)=\sum_{i=1}^{n}a_i^2\,var(U_i)$ als de U_i's onafhankelijk zijn. (5)

Het steekproefgemiddelde (als kansmodel) noteer je als $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ waarbij

- $E(\bar{X})=\mu$ [gebruik (3) en (1)] (6)
- $var(\bar{X})=\frac{\sigma^2}{n}$ [gebruik (2), (5) en (1)]. (7)

Bemerk dat $\sum_{i=1}^{n}X_i=n\,\bar{X}$ zodat $\sum_{i=1}^{n}(X_i-\mu)=n\,\bar{X}-n\,\mu=n\left(\bar{X}-\mu\right)$. (8)

3.2 Het gemiddelde van de steekproefvariantie is de populatievariantie

De steekproefvariantie (als kansmodel) definieer je als $S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$.

Bemerk vooreerst dat

$$\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 = \sum_{i=1}^{n} \left[\left(X_i - \mu \right) - \left(\bar{X} - \mu \right) \right]^2 = \sum_{i=1}^{n} \left[\left(X_i - \mu \right)^2 - 2 \left(X_i - \mu \right) \cdot \left(\bar{X} - \mu \right) + \left(\bar{X} - \mu \right)^2 \right]$$

$$= \sum_{i=1}^{n} \left(X_i - \mu \right)^2 - 2\, n \left(\bar{X} - \mu \right)^2 + n \left(\bar{X} - \mu \right)^2 \text{ [gebruik (8)]}$$

$$= \sum_{i=1}^{n} \left(X_i - \mu \right)^2 - n \left(\bar{X} - \mu \right)^2 = \sum_{i=1}^{n} \left(X_i - E(X_i) \right)^2 - n \left(\bar{X} - E(\bar{X}) \right)^2 \text{ [gebruik (1) en (6)]}$$

zodat

$$E\left[\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \right] = \sum_{i=1}^{n} E\left[\left(X_i - E(X_i) \right)^2 \right] - n\, E\left[\left(\bar{X} - E(\bar{X}) \right)^2 \right] \text{ [gebruik (3)]}$$

$$= \sum_{i=1}^{n} var\,(X_i) - n\, var\,(\bar{X}) \text{ [gebruik (4)]}$$

$$= n\,\sigma^2 - \sigma^2 = (n-1)\sigma^2 \text{ [gebruik (1) en (7)].}$$

Hieruit volgt dat $E\left[\frac{1}{(n-1)} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \right] = \frac{1}{(n-1)} E\left[\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 \right] = \sigma^2$ of dat $E\left(S^2 \right) = \sigma^2$.

$S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$ is dus een goed kansmodel dat gemiddeld op de (vaste maar niet gekende) populatievariantie σ^2 terechtkomt (S^2 is een onvertekende schatter voor σ^2).

3.3 Getalwaarden van kansmodellen

Een waarde van een kansmodel stel je voor door de overeenkomstige kleine letter.

Na het trekken van een steekproef beschik je over jouw toevallig gevonden steekproefwaarden $\left(x_1, x_2, x_3, \ldots, x_n \right)$. Die waarden gebruik je om een waarde van een kansmodel, gebaseerd op steekproefresultaten, te berekenen.

Als je voor de steekproefvariantie (als model) de formule $S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$ gebruikt, dan noteer je een waarde van dit kansmodel (= de variantie van je waarnemingsgetallen) als $s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$. Voor de standaardafwijking heb je: $s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}$.

4 De standaardafwijking van een populatie: voorbeelden

"*Een gegeven verzameling getallen beschouw ik als een populatie en daarom deel ik door n*" is een uitspraak die je niet zomaar doet. Als je echt met een populatie werkt, dan moet je dat in het juiste kader plaatsen.

In de statistiek bestudeer je een populatie X in het kader van kansmodellen.

Eigenschappen van populaties noteer je met een Griekse letter. Dit betekent dat je het gemiddelde van een populatie noteert als μ en de standaardafwijking als σ. Om μ en σ te berekenen gebruik je de algemene formules voor kansmodellen. Meer info vind je in onze teksten over kansmodellen op http://www.uhasselt.be/lesmateriaal-statistiek .

Bij de overgrote meerderheid van kansmodellen deel je helemaal niet door *n* bij het berekenen van de variantie (of de standaardafwijking). Dat zie je hieronder bij continue populaties (zoals de normale) of bij discrete populaties (zoals de binomiale).

4.1 Continue populaties

Een continue populatie X heeft uitkomsten die ***in intervallen*** terechtkomen waarbij de kans om in een interval terecht te komen, gestuurd wordt door een ***dichtheidsfunctie*** $f(x)$.

De standaardafwijking van een continu kansmodel X is gelijk aan

$$sd(X) = \sqrt{\text{var}(X)} \quad = \sqrt{\int_{-\infty}^{+\infty} \left(x - E(X)\right)^2 \, f(x)\, dx}$$

Voorbeeld. *Normaal verdeeld kansmodel*

Voor elke vaste waarde van de parameters μ en σ heb je een welbepaald normaal verdeeld kansmodel X, vastgelegd door de dichtheidsfunctie

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{voor} \quad -\infty < x < +\infty \quad \text{en met} \quad \begin{cases} -\infty < \mu < +\infty \\ 0 < \sigma \end{cases}.$$

Door de algemene formules voor continue populaties toe te passen vind je dat $E(X) = \mu$ en dat

$$\text{var}(X) = \int_{-\infty}^{+\infty} \left(x-\mu\right)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2 \cdot \frac{2}{\sqrt{\pi}} \int_{-\infty}^{+\infty} t^2 e^{-t^2} dt = \sigma^2 \quad \text{met} \quad t = \frac{x-\mu}{\sqrt{2}\cdot\sigma}$$

zodat $sd(X) = \sigma$.

4.2 Discrete populaties

Een discrete populatie X heeft ***discrete uitkomsten*** met welbepaalde kansen, vastgelegd in een ***kansverdeling*** [= de uitkomsten x_i samen met hun kansen $P(X = x_i)$].

De standaardafwijking van een discreet kansmodel X is gelijk aan

$$sd(X) = \sqrt{\text{var}(X)} = \sqrt{\sum_{i=1}^{n} \left(x_i - E(X)\right)^2 \cdot P(X = x_i)}$$

Voorbeeld. *Binomiaal kansmodel*

Bij n onafhankelijke herhalingen van een $0-1$ experiment met succeskans π krijg je het binomiale kansmodel X met uitkomsten 0, 1, 2, … , n en kansen $P(X = k) = \binom{n}{k} \pi^k \left(1-\pi\right)^{(n-k)}$.

De algemene formules voor discrete populaties leveren hier dat het populatiegemiddelde μ gelijk is aan $E(X) = \sum_{k=0}^{n} k \cdot P(X = k) = \sum_{k=0}^{n} k \cdot \binom{n}{k} \pi^k \left(1-\pi\right)^{(n-k)} = n\,\pi$.

De populatievariantie σ^2 is gelijk aan

$$\text{var}(X) = \sum_{k=0}^{n} \left(k-\mu\right)^2 \cdot P(X = k) = \sum_{k=0}^{n} \left(k - n\pi\right)^2 \cdot \binom{n}{k} \pi^k \left(1-\pi\right)^{(n-k)} = n\,\pi\,(1-\pi)$$

zodat de standaardafwijking σ gegeven wordt door $sd(X) = \sqrt{n\,\pi\,(1-\pi)}$.

Voorbeeld. *Discreet uniform kansmodel*

Wanneer een populatie X een eindig aantal uitkomsten $x_1, x_2, x_3, ..., x_n$ heeft waarbij iedere uitkomst dezelfde kans heeft (dan moet $P(X = x_i) = \frac{1}{n}$ voor elke i), dan heb je te maken met een discreet uniform kansmodel.

Het populatiegemiddelde μ vind je uit $E(X) = \sum_{i=1}^{n} x_i \cdot P(X = x_i) = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

De populatievariantie σ^2 is gelijk aan $\text{var}(X) = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot P(X = x_i) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

Om $\text{var}(X)$ te bepalen, maak je de som (over alle uitkomsten) van:
[(*de kwadratische afwijking* van de uitkomst tot het gemiddelde) **maal** (*de kans* van de uitkomst)]

Als je zegt dat je een verzameling getallen beschouwt als een populatie en dus (minstens impliciet) aangeeft dat je te maken hebt met een discreet uniform kansmodel, dan betekent de $\frac{1}{n}$ in de formule $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$ **NIET** dat je *deelt* door het aantal termen, maar **WEL** dat je *vermenigvuldigt* met de kans $\frac{1}{n}$ van elke uitkomst.

Houd dan ook goed **de notatie** in het oog want voor een ***discreet uniforme populatie*** X geldt:

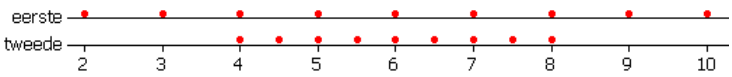
- het gemiddelde $E(X)$ noteer je als μ en bereken je met $\frac{1}{n} \sum_{i=1}^{n} x_i$
- de variantie $\text{var}(X)$ noteer je als σ^2 en bereken je met $\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$. Voor de standaardafwijking volgt dan: $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$.

5 Aanbeveling

5.1 De standaardafwijking van een dataset

In de tweede graad en in de meeste studies in de derde graad onderzoek je de structuur van een verzameling getallen (zonder verdere context of in de context van steekproefresultaten). Daarbij komt het gemiddelde (als een kengetal voor het centrum) aan bod. De notatie die je voor dit gemiddelde gebruikt is $\bar{x}$ en de waarde vind je uit de formule $\frac{1}{n}\sum_{i=1}^{n} x_i$.

De variabiliteit van getallen rond hun gemiddelde kan je grafisch voorstellen door bijvoorbeeld 2 verzamelingen te vergelijken die eenzelfde gemiddelde hebben maar een verschillende variabiliteit.



Zowel de eerste als de tweede dataset heeft een gemiddelde dat gelijk is aan $\bar{x} = 6$. Op zicht zie je dat de eerste dataset een grotere variabiliteit vertoont dan de tweede.

Als je op een numerieke manier, in één getal, de variabiliteit wil karakteriseren, dan wil je een maat die een grotere uitkomst oplevert naarmate de variabiliteit groter is. Als je hier de standaardafwijking, genoteerd als s, berekent met de formule $\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$, dan vind je $s = 2.74$ voor de eerste dataset en $s = 1.37$ voor de tweede. In deze zin voldoet de formule aan de verwachting.

Voor de leerling is een motivatie “ten gronde” op het niveau van het secundair onderwijs moeilijk. Waarom werk je voor afstand niet met absolute waarden $\left|x_i-\bar{x}\right|$? Waarom gebruik je kwadraten $\left(x_i-\bar{x}\right)^2$ als bouwstenen in een som? Waarom deel je door (n-1) om te standaardiseren?

Antwoorden op deze vragen komen uit “statistiek op een hoger niveau”. Als leerlingen in het hoger onderwijs statistische methoden (toetsen van hypothesen, lineaire regressie,…) gebruiken, dan zullen zij daar de standaardafwijking $s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$ ontmoeten waarmee zij van in het middelbaar vertrouwd zijn.

5.2 Besluit

In statistische studies kan je bij een verzameling getallen zo goed als altijd dezelfde methoden en formules gebruiken als waarmee je werkt bij steekproeven. Zelden bestudeer je een verzameling getallen in het kader van een discreet uniforme populatie. Dat doe je alleen maar bij de (theoretische) studie van kansmodellen.

Bij steekproeven gebruik je $\bar{x}$ voor het gemiddelde en s voor de standaardafwijking.
Bij populaties gebruik je μ voor het gemiddelde en σ voor de standaardafwijking.

Voor de standaardafwijking van een verzameling getallen werk je met:

ZO GOED ALS ALTIJD:

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2} .$$

Dit is de klassieke formule in het kader van steekproeven.

ZELDEN:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(x_i - \mu \right)^2} .$$

Dit is de formule in het kader van een discreet uniforme populatie.

NOOIT:

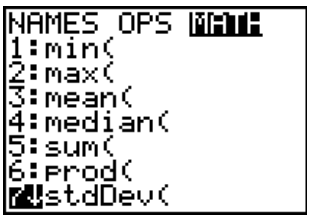
$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2} .$$

Dit is een formule waarbij σ zegt dat het over een populatie gaat en $\bar{x}$ aangeeft dat het een steekproef is.

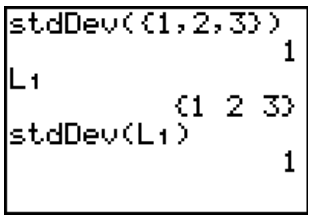
6 Nota: software, GRM en PC

Als je voor een dataset de standaardafwijking vraagt, dan krijg je bij heel wat statistische software maar één antwoord. Je kan er vanuit gaan dat dit antwoord $s=\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$ is.

Voorbeeld (bij de TI-84 Plus)



Druk [2nd][LIST], loop naar MATH en druk 7:stdDEV(. Dit is een manier om de standaardafwijking (stdDev = **st**an**d**ard **Dev**iation) van een lijst getallen op te vragen. Je kan de getallen ter plaatse intikken (zoals {1,2,3}) of je kan de getallen vooraf in een lijst (zoals [L1]) inbrengen. Als antwoord krijg je het getal 1. Dit betekent dat de standaardafwijking hier berekend is met de formule waarbij er gedeeld wordt door $(n-1)$.



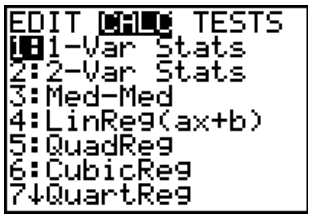
Voor de getallenset {1,2,3} is $\bar{x}=2$ zodat $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=1+0+1=2$. Als je dit resultaat deelt door $(n-1)=2$, dan krijg je dat $s^2=1$ zodat $s=1$.

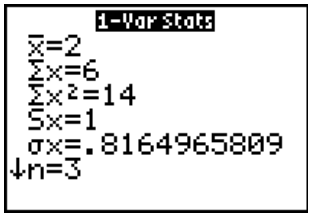
Als je zou delen door $n=3$ dan zou je niet de waarde 1 krijgen maar wel $\sqrt{\frac{2}{3}}\cong 0.82$.

Als je zowel een s als een σ ziet verschijnen dan is:

- $s=\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$
- $\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2}$ want de dataset wordt nu beschouwd als een discreet uniforme populatie, waarbij de notatie voor de standaardafwijking de Griekse letter σ is en de notatie voor het populatiegemiddelde de Griekse letter μ.
  Ken je een pakket waar je tussen de berekende kengetallen naast een σ ook een μ ziet staan? Eigenaardig nietwaar.

Voorbeeld (bij de TI-84 Plus)





De getallenset {1,2,3} staat in de lijst [L1]. Druk [STAT], loop naar CALC en druk 1:1-Var Stats. Zorg ervoor dat de lijst [L1] bij List: staat, loop naar Calculate en druk [ENTER]. Het resultaat Sx=1 verwijst naar de standaardafwijking van de getallenset / steekproef {1,2,3} en de correcte notatie hierbij is een kleine letter s. Het resultaat σx=.81649.. verwijst naar de standaardafwijking σ van een discreet uniforme populatie waarbij de notatie μ van het populatiegemiddelde niet in de lijst voorkomt. Voor veel leerlingen en leerkrachten is σx verwarrend (en totaal overbodig).